TISSUE-SPECIFIC FUNCTIONAL EFFECT PREDICTION OF GENETIC VARIATION AND APPLICATIONS TO COMPLEX TRAIT GENETICS

DANIEL BACKENROTH[1], KRZYSZTOF KIRYLUK[2], BIN XU[3], LYNN PETHUKOVA[4,5], BADRI VARDARAJAN[6], EKTA KHURANA[7], ANGELA CHRISTIANO[5,8], JOSEPH D. BUXBAUM[9,10], IULIANA IONITA-LAZA[1]

[1] Department of Biostatistics, Columbia University, New York, NY 10032
[2] Department of Medicine, Columbia University, New York, NY 10032
[3] Department of Psychiatry, Columbia University, New York, NY 10032
[4] Department of Epidemiology, Columbia University, New York, NY 10032
[5] Department of Dermatology, Columbia University, New York, NY 10032
[6] Department of Neurology, Columbia University, New York, NY 10032
[7] Department of Physiology and Biophysics, Weill Medical College, Cornell University, New York, NY 10021
[8] Department of Genetics and Development, Columbia University, New York, NY 10032
[9] Departments of Psychiatry, Neuroscience, and Genetics and Genomic Sciences,
Icahn School of Medicine at Mount SInai, New York, NY 10029
[10] Friedman Brain Institute and Mindich Child Health and Development Institute,
Icahn School of Medicine at Mount Sinai, New York, NY 10029

Correspondence to Iuliana Ionita-Laza: ii2135@columbia.edu

Date: August 11, 2016.

Abstract

Over the past few years, large scale genomics projects such as the ENCODE and Roadmap Epigenomics have produced genome-wide data on a large number of biochemical assays for a diverse set of human cell types and tissues. Such data play an important role in predicting the functional effects of noncoding genetic variation. We discuss here unsupervised approaches to integrate these diverse annotations for specific tissues and cell types into a single predictor of tissue-specific functional importance. We provide a global view of the sharing of functional variants across large number of tissues and cell types, and demonstrate that functional variants in promoters are more likely to be shared across many tissues compared with enhancers. A multidimensional scaling analysis based on functional scores in multiple tissues reveals clear patterns of similarity between certain tissue types, including similarity between primary tissues derived from the same embryonic tissue of origin. Using eQTL data from the Genotype-Tissue Expression (GTEx) project we show that eQTLs in specific GTEx tissues tend to be most enriched among the functional variants in relevant tissues in Roadmap. Furthermore, we show how these integrated functional scores can be used to derive the most likely tissue-/cell-type for a complex trait using summary statistics from genome-wide association studies (GWAS), and derive a tissue-based correlation map of various complex traits. Finally, we show how the tissue-specific functional scores in conjunction with GWAS summary statistics can shed light on genes and biological processes implicated in a complex trait. Functional scores are available for every possible position in the hg19 human reference genome for 127 tissues and cell types assayed in Roadmap Epigenomics.

1. Introduction

Understanding the functional consequences of noncoding genetic variation is one of the most important problems in human genetics. Comparative genomics studies suggest that most of the mammalian conserved and recently adapted regions consist of noncoding elements [1, 2, 3]. Furthermore, most of the loci identified in genome-wide association studies fall in noncoding regions [4] and are likely to be involved in gene regulation in a tissue- and cell-type specific manner. Noncoding variants are also known to play an important role in cancer. Somatic variants in noncoding regions can act as drivers of tumor progression and germline noncoding variants can act as risk alleles [5]. Thus, improved understanding of tissue-specific functional effects of noncoding variants will have implications for multiple diseases and traits.

Prediction of the functional effects of genetic variation is difficult for several reasons. To begin with, there is no single definition of function. As discussed in [6] there are several possible definitions, depending on whether one considers genetic, evolutionary conservation or biochemical perspectives. These different approaches each have limitations and vary substantially with respect to the specific regions of the human genome that they predict to be functional. In particular, the genetic approach, based on evaluating the phenotypic consequence of a sequence alteration, tends to have low throughput and may miss elements that lead to phenotypic effects manifest only in rare cells or specific environmental contexts. The evolutionary approach relies on accurate multispecies alignment which makes it challenging to identify certain functional elements, such as distal regulatory elements, although recently several approaches have been developed for primate- or even human-specific elements [7]. Another limitation of the evolutionary approach is that it is not sensitive to tissue- and cell-type. Finally, the biochemical approach, although helpful in identifying potentially regulatory elements in specific contexts, does not provide definitive proof of function since the observed biochemical signatures can occur stochastically. Another main difficulty for computational methods is the lack of gold-standard labeled data, especially in a tissue- and cell-type specific manner, that could be used as training examples in supervised learning [8]. We have previously shown that unsupervised approaches can perform well in this setting [9]. The approach in [9], like other functional prediction tools such as CADD [10] and FunSeq [2, 11], are not specific to particular tissues or cell types.

Here we are interested in predicting functional effects of genetic variants in a particular tissue- or cell-type. The ENCODE Project [3] and the NIH Roadmap Epigenomics Program [12] have profiled various epigenetic features in more than 100 different tissues and cell types (Supplemental Tables S1 and S2). We use the Roadmap tissue-specific resource to predict the functional effects of genetic variants in a tissue-specific manner. Most of the epigenetic features profiled by Roadmap are histone modifications, and we focus on these marks here. Histone modifications are chemical modifications of the DNA-binding histone proteins that influence transcription as well as other DNA processes. Particular histone modifications have characteristic genomic distributions [13]. For example, trimethylation of histone H3 at lysine 4 (H3K4me3) is associated with promoter regions, monomethylation at H3K4 (H3K4me1) is associated with enhancer regions, and acetylation of H3K27 (H3K27ac) and of H3K9 (H3K9ac) are associated with increased activation of enhancer and promoter regions, respectively [12]. There are dozens of chromatin marks assayed in large numbers of different tissues and cell types, and studying them individually is inefficient. Trynka et al. [14] have shown that the most phenotypically cell-type specific chromatin marks are H3K4me3, along with H3K4me1, and H3K9ac; in particular they report H3K4me3 as the mark most phenotypically cell-type specific. We focus here on four of these most informative marks, namely H3K4me3, H3K4me1, H3K9ac, and H3K27ac. We introduce an integrated functional score that combines these different epigenetic chromatin marks in specific tissues and cell types. We present results from a flexible nonparametric mixture model, assuming that the variants can be partitioned into two groups, functional and nonfunctional in a particular tissue, and modeling the multivariate annotation data using two-component mixture models, with components corresponding to the two groups.

Here we (1) provide tissue- and cell-type specific functional scores for every possible position in the hg19 human genome for 127 tissues and cell types in Roadmap, (2) provide a global view of the sharing of functional variants across large number of tissues and cell types, and demonstrate that functional variants in promoters are more likely to be shared across many tissues as compared with enhancers, (3) show that eQTLs identified in a specific GTEx tissue tend to be most enriched among the functional variants in a relevant Roadmap tissue, (4) use these tissue- and cell-type specific scores in conjunction with summary statistics from 21 genome-wide association studies (GWAS) to identify the most likely causal tissue-/cell-type for a particular trait, and build a tissue-based correlation matrix among these complex traits, and (5) use tissue-specific functional variants and GWAS summary statistics to identify genes and biological processes important to complex traits.

2. Results

2.1. **Tissue- and cell-type specific scores.**

Here we use data for the histone modification marks H3K4me1, H3K4me3, H3K9ac, and H3K27ac for 127 different tissue- and cell-types represented in the Roadmap datasets (see Supplemental Tables S1 and S2; see also Supplemental Material and Supplemental Figure S1 for more information on these four annotations). Not all of these marks were profiled for each of the 127 different cell types and tissues. However, using the relationships between different marks within and across tissues, signal tracks have been predicted for each of these marks across all tissues [12]. We make use of these predicted signal tracks to compute integrated scores for every possible variant in the human sequence, for 127 cell types and tissues.

Figure 1(A) is a multi-dimensional scaling (MDS) visualization of the correlations between the functional scores for different tissues. General tissue groupings are indicated in different colors. As expected, tissues that are functionally related tend to cluster together. The first dimension clearly separates blood cells (indicated in red), including various primary immune cell subtypes, from solid organ tissues. The second dimension separates stem cells (indicated in blue) from other tissues. When only primary cells and tissues are included in the MDS analysis (Figure 1(B)), blood cells continue to co-cluster and separate from solid tissues in the first dimension, but the second dimension becomes reflective of the embryonic tissue of origin. Along this dimension, the tissues follow the approximate order from ectoderm-derived, mesoderm-derived, to endoderm-derived, suggesting that epigenetic marks are more similar in tissues of the same embryonic origin. These results also indicate that ectoderm- and endoderm-derived tissues are most distant from each other, while mesoderm-derived tissues are more likely to share regulatory elements with ectoderm- and endoderm-derived tissue types.

In Figure 2 we provide a global picture of the sharing of functional variants across tissues in Roadmap, using the generalized Jaccard similarity index, a measure of overlap between functional variants in two tissues (see Methods). As shown there is generally low overlap between functional variants in different tissues, especially those in blood and other tissues. Overall, the median Jaccard index across all pairs of tissues is 0.37. As a comparison, we have computed these overlap indices using only functional variants that fall in promoters (Methods; Supplemental Figure S2) and enhancers (Methods; Supplemental Figure S3). The median Jaccard index for variants falling in promoters is 0.42, and 0.22 for variants falling in enhancers, concordant with the expectation that there is more sharing across tissues for functional variants in the promoters vs. those in enhancers. The patterns of sharing vary across pairs of tissues. For example, the median Jaccard index for the overlap between any blood tissue and the rest of the tissues is 0.32; 0.17 when considering only variants that fall in enhancers, and 0.37 for variants that fall in promoters. Also, concordant with recent analyses [16] investigating the overlap among eQTLs across four tissues (skin, fat, whole blood and lymphoblastoid cell lines), we find considerable overlap of functional variants in fat and skin tissues. For example, the Jaccard index of overlap between functional variants in 'Adipose Derived Mesenchymal Stem Cells' and functional variants in 'Foreskin Fibroblast Primary Cells' is 0.61; 0.66 when considering only variants that fall in promoters, and 0.62 for variants in enhancers. These overlap indices for fat and skin tissues are substantially higher than the median overlap among the tissues in Roadmap, as described above. Because of limitations in sample sizes, most existing studies, including GTEx, focus on the detection of cis-eQTLs, and are able to show a high degree of eQTL sharing across tissues. However, most of the genetic variants that contribute to the heritability of gene expression remain to be identified, and variance-component methods point to limited sharing of trans effects [16], concordant with our empirical findings using functional variants across the entire genome.

Similarly, we show in Figure 3 the distribution of composite functional score (summed across all 127 tissues) for common variants in the 1000 Genomes data, only for variants in promoters, and only for variants in enhancers. As shown before in studies using cis-eQTLs identified in the GTEx project [17], this distribution is bimodal for variants that fall in promoters, showing a group of SNPs functional across all tissues. In contrast, variants that fall in enhancers show a distribution skewed towards less sharing across all tissues, but rather sharing between a few tissues.

2.2. **Enrichment analyses using eQTLs from the Genotype-Tissue Expression project.** The Genotype-Tissue Expression (GTEx) project is designed to establish a comprehensive data resource on genetic variation, gene expression and other molecular phenotypes across multiple human tissues [17]. We focus here on the cis-eQTL results from the GTEx V6 release comprising RNA-seq data on 7051 samples in 44 tissues, each with at least 70 samples (Supplemental Table S3). We are interested in identifying for each GTEx tissue the Roadmap tissues that are most enriched in eQTLs from that GTEx tissue relative to other GTEx tissues (see Methods). In Table 1 we show the top Roadmap Tissue for each GTEx tissue. In most cases, eQTLs from a GTEx tissue show the most enrichment in the functional component of a relevant Roadmap tissue. For example, for the liver tissue in GTEx, liver is the Roadmap tissue with the highest enrichment, for the pancreas tissue in GTEx, the Roadmap tissue with the highest enrichment is pancreas. However, there are also some cases where the top tissue is not necessarily the most intuitive one, and this is especially apparent for tissues of male and female reproductive system. This may be due to small sample sizes for eQTL discovery in GTEx for these tissues (only one sex is contributing). Additionally, some GTEx tissues with unexpected pairings do not have representative counterparts in Roadmap (e.g. thyroid, peripheral nerve, testis, and prostate). Nevertheless, most tissues that are analyzed by GTEx in sufficient numbers can be paired precisely with their existing Roadmap counterparts using our simple enrichment test.

2.3. **Prediction of causal tissues for 21 complex traits.** As an application of our scores to complex trait genetics, we use the recently-developed LD score regression framework [18] to identify the most relevant tissues and cell types for 21 complex traits for which moderate to large GWAS studies have been performed (Table 2; [19]-[39]). The stratified LD score regression approach uses information from all SNPs and explicitly models linkage disequilibrium (LD) to estimate the contribution to heritability of different functional elements. We modify this method to weight SNPs by their tissue specific functional score, and in this way we assess the contribution to heritability of functional SNPs in a particular Roadmap tissue- or cell-type (see Supplemental Material for more details).

In Table 2 we show the top Roadmap tissue-/cell- type for each of the 21 complex traits. For most disorders, the top tissue has previously been implicated in their pathogenesis. Notably, recent data indicates that BMI-associated loci are enriched for expression in the brain and central nervous system [40]. Consistent with these findings, our analysis points to the highest tissue association of BMI with cells of the Brain Germinal Matrix. Similarly, brain represents the top tissue for most neuropsychiatric disorders, education levels, age at menarche and smoking. Blood-derived and immune cells represent the top tissue for virtually all of the autoimmune conditions available for analysis. For example, GWAS findings for ulcerative colitis and Crohn's disease both map specifically to the regulatory elements in Th17 cells, whereas lymphoblastoid cell lines represent the top cell type for IgA nephropathy and rheumatoid arthritis. The most unexpected findings include small intestine as the top tissue for coronary artery disease, primary hematopoietic stem cells for Alzheimer's disease (although bone marrow-derived hematopoietic stem cell therapy has been previously explored in Alzheimer's disease [41]), and CD14+ monocytes for bipolar disorder.

In Figure 4 we show the correlation matrix for the 21 traits based on the Z-scores from the LD score regressions. This correlation structure reflects the extent to which traits share the same causal tissues, rather than the genetic correlation [42]. Three large phenotypic clusters are clearly evident. The most tightly correlated cluster contains autoimmune and inflammatory conditions, including inflammatory bowel diseases, alopecia areata, rheumatoid arthritis and IgA nephropathy. As expected, these conditions share highest functional scores in blood-derived immune cells. The second most strongly inter-correlated cluster is driven by scores in neuronal tissues, and consists of BMI, age at menarche, educational attainment, schizophrenia, and smoking history, with somewhat weaker correlations with epilepsy and bipolar disorder. Lastly, there is a clear co-clustering of cardio-metabolic traits that map to the tissues of liver, pancreas, and small intestine. Interestingly, height, coronary artery disease, and type 2 diabetes all share high functional scores in similar tissues, but are negatively correlated with autoimmune and neuronal disease clusters, suggesting a distinct set of tissues and regulatory elements that are specific to these traits.

2.4. **Gene-based analyses using tissue specific functional SNPs and enhancers-target genes maps.** We have applied a Bayesian gene-based test to assess the association between individual genes and each of the 21 complex traits. Specifically, we first identified for each trait the variants predicted to be functional in the top tissue in Table 2 (namely those variants with posterior probability to be in the functional class greater than 0.5). This reduces the number of variants considered for each trait considerably (on average about 7.7% of variants are retained). For each gene we then included all the functional variants assigned to the gene (see Methods; see also Supplemental Table S4 for a breakdown of variants into different positional classes). We computed for each gene a Bayes factor (BF; Methods).

In Figure 5 we show for each of the 21 complex traits the log(BF) for those genes with log(BF)>0, namely those genes with positive support of an association with trait. Of these genes, those with log(BF)>5 provide very strong evidence of an association. As shown, highly polygenic traits, such as schizophrenia and height, have the largest number of genes with log(BF)>5 (namely, 74 for schizophrenia and 118 for height). Similarly LDL, HDL, triglycerides, inflammatory bowel disease and rheumatoid arthritis have large number of genes ($57-92$) with large BFs, followed by age at menarche with 49 genes, alopecia areata with 33 genes and Alzheimer's disease with 27 genes. Because of small sample sizes for autism and epilepsy GWAS, these diseases have the weakest gene-based associations results. We provide genes ranked based on their corresponding log(BF) for each of the 21 traits in the Supplemental Material.

