Supplementary material for: Joint analysis of functional genomic data and genome-wide association studies of 18 human traits

Joseph K. Pickrell[1,2]

[1] New York Genome Center, New York, NY

[2] Department of Biological Sciences, Columbia University, New York, NY

Correspondence to: jkpickrell@nygenome.org

January 21, 2014

Contents

1 GWAS data

1.1 GIANT data

We downloaded summary statistics from large GWAS of height [Lango-Allen et al., 2010] and BMI [Speliotes et al., 2010] from `http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium`. The height summary statistics consisted of 2,469,635 SNPs either directly genotyped or imputed in an average of 129,945 individuals. We removed all SNPs with a sample size of less than 120,000 individuals. The BMI summary statistics consisted of 2,471,516 summary statistics either directly genotyped or imputed in an average of 120,569 individuals. We removed all SNPs with a sample size of less than 110,000 individuals. We then imputed summary statistics at SNPs identified in the 1000 Genomes Project as described in Section 3.

1.2 GEFOS data

We downloaded summary statistics from large GWAS of bone mineral density [Estrada et al., 2012] from `http://www.gefos.org/?q=content/data-release`. There are two traits in these data: bone density measured in the femoral neck and bone density measured in the lumbar spine. The femoral neck bone density GWAS consisted of 2,478,337 SNPs, and the lumbar spine bone density consisted of 2,468,080 SNPs. Because the sample size at each SNP was not reported, we used the overall study sample sizes of 32,961 and 31,800 as approximations of the sample size at each SNP, and imputed summary statistics as described in Section 3.

1.3 IIBDGC data

We downloaded summary statistics from a large GWAS of Crohn's disease [Jostins et al., 2012] from `http://www.ibdgenetics.org/downloads.html`. The downloaded data consisted of 953,242 SNPs. Because the sample size at each SNP was not reported, we used the overall study sample sizes of 6,299 cases and 15,148 controls as approximations of the sample size at each SNP, and imputed summary statistics as described in Section 3. Note that summary statistics from a GWAS of ulcerative colitis were also available from this site; however, these data contain a number of false positive associations that were filtered by Jostins et al. [2012] using criteria that were not available to us. We thus only used the Crohn's disease association study.

1.4 MAGIC data

We downloaded summary statistics from a large GWAS of fasting glucose levels [Manning et al., 2012] from `http://www.magicinvestigators.org/downloads/`. The downloaded data consisted of 2,628,880 SNPs. Because the sample size at each SNP was not reported, we used the overall study sample size of 58,074 as an approximation of the sample size at each SNP, and imputed summary statistics as described in Section 3.

1.5 Global lipid genetics consortium data

We downloaded summary statistics from a large GWAS of lipid traits [Teslovich et al., 2010] from `http://www.sph.umich.edu/csg/abecasis/public/lipids2010/`. These data consist of summary statistics for association studies of four traits: LDL cholesterol, HDL cholesterol, trigylcerides, and total cholesterol. The HDL data consisted of 2,692,429 SNPs genotyped or imputed in an average of 88,754 individuals, the LDL data consisted of 2,692,564 SNPs genotyped or imputed in an average of 84,685 individuals, the total cholesterol data consisted of 2,692,413 SNPs genotyped or imputed in an average of 89,005 individuals, and the triglycerides data consisted of 2,692,560 SNPs genotypes or imputed in an average of 85,691 individuals. For all traits, we removed SNPs with a sample size less than 80,000 individuals, and imputed summary statistics as described in Section 3.

1.6 Red blood cell trait data

We obtained summary statistics from a large GWAS of red blood cell traits [van der Harst et al., 2012] from the European Genome-Phenome Archive (accession number EGAS00000000132). We downloaded summary statistics from association studies of six traits: hemoglobin levels, mean cell hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), mean cell volume (MCV), packed cell volume (PCV), and red blood cell count (RBC). The hemoglobin level data consisted of 2,593,078 SNPs genotyped or imputed in 50,709 individuals, the MCH data consisted of 2,586,785 SNPs genotyped or imputed in an average of 43,127 individuals, the MCHC data consisted of 2,588,875 SNPs genotyped or imputed in an average of 46,469 individuals, the MCV data consisted of 2,591,132 SNPs genotyped or imputed in an average of 47,965 individuals, the PCV data consisted of 2,591,079 SNPs genotyped or imputed in an average of 44,485 individuals, and the RBC data consisted of 2,589,454 SNPs genotyped or imputed in an average of 44,851 individuals. We removed all SNPs with a sample size of less than 50,000 individuals (for hemoglobin levels) or 40,000 individuals (for the other traits), and imputed summary statistics as described in Section 3.

