bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

Joint analysis of functional genomic data and genome-wide association studies of 18 human traits

Joseph K. Pickrell
Department of Genetics, Harvard Medical School
Correspondence to: joseph_pickrell@hms.harvard.edu

November 21, 2013

bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

Abstract

Annotations of gene structures and regulatory elements can inform genome-wide association studies (GWAS). However, choosing the relevant annotations for interpreting an association study of a given trait remains challenging. We describe a statistical model that uses association statistics computed across the genome to identify classes of genomic element that are enriched or depleted for loci that influence a trait. The model naturally incorporates multiple types of annotations. We applied the model to GWAS of 18 human traits, including red blood cell traits, platelet traits, glucose levels, lipid levels, height, BMI, and Crohn's disease. For each trait, we evaluated the relevance of 450 different genomic annotations, including protein-coding genes, enhancers, and DNase-I hypersensitive sites in over a hundred tissues and cell lines. We show that the fraction of phenotype-associated SNPs that influence protein sequence ranges from around 2% (for platelet volume) up to around 20% (for LDL cholesterol); that repressed chromatin is significantly depleted for SNPs associated with several traits; and that cell type-specific DNase-I hypersensitive sites are enriched for SNPs associated with several traits (for example, fibroblasts in Crohn's disease and muscle tissue in bone density). Finally, by re-weighting each GWAS using information from functional genomics, we increase the number of loci with high-confidence associations by around 5%.

bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

1 Introduction

A fundamental goal of human genetics is to create a catalogue of the genetic polymorphisms that cause phenotypic variation in our species and to characterize the precise molecular mechanisms by which these polymorphisms exert their effects. An important tool in the modern human genetics toolkit is the genome-wide association study, in which hundreds of thousands or millions of single nucleotide polymorphisms (SNPs) are genotyped in large cohorts of individuals and each polymorphism tested for a statistical association with some trait of interest. In recent years, GWAS have identified thousands of genomic regions that show reproducible statistical associations with a wide array of phenotypes and diseases [Visscher et al., 2012].

In general, the loci identified in GWAS of multifactorial traits have small effect sizes and are located outside of protein-coding exons [Hindorff et al., 2009]. This latter fact has generated considerable interest in annotating other types of genomic elements apart from exons. For example, the ENCODE project has generated detailed maps of histone modifications and transcription factor binding in six human cell lines, partially motivated by the goal of interpreting GWAS signals that may act via a mechanism of gene regulation [ENCODE Project Consortium et al., 2012]. Methods for combining potentially rich sources of functional genomic data with GWAS could in principle lead to important biological insights. The development of such a method is the aim of this paper.

There are two lines of research that motivate our work on this problem. The first is what are often called "enrichment" analyses. In this type of analysis, the most strongly associated SNPs in a GWAS are examined to see if they fall disproportionately in specific types of genomic region. These studies have found, for example, that SNPs identified in GWAS are enriched in protein-coding exons, promoters, and untranslated regions (UTRs) [Hindorff et al., 2009; Schork et al., 2013] and among those that influence gene expression [Lappalainen et al., 2013; Nicolae et al., 2010]. Further, in some cases, SNPs associated with a trait are enriched in gene regulatory regions in specific cell types [Cowper-Sal lari et al., 2012; Ernst et al., 2011; Gerasimova et al., 2013; Global Lipids Genetics Consortium et al., 2013; Hnisz et al., 2013; Karczewski et al., 2013; Maurano et al., 2012; Parker et al., 2013; Paul et al., 2013, 2011; Trynka et al., 2013; van der Harst et al., 2012] or near genes expressed in specific cell types [Hu et al., 2011; Lui et al., 2012]. However, the methods in these studies are generally not able to consider more than a single annotation at a time. Further, they are not set up to answer a question that we find important: consider two independent SNPs with equivalent P-values of 1×10^{-7} in a GWAS for some trait (note that this P-value does not reach the standard threshold of 5×10^{-8} for "significance"), the first of which is a nonsynonymous SNP and the second of which falls far from any known gene. What is the probability that the first SNP is truly associated with the trait, and how does this compare to the probability for the second?

A potential answer to this question comes from the second line of research that motivates this work. In association studies where the phenotype being studied is gene expression ("eQTL" studies, for "expression quantitative trait locus"), statistical models have been developed to identify shared characteristics of SNPs that influence gene expression [Gaffney et al., 2012; Lee et al., 2009;

Veyrieras et al., 2008]. In a hierarchical modeling framework, the probability that a given SNP influences gene expression can then depend on these characteristics. The key fact that makes these models useful in the context of eQTL mapping is that there is a large number of unambiguous eQTLs in the genome on which a model can be trained. In the GWAS context, the number of loci unambiguously associated with a given trait has historically been very small; learning the shared properties of two or three loci is not a job well-suited to statistical modeling. However, large meta-analyses of GWAS now regularly identify tens to hundreds of independent loci that influence a trait (e.g. Lango-Allen et al. [2010]; Teslovich et al. [2010]). The merits of hierarchical modeling in this context [Chen and Witte, 2007; Heron et al., 2011; Lewinger et al., 2007] are thus worth revisiting. Indeed, Carbonetto and Stephens [2013] have reported success in identifying loci involved in autoimmune diseases using a hierarchical model that incorporates information about groups of genes known to interact in a pathway.

In this paper we present a hierarchical model for jointly analyzing GWAS and genomic annotations. We applied this model to GWAS of 18 diseases and traits; for each trait, we learned the relevant types of genomic information from a set of 450 genome annotations.

2 Results

We assembled a set of 18 GWAS with publicly available summary statistics and a large number (at least around 20) of loci associated with the trait of interest. These included studies of red blood cell traits [van der Harst et al., 2012], platelet traits [Gieger et al., 2011], Crohn's disease [Jostins et al., 2012], BMI [Speliotes et al., 2010], lipid levels [Teslovich et al., 2010], height [Lango-Allen et al., 2010], bone mineral density [Estrada et al., 2012] and fasting glucose levels [Manning et al., 2012]. We used ImpG [Pasaniuc et al., 2013] to impute the summary statistics from each study for all common SNPs identified in European populations by the 1000 Genomes Project [Abecasis et al., 2010]. Overall we successfully imputed association statistics for around 80% of common SNPs (Supplementary Figure 1). We then assembled a set of genome annotations, paying specific attention to annotations available for many cell types since important regulatory elements may be active only in specific cell types. The main sources of genome annotations were 402 maps of DNase-I hypersensitivity in a wide range of primary cell types and cell lines [Maurano et al., 2012; Thurman et al., 2012]. We also included as annotations the output from "genome segmentation" of the six main ENCODE cell lines [Hoffman et al., 2013]; for each section of the genome in each cell line, Hoffman et al. [2013] report whether the histone modifications in the region are consistent with enhancer activity, transcription start sites, promoter-flanking regions, CTCF binding sites, or repressed chromatin. Finally, we included elements of gene structures (protein-coding exons and 3' and 5' UTRs). In total, we used data from 18 traits and 450 genomic annotations.