We next tested whether the genes with log(BF)>0, i.e. those showing positive support for association with the trait, are enriched in specific KEGG pathways by GSEA [43] (see Methods). All significant (FDR$<5\%$) KEGG pathways for these traits are listed in the Supplemental Material. This analysis replicates several previously reported pathway associations. For example, for height we recapitulate previously reported enrichment for "Pathways in Cancer", "TGF-beta Signaling", and "WNT Signaling" [44]. For IgAN, we confirm highly significant enrichment for the pathways of "Intestinal Immune Network for IgA Production" and "Leishmania Infection" [35]. Similarly, the top most significant pathways common for IBD include previously implicated "JAK-STAT Signaling Pathway", "Chemokine Signaling Pathway", and "Cytokine-Cytokine Receptor Interaction" [26]. For alopecia areata, there is extensive overlap with the results from a previous pathway analysis of GWAS loci, including the identification of immune pathways and known disease comorbidities [27]. Most notably, our analysis placed "JAK-STAT signaling" among the most significant pathways, which has high clinical relevance due to our work validating etiological contributions with immunological and pharmacological studies in the mouse model and human patients [28]. For Alzheimer's Disease we found significant enrichments in several immune pathways for Alzheimer's Disease, including "Systemic Lupus Erythematosus", "Asthma", and "Allograft Rejection", confirming previous findings about a role for immune response and inflammation genes in Alzheimer's Disease [29].

Among novel findings, we observed "TGF-beta Signaling Pathway" enrichment for Type 2 Diabetes ($P = 4.8 \times 10^{-7}$); and enrichments in the pathways of "Axon Guidance" ($P = 4.2 \times 10^{-6}$), "MAPK Signaling" ($P = 8.8 \times 10^{-6}$), and "Fc-gamma Receptor Mediated Phogocytosis" ($P = 9.7 \times 10^{-6}$) for Age at Menarche. In addition, we observed "KEGG Lysosome" pathway enrichment for Coronary Artery Disease ($P = 8.9 \times 10^{-5}$); "Adherens Junctions" ($P = 6.7 \times 10^{-5}$), "Proteosome" ($P = 1.6 \times 10^{-4}$), and "Glycerolipid Metabolism" ($P = 1.7 \times 10^{-4}$) for HDL; and "Complement and Coagulation Cascade" ($P = 1.0 \times 10^{-5}$) for triglycerides levels. These findings provide several novel insights into disease-associated biological processes and generate new hypotheses and can be tested in focused experimental studies. Moreover, our results demonstrate that the information on tissue-specific functional variants combined with GWAS summary statistics can be used effectively to identify candidate disease genes and highlight specific pathogenic pathways for complex traits.

3. Discussion

We have presented here an unsupervised approach for the functional prediction of genetic variation in specific tissue-/cell-types using histone modification data from the Roadmap Epigenomics project. Such context specific functional prediction of genetic variation is essential for the interpretation of genetic variants implicated by GWAS of complex traits. Although one could use individual tissue- and cell-type specific histone marks, the large number of such epigenetic markers and the large number of tissue- and cell-types make the individual analysis of these features inefficient, and the resulting results difficult to interpret. Therefore we have proposed here an integrated functional score.

Our approach is nonparametric, and as such is very flexible and can accommodate any type of continuous distributions. We note that since the histone modifications we consider in our applications can be dichotomized, for example, with peak calling methods [46], alternative parametric methods are also possible [47]. In the Methods section and the Supplement we also consider two parametric approaches based on mixtures of multivariate Bernoulli distributions. As shown, these parametric methods perform similarly to our nonparametric approach. The parametric approaches, although simple, lack flexibility and require the data to be dichotomized (more generally, conform to a parametric distribution). As discussed in [48], chromatin immunoprecipitation sequencing (ChIP-seq) experiments targeting histone modifications require difficult experimental conditions, and as such the resulting data quality can be low, especially in low input samples, making the conversion to dichotomous variables problematic.

We show that the overlap of functional variants across tissues shows an almost block diagonal structure, with fairly large overlap among closely related tissues and low overlap among different types of tissues. As shown before in the context of eQTL studies, functional variants in blood have the least overlap with functional variants in other tissues. Furthermore we show that functional variants in enhancer elements are less likely to be shared across many tissues, compared with functional variants in promoters. Our results on the overlap of functional variants across tissues can be used to select a surrogate tissue for a trait of interest when data on the true tissue is not available.

Tissue- and cell-type specific scores have important applications to complex trait genetics. As shown before [18], they can be used to infer the most relevant tissue for a trait of interest, and can help focus the search for causal variants in complex traits by restricting the set of candidate variants to only those that are predicted to be functional in a tissue relevant for the trait.

These context specific scores can also be used in improving power of eQTL studies, especially when testing large number of trans-eQTLs-gene pairs. Due to the very large number of possible SNP-gene pairs, trans-eQTL studies have had limited success so far, especially with modest sample sizes available at this time in projects such as GTEx. In such cases, using the tissue specific functional scores as prior information in a weighted-FDR framework can be helpful.

Related methods such as ChromHMM [49] and Segway [50] differ in important ways from our approach. While such methods integrate multiple chromatin datasets just as our own method, their goal is to segment the genome into non-overlapping segments, representing major patterns of chromatin marks, and label these segments using a small set of labels such as transcription start site, promoter flanking, enhancer etc.. Our approach aims to provide variant resolution functional predictions across multiple tissues and cell types, recognizing that not all variants falling into a functional segment have the same functional effect. For the 127 different tissues and cell types, our approach labels on average 4% of the genome as functional in a specific tissue or cell type, while ChromHMM and Segway tend to label a larger proportion of variants sitting in functional segments.

The accuracy of the results presented in this work is dependent on having sufficiently large sample sizes for the GWAS studies, and biochemical assays for a comprehensive set of tissues and cell types. This is especially true for identifying specific genes associated with traits of interest. Our results on pinpointing genes with strong evidence for association with a trait mirror results from GWAS studies, with traits such as schizophrenia and height showing large numbers of strong associations, while autism and epilepsy resulting in no strong candidates due to small sample sizes for the corresponding GWAS studies.

Precomputed scores for every possible variant in the human genome, for 127 tissue- and cell-types available in Roadmap are available for download at our website.

Acknowledgments We gratefully acknowledge support by National Institutes of Health grants MH106910 (II-L, BX), AR065963 (LP, AC), DK105124 and the Herbert Irving Scholars Award (KK), MH100233 (to JDB), and the Seaver Foundation. Data on coronary artery disease/myocardial infarction have been contributed by CARDIoGRAMplusC4D investigators and have been downloaded from www.cardiogramplusc4d.org. Data on glycaemic traits have been contributed by MAGIC investigators and have been downloaded from www.magicinvestigators.org.

Web-based resources

Eigen: http://www.columbia.edu/∼ ii2135/eigen.html
ENCODE: https://www.encodeproject.org/
GSEA: http://software.broadinstitute.org/gsea/index.jsp
GTEx: http://www.gtexportal.org/home/
Reg2Map: http://www.broadinstitute.org/∼meuleman/reg2map/HoneyBadger2-intersect_release/
Roadmap Epigenomics: http://www.roadmapepigenomics.org/
Roadmap Stringent enhancers: http://www.broadinstitute.org/ meuleman/reg2map/HoneyBadger2-intersect_release/
1000 Genomes: http://www.1000genomes.org/
UCSC genome browser: https://genome.ucsc.edu/
Tissue-specific functional scores (npEM): https://xioniti01.u.hpc.mssm.edu/npEM/

GWAS summary statistics:
Age at menarche: http://www.reprogen.org/Menarche_Nature2014_GWASMetaResults_17122014.zip
Alopecia areata: http://www.broadinstitute.org/∼sripke/share_links/sRSxpynHPaYRJ1SnYXD17eo3qK8IE6
_daner_ALO4_1011b_mdsex/
Alzheimer's disease: http://web.pasteur-lille.fr/en/recherche/u744/igap/igap_download.php
Autism: http://www.med.unc.edu/pgc/files/resultfiles/pgcasdeuro.gz
Bipolar Disorder: http://www.med.unc.edu/pgc/files/resultfiles/pgc.bip.2012-04.zip
BMI, Height: http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
Coronary Artery Disease: ftp://ftp.sanger.ac.uk/pub/cardiogramplusc4d/cardiogram_gwas_results.zip
Crohn's Disease: ftp://ftp.sanger.ac.uk/pub/consortia/ibdgenetics/cd-meta.txt.gz
Educational Attainment: http://ssgac.org/documents/SSGAC_Rietveld2013.zip
Epilepsy: http://www.epigad.org/gwas_ilae2014/ILAE_All_Epi_11.8.14.txt.gz
Ever Smoked: http://www.med.unc.edu/pgc/files/resultfiles/tag.evrsmk.tbl.gz
Fasting Glucose: ftp://ftp.sanger.ac.uk/pub/magic/MAGIC_Manning_et_al_FastingGlucose_MainEffect.txt.gz
HDL: http://www.broadinstitute.org/mpg/pubs/lipids2010/HDL_ONE_Eur.tbl.sorted.gz
IGAN: dbGaP Study Accession: phs000431.v2.p1
LDL: http://www.broadinstitute.org/mpg/pubs/lipids2010/LDL_ONE_Eur.tbl.sorted.gz
Rheumatoid Arthritis: http://plaza.umin.ac.jp/yokada/datasource/files/GWASMetaResults/
RA_GWASmeta_European_v2.txt.gz

Schizophrenia: http://www.med.unc.edu/pgc/files/resultfiles/scz2.snp.results.txt.gz
Triglycerides: http://www.broadinstitute.org/mpg/pubs/lipids2010/TG_ONE_Eur.tbl.sorted.gz
Type 2 Diabetes: http://www.diagram-consortium.org/downloads.html
Ulcerative Colitis: ftp://ftp.sanger.ac.uk/pub/consortia/ibdgenetics/ucmeta-sumstats.txt.gz

4. Methods

4.1. **Unsupervised methods for the integration of functional predictions.**

4.1.1. *Nonparametric (multivariate) mixture models.* Assume we have a set of m genetic variants. For each variant i, $1 \leq i \leq m$, we have k functional annotations and the vector of scores is denoted by $\mathbf{Z_i} = (Z_{i1}, \ldots, Z_{ik})$. Let $\mathbf{Z} = (\mathbf{Z_1}, \ldots, \mathbf{Z_m})$ be the set of functional scores for all the variants. We assume all the scores are continuous. We assume there exists an underlying two-component mixture model, corresponding to two possible functional classes for the variants: functional and non-functional. We also let $\mathbf{C} = (C_1, \ldots, C_m)$ denote the set of indicator variables for all the variants, where $C_i = 1$ if variant i is functional and $C_i = 0$ if variant i is non-functional. We are not able to observe $\mathbf{C}$. Our primary goal here is to fit a nonparametric mixture model ψ with two components: $\psi = (\pi, f_0, f_1)$, where f_0 and f_1 are the probability densities for each of the components and π is a mixing parameter, and to calculate posterior probabilities for each variant to be functional given the observed scores $\mathbf{Z}$, i.e. $P_\psi(C_i = 1|\mathbf{Z})$.

We use an EM-like algorithm and a kernel method for the nonparametric density estimation of the component densities in the mixture model. This algorithm is similar to that described in [51] for fitting nonparametric multivariate mixtures; however whereas they assume the individual scores to be conditionally independent, we relax the assumption and only require blockwise conditional independence, and scores within a block having arbitrary distributions and correlation structures.

Let B be the number of conditionally independent blocks, containing $n_1, \ldots, n_B$ annotations respectively. The observations are then contained in $\mathbf{R}^{n_1} \times \mathbf{R}^{n_2} \times \cdots \times \mathbf{R}^{n_B} = \mathbf{R}^k$. Because of the assumption of conditional independence among scores in different blocks, for any $\mathbf{u} = (u_1, \ldots, u_k)$ we can write:

$$f_0(\mathbf{u}) = \prod_{j=1}^{B} f_{0j}(\tilde{\mathbf{u}}_\mathbf{j}), \quad \text{and} \quad f_1(\mathbf{u}) = \prod_{j=1}^{B} f_{1j}(\tilde{\mathbf{u}}_\mathbf{j}), \tag{4.1}$$

where f_{0j} and f_{1j} are multivariate densities corresponding to scores in block j, continuous on $\mathbf{R}^{n_j}$. Also, $\tilde{\mathbf{u}}_\mathbf{j}$ corresponds to the scores in the jth block.

We propose the following EM-like algorithm for the nonparametric estimation of the mixture model $\psi = (\pi, f_0, f_1)$:

Step 1. Use k-means clustering to obtain initial component membership probabilities, i.e. initialize the posterior probability p_i to be 1 for each variant i in the functional group and 0 for the variants in the non-functional group.

Step 2. Fit a multivariate kernel density estimate for each block and component separately: f_{1j}^{new} and f_{0j}^{new} for each block $j = 1 \ldots B$, weighting variants by component membership probability. More explicitly, for any $\mathbf{u} = (\tilde{\mathbf{u}}_\mathbf{1}, \ldots, \tilde{\mathbf{u}}_\mathbf{B}) \in \mathbf{R}^k$ and $j = 1, \ldots, B$, we let

$$f_{0j}^{\text{new}}(\tilde{\boldsymbol{u}}_{\boldsymbol{j}}) = \frac{\sum_{i=1}^{m}(1-p_i)\mathrm{K}_{\mathbf{H}_{0j}}(\tilde{\boldsymbol{u}}_{\boldsymbol{j}} - \tilde{\boldsymbol{Z}}_{\boldsymbol{ij}})}{\sum_{i=1}^{m}(1-p_i)} = \frac{1}{m(1-\pi)}\sum_{i=1}^{m}(1-p_i)\mathrm{K}_{\mathbf{H}_{0j}}(\tilde{\boldsymbol{u}}_{\boldsymbol{j}} - \tilde{\boldsymbol{Z}}_{\boldsymbol{ij}}),$$

and

$$f_{1j}^{\text{new}}(\tilde{\mathbf{u}}_{\boldsymbol{j}}) = \frac{\sum_{i=1}^{m} p_i K_{\mathbf{H}_{1j}}(\tilde{\mathbf{u}}_{\boldsymbol{j}} - \tilde{\mathbf{Z}}_{\boldsymbol{ij}})}{\sum_{i=1}^{m} p_i} = \frac{1}{m\pi} \sum_{i=1}^{m} p_i K_{\mathbf{H}_{1j}}(\tilde{\mathbf{u}}_{\boldsymbol{j}} - \tilde{\mathbf{Z}}_{\boldsymbol{ij}}).$$

Then

$$f_0^{\text{new}}(\mathbf{u}) = \prod_{j=1}^{B} f_{0j}^{\text{new}}(\tilde{\mathbf{u}}_{\mathbf{j}}), \quad \text{and} \quad f_1^{\text{new}}(\mathbf{u}) = \prod_{j=1}^{B} f_{1j}^{\text{new}}(\tilde{\mathbf{u}}_{\mathbf{j}}).$$

For each block j, we take both $\mathbf{H}_{0j}$ and $\mathbf{H}_{1j}$ to be the same. The kernel is taken to be the probability density function of $MVN(0, \mathbf{H})$, where $\mathbf{H} = \text{diag}(h_1, \ldots, h_k)$ is a diagonal bandwidth matrix (hence a product kernel). We choose the bandwidth parameter h_i to be

$$h_i = 0.9 \min\{\text{SD}_i, \text{IQR}_i/1.34\} m^{-1/5}$$

according to a rule of thumb due to Silverman [52], where SD_i and IQR_i are the standard deviation and interquartile range of score i, respectively. Choosing an appropriate bandwidth is very important for the accuracy of the estimation. The bandwidth we chose will work well if the true density resembles the normal distribution, but can be quite misleading when the true density deviates from the normal distribution (e.g. for multimodal distributions).

Step 3. Update component membership probabilities based on the fitted densities. Given current estimates of $\psi = (\pi, f_0^{\text{new}}, f_1^{\text{new}})$, update the posterior probability for each variant i to be functional given the scores, i.e. $P_\psi(C_i = 1 \mid \mathbf{Z})$, where $i = 1, \ldots, m$, by

$$p_i^{\text{new}} = P_\psi(C_i = 1 \mid \mathbf{Z_i}) = \frac{\pi f_1^{\text{new}}(\mathbf{Z_i})}{\pi f_1^{\text{new}}(\mathbf{Z_i}) + (1 - \pi) f_0^{\text{new}}(\mathbf{Z_i})}.$$

Set $\pi^{\text{new}} = (1/m) \sum_{i=1}^{m} p_i^{\text{new}}$.

Repeat Steps 2-3 until convergence criterium is met.

We refer to this approach as the nonparametric (multivariate) mixture model with block structure.

4.1.2. *Nonparametric (multivariate) mixture model with conditional independence assumptions.* We also consider a particular model that assumes conditional independence among annotations, i.e., that ignores the block structure mentioned above. By taking the bandwidth matrix $\mathbf{H}$ in the multivariate mixture model to be $\mathbf{H} = \text{diag}(h_1, \ldots, h_k)$, where h_i is specified according to the Silverman's rule, and assuming each annotation is in its own block, we get a multivariate mixture model with a conditional independence covariance structure. This is the same *npEM* model as in [51], and is implemented in the R package *mixtools* [53].