1.7 Platelet traits

Summary statistics from a large GWAS of platelet traits [Gieger et al., 2011] were generously provided to us by Nicole Soranzo. The data consist of summary statistics from association studies of two traits: platelet counts and mean platelet volume. The platelet count data consisted of 2,705,636 SNPs genotyped or imputed in an average of 44,217 individuals, and the platelet volume data consisted of 2,690,858 SNPs genotyped or imputed in an average of 16,745 individuals. We removed all SNPs with sample sizes less than 40,000 (for platelet counts) or 15,000 (for platelet volume), and imputed summary statistics as described in Section 3.

2 Functional genomic data

2.1 DNase-I hypersensitivity data

We downloaded DNase-I hypersensitivity data from two sources. The first was a set of regions defined as DNase-I hypersensitive by Maurano et al. [2012] in 349 samples. We downloaded .bed files for 349 samples from `http://www.uwencode.org/proj/Science_Maurano_Humbert_et_al/` on February 13, 2013. These samples include 116 samples from cell lines or sorted blood cells, and 333 samples from primary fetal tissues. These latter samples were sampled from several tissues at various time points; we treated each track as independent rather than pooling data from tissues, since different experiments may have slightly different properties. The tissues in this latter group are fetal heart, fetal brain, fetal lung, fetal kidney, fetal intestine (large and small), fetal muscle, fetal placenta, and fetal skin.

The second was a set of regions defined as DNase-I hypersensitive by the Crawford lab in the context of the ENCODE project [Thurman et al., 2012]. We downloaded .bed files for 53 samples from `http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/fdrPeaks/` on March 29, 2013. We restricted ourselves to the files labeled as being generated at Duke University. Each experiment defined a set of regions of open chromatin in a particular cell type or cell line.

The "Duke" DNase-I hypersensitive sites are all of exactly 150 bases in length, and each annotation covers approximately 1% of the genome (range: 0.4 - 1.9 % of the genome). The "Maurano" DNase-I hypersensitive sites are on average 514 bases long, and each covers on average 2.7% of the genome (range: 0.9-5.1 % of the genome).

2.2 Chromatin state data

We downloaded the "genome segmentations" of the six ENCODE cell lines [Hoffman et al., 2013] from `http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/segmentations/jan2011/hub/` on December 18, 2012. We used the "combined" segmentation from two algorithms. This segmentation splits the genome into non-overlapping regions described as CTCF binding sites, enhancers, promoter-flanking regions, repressed chromatin, transcribed regions, transcription start sites, and weak enhancers. This segmentation was done independently in each of six cell lines, for a total of 42 annotations.

Overall the "repressed chromatin" mark covers the largest fraction of the genome, on average 66% (ranging from 60% for HUVEC cells to 70% for H1 ES cells). The "transcribed" mark covers on average 13% of the genome, the "CTCF" mark 1% of the genome, the "enhancer" mark 0.9% of the genome, the "TSS" mark 0.7% of the genome, the "weak enhancer" mark 0.4% of the genome, and the "promoter-flanking" mark 0.2% of the genome. The remainder of the genome is not mappable by short reads and it thus excluded from these annotations.

2.3 Gene models

We downloaded the Ensembl gene annotations from the UCSC genome browser on May 21. Annotations of nonsynonymous and synonymous status for all SNPs in phase 1 of the 1000 Genomes Project were obtained from `ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase1/analysis_results/functional_annotation/annotated_vcfs/`. Coding exons cover about 3% of the genome, while 3' UTRs and 5' UTRs cover 2% and 0.6% of the genome, respectively.

3 Imputation of summary statistics

We used ImpG v1.0 [Pasaniuc et al., 2013] under the default settings to impute summary statistics from all GWAS. As a reference panel, we used all haplotypes from European individuals in phase 1 of the 1000 Genomes Project, and only used SNPs with a minor allele frequency greater than 2%. The reference haplotype files were derived from the 1000 Genomes integrated phase 1 v3.20101123 calls, downloaded from `ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase1/analysis_results/integrated_call_sets/`. We used all 379 individuals labeled as "European". After imputation, we removed all imputed SNPs with a predicted accuracy (in terms of correlation with the true summary statistics) less than 0.8. Overall, for each GWAS, we successfully imputed about 75-80% of SNPs with a minor allele frequency over 10% (Figure 1).