For each trait, we set out to identify which of the 450 annotations (if any) were enriched in genetic variants influencing the trait. To do this, we developed a hierarchical model that learns the shared properties of loci influencing a trait. The full details of the model are presented in the Methods, but can be summarized briefly. Conceptually, we break the genome into large, non-

overlapping blocks (with an average size of 2.5 Mb). Let the prior probability that any block k contains an association be Π_k. *If* there is an association in block k, then let the prior probability that any SNP i is the causal SNP be π_{ik}. We then allow both Π_k and π_{ik} to depend on annotations of the region and SNP, respectively, and estimate these quantities based on the patterns of enrichment across the whole genome. We tested this approach using simulations based on real data from a GWAS of height (Supplementary Material).

The methodology is best illustrated with an example. We started with an analysis of a GWAS of high-density lipoprotein (HDL) levels [Teslovich et al., 2010]. We first took each genomic annotation individually, and estimated its level of enrichment (or depletion) for loci that influence HDL (in the model, we additionally included a regional effect of gene density and and a SNP effect of distance to the nearest transcription start site, see the Methods for details). In Figure 1A, we show the top 40 annotations, ordered by how well each improves the fit of the model. Loci that influence HDL are most strongly enriched in enhancers identified in the HepG2 cell line, and most strongly depleted from genomic regions repressed in that same cell line. HepG2 cells are derived from a liver cancer; the relevance of this cell line to a lipid phenotype makes intuitive sense. However, there are many other additional (correlated) genome annotations that are enriched for loci that influence HDL (Figure 1A). We thus built a model including multiple annotations; to mitigate over-fitting in this situation we used a cross-validation approach (Methods). The best-fitting model is shown in Figure 1B. It include both enhancers and repressed chromatin identified in HepG2 cells, as well as coding exons and chromatin repressed in K562 cells.

A convenient side effect of fitting an explicit statistical model relating properties of SNPs to the probability of association is that we can use the functional information to re-weight the GWAS (Methods). We used the combined model for HDL to re-weight the association statistics across the genome (Figure 1C). There are several regions of the genome with strong evidence for association with HDL (posterior probability of association [PPA] over 0.9) only when using the model incorporating functional information. In Figure 2, we show one such region, near the gene NR0B2. The model identifies the SNP rs6659176 as the most likely candidate to be the causal polymorphism in this region. This SNP has a P-value of 1.5×10^{-6}. However, this SNP falls falls in a coding exon (in fact it is nonsynonymous), leading the model to conclude that this P-value is in fact strong evidence for association. Indeed, larger studies of HDL have confirmed the evidence for association in this region (P-value of 9.7×10^{-16} at rs12748152, which has $r^2 = 0.85$ with rs6659176 [Global Lipids Genetics Consortium et al., 2013]). This region, though not this particular SNP, was also identified in a scan for SNPs influencing multiple lipid phenotypes [Stephens, 2013].

We applied this method to all 18 traits. We were first interested in estimating the fraction of associations for each trait that can be explained by nonsynonymous polymorphisms versus polymorphisms that do not influence protein sequences. For each trait, we fit a model including promoters (SNPs within 5kb of a transcription start site) as well as nonsynonymous polymorphisms. For all traits, nonsynonymous polymorphsims are enriched among those that influence the trait, though this enrichment is not statistically significant for all traits (Figure 3A). We then used these enrichments to estimate the fraction of associations for each trait that are driven by nonsynonymous

bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

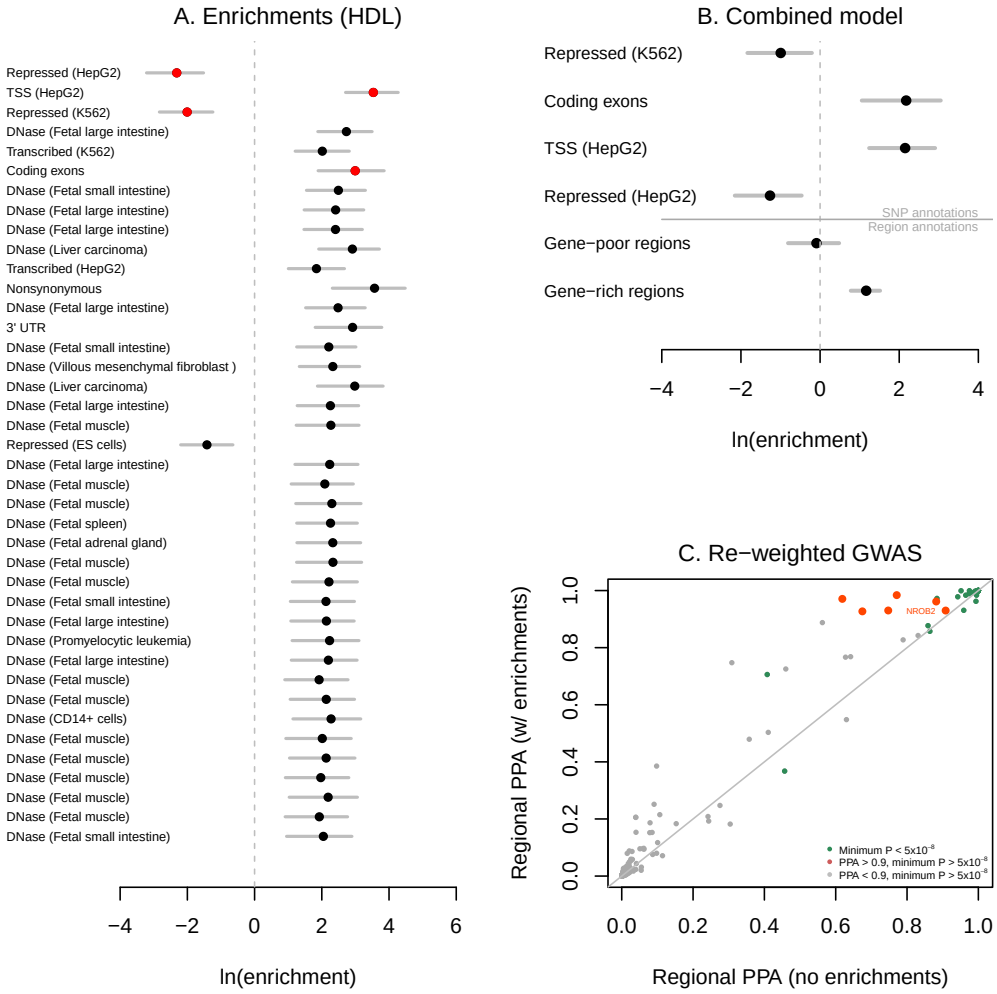

Figure 1. **Application of the model to HDL cholesterol. A. Single-annotation models.** We fit the model to each annotation individually, including a SNP-level effect for SNPs 0-5kb from a TSS, a SNP-level effect for SNPs 5-10kb from a TSS, and region-level effects for regions in the top third and bottom third of gene density. Plotted are the maximum likelihood estimates and 95% confidence intervals of the enrichment parameter for each annotation. Annotations are ordered according to how much they improve the likelihood of the model (at the top are those that improve the likelihood the most), and in red are those included in the joint model. **B. Joint model.** Using the algorithm described in the Methods we built a model combining multiple annotations. Shown are the maximum likelihood estimates and 95% confidence intervals of the enrichment effects of each annotation. Note that though these are the maximum likelihood estimates, model choice was performed using a penalized likelihood. **C. Re-weighted GWAS.** We re-weighted the GWAS using the model with all the annotations in **B** (under the penalized enrichment parameters from Supplementary Table 7). Each point represents a region of the genome, and shown are the posterior probabilities of association (PPA) of the region in the models with and without the annotations.

bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

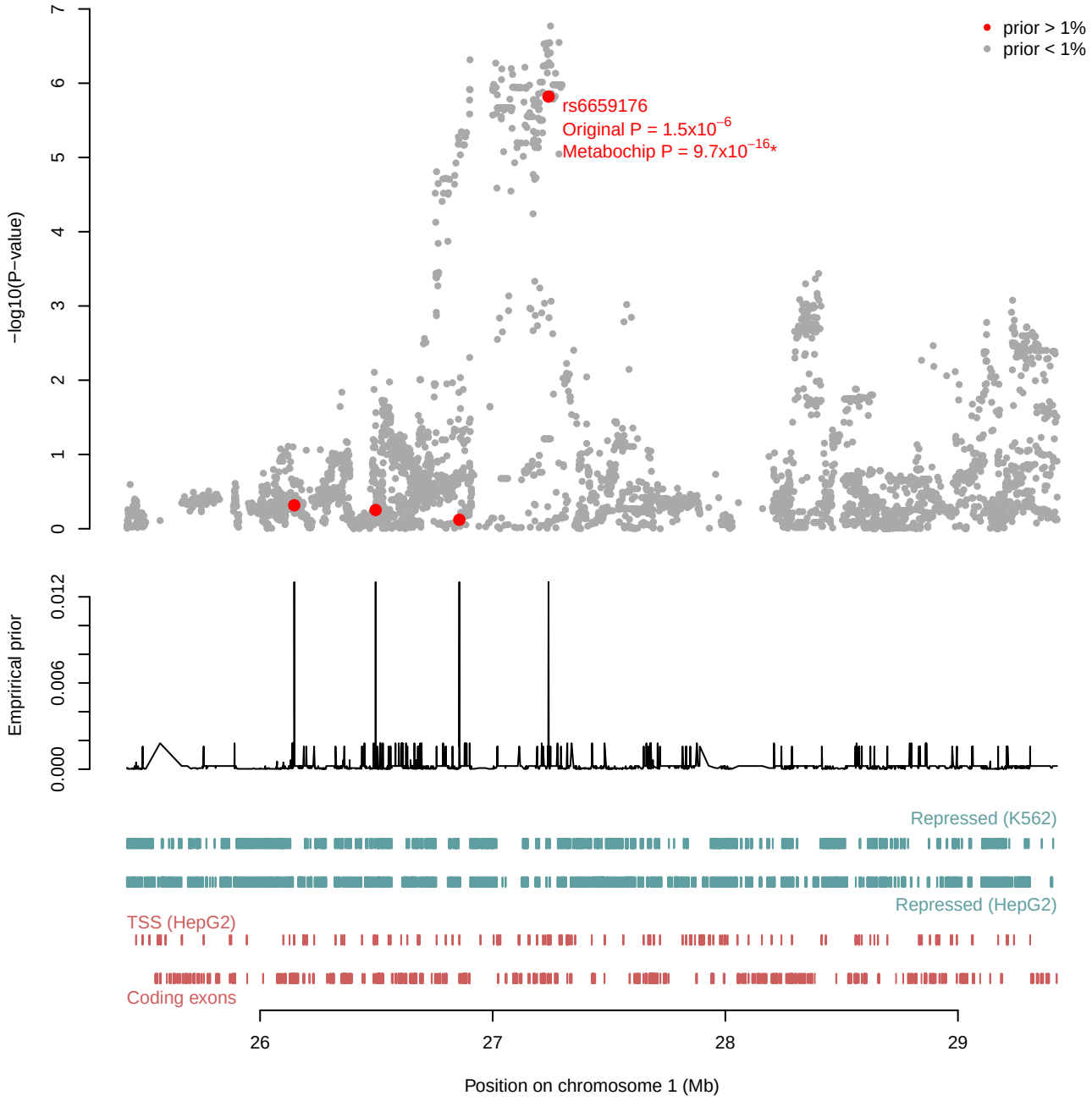

Figure 2. **Regional plot surrounding NR0B2.** In the top panel we plot the P-values for association with HDL levels at each SNP in this region. In the middle panel is the fitted empirical prior probability that each SNP is the causal one using the model with all the annotations in Figure 1B. In the lower panel are the positions of the annotations included in the model. * The reported P-value is for rs12748152, which has $r^2 = 0.85$ with rs6659176.

bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

polymorphisms (Supplementary Material). This fraction varies from around 2% to around 20%, with an average of 10% (Figure 3B). We conclude that the relative importance of changes in protein sequence versus gene expression likely varies across traits.

We then used all 450 genome annotations to build models of enrichment for each trait. As for HDL, we first estimated enrichment levels individually for each annotation (Supplementary Figures 3-11), then generated a combined model for each trait. The parameters of the combined models are shown in Figure 4 and Supplementary Figure 12, and details of the exact annotations are in Supplementary Tables 1-18. In general, the models generated with this method are sparse and biologically interpretable. A few general patterns emerge from this analysis. Apart from the repeated occurrence of annotations related to protein-coding genes, marks of repressed chromatin are often significantly depleted for SNPs influencing traits. For example, SNPs influencing Crohn's disease are depleted from repressed chromatin identified in a lymphoblastoid cell line (Figure 4D; log enrichment of -1.27, 95% CI [-2.12, -0.54]), SNPs influencing height are significantly depleted from repressed chromatin in HeLa cells (Figure 4I; log enrichment of -1.04, 95% CI [-1.66, -0.49]), and SNPs influencing red blood cell volume are significantly depleted from repressed chromatin in an erythroblast-derived cell line (Figure 4F; log enrichment of -2.71, 95% CI [-4.33, -1.65]).

We additionally observed cell type-specific enrichments in enhancer elements and DNase hypersensitive sites for SNPs that influence traits. Most of the observed enrichments are readily interpreted in light of the known biology of the trait. For example, SNPs that influence platelet volume and platelet count are enriched in open chromatin identified in $CD34^+$ cells, known to be on the cell lineage that leads to platelets [Deutsch and Tomer, 2006] (Figure 4A,B; log enrichment of 1.26, 95% CI [0.41, 1.98] for platelet count; log enrichment of 2.09, 95% CI [1.17, 2.95] for platelet volume); and SNPs that influence corpuscular hemoglobin concentration are enriched in open chromatin identified in K562 cells, a cell line derived from a cancer of erythroblasts (Supplementary Figure 12E; log enrichment of 1.85, 95% CI [0.42, 3.08]). For some traits, however, the connection between the trait and the tissues identified is not immediately obvious. For example, SNPs associated with bone density in the lumbar spine are enriched in open chromatin in myoblasts (Figure 4C; log enrichment of 2.93, 95% CI [1.55, 4.14]) and SNPs associated with red blood cell count are enriched in open chromatin in the fetal stomach (Supplementary Figure 12G; log enrichment of 3.68, 95% CI [2.69, 4.79]).