We fit two different *npEM* models using annotation data from two different sets of variants. For one *npEM* model, we use all the variants whose posterior probabilities we wish to calculate. Since the kernel density estimation step implemented in the R package *mixtools* is too slow for use with millions of variants, for this model we replace that step with a binned kernel density estimation using the R package *KernSmooth* [54]. This method is based on a regular grid with equally spaced points at which the density is estimated. We call this method *npEM-binned*. For the other *npEM* model, we use a different method to select a set of $20,000$ variants to use to train the *npEM* model for each tissue, with overrepresentation of variants with low and high functional annotation scores. For each functional annotation (in practice we use $k = 4$ functional annotations), we randomly select $2,500$ variants equally distributed among the following quantiles of the functional annotation scores for all variants for that annotation: 0, 0.00001, 0.0001, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99, 0.999, 0.9999, 0.99999. We then randomly select enough additional variants so that we have $20,000$ variants in our training set. We then use this training set to train a different *npEM* model for each tissue. We can then use these models to calculate tissue-specific posterior probabilities for any variant (even one not in the training set), using our estimates of the component-specific densities f_0 and f_1 and the mixing parameter π. We call this method *npEM-quantile*. The results in the Results section are those from the *npEM-quantile* method, since, as we show in the Supplemental Material, this method is robust and tends to perform well relative to alternative approaches we considered.

4.1.3. *Identifiability issues.* Nonparametric mixture models are not identifiable in general. However, under certain conditions nonparametric mixtures are identifiable. This has been shown to be the case for multivariate mixtures with two components in $\mathbf{R}^k$ for $k \geq 3$, if the coordinates are conditionally independent given the component [55]. In particular, the nonparametric mixture model with conditional independence is identifiable as long as there are at least 3 annotations. This result generalizes to the setting of blockwise conditional independence, as long as there are at least three blocks that are independent, conditionally on the latent functional class [56].

4.1.4. *Convergence issues.* Nonparametric density estimation in high dimensions is challenging. Kernel density estimation, as used here, works well for small dimensions (e.g. $d \leq 3$). For dimension d, the best bandwidth selection method leads to a mean integrated squared error of $O(m^{-4/(4+d)})$ [54], which grows very slowly with m as d increases. In this paper we use blocks of small dimension (max d is 4). When $d = 4$, the *best* convergence rate is $O(m^{-0.5})$. Hence, at this rate, to reduce estimation error in half, one would need a four-fold increase in the size (i.e. number of variants) of the dataset.

4.1.5. *Eigen-PC.* A related approach for integrating the different annotations for a genetic variant has been introduced in Ionita-Laza et al. [9]. The underlying model is also a two component nonparametric mixture model. This approach is based on the eigendecomposition of the annotation covariance matrix, and using the lead eigenvector to weight the individual annotations. More precisely, if we denote by $\mathbf{Q}$ the variance-covariance matrix of the annotation scores $\mathbf{Z}$, we take the eigendecomposition of $\mathbf{Q}$, and the score for a variant is the weighted sum of the annotations, with the vector of weights equal to the first eigenvector of $\mathbf{Q}$. We have shown in [9] that the Eigen-PC score performs favorably against existing methods such as CADD [10], especially for noncoding variants.

4.1.6. *Parametric (multivariate Bernoulli) mixture models.* In our particular application to histone modification data from Roadmap Epigenomics, it is possible to discretize the data since in theory a histone modification is either present or absent at a specific genomic position (we make use of the results of gapped peak calling on the signal tracks). Using the same notations as before, we now assume the k annotations for a variant, $\mathbf{Z_i} = (Z_{i1}, \ldots, Z_{ik})$, are binary. The simplest model (*mvB*) assumes conditional independence among all k annotations. This model is similar to the one in [47].

For any $\mathbf{u} = (u_1, \ldots, u_k) \in \{0,1\}^k$ we have

$$P(\mathbf{u}|C=1) = \prod_{j=1}^{k} p_{1j}^{u_j}(1-p_{1j})^{1-u_j}, \quad \text{and} \quad P(\mathbf{u}|C=0) = \prod_{j=1}^{k} p_{0j}^{u_j}(1-p_{0j})^{1-u_j}. \tag{4.2}$$

Let $\mathbf{p_1} = (p_{11}, \ldots, p_{1k})$ and $\mathbf{p_0} = (p_{01}, \ldots, p_{0k})$. Then the EM algorithm for estimating the mixture of multivariate Bernoulli model $\boldsymbol{\psi} = (\pi, \mathbf{p_1}, \mathbf{p_0})$ is as follows:

Step 1. Use k-means clustering to obtain initial estimates for the parameters.

Step 2.

$$p_i^{\text{new}} = P_\psi(C_i = 1 \,|\, \mathbf{Z_i}) = \frac{\pi P(\mathbf{Z_i}|C_i = 1)}{\pi P(\mathbf{Z_i}|C_i = 1) + (1-\pi) P(\mathbf{Z_i}|C_i = 0)}$$

Step 3. Set $\pi^{\text{new}} = (1/m)\sum_{i=1}^m p_i^{\text{new}}$ and

$$\begin{aligned}\mathbf{p_1} &= \frac{1}{\sum_{i=1}^m P_\psi(C_i = 1 \,|\, \mathbf{Z}_i)} \sum_{i=1}^m P_\psi(C_i = 1 \,|\, \mathbf{Z}_i)\mathbf{Z}_i \\ \mathbf{p_0} &= \frac{1}{\sum_{i=1}^m P_\psi(C_i = 0 \,|\, \mathbf{Z}_i)} \sum_{i=1}^m P_\psi(C_i = 0 \,|\, \mathbf{Z}_i)\mathbf{Z}_i\end{aligned}$$

where $\mathbf{p_j} = (p_{j1}, \ldots, p_{jk})$ for $j = 0, 1$.

Repeat Steps 2-3 until convergence criterium is met.

4.1.7. *Identifiability issues.* It is well known that finite mixtures of Bernoulli products are not identifiable in a strict sense [57]. However, Allman et al. [56] have shown that finite mixtures of Bernoulli products are in fact *generically* identifiable, which means that only a subset of parameters of measure zero may not be identifiable. In other words, any observed dataset has probability one of being drawn from a distribution with identifiable parameters. This explains why in practice it makes sense to estimate these models, despite their lack of strict identifiability. Allman et al. [56] show that an r component mixture of products of p independent Bernoulli is generically identifiable if $p \geq 2\lceil log_2(r)\rceil + 1$. In our case, since we assume $r = 2$ we need to have at least 3 independent Bernoulli variables.

4.1.8. *Mixture of multivariate Bernoulli, with dependence.* We also fit a two-component multivariate Bernoulli mixture that accounts for within-component correlations among the annotations. Since we only have $k = 4$ binary annotations there are only 2^k possible functional annotation vectors. Let these vectors be $\mathbf{X_j}$ for $1 \leq j \leq 2^k$, and $\mathbf{X}$ be the $2^k \times k$ matrix of these annotation vectors. Let B_j be the number among the m variants whose values of the functional annotations is $\mathbf{X_j}$, and let $\mathbf{B} = (B_1, \ldots, B_{2^k})$.

To initialize our EM algorithm that accounts for within-component correlations, we use a two-component multivariate binary mixture model that assumes conditional independence of the annotations within each component. Fitting this mixture model yields the mixing coefficient π_1 for the first component (the mixing coefficient for the second component is $1 - \pi_1$) as well as, for each possible vector of functional annotation values $\mathbf{X_j}$ with $1 \leq j \leq 2^k$, the posterior probability p_{j1} that a variant with this vector of functional annotations is in the first component. Let $\mathbf{p_1} = (p_{11}, \ldots, p_{2^k 1})$, and $\mathbf{p_2} = 1 - \mathbf{p_1}$. Using these mixing coefficients and posterior probabilities, we use the following EM algorithm to fit our model:

Step 1. Calculate the expected number of variants assigned to the two components using the current estimates of the posterior probabilities. Let $\mathbf{C_1} = (B_1 p_{11}, \ldots, B_{2^k} p_{2^k 1})$ be the vector of number of variants assigned to the first component, and $\mathbf{C_2} = (B_1 p_{12}, \ldots, B_{2^k} p_{2^k 2})$ be the vector of the number of variants assigned to the second component.

Step 2. Fit two log-linear models, using the vectors $\mathbf{C_1}$ and $\mathbf{C_2}$ (rounded, so that we can use a Poisson response distribution) as weights. The design matrix for each of these two models consists of the matrix $\mathbf{X}$ to which columns consisting of products of appropriate columns of $\mathbf{X}$ are adjoined (see discussion on interaction terms below). These products induce within-component dependence between the functional annotations. Let $\mathbf{D_1}$ and $\mathbf{D_2}$ be the fitted values resulting from fitting these log-linear models.

Step 3. Reestimate the posterior probabilities with the equations

$$p_{j1} = \frac{\frac{\pi_1 D_{j1}}{\sum_{i=1}^{2^k} D_{i1}}}{\frac{\pi_1 D_{j1}}{\sum_{i=1}^{2^k} D_{i1}} + \frac{\pi_2 D_{j2}}{\sum_{i=1}^{2^k} D_{i2}}}$$

and $p_{j2} = 1 - p_{j1}$.

Step 4. Reestimate the mixing coefficients π_1 and π_2 with the equations:

$$\pi_1 = \frac{\sum_{i=1}^{2^k} B_i p_{i1}}{m}$$

and $\pi_2 = 1 - \pi_1$.

Repeat steps 1 through 4 until convergence.

We wish to account for the most important within-component correlations between the different functional annotations. In our applications below, we use four different functional annotations, so there are six different pairs of annotations whose correlation we could model (we do not model higher-order correlations between annotations). To avoid overfitting, we pick two correlations for each component. In our application, where we use the annotations H3K4me1, H3K4me3, H3K9ac and H3K27ac, we include interaction terms between H3K9ac and H3K4me3 and H3K9ac and H3K27ac, since these pairs have the highest marginal correlations, as illustrated by Supplemental Figure S1.

4.2. **Generalized Jaccard index of overlap.** If $\mathbf{X} = (x_1, \ldots, x_k)$ and $\mathbf{Y} = (y_1, \ldots, y_k)$ are two vectors with $x_i, y_i \geq 0$ (e.g. vector of posterior probabilities for variants to be in the functional components for two different tissues), then the generalized Jaccard index of overlap is defined as:

$$J(\mathbf{X}, \mathbf{Y}) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)}.$$

When $\mathbf{X}$ and $\mathbf{Y}$ are binary vectors, then the Jaccard index of overlap is simply the size of the intersection divided by the size of the union of the two sets. The closer it is to 1, the more overlap there is between the two sets. A Jaccard index of 0 means no overlap.

4.3. **Promoter and tissue-specific enhancer regions.** The promoter region of a protein-coding gene is defined as the union of the regions 2,500 bases upstream of any protein-coding transcripts for the gene, as defined by GENCODE version 24. For enhancer regions we use SNPs within any enhancer region, with respect to any tissue.

4.4. **GTEx enrichment.** Let $G_1, \ldots, G_{44}$ be the 44 GTEx tissues with at least 70 samples (Supplemental Table S3), and $R_1, \ldots, R_{127}$ be the 127 Roadmap tissues. For a given tissue in GTEx G_i we are interested in identifying the Roadmap tissue R_j with the highest enrichment in eQTLs from G_i relative to other tissues in GTEx. Let

$$p_{R_j | \, G_i} = \frac{\text{\#eQTLs in tissue } G_i \text{ in functional component } R_j}{\text{\#eQTLs in tissue } G_i}.$$

Note that the number of eQTLs in tissue G_i is a weighted count, with an eQTL weighted by the inverse of the number of GTEx tissues in which the variant is eQTL. This way eQTLs that are unique to tissue G_i are given higher weight relative to eQTLs that are shared across many tissues. We cannot use $p_{R_j | \, G_i}$ to rank Roadmap tissues because of the different sizes of the functional component for the Roadmap tissues. Instead we need to normalize $p_{R_j | \, G_i}$ as follows:

$$\widetilde{p}_{R_j | \, G_i} = \frac{p_{R_j | \, G_i}}{\sum_{i=1}^{44} p_{R_j | \, G_i}}$$

The eQTLS that we used in these analyses are all significantly associated SNP-gene pairs in each of these 44 GTEx tissues, produced using a permutation threshold-based approach as described by the GTEx Consortium [17].

4.5. **Gene-based tests using tissue specific functional SNPs and enhancers-target genes maps.** For each gene G and trait D we want to test H_0: G is independent of D. We assign SNPs to genes annotated as protein-coding genes in GENCODE version 24. A SNP is assigned to a gene with respect to a particular tissue if 1) it is in the region from the first to the last exon for the gene, as delineated by GENCODE (region 1), 2) if it is within the promoter region for the gene (defined as the union of the regions 2,500 bases upstream of any protein-coding transcripts for the gene, as defined by GENCODE) (region 2), 3) if it is within an enhancer region assigned to the gene with respect to that particular tissue [12] (region 3) or 4) if it is within 200,000 bases of the gene and not within regions 1, 2 or 3 with respect to any other gene. We consider a SNP to be functional in the context of a trait D if the posterior probability for the SNP to be functional in the top tissue (i.e., the tissue with the highest Z score in the LD score regression for that trait) for the trait D is greater than 0.5. For each of these functional SNPs we have a Z score statistic from the association test of the SNP and trait D in a GWAS study. Let Z_i be the Z score for the ith SNP connected to gene G.

Let A be a latent variable, with $A = 1$ if G is associated with D and 0 otherwise. For each SNP i we compute a Bayes factor (BF):

$$BF_i = \frac{P(Z_i|A=1)}{P(Z_i|A=0)}.$$

Then for a gene G the log(BF) can be approximated by the average log(BF) for the SNPs connected to G. We now show how we can compute the BF for a given SNP. We follow closely the approach in He et al. [58].

Let V_i be a second latent variable such that $V_i = 1$ if SNP i is associated with trait D and 0 otherwise. Then we have:

$$\begin{aligned} P(Z_i|A=1) &= P(V_i=1|A=1)P(Z_i|V_i=1) + P(V_i=0|A=1)P(Z_i|V_i=0), \\ P(Z_i|A=0) &= P(V_i=1|A=0)P(Z_i|V_i=1) + P(V_i=0|A=0)P(Z_i|V_i=0). \end{aligned}$$

We take $P(V_i = 1|A = 0) = 0$, as we assume that if a gene is not associated with a trait then no SNP connected to the gene is associated with the trait. We also take $P(V_i = 1|A = 1) = 0.8$, reflecting the high probability that a functional SNP in a trait associated gene is likely to be associated with the trait.

When $V_i = 0$ then $Z_i \sim N(0, 1)$. Let us now consider the case $V_i = 1$. For a quantitative trait, if the Z score comes from a linear regression of trait on genotype, He et al. [58] show, assuming Hardy-Weinberg equilibrium, that when $V_i = 1$,

$$Z_i \sim N(0, 1 + 2Mp(1-p)\sigma_a^2),$$

where M is the sample size and p is the SNP population allele frequency. Following He et al. [58], we take $\sigma_a = 0.5$, based on earlier Bayesian studies for quantitative traits [59].

For a binary trait, assuming the Armitage trend statistic was used to derive the Z scores, He et al. [58] show that:

$$P(Z_i|V_i=1) = \int N\left(Z_i|\frac{\mu_1(\beta)}{\sigma_0}, \frac{\sigma_1^2(\beta)}{\sigma_0}\right) N(\beta|0,\sigma_a^2)d\beta,$$

where β is the effect size (the logarithm of the odds ratio of the risk allele), σ_a^2 is the prior variance of the effect size for functional SNPs and μ_1, σ_0^2 and σ_1^2 are defined below.

Based on prior studies, we take $\sigma_a^2 = 0.2$ [59]. To calculate μ_1, σ_0^2 and σ_1^2 we must estimate genotype frequencies among cases and controls; we do not know these since we only have summary statistics from the GWAS studies. Let the genotypes be indexed by i, $i \in \{0, 1, 2\}$ and let x_i be the number of high-risk alleles in genotype i. Let the frequency of the ith genotype in cases be p_i and in controls q_i. Also, let f_i be the penetrance of the ith genotype, its probability of causing the trait, and g_i its frequency in the population. Given any effect size β, the frequencies p_i and q_i can be estimated assuming Hardy-Weinberg equilibrium, the prevalence of the trait K, the population allele frequency for the high-risk allele p and the multiplicative genetic model. Specifically, we have the following equations that can be used to derive p_i and q_i, under a multiplicative model for the penetrances, where $\gamma = e^{\beta}$:

$$\begin{aligned}
f_0 &= \frac{K}{[\gamma p + (1-p)]^2} \\
f_1 &= \gamma f_0 \\
f_2 &= \gamma^2 f_0 \\
p_i &= \frac{f_i g_i}{K} \\
q_i &= \frac{(1-f_i) g_i}{1-K},
\end{aligned}$$

and the following equations for μ_1, σ_1^2 and σ_0^2:

$$\begin{aligned}
\mu_1 &= M\phi(1-\phi) \sum_i x_i (p_i - q_i) \\
\sigma_1^2 &= M\phi(1-\phi)^2 \left[\sum_i x_i^2 p_i - \left(\sum_i x_i p_i \right)^2 \right] + M\phi^2(1-\phi) \times \left[\sum_i x_i^2 q_i - \left(\sum_i x_i q_i \right)^2 \right] \\
\sigma_0^2 &= M\phi(1-\phi) \left[\sum_i x_i^2 q_i - \left(\sum_i x_i q_i \right)^2 \right].
\end{aligned}$$

where ϕ is the proportion of cases in the GWAS sample. The prevalences we assume for the binary traits are in Supplemental Table S5.