To verify that imputation did not induce inflation of the test statistics, we computed the genomic control inflation factor λ_{GC} [Bacanu et al., 2002] before and after imputation (Supplementary Table 1). In all studies, inflation decreased after imputation, sometimes leading to a marked deflation in the test statistics. This is consistent with previous observations using this software [Pasaniuc et al., 2013]. The reason for this deflation is the shrinkage prior used in the imputation, which leads to conservative estimates of significance (imposed to strictly avoid false positive associations).

4 Details of application of the hierarchical model

4.1 Simulations

To test the performance of the model, we performed simulations using a GWAS of height [Lango-Allen et al., 2010]. Using the imputed summary statistics, we split the genome into blocks of 5,000 SNPs, then extracted the blocks with a genome-wide significant SNP reported in Lango-Allen et al. [2010]. In each block, we had a reported Z-score for each SNP. To simulate annotations, we called the SNP with the smallest P-value in the region the "causal" SNP. We then simulated annotations by placing all non-"casual" SNPs in an annotation with rate r_1, and all "casual" SNPs in the annotation with rate r_2. We also varied the numbers of blocks included in the model. In each simulation, we randomly assigned SNPs to annotations according to determined rates, then ran our model under the assumption that $\Pi_k = 1$, that is, all blocks contain a causal SNP. We then calculated power as the fraction of simulations in which the confidence intervals of the annotation effect did not overlap zero.

We chose parameter settings of r_1 and r_2 such that the enrichment factors were similar to those in observed data (log-enrichment of 0.98 and 1.80). We chose r_1 to be either 0.2 and 0.1. For each set of parameters, we simulated 100 annotations and ran the model separately on each. Shown in Figure 2 is the power of the model. As expected, power increased as r_1 or the effect size increased, and as the number of loci increased.

4.2 Robustness to choice of prior and window size

There are two parameters in the model that are set by the user–the prior variance W on the effect size and the window size defining "independent" blocks of the genome. We empirically tested the robustness of the model to variation in these parameters using the Crohn's disease dataset. We ran the model on each annotation using $W = 0.1$ and $W = 0.5$, additionally including–as in our main analyses–region-level parameters for regions in the top third and bottom third of gene density and SNP-level parameters for SNPs located from 0-5kb from a transcription start site and SNPs 5-10kb from a transcription start site. Plotted in Figure 16A are these annotation parameter estimates for all annotations where the 95% confidence intervals did not overlap 0 in at least one run. The estimates from the two runs with different priors are highly correlated. We additionally tested window sizes of 5,000 SNPs and 10,000 SNPs (both with $W = 0.1$). The annotation effect estimates from these two window sizes are plotted in Figure 16B, and again are highly correlated.

4.3 Quantifying the relative roles of coding versus non-coding changes in each phenotype

To generate Figure 3 in the main text, we fit a model to each GWAS where we included region-level annotations for regions in the top third and bottom third of the distribution of gene density, and SNP-level annotations for non-synonymous SNPs and SNPs within 5kb of a transcription start site. Shown in Figure 3A in the main text are the estimates of the enrichment parameter for non-synonymous SNPs. At each SNP, the result of this model is the posterior probability that the SNP is the causal one in its region. If we let this posterior probability at SNP i be PPA_i, then the fraction of causal SNPs that are non-synonymous, f_{NS} is:

$$f_{NS} = \frac{\sum_i PPA_i I_i^{NS}}{\sum_i PPA_i}, \tag{1}$$

where I_i^{NS} is an indicator variable that takes value one if SNP i is non-synonymous and zero otherwise. To get error bars on this fraction, we performed a block jackknife. We split the genome into 20 blocks with equal numbers of SNPs. If f_{NS}^j is the estimate of the fraction of casual SNPs that are non-synonymous excluding block j, then:

$$SE = \sqrt{\frac{19}{20}\sum_{j=1}^{20}(f_{NS}^j - \bar{f}_{NS})^2}, \tag{2}$$

where $\bar{f}_{NS} = \frac{1}{20}\sum_{i=1}^{20} f_{NS}^{i}$. In Supplementary Figure 3, we show the corresponding results for synonymous SNPs.