For two traits–Crohn's disease (Figure 4D) and red blood cell count (Supplementary Figure 12G)–we noticed that annotations initially identified as enriched for SNPs influencing the trait ended up in the combined model as being depleted for SNPs influencing the trait. On further examination (Supplementary Material), we found that these effects are due to statistical interactions. For example, when treated alone, SNPs that influence Crohn's disease are enriched in DNase-I hypersensitive sites identified in fetal fibroblasts from the abdomen (log enrichment of 2.20, 95% CI [1.15, 3.02]). However, DNase-I hypersensitive sites identified in fetal fibroblasts from the back show an even stronger enrichment (log enrichment of 2.92, 95% CI [2.07, 3.67]), and sites in common between the two annotation are intermediate (log enrichment of 2.59, 95% CI [1.59, 3.39]). This leads to an interaction where in the joint model the contribution of the DNase-I hy-

bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

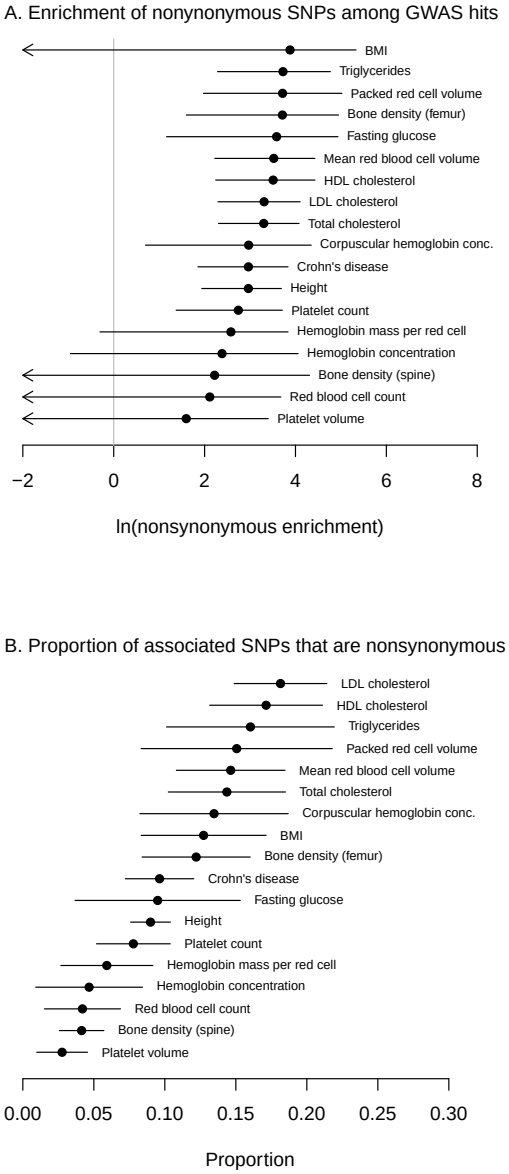

Figure 3. **Estimated role of protein-coding changes in each trait. A. Estimated enrichment of non-synonymous SNPs.** For each trait, we fit a model including an effect of non-synonymous SNPs and an effect of SNPs within 5kb of a TSS. Shown are the estimated enrichments parameters and 95% confidence intervals for the non-synonymous SNPs. **B. Estimated proportion of GWAS hits driven by non-synonymous SNPs.** For each trait, using the model fit in **A.**, we estimated the proportion of GWAS signals driven by non-synonymous SNPs. Shown is this estimate and its standard error (Supplementary Material).

bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

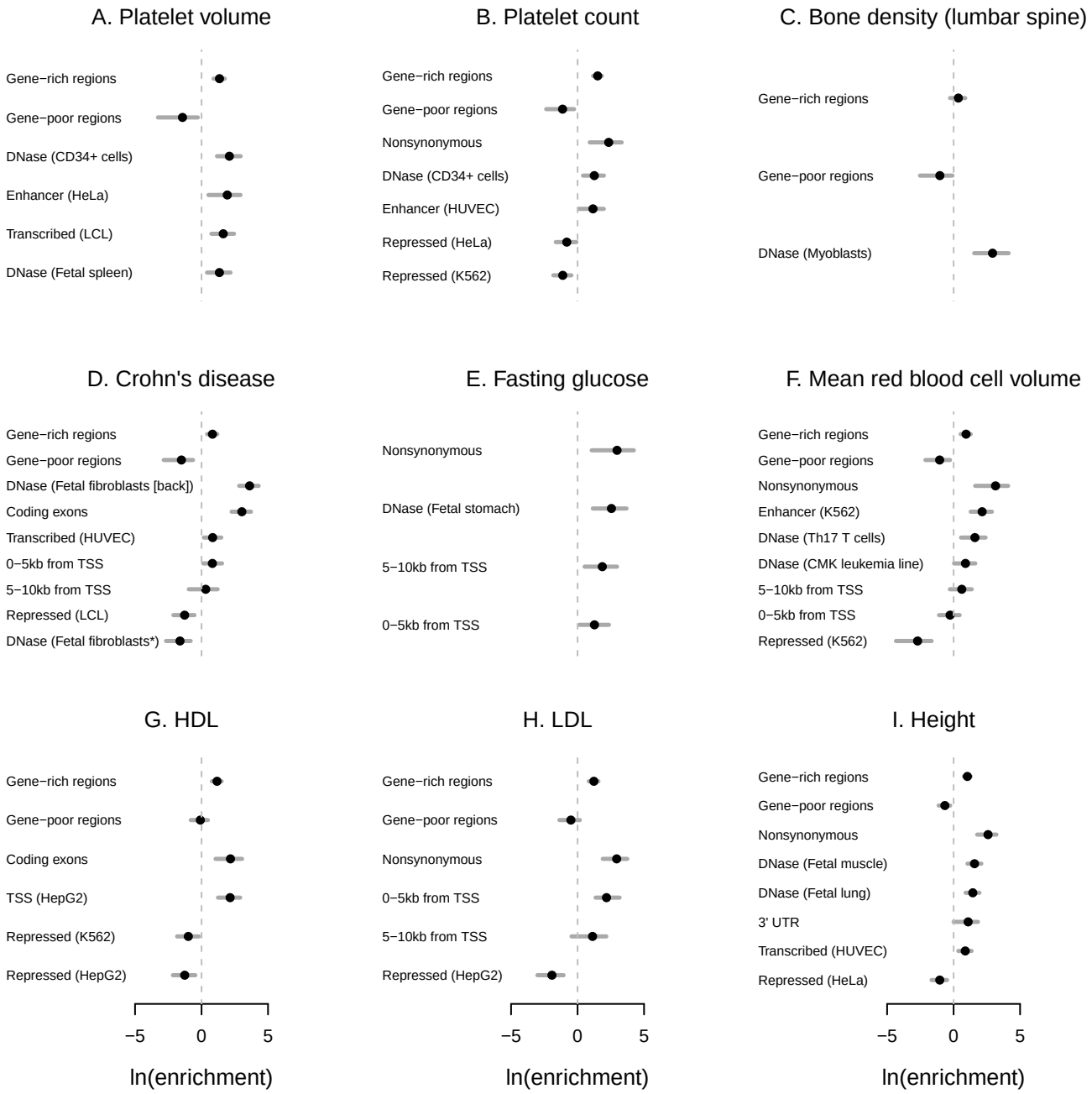

Figure 4. **Combined models for nine traits.** For each trait, we built a combined model of annotations using the algorithm presented in the Methods. Shown are the maximum likelihood estimates and 95% confidence intervals for all annotations included in each model. Note that though these are the maximum likelihood estimates, model choice was done using a penalized likelihood (Methods). For the other nine traits, see Supplementary Figure 12. *This annotation of DNase-I hypersensitive sites in fetal fibroblasts from the abdomen has a positive effect when treated alone; see the text for discussion.

persensitive sites identified in fetal abdominal fibroblasts is negative. Though this is a statistical explanation for this observation, the biological explanation is not immediately clear. It seems likely that DNase-I hypersensitive sites are a heterogeneous set of different classes of elements, and that different experiments are more sensitive, for either technical or biological reasons, to subsets of these elements.

Finally, we explored the potential of this model to identify new loci (as in Figure 2). In order to do this, one needs a threshold for "significance" in this model, ideally with similar properties as the standard P-value threshold of 5×10^{-8}. To calibrate the method, we used the fact that we initially applied our method to a study of four lipid traits that identified about 100 loci in sample sizes of around 90,000 individuals [Teslovich et al., 2010]. Since then, larger studies have raised the number of loci associated with lipid traits to 157 [Global Lipids Genetics Consortium et al., 2013]. If we treat a locus with a P-value of 5×10^{-8} in the larger study as a "true positive" and a locus that does not reach this threshold in the larger study as a "true negative", we can calibrate a threshold for posterior probability of association using the replication data (Supplementary Material). We found that a threshold of a regional PPA of 0.9 performed similarly to a stringent P-value threshold (Supplementary Figure 13). Combining the loci from both the standard P-value approach and our approach resulted in an approximately 5% increase in the number of identified loci while still maintaining a false positive rate close to zero (Supplementary Figure 13, Supplementary Table 19). This is only a modest gain in power; that said, by applying this method to all 18 traits we identified 49 loci that did not reach a standard statistical significance threshold of 5×10^{-8} but have a PPA over 0.9 (Supplementary Table 20). Based on the above results for lipids, the level of evidence that these loci are true positives is approximately the same as those that have $P = 5 \times 10^{-8}$ in a standard GWAS. Indeed, the majority of these loci have since been identified in larger cohorts than those used in this paper (Supplementary Table 20).

3 Discussion

In this paper, we have developed a statistical model for identifying genomic annotations that are most relevant to the biology of a given phenotype. We have shown that this model is able to scan through hundreds of genomic annotations to identify a sparse set of biologically-interpretable annotations without prior knowledge of the biology of the phenotype.

Linking GWAS to biology

Perhaps the most striking observation is that chromatin annotated as repressed in a given cell type is often depleted for SNPs that influence traits. Since approximately 60-70% of the genome falls in this annotation in any given cell type (Supplementary Material), this information could dramatically limit the number of SNPs considered when fine-mapping loci identified in GWAS. Additionally, we identified several non-obvious connections between tissue and phenotypes. For example, SNPs that influence bone density in the spine and the femur are both enriched in DNAse-I hypersensitive sites

identified in muscle-related tissues (myoblasts and the fetal heart; Figure 4D and Supplementary Figure 12B). Though it is well-known that muscle function correlates with bone mineral density, twin studies have suggested that this correlation is not due to genetic factors [Arden and Spector, 1997]. However, a modest shared genetic component between muscle function and bone mineral density cannot be excluded by these studies, and our results suggest that this possibility is worth re-examining. Additionally, we have shown that SNPs that influence Crohn's disease risk are enriched in DNase-I hypersensitive sites identified in fibroblasts (Figure 4D). Though Crohn's disease is an autoimmune disease, a major complication of the disease is intestinal stricture mediated through abnormal fibroblast activity [Burke et al., 2007]. Our results suggest that the intestinal response to inflammation, and not just the autoimmune activity causing inflammation, may be modulated by genetic variation.

Modeling assumptions

We have made several modeling assumptions that merit discussion. First, by splitting the genome into blocks based on numbers of SNPs, we are making the implicit assumption that the probability that a genomic region contains a SNP associated with a given phenotype depends on the SNP density rather than the physical size–that is, a short genomic region with a large number of SNPs is *a priori* as likely to have an association as a long genomic region with few SNPs. We have also made a more restrictive assumption that there can be only a single causal SNP in a given genomic region. This assumption is a natural starting point, but as GWAS sample sizes increase even more it will begin to be untenable. Advances in methods for joint analysis of multiple SNPs (e.g. Yang et al. [2012]) may provide a way forward in this situation. Finally, we note that the model is limited by the types of genomic annotations that are available, and the best annotations identified in the model may be "proxies" for the truly-relevant annotations. For example, SNPs associated with height are enriched in DNase-I hypersensitive sites identified in the fetal lung (Figure 4I); taken literally, this would suggest that some SNPs influence height though lung development. An alternative possibility, however, is that patterns of open chromatin in the lung (which is of course a heterogeneous tissue) are useful proxies for patterns of open chromatin in a cell type that has not been profiled; this hypothetical cell type could in principle be present in any tissue.

Prospects for fine-mapping GWAS loci using functional genomic data

We have primarily focused on using our model to identify annotations relevant to a trait of interest, though we have also explored using this information to identify novel loci. A third natural application, which we have not explored, is the possibility for fine-mapping GWAS loci using functional genomic information [Maller et al., 2012]. Indeed, the posterior probability that each SNP in a given genomic region is the causal one is explicitly included in our model. However, in current applications around 20% of common SNPs are neither genotyped nor successfully imputed; this is a major limitation to fine-mapping that cannot be overcome with statistical means. As GWAS move to even denser genotyping or sequencing, we expect that re-visiting this issue will be fruitful.

bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

Methods

In this section we detail the specifics of the hierarchical model; a description of the data used is in the Supplementary Material. The model we propose is most closely related to that developed by Veyrieras et al. [2008] in the context of eQTL mapping. Conceptually, we split the genome into independent blocks, such that the blocks are larger than the extent of LD in the population. We then allow each block to contain either a single polymorphism that causally influences the trait or none. We model the prior probability that any given block contains an association and the conditional prior probability that any given SNP in the block is the causal one. The key is that we allow these probabilities to vary according to functional annotations–for example, gene-rich regions might be more likely to contain associations, and if there is an association, the causal polymorphisms may be more likely to fall in a transcription factor binding site. We then estimate these priors using an empirical Bayes approach. Software implementing this model is available at `http://gwas.googlecode.com`.