References

[1] Lindblad-Toh K et al. (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478: 476–482.

[2] Khurana E et al. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 342: 1235587.

[3] ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57–74.

[4] Altshuler D, Daly MJ, Lander ES (2008) Genetic mapping in human disease. *Science* 322: 881–888.

[5] Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, Gerstein M (2016) Role of non-coding sequence variants in cancer. *Nat Rev Genet* 17: 93–108.

[6] Kellis M et al. (2014) Defining functional DNA elements in the human genome. *Proc Natl Acad Sci USA* 111: 6131–6138.

[7] Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 9: e1003709.

[8] Mohri M, Rostamizadeh A, Talwalkar A (2012) Foundations of Machine Learning. The MIT Press ISBN 9780262018258.

[9] Ionita-Laza I, McCallum K, Xu B, Buxbaum JD (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 48: 214–220.

[10] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46: 310–315.

[11] Fu Y, Liu Z, Lu S, Bedford J, Mu X, Yip K, Khurana E, Gerstein M (2014) FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biology* 15: 480

[12] Roadmap Epigenomics Consortium (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518: 317–330.

[13] Bannister AJ, Kouzarides T (2011) Regulation of chromatin by histone modifications. *Cell Res.* 21(3): 381–395.

[14] Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, Raychaudhuri S (2013) Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet* 45: 124–130.

[15] 1,000 Genomes Project Consortium (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491: 56–65.

[16] Buil A, Viuela A, Brown A, Davies M, Padioleau I, Bielser D, Romano L, Glass D, Di Meglio P, Small K, Spector T, Dermitzakis ET (2016) Quantifying the degree of sharing of genetic and non-genetic causes of gene expression variability across four tissues. doi: http://dx.doi.org/10.1101/053355

[17] The GTEx Consortium (2015) Science. 348: 648–660.

[18] Finucane HK et al. (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 47: 1228–1235.

[19] Perry JR et al. (2014) Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* 514: 92–97.

[20] Betz RC et al. (2015) Genome-wide meta-analysis in alopecia areata resolves HLA associations and reveals two new susceptibility loci. *Nat Commun* 6: 5966.

[21] Lambert JC et al. (2013) Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet* 45: 1452–1458.

[22] Cross-Disorder Group of the Psychiatric Genomics Consortium (2013) Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* 381: 1371–1379.

[23] Psychiatric GWAS Consortium Bipolar Disorder Working Group (2011) Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. *Nat Genet* 43: 977–983.

[24] Speliotes EK et al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 42: 937–948.

[25] Schunkert H et al. (2011) Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* 43: 333–338.

[26] Jostins L et al. (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* 491: 119–124.

[27] Petukhova L, Christiano AM (2016) Functional Interpretation of Genome-Wide Association Study Evidence in Alopecia Areata. *The Journal of investigative dermatology* 136: 314–317.

[28] Xing L et al. (2014) Alopecia areata is driven by cytotoxic T lymphocytes and is reversed by JAK inhibition. *Nature medicine* 20: 1043–1049.

[29] Yokoyama JS et al. (2016) Association Between Genetic Traits for Immune-Mediated Diseases and Alzheimer Disease. *JAMA Neurol* 73: 691-697.

[30] Rietveld CA et al. (2013) GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* 314: 1467–1471.

[31] International League Against Epilepsy Consortium on Complex Epilepsies (2014) Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. *Lancet Neurol* 13: 893–903.

[32] Tobacco and Genetics Consortium (2010) Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nat Genet* 42:441–447.

[33] Manning AK et al. (2012) A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* 44: 659–669.

[34] Teslovich TM et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466: 707–713.

[35] Kiryluk K et al. (2014) Discovery of new risk loci for IgA nephropathy implicates genes involved in immunity against intestinal pathogens. *Nat Genet* 46: 1187–1196.

[36] Okada Y et al. (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506: 376–381.

[37] Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 511: 421–427.

[38] Morris AP et al. (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* 44: 981–990.

[39] Lango AH et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467: 832–838.

[40] Locke AE et al. (2015) Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518: 197–206.

[41] Magga J et al. (2012) Production of monocytic cells from bone marrow stem cells: therapeutic usage in Alzheimer's disease. *J Cell Mol Med* 16: 1060–1073.

[42] Bulik-Sullivan B et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet* 47: 1236–1241.

[43] Subramanian A et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 102: 15545–15550.

[44] Chan Y et al. (2015) Genome-wide Analysis of Body Proportion Classifies Height-Associated Variants by Mechanism of Action and Implicates Genes Important for Skeletal Development. *Am J Hum Genet* 96: 695–708.

[45] Jostins L et al. (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 491: 119–124.

[46] Zhang Y et al. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9: R137.

[47] Lu Q, Powles RL, Wang Q, He BJ, Zhao H (2016) Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies. *PLoS Genet* 12: e1005947.

[48] Wei Koh Pang, Pierson Emma, Kundaje Anshul (2016) Denoising genome-wide histone ChIP-seq with convolutional neural networks. doi: http://dx.doi.org/10.1101/052118

[49] Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* 9: 215–216.

[50] Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes J, Noble WS (2012) Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat Methods* 9: 473–476.

[51] Benaglia, T., Chauveau, D., and Hunter, D. R. (2009a), An EM-like algorithm for semi- and non-parametric estimation in multivariate mixtures, Journal of Computational and Graphical Statistics, 18, 505-526.

[52] Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis, Chapman & Hall, London

[53] Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009b), mixtools: An R package for analyzing finite mixture models. Journal of Statistical Software, 32(6):1-29.

[54] Wand, M.P; Jones, M.C. (1995). Kernel Smoothing. London: Chapman & Hall/CRC. ISBN 0-412-55270-1.

[55] Hall, P. and Zhou, X-H. (2003), Nonparametric estimation of component distributions in a multivariate mixture. Ann. Statist., 31: 201-224.

[56] E. S. Allman, C. Matias, and J. A. Rhodes (2009) Identifiability of parameters in latent structure models with many observed variables Ann. Statist., Volume 37, Number 6A (2009), 3099-3132.

[57] Gyllenberg MK, Oski TR, Eilink E, Verlaan M (1994) Nonuniqueness in probabilistic numerical identification of bacteria. *J Appl Probab* 31: 542–548.

[58] He X et al. (2013) Sherlock: Detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am J Hum Genet* 95: 667-680.

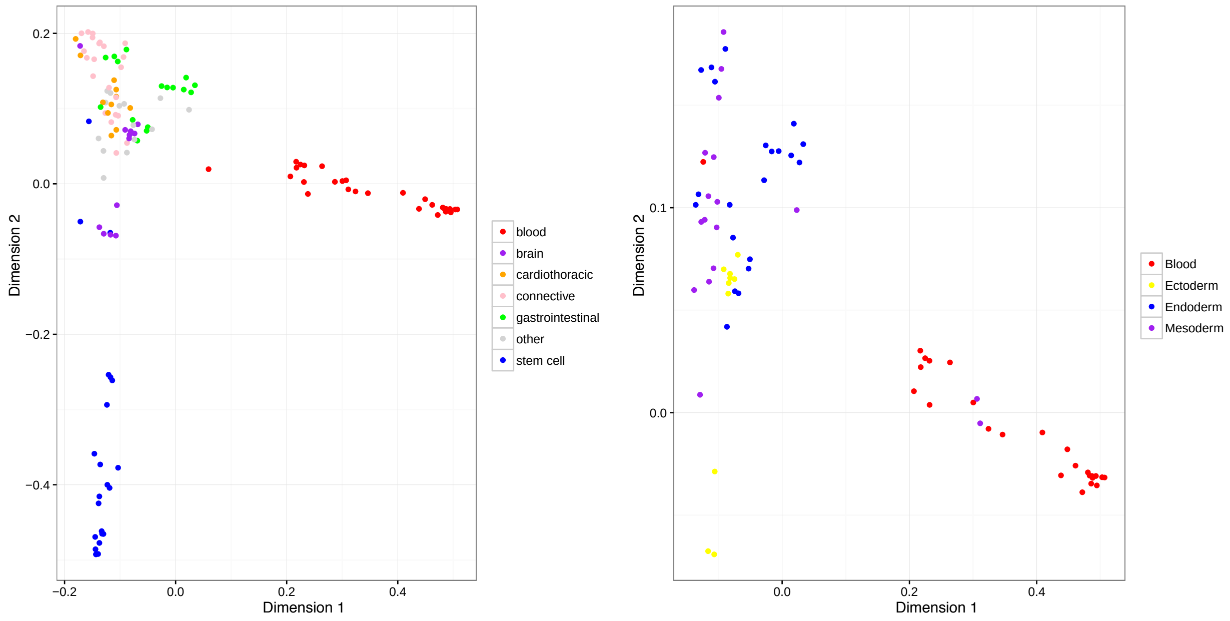
[59] Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* 10: 681–690.

Table 1. Enrichment of GTEx eQTLs among functional variants in tissues and cell types in Roadmap Epigenomics. The top Roadmap tissue is given for each GTEx tissue. The number of GTEx samples for each tissue is also reported.

GTEx Tissue	n_{GTEx}	Roadmap Epigenome Name
Muscle - Skeletal	361	Skeletal Muscle Female
Whole Blood	338	Primary neutrophils from peripheral blood
Skin - Sun Exposed (Lower leg)	302	NHEK-Epidermal Keratinocyte Primary Cells
Adipose - Subcutaneous	298	Mesenchymal Stem Cell Derived Adipocyte Cultured Cells
Artery - Tibial	285	Aorta
Lung	278	ES-I3 Cells
Thyroid	278	Fetal Heart
Cells - Transformed fibroblasts	272	Osteoblast Primary Cells
Nerve - Tibial	256	Osteoblast Primary Cells
Esophagus - Mucosa	241	NHEK-Epidermal Keratinocyte Primary Cells
Esophagus - Muscularis	218	Stomach Mucosa
Artery - Aorta	197	Bone Marrow Derived Cultured Mesenchymal Stem Cells
Skin - Not Sun Exposed (Suprapubic)	196	NHEK-Epidermal Keratinocyte Primary Cells
Heart - Left Ventricle	190	Fetal Heart
Adipose - Visceral (Omentum)	185	Monocytes-CD14+ RO01746 Primary Cells
Breast - Mammary Tissue	183	ES-WA7 Cells
Stomach	170	Gastric
Colon - Transverse	169	Rectal Mucosa Donor 31
Heart - Atrial Appendage	159	Fetal Heart
Testis	157	Osteoblast Primary Cells
Pancreas	149	Pancreas
Esophagus - Gastroesophageal Junction	127	Ovary
Adrenal Gland	126	Fetal Adrenal Gland
Colon - Sigmoid	124	Colon Smooth Muscle
Artery - Coronary	118	Mesenchymal Stem Cell Derived Adipocyte Cultured Cells
Cells - EBV-transformed lymphocytes	114	GM12878 Lymphoblastoid Cells
Brain - Cerebellum	103	Fetal Brain Male
Brain - Caudate (basal ganglia)	100	Brain Substantia Nigra
Liver	97	Liver
Brain - Cortex	96	H1 Derived Neuronal Progenitor Cultured Cells
Brain - Nucleus accumbens (basal ganglia)	93	ES-WA7 Cells
Brain - Frontal Cortex (BA9)	92	HUES64 Cells
Brain - Cerebellar Hemisphere	89	Fetal Brain Male
Spleen	89	Primary B cells from cord blood
Pituitary	87	Ganglion Eminence derived primary cultured neurospheres
Prostate	87	Liver
Ovary	85	NHDF-Ad Adult Dermal Fibroblast Primary Cells
Brain - Putamen (basal ganglia)	82	Brain Substantia Nigra
Brain - Hippocampus	81	Brain Inferior Temporal Lobe
Brain - Hypothalamus	81	ES-WA7 Cells
Vagina	79	Primary B cells from cord blood
Small Intestine - Terminal Ileum	77	Fetal Intestine Large
Brain - Anterior cingulate cortex (BA24)	72	Brain Dorsolateral Prefrontal Cortex
Uterus	70	Primary T CD8+ memory cells from peripheral blood

Table 2. Top Tissues or Cell Types in Roadmap for 21 GWAS traits. The p-value from the LD score regression, as well as the GWAS sample size are reported for each trait.

Trait	Roadmap Epigenome Name	-log10(p)	n_{GWAS}
Schizophrenia	Fetal Brain Male	15.477	82,315
Height	Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells	12.656	133,653
Rheumatoid Arthritis	GM12878 Lymphoblastoid Cells	10.713	58,284
Age at Menarche	Brain Germinal Matrix	8.169	132,989
Crohn's Disease	Primary T helper 17 cells PMA-I stimulated	6.504	20,883
Educational Attainment	Fetal Brain Female	6.307	101,069
BMI	Brain Germinal Matrix	5.829	123,865
HDL	Liver	5.316	99,900
Triglycerides	Liver	5.047	96,598
LDL	Liver	4.985	95,454
Coronary Artery Disease	Small Intestine	4.825	86,995
Ulcerative Colitis	Primary T helper 17 cells PMA-I stimulated	4.805	27,432
Alopecia Areata	Primary mononuclear cells from peripheral blood	4.630	7,776
IGAN	GM12878 Lymphoblastoid Cells	4.572	11,946
Epilepsy	Brain Angular Gyrus	4.222	34,853
Alzheimer's	Primary hematopoietic stem cells short term culture	3.659	54,162
Type2 Diabetes	Pancreatic Islets	3.626	69,033
Bipolar Disorder	Monocytes-CD14+ RO01746 Primary Cells	3.219	16,731
Ever Smoked	Brain Germinal Matrix	2.985	74,035
Fasting Glucose	Pancreatic Islets	2.746	58,074
Autism	HUES48 Cells	1.976	10,263



(A)

(B)

Figure 1. (A) Multidimensional scaling plot of the correlations between the functional scores for the different tissues. (B) Multidimensional scaling plot of the correlations between the functional scores for the primary cells and primary tissues, along with their embryonic tissue of origin.

Supplemental Material

Histone Marks. Supplemental Figure S1 shows various features of the four histone marks (H3K4me3, H3K4me1, H3K9ac, and H3K27ac) in the 127 cell types. The top left panel is a boxplot that shows the percent of the reference SNPs that lie within gapped peaks for the four annotations, across the different cell types. H3K4me1 appears to be a much more common mark than the others, followed by H3K27ac. The top middle panel shows the distribution of raw signal track values for variants that lie inside and outside gapped peaks for female fetal brain tissue. As they should, variants within gapped peaks tend to have higher raw signal track values. The top right panel is a boxplot that shows the mean signal track value for variants inside and outside gapped peaks for the 4 annotations, across the different cell types. H3K4me3 shows the largest difference in mean value between variants inside and outside gapped peaks; H3K4me1 the least. The bottom left panel shows the correlation matrix for signal track values for the 4 annotations for female fetal brain tissue. Pairwise correlations are highest for H3K27ac and H3K9ac, and for H3K9ac and H3K4me3. The bottom right panel is a boxplot that shows these pairwise correlations across the different cell types; the pairwise correlations for H3K27ac and H3K9ac, and for H3K9ac and H3K4me3 are generally the highest for all tissues.