4.4 Interaction effects in annotation models

As noted in the main text, there were two cases in which the sign of the annotation effect flipped between the single annotation models and the combined models. These were Crohn's disease (Supplementary Table 6) and red blood cell count (Supplementary Table 18). In the main text we discuss the Crohn's disease example. For the red blood cell count example, note that SNPs influencing this trait are enriched in the annotation of DNAse-I hypersensitive sites in the fetal renal pelvis when this annotation is considered alone (log enrichment of 1.72, 95% CI [0.03, 3.89]). This annotation is correlated with the fetal stomach annotation, which has a log enrichment of 3.35 (95% CI [2.29, 4.47]) when treated alone. The SNPs in both of these annotations have a log enrichment of 1.67 (95% CI [-1.27, 2.93]), which leads to the interaction effect. Essentially the signal in the fetal stomach is driven by those SNPs that fall in DNase-I hypersensitive sites in the fetal stomach but *not* the fetal renal pelvis. This suggests that there are a subset of DNase-I hypersensitive sites that are of particular interest for this phenotype. The interpretation of the Crohn's disease example is similar.

4.5 Calibrating a "significance" threshold

For each genomic region, our method estimates the posterior probability that the region contains a SNP associated with a trait. If the model were a perfect description of reality, this probability could be interpreted literally. Since the model is not perfect, however, we sought a more empirical calibration. We used the fact that we initially ran the method on the GWAS data reported by Teslovich et al. [2010] on four lipid traits. Since then, a GWAS with more individuals (though at a considerably smaller number of SNPs) has been reported for these four traits [Global Lipids Genetics Consortium et al., 2013]. This latter study contains many of the individuals from the former (which had approximately 90,000 individuals), as well as about 80,000 more individuals. However, the additional individuals were genotyped in the Metabochip [Voight et al., 2012], which has less than 200,000 markers, rather than the more dense standard GWAS arrays. This means that some regions of the genome do not benefit from the larger sample size.

For each region of the genome for each of the four traits, we built a table containing the minimum P-value from Teslovich et al. [2010], the posterior probability of association in the region (computed using the data from Teslovich et al. [2010]), the minimum P-value from Global Lipids Genetics Consortium et al. [2013], and the sample size used to get this minimum P-value (from Global Lipids Genetics Consortium et al. [2013]). We discarded regions where sample size at the SNP with the minimum P-value in the replication data set was smaller than 120,000 (since in these regions there is essentially no new data). We then coded each region as a "true positive" if the minimum P-value from Global Lipids Genetics Consortium et al. [2013] was less than 5×10^{-8}

and a "true negative" otherwise. In Figure 15, we plot the number of "true positives" and "false negatives" that exceed various P-value and PPA thresholds. Note that since the data in Global Lipids Genetics Consortium et al. [2013] is not independent of that in Teslovich et al. [2010], this comparison is not appropriate for evaluating the relative performance of P-values versus the PPA. Our goal was simply to find a PPA threshold with similar performance in terms of reducing the number of false positives as the standard P-value threshold of 5×10^{-8}.

By visual inspection we set a PPA threshold at 0.9 (Figure 15). At this threshold, we identify 45 "true positives" and zero "false positives" for HDL, 43 and 1 for LDL, 47 and zero for total cholesterol, and 27 and zero for triglycerides. These are similar to the numbers for a P-value threshold of 5×10^{-8} (Supplementary Table 21). Combining the loci identified by both methods leads to 48 loci for HDL (versus 43 using a P-value threshold), 44 for LDL (versus 40), 51 for TC (versus 51) and 30 for TG (versus 29). This is on average an increase of 6% in the number of loci identified. Note that this number is likely a lower bound, since the P-values in the replication study are naturally highly correlated to those in the initial study since they use many of the same individuals. A proper comparison would use a completely separate, large set of individuals to determine "true positives" and "true negatives", but such samples are not yet available.