Computing the Bayes factor

The basic building block of the model is a linear regression model. Consider a single SNP genotyped in N phenotyped individuals. Assume each individual has an associated measurement of a quantitative trait (we describe a slight modification for case-control studies later), and let $\vec{y}$ be the vector of phenotypes. Let $\vec{g}$ be the vector of genotypes (coded 0, 1, or 2 according to counts of an arbitrarily-defined allele). We use a standard additive linear model:

$$E[y_i] = \alpha + \beta g_i. \tag{1}$$

We would like to compare two models: one where $\beta = 0$ and one where $\beta \neq 0$. A natural way to compare these two models is the Bayes factor:

$$BF = \frac{\int P(\vec{y}|\vec{g}, H_1)}{\int P(\vec{y}|\vec{g}, H_0)}, \tag{2}$$

where H_1 and H_0 represent the parameters of the the alternative and null models, respectively, and which are integrated out.

To compute Equation 2, we use the approximate Bayes factor from Wakefield [2008]. This Bayes factor has the practically important property that is can be calculated from a summary of the linear regression, without access to the underlying genotype vector $\vec{g}$. For completeness, we re-iterate here the underlying model. If $\hat{\beta}$ is the maximum likelihood estimator of β and $\sqrt{V}$ is the standard error of $\hat{\beta}$, Wakefield [2008] suggests a model in which:

$$\hat{\beta} \sim N(\beta, V). \tag{3}$$

Wakefield [2008] places a normal prior on β, such that $\beta \sim N(0, W)$. Under this model Equation 2 becomes:

$$BF = \frac{\sqrt{1-r}}{\exp[-\frac{Z^2}{2}r]}, \tag{4}$$

where $r = \frac{W}{V+W}$ and $Z = \frac{\hat{\beta}}{\sqrt{V}}$ (a standard Z-score). Thus, from a Z-score, an estimate of V, and the prior variance W, we can obtain a Bayes factor measuring the statistical support for a model in which a SNP is associated with a trait as compared to a model in which a SNP is not associated with a trait. Note, however, that because of linkage disequilibrium in the genome, any true causal association will lead to multiple true statistical associations. In all applications, we set $W = 0.1$ as the prior, such that the majority of the weight of the prior is on small effect sizes (results are robust to some variation in this prior, see Supplementary Material and Supplementary Figure 14).

Hierarchical model

Now consider a set of M SNPs, each of which has been genotyped in N individuals in a GWAS. Our goal is to build a model to identify the shared characteristics of SNPs that causally influence a trait. Because of LD, there will be many associations in the genome that are not causal; however, these will all be restricted to a block around the truly causal site. We thus split the genome into contiguous blocks of size K SNPs (in all of our applications, we set $K = 5,000$, though doubling this block size had little effect on the results; see Supplementary Materials and Supplementary Figure 14), such that there are M/K blocks. We choose the block size to be much larger than the extent of linkage disequilibrium in the population. Let Π_k be the prior probability that block k contains a causal SNP associated with the trait. The probability of the data (the set of observed phenotypes) is then:

$$P(\vec{y}) = \sum_{k=1}^{M/K} (1-\Pi_k)P_k^0 + \Pi_k P_k^1, \tag{5}$$

where P_k^0 is the probability of the data in block k under the model where there are no SNPs associated with the trait in the block, and P_k^1 is the probability of the data in block k under the model where there is one SNP associated with the trait in the block. Further,

$$P_k^1 = \sum_{i \in S_k} \pi_{ik} P_{ik}^1, \tag{6}$$

where S_k is the set of SNPs in block k, π_{ik} is the prior probability that SNP i is the causal SNP in the region conditional on there being an association in block k, and P_{ik}^1 is the probability of the data under the model where this SNP is associated with the trait. Note that this is not a multiple regression model where we jointly model the effects of multiple SNPs on a trait (as in, for example, Carbonetto and Stephens [2013]).

We can now allow the prior probabilities–both Π_k (the prior on the block of SNPs containing an association) and π_{ik} (the prior probability that SNP i is the causal SNP assuming there is a single association in block k)–to depend on external information. We would also like to avoid subjective

variation in Π_k and π_i, but instead learn from the data itself to which genomic annotations are most important. Specifically, we model the regional prior probability as:

$$\ln\left(\frac{\Pi_k}{1-\Pi_k}\right) = \kappa + \sum_{l=1}^{L_1} \gamma_l I_{kl}, \tag{7}$$

where L_1 is the number of region-level annotations in the model, γ_l is the effect associated with annotation l and I_{kl} takes the value 1 if region k is annotated with annotation l and 0 otherwise. For example, in practice we will estimate a γ parameter for regions of high or low gene density. We then model the SNP prior probability as:

$$\pi_{ik} = \frac{e^{x_i}}{\sum\limits_{j \in S_k} e^{x_j}}, \tag{8}$$

where

$$x_i = \sum_{l=1}^{L_2} \lambda_l I_{il}, \tag{9}$$

where L_2 is the number of SNP-level annotations in the model, λ_l is the effect of SNP annotation l and I_{il} takes the value 1 if SNP i falls in annotation l and 0 otherwise. For example, in practice we will estimate a λ parameter for nonsynonymous SNPs.

Fitting the model

Combining terms above, we see that the likelihood of the data can be written down as:

$$L(\vec{y}|\theta) = \prod_{k=0}^{M/K} (1-\Pi_k)P_k^0 + \Pi_k \sum_{i=0}^{K} \pi_{ij} P_{ik}^1 \tag{10}$$

$$= \prod_{k=0}^{M/K} P_k^0[(1-\Pi_k) + \Pi_k \sum_{i=0}^{K} \pi_{ik} BF_i], \tag{11}$$

where θ is contains all the parameters of the model, most notably the set of annotation parameters. We maximize this function using the Nelder-Mead algorithm implemented in the GNU Scientific Library.

Shrinkage estimators of the annotation parameters

While maximizing Equation 10 gives the maximum likelihood estimates of all parameters, one concern is that there may be some level of overfitting. When comparing models, we instead shrink these parameters towards zero. Specifically, we define a penalized likelihood function:

$$L^*(\vec{y}|\theta) = L(\vec{y}|\theta) - p\left(\sum_{l=1}^{L_1} \gamma_l^2 + \sum_{l=1}^{L_2} \lambda_l^2\right). \tag{12}$$

The penalty p on the sum of the squared annotation parameters is the one used in ridge regression [Hastie et al., 2001]. In ridge regression, parameter estimates under this penalty are equivalent to estimating the posterior mean of the parameter if the prior distribution of the parameter is Gaussian [Hastie et al., 2001]; changing the tuning parameter p is equivalent to changing the prior. We suspect that the interpretation in this model is similar. Since this penalized likelihood can not be used for formal statistical tests, we tune the p parameter by cross-validation. An alternative approach here would be to explicitly put a prior on the enrichment parameters, but in the absence of a conjugate prior this would likely add substantially to the computational burden for little practical benefit.