Comparison of five different methods. Supplemental Figure S4 shows various features of the Eigen-PC and four different mixture models discussed in the Methods section, fit using these four annotations in the 127 cell types. The top left panel shows the weights in the 127 tissue-specific Eigen-PC models for each of the 4 histone modification marks. The weights are quite consistent across tissues, and are highest for H3K9ac, the annotation most highly correlated with the others, and lowest for H3K4me1, the annotation least correlated with the others. For the four

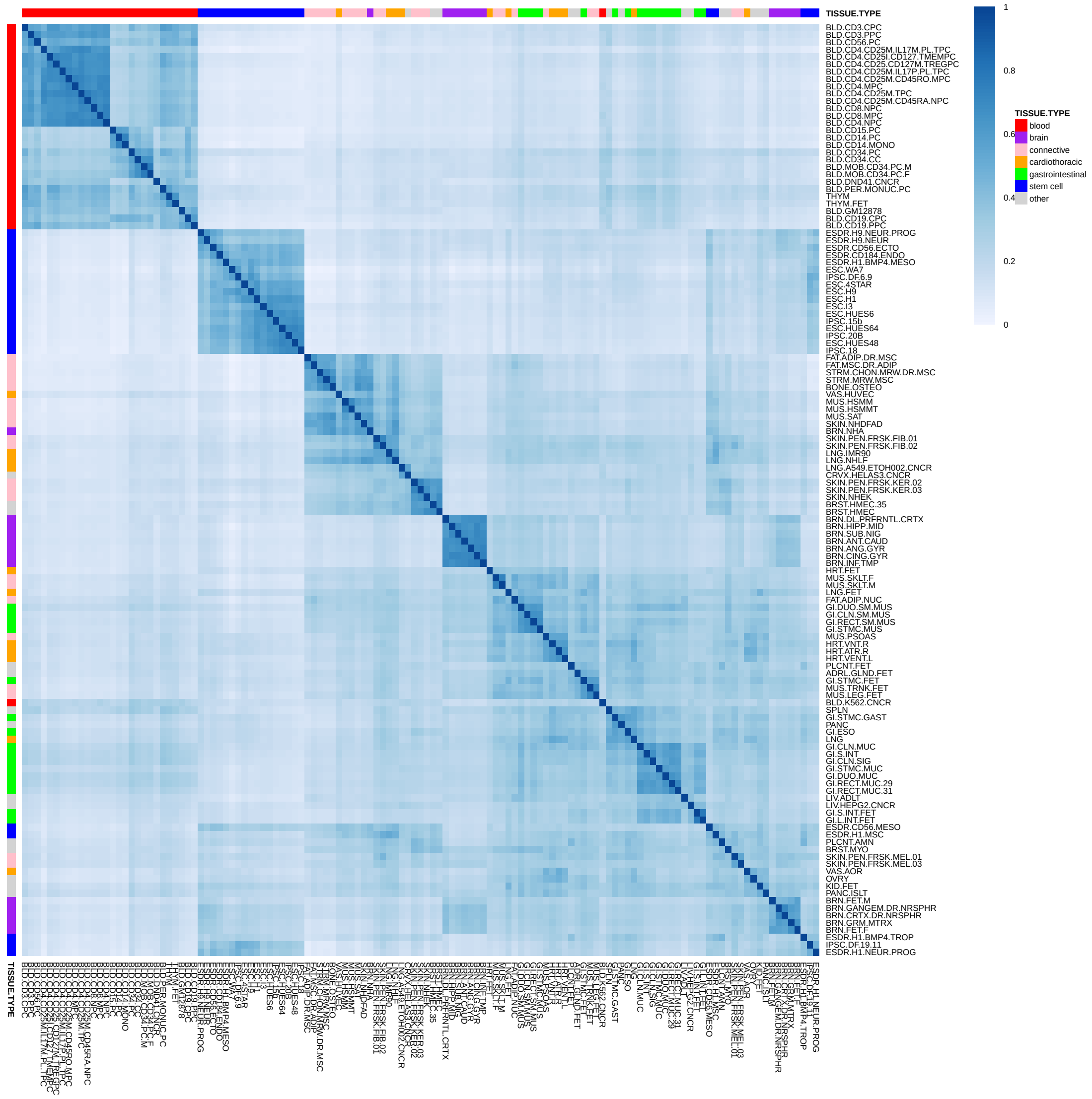


FIGURE 2. Jaccard index of overlap among functional variants in different cell types and tissues in Roadmap Epigenomics.

mixture methods with probabilistic interpretation, the top middle panel gives the distribution of the functional component probability for the 127 tissue types. The probability of the functional component is similar for the two binary methods and the npEM-quantile method, and much higher for the npEM-binned method. The top right panel shows the proportion of variants with posterior probabilities between 0.1 and 0.9. Only for the npEM-binned method is this proportion appreciable, indicating that the three other mixture models resemble binary classifiers, either classifying variants as functional or non-functional, for the most part without much uncertainty. The bottom left panel gives the distribution of the mean signal track value for each annotation for the two different components of the probabilistic mixture models across the 127 cell types. The functional and

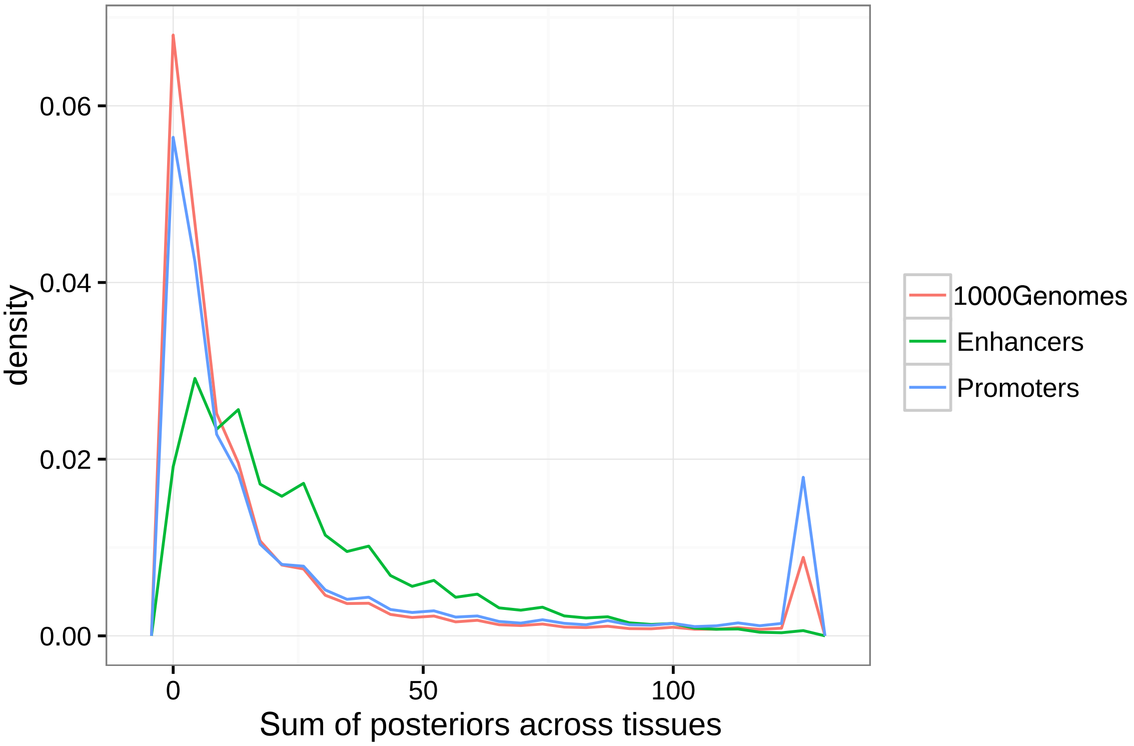


FIGURE 3. Distribution of total posterior probability (summed across all 127 tissues in Roadmap) for common variants in the 1000 Genomes project. Also shown are the distributions for variants in enhancers only, and promoters only.

non-functional components are much less separated in the npEM-binned models than they are in the other three models. The bottom right panel shows the distribution of pairwise correlations between the functional scores for the reference SNPs for the five different functional scores (posterior probabilities, in the case of the mixture models) across the different cell types. The binary methods and the npEM-quantile method give highly correlated posterior probabilities; the Eigen-PC score is less highly correlated with those three methods. The posterior probabilities from the npEM-binned method have low to modest correlation with the scores from the other methods. Overall, npEM-binned tends to perform poorly. We speculate that this is because there are technical artefacts, like chromatin accessibility, that cause weak correlations between histone modification measurements that are not related to function, in the range close to zero. With the regular binning used for the npEM-binned method, variants in this range close to zero represent most of the variants in the training set (see Figure S1) and so the npEM-binned mixture models largely reflect these artefactual correlations. On the other hand, with the binning procedure we use for the npEM-quantile method, functional variants with high histone modification measurements are enriched, and so the mixture models reflect the correlations caused by functional status reflected in these high modification measurements.

In Supplemental Figure S5 we show a multidimensional-scaling visualization of the correlations between the functional scores for a set of reference SNPs for 127 different tissues for the five methods. As shown, the two binary models, Eigen-PC and npEM produce similar patterns of correlations, while npEM-binned does not seem to perform well in separating the different types of tissues (see

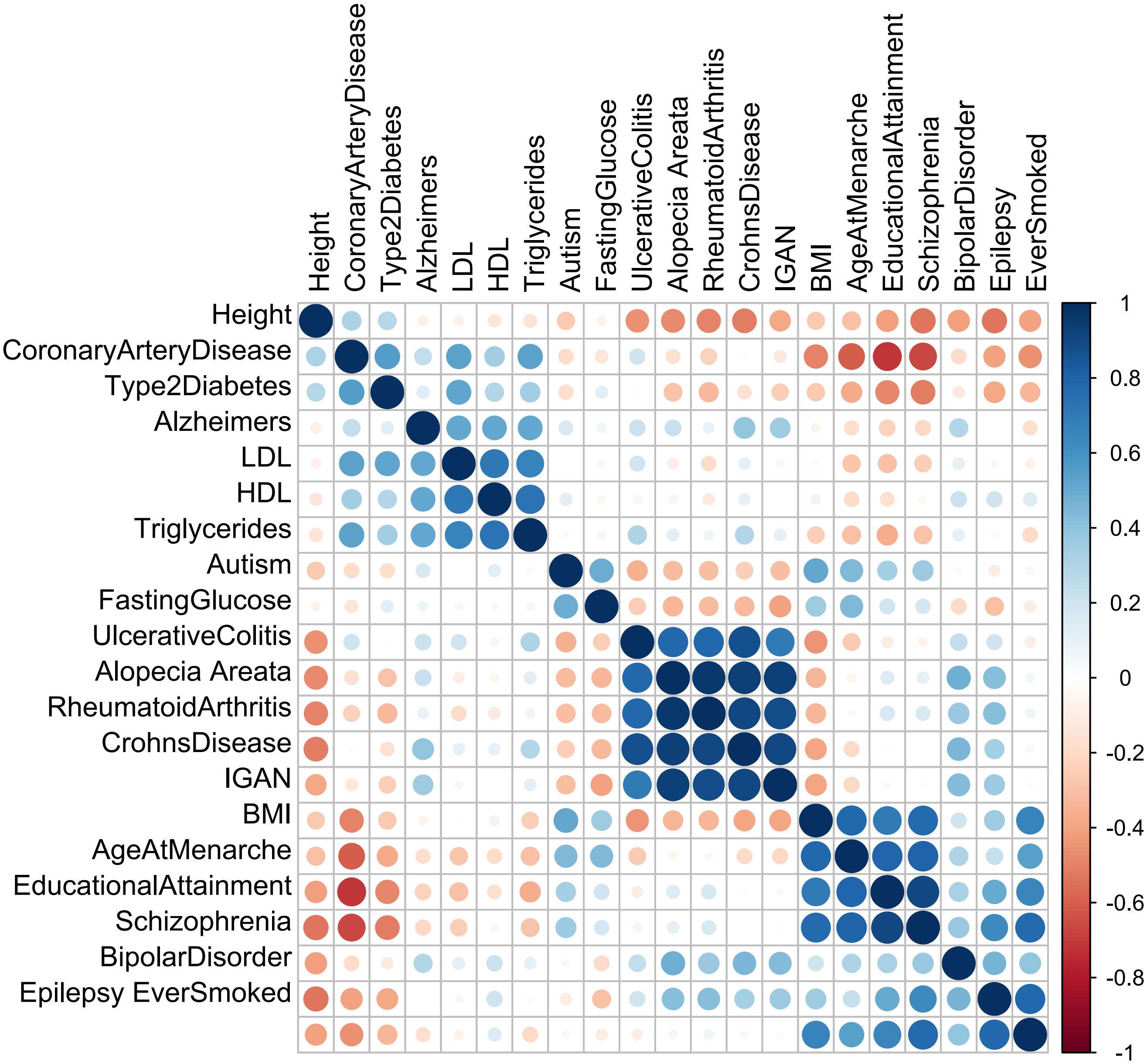


FIGURE 4. Tissue Correlations for 21 common traits.

also Supplemental Figure S6 showing correlations for the primary cells and primary tissues, along with their embryonic tissue of origin).

We have performed LD score regression by using functional scores derived from the four methods (we have excluded npEM-binned as it performs poorly in the analyses above). In Supplemental Figure S7 we show the top tissue for 21 complex traits using functional scores derived using each of the four methods, as well as the four individual histone marks. Overall, we find that npEM-quantile

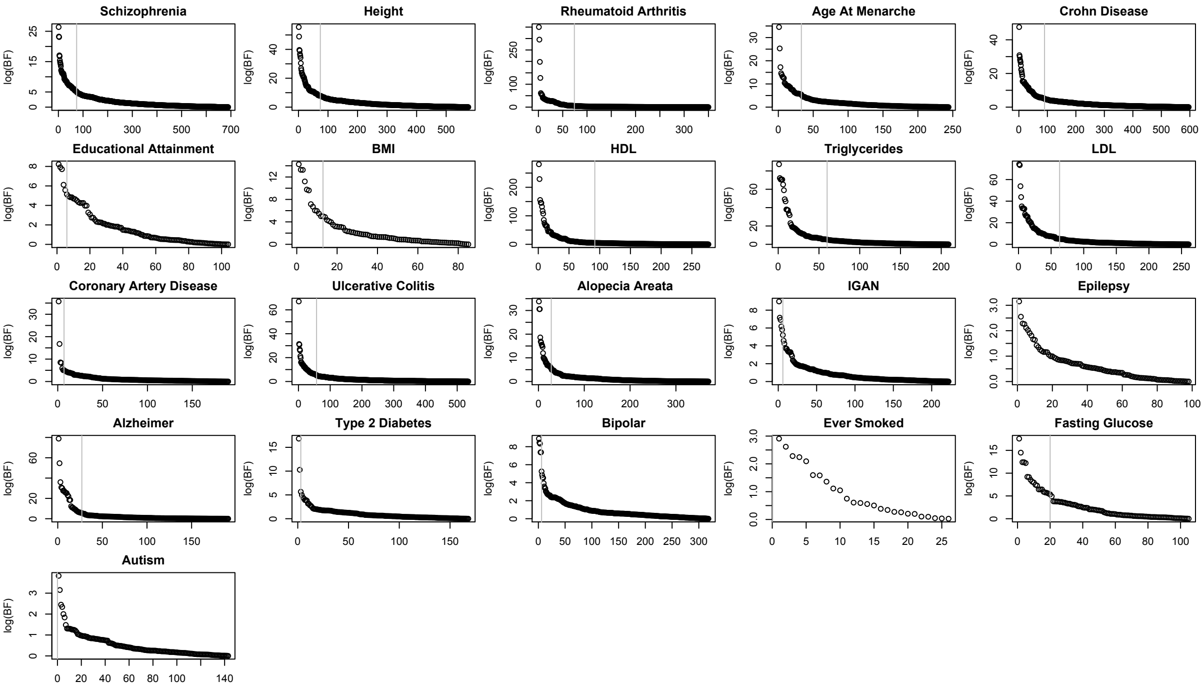


FIGURE 5. log(BF) for those genes with log(BF)> 0, for 21 common traits. The vertical gray line represents the log(BF) threshold of 5.

performs well and identifies plausible candidate tissues for most of the complex traits we analyzed. Although one can analyze the individual histone marks, the results can be difficult to interpret as different histone marks can implicate different tissues. For example, for "T2Diabetes", pancreas is the top tissue only for npEM-quantile, mvB and H3K9ac; the other three histone marks identify different tissues as the top tissue.

LD score regression to identify the tissue of interest. The LD score regression approach [18] uses two sets of SNPs, reference SNPs and regression SNPs. The regression SNPs are SNPs that are used in a regression of χ^2 statistics from GWAS studies against the "LD scores" of those regression SNPs. The LD score of a regression SNP is a numeric score which captures the functional effects of all reference SNPs in LD with that SNP, appropriately weighted to account for the extent of the LD. Here, following [18] we use as regression SNPs HapMap3 SNPs, chosen for their high imputation quality, and as reference SNPs those SNPs with minor allele count greater than 5 in the 379 European samples from the 1000 Genome Project [15]. We first compute tissue-specific scores using each of our methods for the $9,254,335$ SNPs with minor allele count greater than 5 in the 379 European samples from the 1000 Genomes Project, which we will subsequently use as our "reference SNPs" for LD score regression.

In the LD score regression approach, a linear model is used to model a quantitative phenotype y_i for an individual i:

$$y_i = \sum_{j \in G} X_{ij}\beta_j + \epsilon_{ij}.$$

Here G is some set of SNPs, X_{ij} is the genotype of individual i at SNP j, and β_j is the effect size of SNP j. In this framework, $\boldsymbol{\beta}$, the vector of all the β_j, is modeled as a mean-0 random vector with independent entries, and the variance of β_j depends on the functional categories included in the model. We have a set of functional categories $C_1, \ldots, C_C$, and the variance of a SNP's effect size will depend on which functional categories it belongs to:

$$\mathrm{Var}(\beta_j) = \sum_{c:j \in C_c} \tau_c.$$

Here τ_c is the contribution of SNPs in category C_c to the variance of effect sizes of SNPs in that category. In [18], the authors show that τ_c can be estimated through the following equation:

$$E[\chi_j^2] = N \sum_c \tau_c l(j, c) + 1.$$

Here χ_j^2 is the chi-squared statistic for SNP j from a GWAS study, N is the sample size from that study, and $l(j, c)$ is the LD score of SNP j with respect to category c, $l(j, c) = \sum_{k \in C_c} r_{jk}^2$. Here r_{jk}^2 is the correlation between SNP j and a SNP k in category C_c, so that the sum in the above equation is over categories, with τ_c weighted by the sum of correlations between SNP j and all SNPs in category C_c. This equation therefore allows for the estimation of the τ_c via the regression of the chi-squared statistics from a GWAS study on the LD scores of the regression SNPs.

Here, we extend the LD score by allowing SNPs to be assigned to a category C_c probabilistically, that is, we assume a probability p_{kc} that SNP k belongs to category C_c, and therefore that the variance of its effect size is affected by its membership in that category. This only involves minor changes to the above equations, namely, we have that

$$\mathrm{Var}(\beta_j) = \sum_{c:j \in C_c} p_{jc} \tau_c,$$

where p_{jc} is the probability that SNP j belongs to category C_c, and as above

$$E[\chi_j^2] = N \sum_c \tau_c l(j, c) + 1,$$

although now $l(j, c) = \sum_{k \in C_c} p_{kc} r_{jk}^2$, p_{kc} being the probability that SNP k belongs to category C_c. We can therefore still estimate the τ_c via the regression of the chi-squared statistics from a GWAS study on the LD scores of the regression SNPs, but in calculating these LD scores we weight the correlation of a SNP k with a regression SNP j by the probability that SNP k belongs to a particular category.

For each tissue and phenotype, and each of our functional scores, we fit a separate LD score regression model, including the LD score derived using the posterior probability each regression SNP is in the functional component in that tissue, to estimate the contribution of that component to the variability of effect sizes of SNPs that belong to that component. To control for overlap of the tissue-specific functional score with other functional categories, we use the same 54 baseline categories used in [18], which represent various non-tissue-specific annotations, including histone modification measurements combined across tissues, measurements of open chromatin, and super enhancers.

Note that for the Eigen-PC approach, the derivation above does not strictly apply, but we still interpret the coefficient τ estimated in the LD score regression as the contribution of the component represented by the Eigen-PC approach to the variability of effect sizes of SNPs in that category.

References

[1] Regier DA, Narrow WE, Rae DS, Manderscheid RW, Locke BZ, Goodwin FK (1993) The de facto mental and addictive disorders service system. Epidemiologic Catchment Area prospective 1-year prevalence rates of disorders and services. *Archives of General Psychiatry* 2: 85–94.

[2] Myasoedova E, Crowson CS, Kremers HM, Therneau TM, Gabriel SE (2010) Is the incidence of rheumatoid arthritis rising? Results from Olmsted County, Minnesota, 1955-2007 *Arthritis Rheum.* 62: 1576–1582.

[3] Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris B, Chernoff G, Benchimol EI, Panaccione R, Ghosh S, Barkema HW, Kaplan GG (2012) Increasing Incidence and Prevalence of the Inflammatory Bowel Diseases With Time, Based on Systematic Review *Gastroenterology* 142: 46–54.

[4] Fricke ACV, Miteva M (2015) Epidemiology and burden of alopecia areata: a systematic review *Clinical, Cosmetic and Investigational Dermatology* 8: 397–403.

[5] Magistroni R, D'Agati VD, Appel GB, Kiryluk K (2015) New developments in the genetics, pathogenesis, and therapy of IgA nephropathy. *Kidney Int* 88: 974–989.

[6] Institute of Medicine of the National Academies Epilepsy across the spectrum (2012)

[7] Alzheimer's Association (2015) 2015 Alzheimer's Disease Facts and figures *Alzheimer's and Dementia* 11: 332

[8] CDC National Diabetes Statistics Report (2014)

[9] Kessler RC, Chiu WT, Demler O, Walters EE (2005) Prevalence, Severity, and Comorbidity of Twelve-month DSM-IV Disorders in the National Comorbidity Survey Replication (NCS- R) *Archives of General Psychiatry* 62: 617–627.

[10] Christensen DL, Baio J, Braun KV, et al. (2016) Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years ? Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2012. *MMWR Surveill Summ* 2016:1–23.

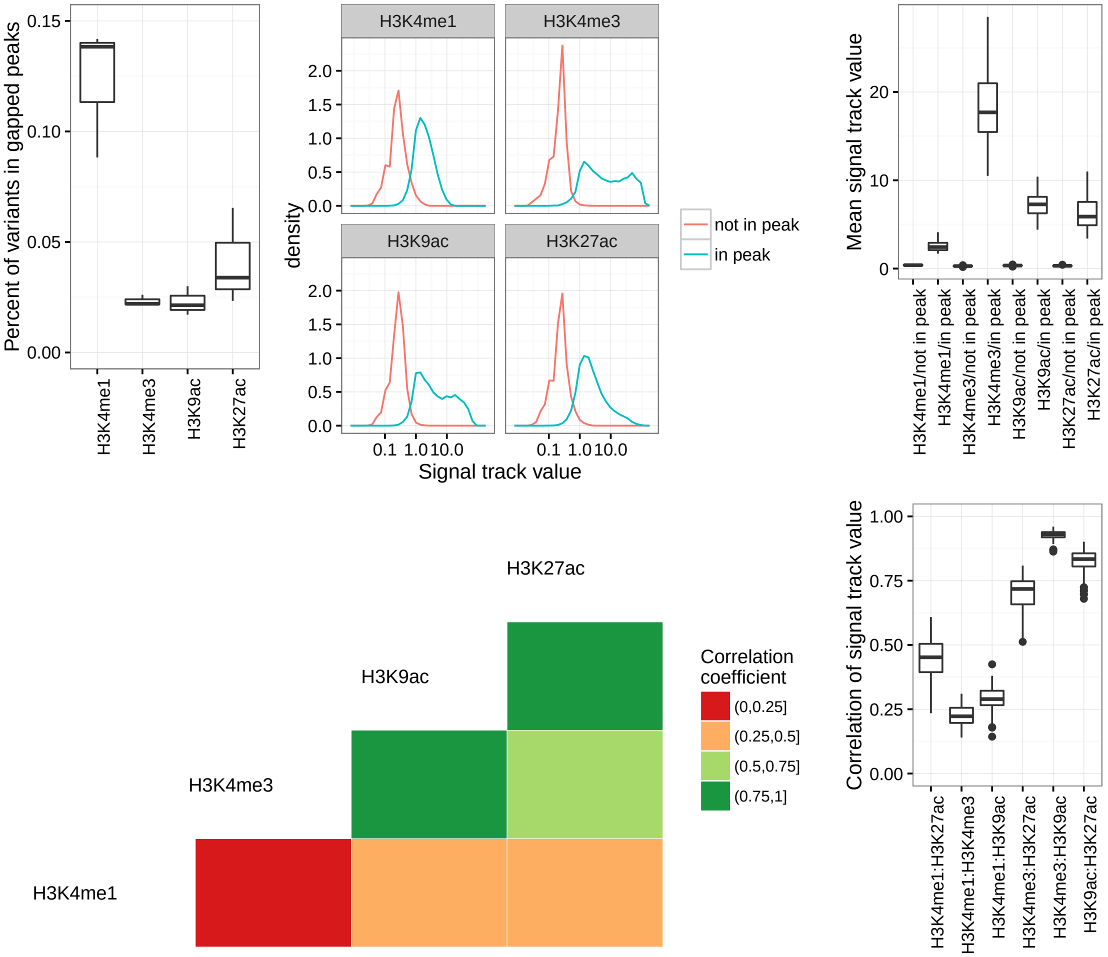


Figure S1. Four histone marks in 127 cell types. The top left panel is a boxplot that shows the percent of the reference SNPs that lie within gapped peaks for the 4 annotations, across the different cell types. The top middle panel shows the distribution of raw signal track values for variants that lie inside and outside gapped peaks for female fetal brain tissue. The top right panel is a boxplot that shows the mean signal track value for variants inside and outside gapped peaks for the 4 annotations, across the different cell types. The bottom left panel shows the correlation matrix for signal track values for the 4 annotations for female fetal brain tissue. The bottom right panel is a boxplot that shows these pairwise correlations across the different cell types.

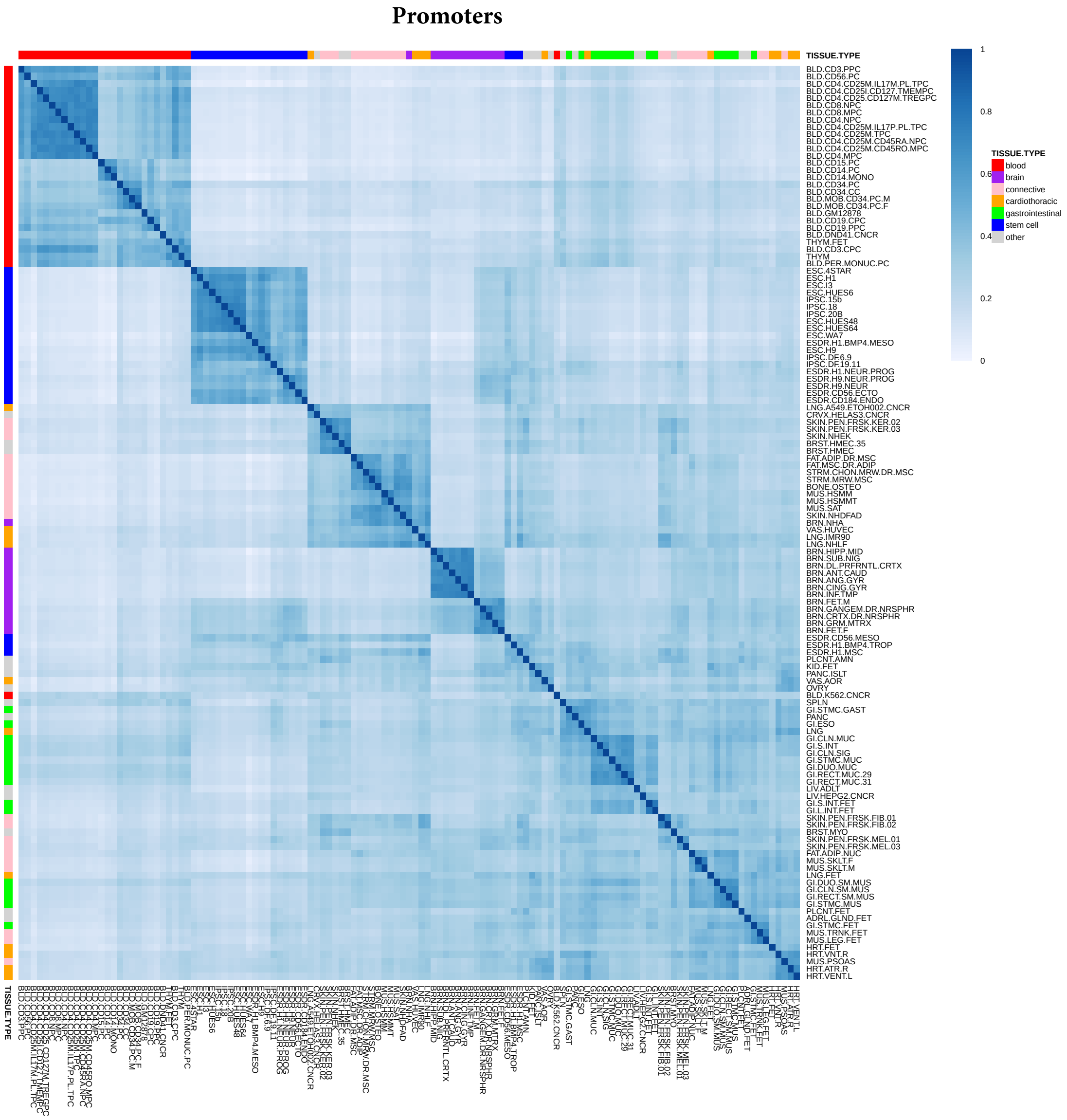


FIGURE S2. Jaccard index of overlap among functional variants falling in promoter regions in different cell types and tissues in Roadmap.

Enhancers

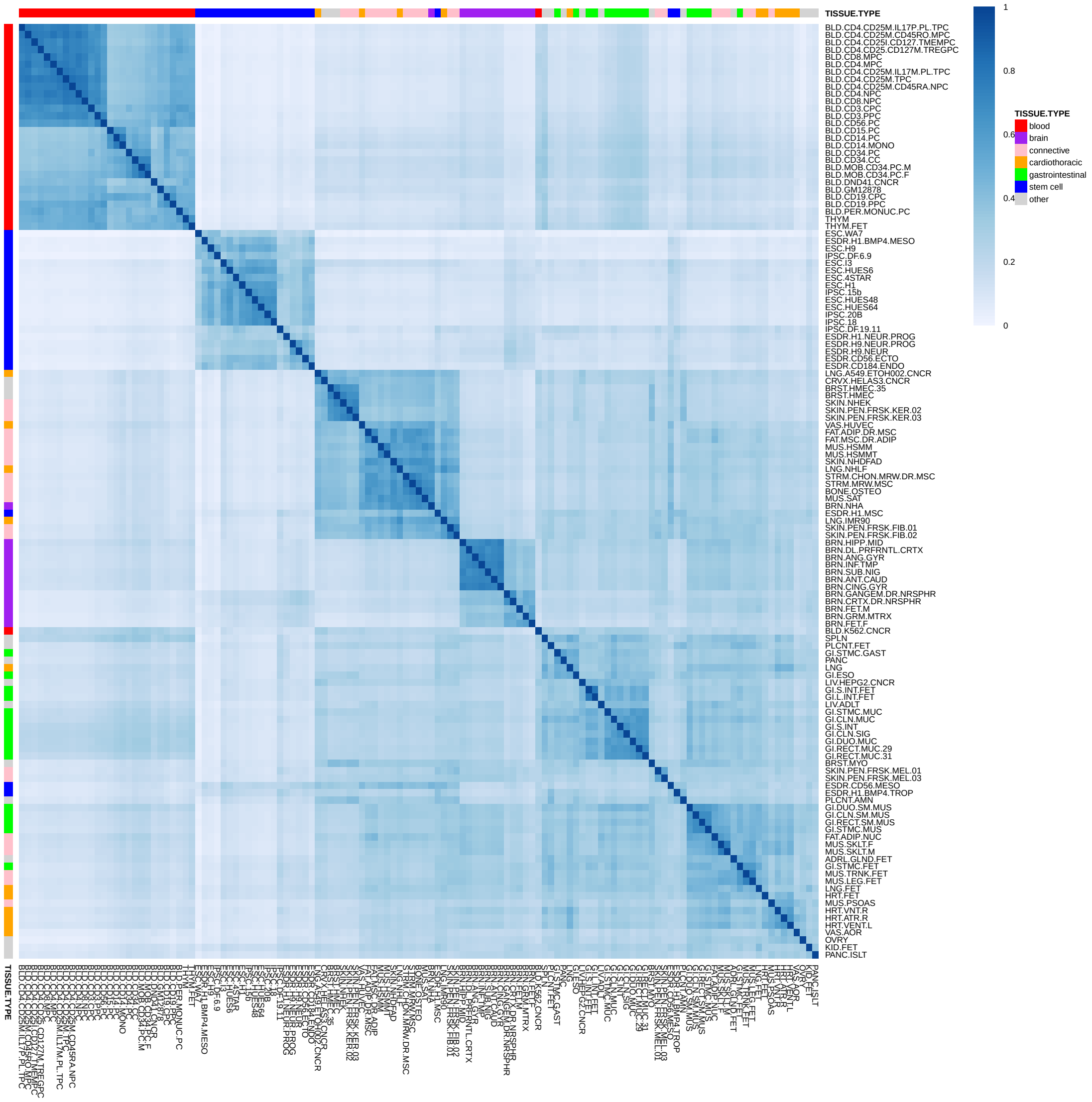


FIGURE S3. Jaccard index of overlap among functional variants falling in enhancer regions in different cell types and tissues in Roadmap.