Cross-validation

To compare models and tune the penalty p in the penalized likelihood above, we used a 10-fold cross-validation approach. We split the chromosomal segments into 10 folds. Let θ^p_{-f} be the parameters of the model estimated while holding out the data from fold f and under penalty p, and $L^*_f(\theta^p_{-f})$ be the penalized likelihood of the data in fold f under the model optimized without using fold f. Then:

$$L^{'}(\theta^p) = \frac{1}{10}\sum_{i=1}^{10} L^*_i(\theta^p_{-i}). \tag{13}$$

Note that the size of the folds used in this cross-validation means that each fold excludes more than an entire chromosome. This means that no individual chromosome can have undue influence on the parameters included in the model.

Model choice

Consider a single phenotype and a set of L functional annotations of SNPs (in our case L is in the hundreds). Including all L SNP annotations in the model is neither biologically interesting nor computationally feasible. We thus set out to choose a relatively sparse model that fits the data. We start with forward selection: for each of the L annotations, we fit a model including region-level parameters for regions in the top and bottom third of gene density, a SNP-level parameter for SNPs from 0-5 kb from a TSS, a SNP-level parameter for SNPs from 5-10kb from a TSS, and a SNP-level parameter for the annotation in question. We then identify the set of annotations that significantly improve the model fit (as judged by a likelihood ratio test using the likelihood from Equation 10). We then:

1. Add the annotation that most significantly improves the likelihood to the model.

2. For each annotation identified as having a significant marginal effect, test a model including the annotation and those that have already been added.

3. If any annotation remains significant, go back to step 1.

At this point, there are generally a small number of annotations in the model, but the model may be over-fit. We then switch to using the 10-fold cross-validation likelihood in Equation 13. We first tune the penalty parameter p by finding the value of p that maximizes the cross-validation likelihood. We then:

1. Drop each annotation from the model in turn, and evaluate the cross-validation likelihood. When dropping annotations, we additionally try dropping the region-level annotations on gene density and the SNP-level annotations on distance to the nearest TSS.

2. If a simpler model has a higher cross-validation likelihood than the full model, drop the annotation from the model and return to step 1.

3. Report the model that has the highest cross-validation likelihood.

Approximating V_i

In order to compute the Bayes factor in Equation 4, we need an estimate of V_i, the standard error of the estimated effect size of SNP i. In principle, this is trivial output from standard regression software; however, it is rarely reported. Instead, let f_i be the minor allele frequency of SNP i computed from an external sample of the same ancestry as the population in which the association study was done (we use data from the 1000 Genomes Project [Abecasis et al., 2010]). Let N_i be the number of individuals in the association study at SNP i (this can vary across SNPs due to missing data). Then:

$$V_i \approx \frac{1}{N_i f_i (1 - f_i)}. \tag{14}$$

Note that this variance is independent of the actual scale of the measurements; this is appropriate because the Z-scores are independent of the scale of the measurements as well.

Case-control studies

For all of the above, we have considered studies of quantitative traits. For a case-control study, we assume that we have summary statistics from logistic regression instead of linear regression. All aspects of the model are identical, with the exception of the approximation of V_i. Define N_{case} and $N_{control}$ as the numbers of cases and controls, respectively. Now, [Wakefield, 2008]:

$$V_i \approx \frac{N_{case} + N_{control}}{N_{case} N_{control} [2 f_i (1 - f_i) + 4 f_i^2 - (2 f_i (1 - f_i) + 2 f_i)^2]}. \tag{15}$$

The variance here is on a log-odds scale.

bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

Posterior probabilities of association

Once the model has been fit, we have empirical estimates of the prior probability that region k contains an association, $\hat{\Pi}_k$ and the prior probability that SNP i is the causal one, $\hat{\pi}_{ik}$ (conditional on there being an association). We define a Bayes factor summarizing the evidence for association in the *region* (see, for example Maller et al. [2012]):

$$BF_k^R = \sum_{i \in S_k} \hat{\pi}_{ik} BF_i, \tag{16}$$

where S_k is the set of SNPs in region k and BF_i is the Bayes factor for SNP i (Equation 4). The posterior probability that region k contains an association is then:

$$PPA_k^R = \frac{\hat{\Pi}_k BF_k^R/(1-\hat{\Pi}_k)}{1+\hat{\Pi}_k BF_k^R/(1-\hat{\Pi}_k)} \tag{17}$$

We can also define the posterior probability that any given SNP i in region k is the causal one under our model:

$$PPA_{ik} = \frac{\hat{\pi}_{ik} BF_i}{\sum_{j \in S_k} \hat{\pi}_{jk} BF_j}. \tag{18}$$

This is similar to the calculation in Maller et al. [2012], except that we allow the prior probability π_{ik} to vary across SNPs.

bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

Acknowledgements. We thank David Reich, Nick Patterson, Alkes Price, and Po-Ru Loh for helpful discussions and suggestions. We thank Nicole Soranzo for providing access to the platelet studies, and Luke Jostins for assistance in obtaining the Crohn's disease data. This work was supported by NIH postdoctoral fellowship GM103098 to JKP.

References

Abecasis, G., Altshuler, D., Auton, A., Brooks, L., Durbin, R., Gibbs, R. A., Hurles, M. E., McVean, G. A., Bentley, D., Chakravarti, A., *et al.*, 2010. A map of human genome variation from population-scale sequencing. *Nature*, **467**(7319):1061–1073.

Arden, N. K. and Spector, T. D., 1997. Genetic influences on muscle strength, lean body mass, and bone mineral density: a twin study. *J Bone Miner Res*, **12**(12):2076–81.

Burke, J. P., Mulsow, J. J., O'Keane, C., Docherty, N. G., Watson, R. W. G., and O'Connell, P. R., 2007. Fibrogenesis in Crohn's disease. *Am J Gastroenterol*, **102**(2):439–48.

Carbonetto, P. and Stephens, M., 2013. Integrated Enrichment Analysis of Variants and Pathways in Genome-Wide Association Studies Indicates Central Role for IL-2 Signaling Genes in Type 1 Diabetes, and Cytokine Signaling Genes in Crohn's Disease. *PLoS Genet*, **9**(10):e1003770.

Chen, G. K. and Witte, J. S., 2007. Enriching the analysis of genomewide association studies with hierarchical modeling. *Am J Hum Genet*, **81**(2):397–404.

Cowper-Sal lari, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., Moore, J. H., and Lupien, M., 2012. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat Genet*, **44**(11):1191–8.

Deutsch, V. R. and Tomer, A., 2006. Megakaryocyte development and platelet production. *Br J Haematol*, **134**(5):453–66.

ENCODE Project Consortium, Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., and Snyder, M., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414):57–74.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., *et al.*, 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**(7345):43–9.