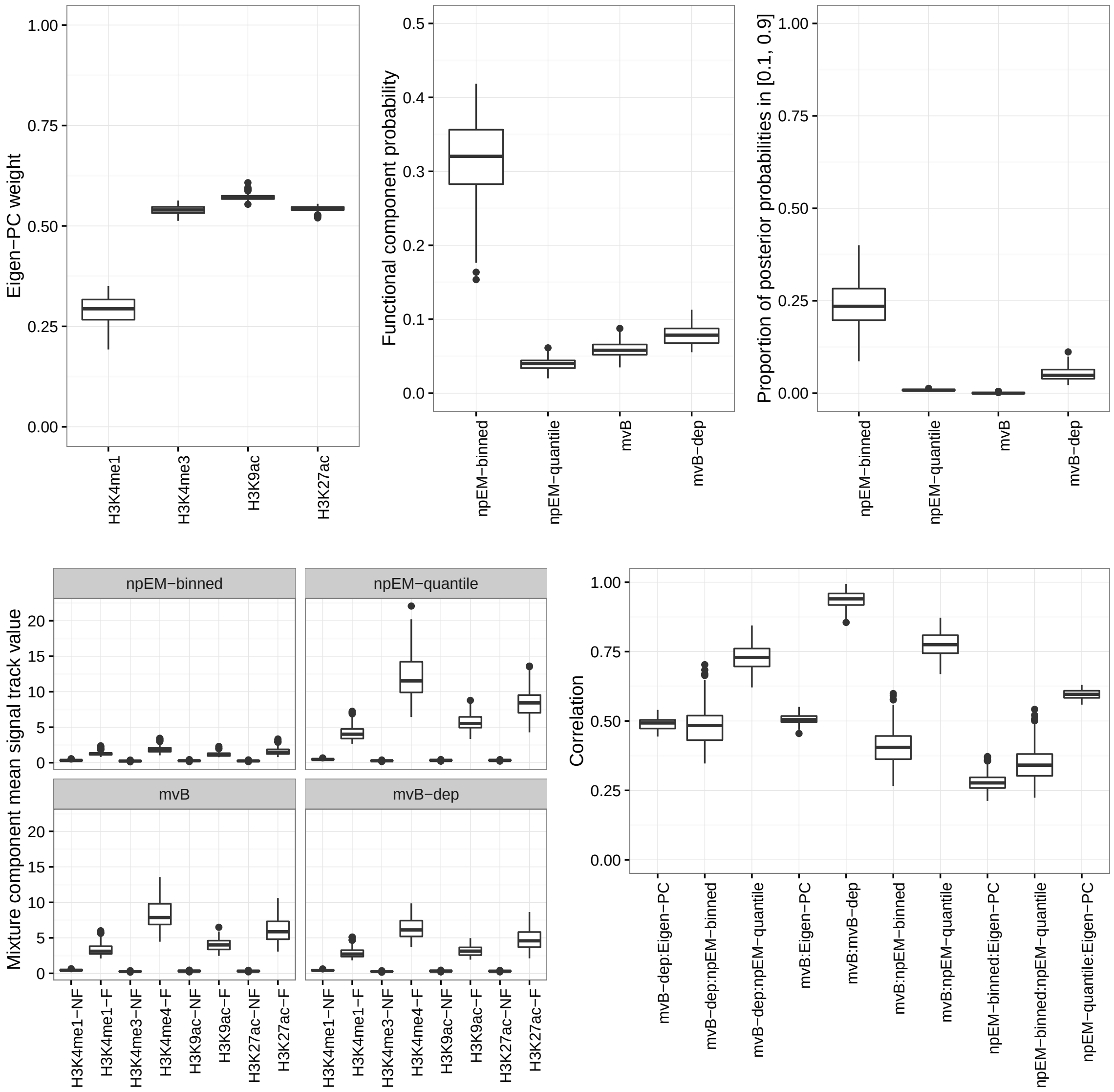


Figure S4. Eigen-PC and 4 mixture models. The top left panel shows the weights in the 127 tissue-specific Eigen-PC models for each of the 4 histone modification marks. For the four mixture methods with probabilistic interpretation, the top middle panel gives the distribution of the functional component probability for the 127 tissue types. The top right panel shows the proportion of variants with posterior probabilities between 0.1 and 0.9. The bottom left panel gives the distribution of the mean signal track value for each annotation for the two different components of the probabilistic mixture models across the 127 cell types. The bottom right panel shows the distribution of pairwise correlations between the functional scores (posterior probabilities, in the case of the mixture models) for the reference SNPs for the five different methods across the different cell types.

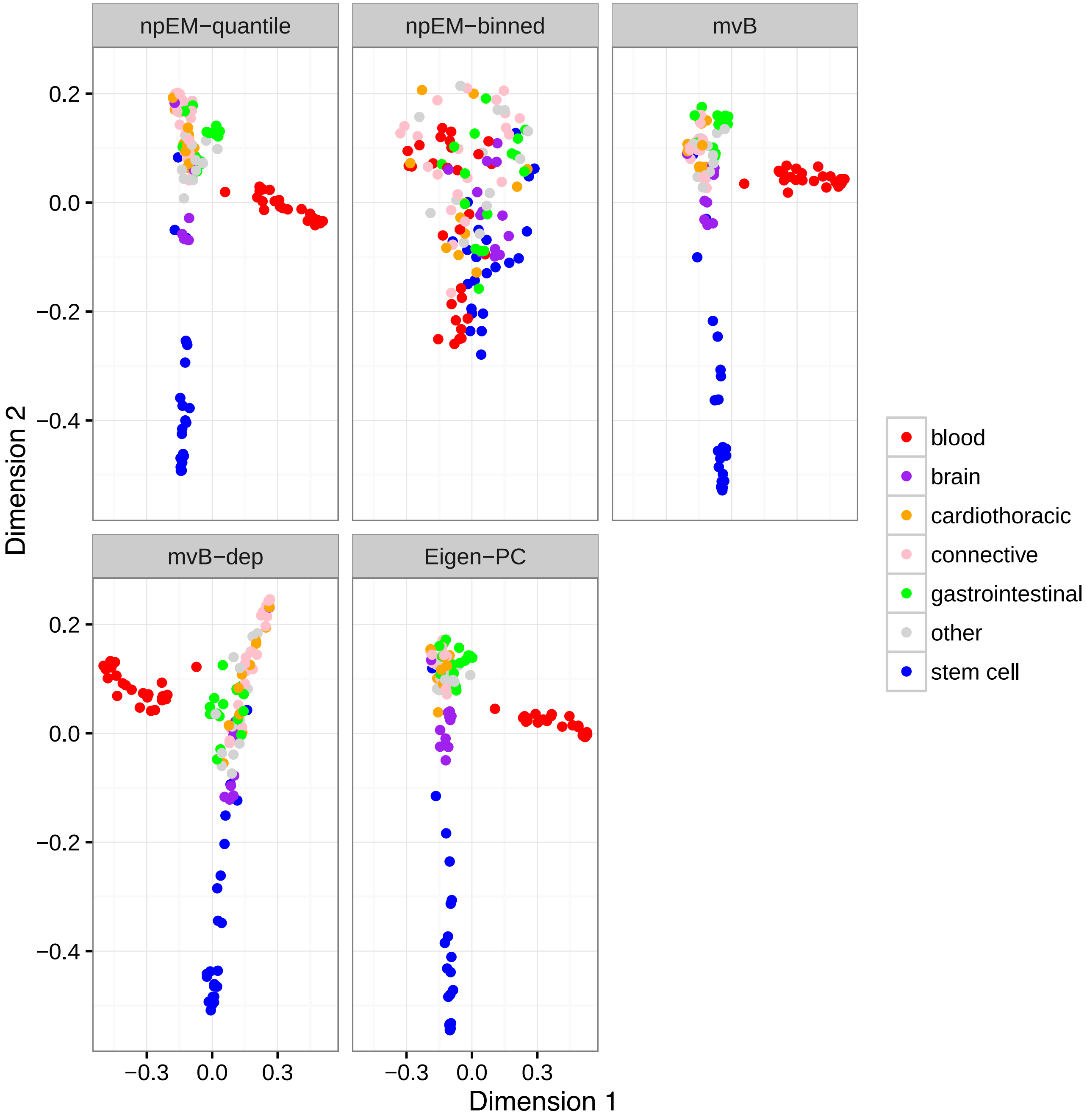


FIGURE S5. Multidimensional scaling plot of the correlations between the functional scores for the different tissues, for five methods.

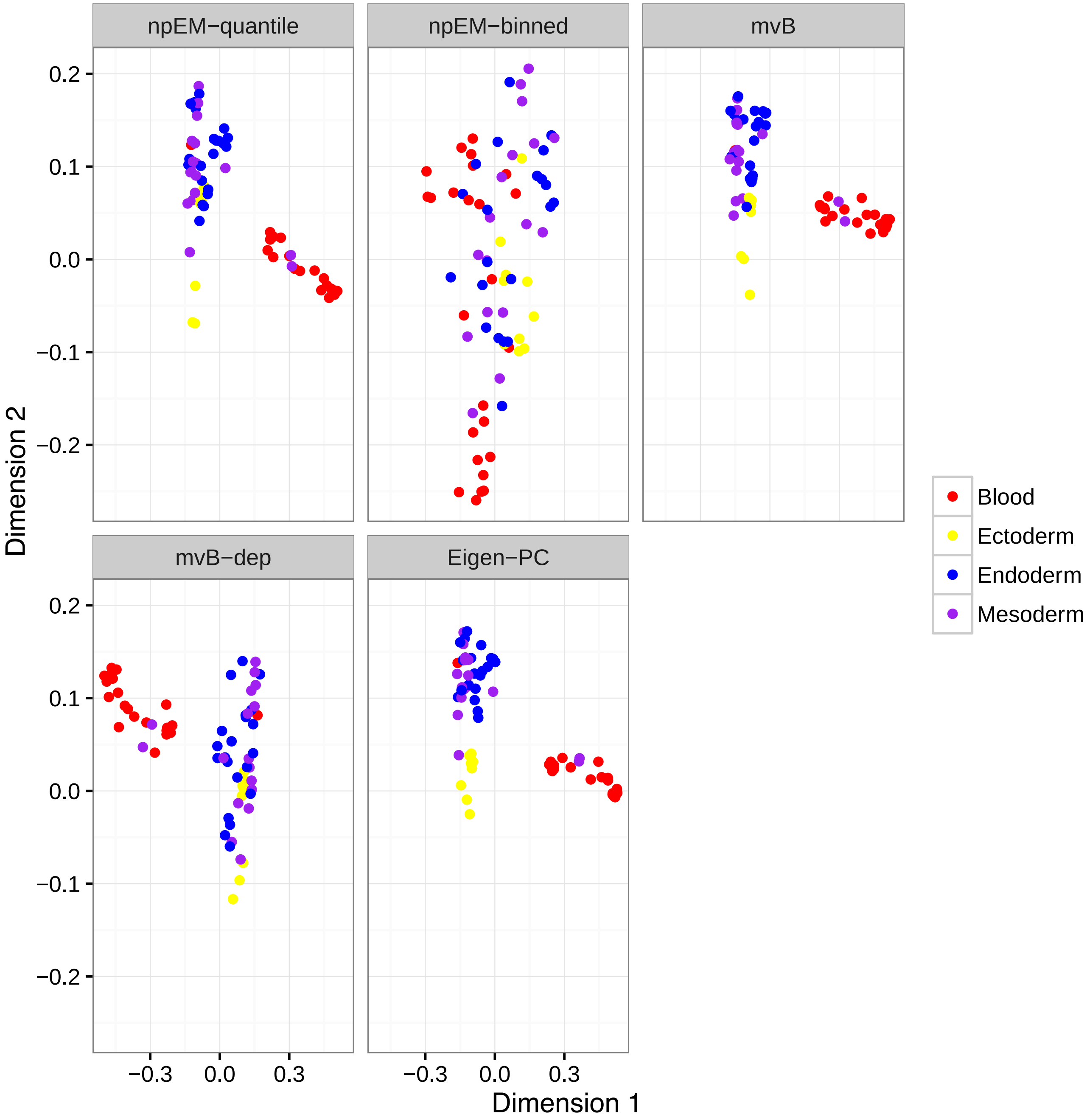


FIGURE S6. Multidimensional scaling plot of the correlations between the functional scores for the primary cells and primary tissues, along with their embryonic tissue of origin, for the five methods.

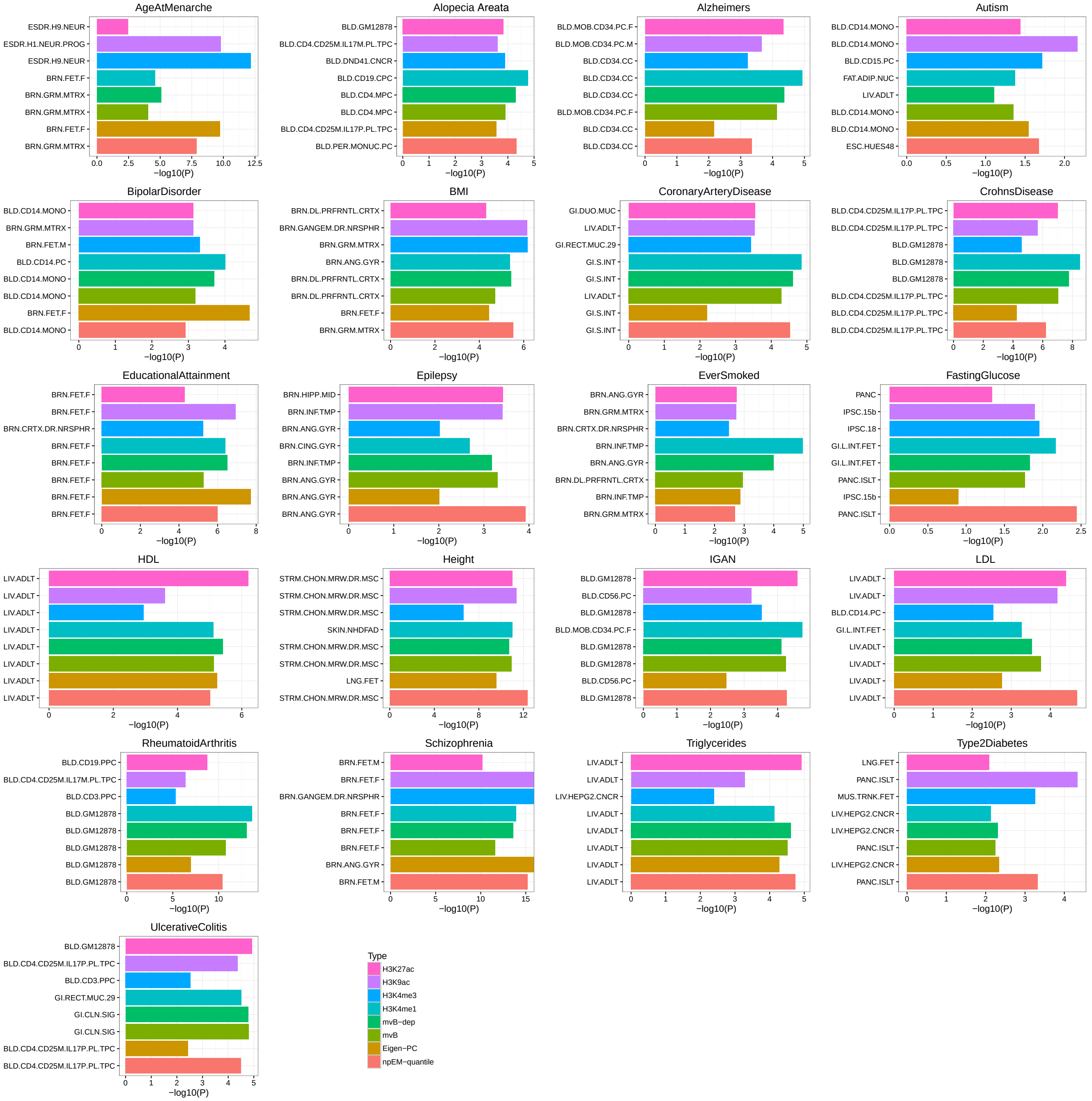


FIGURE S7. Top tissue-/cell-type using each of the four individual histone marks, as well as four different combination methods.

Table S1. Tissues and Cell Types in Roadmap (part 1)

Epigenome.ID	Epigenome.Mnemonic	Standardized.Epigenome.name
E017	LNG.IMR90	IMR90 fetal lung fibroblasts Cell Line
E002	ESC.WA7	ES-WA7 Cells
E008	ESC.H9	H9 Cells
E001	ESC.I3	ES-I3 Cells
E015	ESC.HUES6	HUES6 Cells
E014	ESC.HUES48	HUES48 Cells
E016	ESC.HUES64	HUES64 Cells
E003	ESC.H1	H1 Cells
E024	ESC.4STAR	ES-UCSF4 Cells
E020	IPSC.20B	iPS-20b Cells
E019	IPSC.18	iPS-18 Cells
E018	IPSC.15b	iPS-15b Cells
E021	IPSC.DF.6.9	iPS DF 6.9 Cells
E022	IPSC.DF.19.11	iPS DF 19.11 Cells
E007	ESDR.H1.NEUR.PROG	H1 Derived Neuronal Progenitor Cultured Cells
E009	ESDR.H9.NEUR.PROG	H9 Derived Neuronal Progenitor Cultured Cells
E010	ESDR.H9.NEUR	H9 Derived Neuron Cultured Cells
E013	ESDR.CD56.MESO	hESC Derived CD56+ Mesoderm Cultured Cells
E012	ESDR.CD56.ECTO	hESC Derived CD56+ Ectoderm Cultured Cells
E011	ESDR.CD184.ENDO	hESC Derived CD184+ Endoderm Cultured Cells
E004	ESDR.H1.BMP4.MESO	H1 BMP4 Derived Mesendoderm Cultured Cells
E005	ESDR.H1.BMP4.TROP	H1 BMP4 Derived Trophoblast Cultured Cells
E006	ESDR.H1.MSC	H1 Derived Mesenchymal Stem Cells
E062	BLD.PER.MONUC.PC	Primary mononuclear cells fromperipheralblood
E034	BLD.CD3.PPC	Primary T cells fromperipheralblood
E045	BLD.CD4.CD25I.CD127.TMEMPC	Primary T cells effector/memory enriched from peripheral blood
E033	BLD.CD3.CPC	Primary T cells from cord blood
E044	BLD.CD4.CD25.CD127M.TREGPC	Primary T regulatory cells fromperipheralblood
E043	BLD.CD4.CD25M.TPC	Primary T helper cells fromperipheralblood
E039	BLD.CD4.CD25M.CD45RA.NPC	Primary T helper naive cells fromperipheralblood
E041	BLD.CD4.CD25M.IL17M.PL.TPC	Primary T helper cells PMA-I stimulated
E042	BLD.CD4.CD25M.IL17P.PL.TPC	Primary T helper 17 cells PMA-I stimulated
E040	BLD.CD4.CD25M.CD45RO.MPC	Primary T helper memory cells from peripheral blood 1
E037	BLD.CD4.MPC	Primary T helper memory cells from peripheral blood 2
E048	BLD.CD8.MPC	Primary T CD8+ memory cells from peripheral blood
E038	BLD.CD4.NPC	Primary T helper naive cells from peripheral blood
E047	BLD.CD8.NPC	Primary T CD8+ naive cells from peripheral blood
E029	BLD.CD14.PC	Primary monocytes fromperipheralblood
E031	BLD.CD19.CPC	Primary B cells from cord blood
E035	BLD.CD34.PC	Primary hematopoietic stem cells
E051	BLD.MOB.CD34.PC.M	Primary hematopoietic stem cells G-CSF-mobilized Male
E050	BLD.MOB.CD34.PC.F	Primary hematopoietic stem cells G-CSF-mobilized Female
E036	BLD.CD34.CC	Primary hematopoietic stem cells short term culture
E032	BLD.CD19.PPC	Primary B cells from peripheral blood
E046	BLD.CD56.PC	Primary Natural Killer cells fromperipheralblood
E030	BLD.CD15.PC	Primary neutrophils fromperipheralblood
E026	STRM.MRW.MSC	Bone Marrow Derived Cultured Mesenchymal Stem Cells
E049	STRM.CHON.MRW.DR.MSC	Mesenchymal Stem Cell Derived Chondrocyte Cultured Cells
E025	FAT.ADIP.DR.MSC	Adipose Derived Mesenchymal Stem Cell Cultured Cells
E023	FAT.MSC.DR.ADIP	Mesenchymal Stem Cell Derived Adipocyte Cultured Cells
E052	MUS.SAT	Muscle Satellite Cultured Cells
E055	SKIN.PEN.FRSK.FIB.01	Foreskin Fibroblast Primary Cells skin01
E056	SKIN.PEN.FRSK.FIB.02	Foreskin Fibroblast Primary Cells skin02
E059	SKIN.PEN.FRSK.MEL.01	Foreskin Melanocyte Primary Cells skin01
E061	SKIN.PEN.FRSK.MEL.03	Foreskin Melanocyte Primary Cells skin03
E057	SKIN.PEN.FRSK.KER.02	Foreskin Keratinocyte Primary Cells skin02
E058	SKIN.PEN.FRSK.KER.03	Foreskin Keratinocyte Primary Cells skin03
E028	BRST.HMEC.35	Breast variant Human Mammary Epithelial Cells (vHMEC)
E027	BRST.MYO	Breast Myoepithelial Primary Cells
E054	BRN.GANGEM.DR.NRSPHR	Ganglion Eminence derived primary cultured neurospheres
E053	BRN.CRTX.DR.NRSPHR	Cortex derived primary cultured neurospheres
E112	THYM	Thymus
E093	THYM.FET	Fetal Thymus

Table S2. Tissues and Cell Types in Roadmap (part 2)

Epigenome.ID	Epigenome.Mnemonic	Standardized.Epigenome.name
E071	BRN.HIPP.MID	Brain Hippocampus Middle
E074	BRN.SUB.NIG	Brain Substantia Nigra
E068	BRN.ANT.CAUD	Brain Anterior Caudate
E069	BRN.CING.GYR	Brain Cingulate Gyrus
E072	BRN.INF.TMP	Brain Inferior Temporal Lobe
E067	BRN.ANG.GYR	Brain Angular Gyrus
E073	BRN.DL.PRFRNTL.CRTX	Brain_Dorsolateral_Prefrontal_Cortex
E070	BRN.GRM.MTRX	Brain Germinal Matrix
E082	BRN.FET.F	Fetal Brain Female
E081	BRN.FET.M	Fetal Brain Male
E063	FAT.ADIP.NUC	Adipose Nuclei
E100	MUS.PSOAS	Psoas Muscle
E108	MUS.SKLT.F	Skeletal Muscle Female
E107	MUS.SKLT.M	Skeletal Muscle Male
E089	MUS.TRNK.FET	Fetal Muscle Trunk
E090	MUS.LEG.FET	Fetal Muscle Leg
E083	HRT.FET	Fetal Heart
E104	HRT.ATR.R	Right Atrium
E095	HRT.VENT.L	Left Ventricle
E105	HRT.VNT.R	Right Ventricle
E065	VAS.AOR	Aorta
E078	GI.DUO.SM.MUS	Duodenum Smooth Muscle
E076	GI.CLN.SM.MUS	Colon Smooth Muscle
E103	GI.RECT.SM.MUS	Rectal Smooth Muscle
E111	GI.STMC.MUS	Stomach Smooth Muscle
E092	GI.STMC.FET	Fetal Stomach
E085	GI.S.INT.FET	Fetal Intestine Small
E084	GI.L.INT.FET	Fetal Intestine Large
E109	GI.S.INT	Small Intestine
E106	GI.CLN.SIG	Sigmoid Colon
E075	GI.CLN.MUC	Colonic Mucosa
E101	GI.RECT.MUC.29	Rectal Mucosa Donor 29
E102	GI.RECT.MUC.31	Rectal Mucosa Donor 31
E110	GI.STMC.MUC	Stomach Mucosa
E077	GI.DUO.MUC	Duodenum Mucosa
E079	GI.ESO	Esophagus
E094	GI.STMC.GAST	Gastric
E099	PLCNT.AMN	Placenta Amnion
E086	KID.FET	Fetal Kidney
E088	LNG.FET	Fetal Lung
E097	OVRY	Ovary
E087	PANC.ISLT	Pancreatic Islets
E080	ADRL.GLND.FET	Fetal Adrenal Gland
E091	PLCNT.FET	Placenta
E066	LIV.ADLT	Liver
E098	PANC	Pancreas
E096	LNG	Lung
E113	SPLN	Spleen
E114	LNG.A549.ETOH002.CNCR	A549 EtOH 0.02pct Lung Carcinoma Cell Line
E115	BLD.DND41.CNCR	Dnd41 TCell Leukemia Cell Line
E116	BLD.GM12878	GM12878 Lymphoblastoid Cells
E117	CRVX.HELAS3.CNCR	HeLa-S3 Cervical Carcinoma Cell Line
E118	LIV.HEPG2.CNCR	HepG2 Hepatocellular Carcinoma Cell Line
E119	BRST.HMEC	HMEC Mammary Epithelial Primary Cells
E120	MUS.HSMM	HSMM Skeletal Muscle Myoblasts Cells
E121	MUS.HSMMT	HSMM cell derived Skeletal Muscle Myotubes Cells
E122	VAS.HUVEC	HUVEC Umbilical Vein Endothelial Primary Cells
E123	BLD.K562.CNCR	K562 Leukemia Cells
E124	BLD.CD14.MONO	Monocytes-CD14+ RO01746 Primary Cells
E125	BRN.NHA	NH-A Astrocytes Primary Cells
E126	SKIN.NHDFAD	NHDF-Ad Adult Dermal Fibroblast Primary Cells
E127	SKIN.NHEK	NHEK-Epidermal Keratinocyte Primary Cells
E128	LNG.NHLF	NHLF Lung Fibroblast Primary Cells
E129	BONE.OSTEO	Osteoblast Primary Cells

Table S3. GTEx Tissues

Tissue	Sample size
Muscle - Skeletal	361
Whole Blood	338
Skin - Sun Exposed (Lower leg)	302
Adipose - Subcutaneous	298
Artery - Tibial	285
Lung	278
Thyroid	278
Cells - Transformed fibroblasts	272
Nerve - Tibial	256
Esophagus - Mucosa	241
Esophagus - Muscularis	218
Artery - Aorta	197
Skin - Not Sun Exposed (Suprapubic)	196
Heart - Left Ventricle	190
Adipose - Visceral (Omentum)	185
Breast - Mammary Tissue	183
Stomach	170
Colon - Transverse	169
Heart - Atrial Appendage	159
Testis	157
Pancreas	149
Esophagus - Gastroesophageal Junction	127
Adrenal Gland	126
Colon - Sigmoid	124
Artery - Coronary	118
Cells - EBV-transformed lymphocytes	114
Brain - Cerebellum	103
Brain - Caudate (basal ganglia)	100
Liver	97
Brain - Cortex	96
Brain - Nucleus accumbens (basal ganglia)	93
Brain - Frontal Cortex (BA9)	92
Brain - Cerebellar Hemisphere	89
Spleen	89
Pituitary	87
Prostate	87
Ovary	85
Brain - Putamen (basal ganglia)	82
Brain - Hippocampus	81
Brain - Hypothalamus	81
Vagina	79
Small Intestine - Terminal Ileum	77
Brain - Anterior cingulate cortex (BA24)	72
Uterus	70
Brain - Amygdala	62
Brain - Spinal cord (cervical c-1)	59
Brain - Substantia nigra	56
Minor Salivary Gland	51
Kidney - Cortex	26
Bladder	11
Cervix - Ectocervix	6
Fallopian Tube	6
Cervix - Endocervix	5

Table S4. Breakdown of SNPs with p value $< 10^{-6}$ in different positional categories: intron, exon, promoter, enhancer in top tissue, enhancer in any Roadmap tissue, and within 200 kb of a gene. A SNP is allowed to be in multiple categories, e.g. in an intron for one gene, and in the promoter or enhancer of another gene. More details on the definition of these regions are given in the Methods section.

Trait	Type	n.intron	%.intron	n.exon	%.exon	n.prom	%.prom	n.enh	%.enh	n.enh.any.tissue	%.enh.any.tis	n.closest.gene	%.closest.gene	n.SNPS
AgeAtMenarche	All	1600	0.4282	104	0.0278	94	0.0252	7	0.0019	55	0.0147	1469	0.3931	3737
AgeAtMenarche	Functional	84	0.5833	6	0.0417	10	0.0694	4	0.0278	8	0.0556	41	0.2847	144
Alopecia	All	136	0.0950	44	0.0307	54	0.0377	2	0.0014	19	0.0133	175	0.1222	1432
Alopecia	Functional	23	0.3651	9	0.1429	9	0.1429	2	0.0317	5	0.0794	24	0.3810	63
Alzheimer's	All	752	0.4769	137	0.0869	213	0.1351	33	0.0209	73	0.0463	405	0.2568	1577
Alzheimer's	Functional	140	0.5645	37	0.1492	31	0.1250	30	0.1210	37	0.1492	50	0.2016	248
Autism	All	5	1.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	5
Autism	Functional	0	NA	0	NA	0	NA	0	NA	0	NA	0	NA	NA
BipolarDisorder	All	71	0.4465	10	0.0629	7	0.0440	0	0.0000	4	0.0252	76	0.4780	159
BipolarDisorder	Functional	7	0.6364	1	0.0909	0	0.0000	0	0.0000	0	0.0000	3	0.2727	11
BMI	All	429	0.3679	31	0.0266	37	0.0317	4	0.0034	19	0.0163	502	0.4305	1166
BMI	Functional	25	0.6410	3	0.0769	4	0.1026	1	0.0256	2	0.0513	8	0.2051	39
CoronaryArteryDisease	All	141	0.5090	14	0.0505	10	0.0361	0	0.0000	5	0.0181	115	0.4152	277
CoronaryArteryDisease	Functional	10	0.3030	1	0.0303	1	0.0303	0	0.0000	2	0.0606	22	0.6667	33
CrohnsDisease	All	594	0.4420	75	0.0558	61	0.0454	25	0.0186	64	0.0476	461	0.3430	1344
CrohnsDisease	Functional	108	0.3985	23	0.0849	18	0.0664	20	0.0738	29	0.1070	100	0.3690	271
EducationalAttainment	All	74	0.3978	7	0.0376	10	0.0538	2	0.0108	3	0.0161	68	0.3656	186
EducationalAttainment	Functional	10	0.5263	1	0.0526	1	0.0526	2	0.1053	3	0.1579	8	0.4211	19
Epilepsy	All	100	0.5988	2	0.0120	9	0.0539	0	0.0000	0	0.0000	56	0.3353	167
Epilepsy	Functional	1	1.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0000	1
EverSmoked	All	1	0.5000	1	0.5000	1	0.5000	0	0.0000	0	0.0000	0	0.0000	2
EverSmoked	Functional	0	NA	0	NA	0	NA	0	NA	0	NA	0	NA	NA
FastingGlucose	All	466	0.6332	42	0.0571	45	0.0611	2	0.0027	19	0.0258	206	0.2799	736
FastingGlucose	Functional	34	0.5965	5	0.0877	5	0.0877	1	0.0175	5	0.0877	16	0.2807	57
HDL	All	1444	0.4727	206	0.0674	181	0.0592	55	0.0180	196	0.0642	1148	0.3758	3055
HDL	Functional	325	0.5417	42	0.0700	42	0.0700	48	0.0800	77	0.1283	187	0.3117	600
Height	All	3255	0.4559	317	0.0444	261	0.0366	66	0.0092	206	0.0289	2441	0.3419	7139
Height	Functional	440	0.4949	49	0.0551	53	0.0596	52	0.0585	71	0.0799	312	0.3510	889
IGAN	All	222	0.0921	33	0.0137	35	0.0145	0	0.0000	21	0.0087	60	0.0249	2411
IGAN	Functional	107	0.7868	8	0.0588	9	0.0662	0	0.0000	16	0.1176	12	0.0882	136
LDL	All	648	0.2680	113	0.0467	90	0.0372	41	0.0170	124	0.0513	982	0.4061	2418
LDL	Functional	94	0.2741	36	0.1050	18	0.0525	30	0.0875	49	0.1429	143	0.4169	343
RheumatoidArthritis	All	2048	0.0782	338	0.0129	411	0.0157	108	0.0041	337	0.0129	2802	0.1070	26197
RheumatoidArthritis	Functional	535	0.3677	77	0.0529	131	0.0900	83	0.0570	178	0.1223	691	0.4749	1455
Schizophrenia	All	7718	0.3565	1053	0.0486	1068	0.0493	42	0.0019	371	0.0171	8335	0.3850	21648
Schizophrenia	Functional	619	0.4697	118	0.0895	104	0.0789	32	0.0243	54	0.0410	511	0.3877	1318
Triglycerides	All	1009	0.4055	108	0.0434	123	0.0494	18	0.0072	85	0.0342	893	0.3589	2488
Triglycerides	Functional	146	0.5052	21	0.0727	27	0.0934	18	0.0623	34	0.1176	93	0.3218	289
Type2Diabetes	All	261	0.6658	8	0.0204	13	0.0332	0	0.0000	4	0.0102	107	0.2730	392
Type2Diabetes	Functional	9	0.5000	0	0.0000	1	0.0556	0	0.0000	0	0.0000	6	0.3333	18
UlcerativeColitis	All	493	0.4800	78	0.0759	62	0.0604	18	0.0175	57	0.0555	311	0.3028	1027
UlcerativeColitis	Functional	76	0.5429	10	0.0714	15	0.1071	13	0.0929	18	0.1286	41	0.2929	140

Table S5. Assumed prevalences for 13 binary GWAS traits.

Trait	Assumed prevalence
Schizophrenia	0.01[1]
Rheumatoid Arthritis	0.007 [2]
Crohn's Disease	0.00319 [3]
Coronary Artery Disease	0.06
Ulcerative Colitis	0.00249 [3]
Alopecia Areata	0.0015 [4]
IGAN	0.005 [5]
Epilepsy	0.007 [6]
Alzheimer's	0.11 [7]
Type2 Diabetes	0.07 [8]
Bipolar Disorder	0.026 [9]
Ever Smoked	0.57[1]
Autism	0.015 [10]