Estrada, K., Styrkarsdottir, U., Evangelou, E., Hsu, Y.-H., Duncan, E. L., Ntzani, E. E., Oei, L., Albagha, O. M., Amin, N., Kemp, J. P., *et al.*, 2012. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nature genetics*, **44**(5):491–501.

bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

Gaffney, D. J., Veyrieras, J.-B., Degner, J. F., Pique-Regi, R., Pai, A. A., Crawford, G. E., Stephens, M., Gilad, Y., and Pritchard, J. K., 2012. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biol*, **13**(1):R7.

Gerasimova, A., Chavez, L., Li, B., Seumois, G., Greenbaum, J., Rao, A., Vijayanand, P., and Peters, B., 2013. Predicting cell types and genetic variations contributing to disease by combining GWAS and epigenetic data. *PLoS One*, **8**(1):e54359.

Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A. H., Labrune, Y., *et al.*, 2011. New gene functions in megakaryopoiesis and platelet formation. *Nature*, **480**(7376):201–208.

Global Lipids Genetics Consortium, Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., *et al.*, 2013. Discovery and refinement of loci associated with lipid levels. *Nat Genet*, .

Hastie, T., Tibshirani, R., and Friedman, J. J. H., 2001. *The elements of statistical learning*, volume 1. Springer New York.

Heron, E. A., O'dushlaine, C., Segurado, R., Gallagher, L., and Gill, M., 2011. Exploration of empirical Bayes hierarchical modeling for the analysis of genome-wide association study data. *Biostatistics*, **12**(3):445–461.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, **106**(23):9362–7.

Hnisz, D., Abraham, B. J., Lee, T. I., Lau, A., Saint-André, V., Sigova, A. A., Hoke, H. A., and Young, R. A., 2013. Super-Enhancers in the Control of Cell Identity and Disease. *Cell*, .

Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., Birney, E., *et al.*, 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic acids research*, **41**(2):827–841.

Hu, X., Kim, H., Stahl, E., Plenge, R., Daly, M., and Raychaudhuri, S., 2011. Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. *The American Journal of Human Genetics*, **89**(4):496–506.

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Schumm, L. P., Sharma, Y., Anderson, C. A., *et al.*, 2012. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, **491**(7422):119–124.

Karczewski, K. J., Dudley, J. T., Kukurba, K. R., Chen, R., Butte, A. J., Montgomery, S. B., and Snyder, M., 2013. Systematic functional regulatory assessment of disease-associated variants. *Proceedings of the National Academy of Sciences*, **110**(23):9607–9612.

bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

Lango-Allen, H., Estrada, K., Lettre, G., Berndt, S. I., Weedon, M. N., Rivadeneira, F., Willer, C. J., Jackson, A. U., Vedantam, S., Raychaudhuri, S., *et al.*, 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, **467**(7317):832–838.

Lappalainen, T., Sammeth, M., Friedländer, M. R., 't Hoen, P. A. C., Monlong, J., Rivas, M. A., Gonzàlez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P. G., *et al.*, 2013. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, **501**(7468):506–11.

Lee, S.-I., Dudley, A. M., Drubin, D., Silver, P. A., Krogan, N. J., Pe'er, D., and Koller, D., 2009. Learning a prior on regulatory potential from eQTL data. *PLoS Genet*, **5**(1):e1000358.

Lewinger, J. P., Conti, D. V., Baurley, J. W., Triche, T. J., and Thomas, D. C., 2007. Hierarchical Bayes prioritization of marker associations from a genome-wide association scan for further investigation. *Genet Epidemiol*, **31**(8):871–82.

Lui, J. C., Nilsson, O., Chan, Y., Palmer, C. D., Andrade, A. C., Hirschhorn, J. N., and Baron, J., 2012. Synthesizing genome-wide association studies and expression microarray reveals novel genes that act in the human growth plate to modulate height. *Hum Mol Genet*, **21**(23):5193–201.

Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M., Auton, A., Myers, S., Morris, A., *et al.*, 2012. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nature genetics*, **44**(12):1294–1301.

Manning, A. K., Hivert, M.-F., Scott, R. A., Grimsby, J. L., Bouatia-Naji, N., Chen, H., Rybin, D., Liu, C.-T., Bielak, L. F., Prokopenko, I., *et al.*, 2012. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nature genetics*, **44**(6):659–669.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., *et al.*, 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, **337**(6099):1190–5.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J., 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet*, **6**(4):e1000888.

Parker, S. C. J., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., van Bueren, K. L., Chines, P. S., Narisu, N., NISC Comparative Sequencing Program, *et al.*, 2013. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A*, .

Pasaniuc, B., Zaitlen, N., Shi, H., Bhatia, G., Gusev, A., Pickrell, J., Hirschhorn, J., Strachan, D. P., Patterson, N., and Price, A. L., *et al.*, 2013. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *arXiv preprint arXiv:1309.3258*, .

bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

Paul, D. S., Albers, C. A., Rendon, A., Voss, K., Stephens, J., HaemGen Consortium, van der Harst, P., Chambers, J. C., Soranzo, N., Ouwehand, W. H., *et al.*, 2013. Maps of open chromatin highlight cell type-restricted patterns of regulatory sequence variation at hematological trait loci. *Genome Res*, **23**(7):1130–41.

Paul, D. S., Nisbet, J. P., Yang, T.-P., Meacham, S., Rendon, A., Hautaviita, K., Tallila, J., White, J., Tijssen, M. R., Sivapalaratnam, S., *et al.*, 2011. Maps of open chromatin guide the functional follow-up of genome-wide association signals: application to hematological traits. *PLoS Genet*, **7**(6):e1002139.

Schork, A. J., Thompson, W. K., Pham, P., Torkamani, A., Roddey, J. C., Sullivan, P. F., Kelsoe, J. R., O'Donovan, M. C., Furberg, H., Tobacco and Genetics Consortium, *et al.*, 2013. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet*, **9**(4):e1003449.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Luan, J., Mägi, R., *et al.*, 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature genetics*, **42**(11):937–948.

Stephens, M., 2013. A unified framework for association analysis with multiple related phenotypes. *PLoS One*, **8**(7):e65245.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., *et al.*, 2010. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, **466**(7307):707–713.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., *et al.*, 2012. The accessible chromatin landscape of the human genome. *Nature*, **489**(7414):75–82.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., and Raychaudhuri, S., 2013. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet*, **45**(2):124–30.

van der Harst, P., Zhang, W., Leach, I. M., Rendon, A., Verweij, N., Sehmi, J., Paul, D. S., Elling, U., Allayee, H., Li, X., *et al.*, 2012. Seventy-five genetic loci influencing the human red blood cell. *Nature*, **492**(7429):369–375.

Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K., 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet*, **4**(10):e1000214.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J., 2012. Five years of GWAS discovery. *Am J Hum Genet*, **90**(1):7–24.

bioRxiv preprint doi: https://doi.org/10.1101/000752; this version posted November 22, 2013.

Wakefield, J., 2008. Bayes factors for genome-wide association studies: comparison with P-values. *Genet Epidemiol*, **33**(1):79–86.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Madden, P. A. F., Heath, A. C., Martin, N. G., Montgomery, G. W., *et al.*, 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet*, **44**(4):369–75, S1–3.