4.6 Identification of novel loci

For each fitted model (using the parameters from Supplementary Tables 3-20 estimated using the penalized likelihood), we calculated the posterior probability of association in each genomic region. We then identified all regions with a PPA greater than 0.9 but that had a minimum P-value less than 5×10^{-8}. For each remaining region, we identified the "lead" SNP as the SNP with the largest posterior probability of being the causal SNP in the region. If this SNP was within 500kb of a SNP with $P < 5 \times 10^{-8}$ (this can happen because we use non-overlapping windows and sometimes the best SNP is at the edge of the region), we removed it. We also manually removed two regions (surrounding rs8076131 in Crohn's disease and surrounding rs11535944 in HDL), where the "new" association was in LD with a previously reported SNP over 500kb away. In Supplementary Table 22, we show the remaining SNPs; these regions are high-confidence associations that did not reach traditional genome-wide significance.

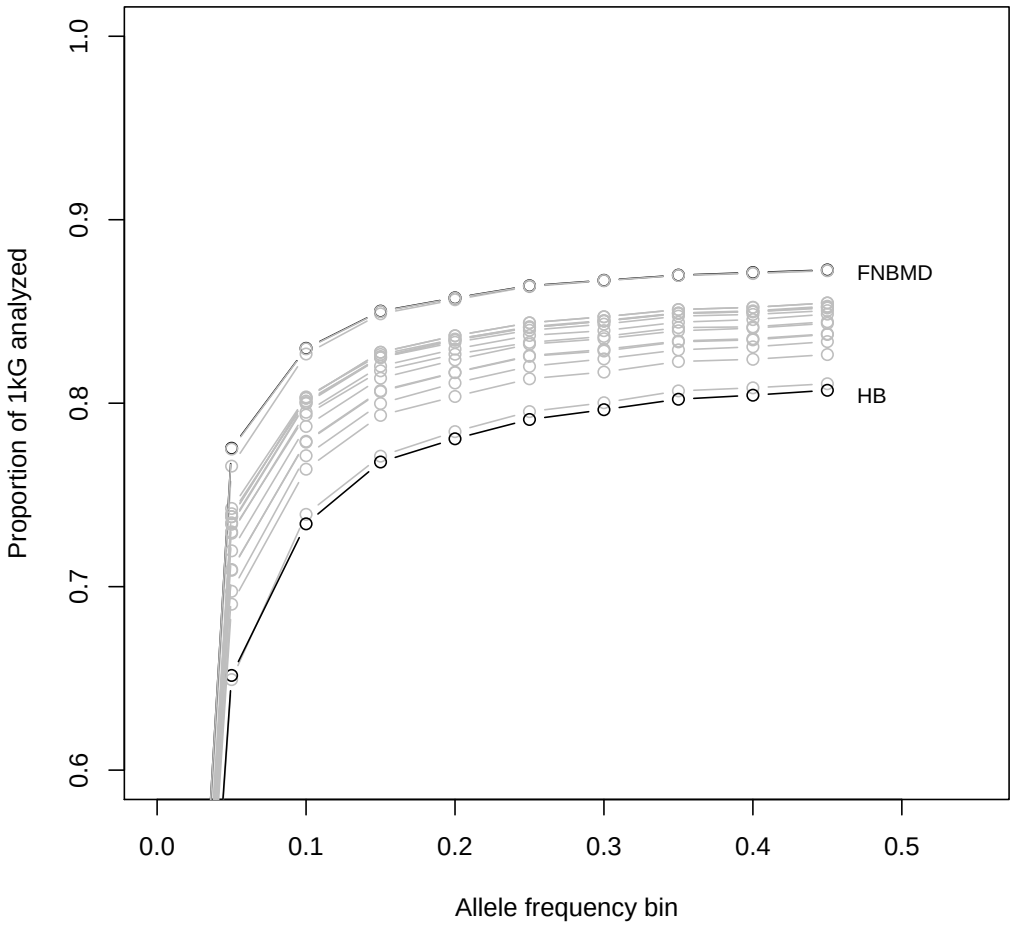

Figure 1. **Proportion of SNPs in the 1000 Genomes Project either genotyped or successfully imputed.** For each trait, we split all SNPs in phase 1 of the 1000 Genomes Project into bins based on their minor allele frequency in the European population. Bin sizes were of 5% frequency. Shown are the proportions of SNPs in each bin that were either genotyped or successfully imputed for each trait (the points are at the lower ends of the bins, such that the point at 45% frequency contains all SNPs from 45%-50% minor allele frequency). Labeled are the traits with the lowest and highest coverage. HB = hemogobin levels, FNBMD = femoral neck bone mineral density.

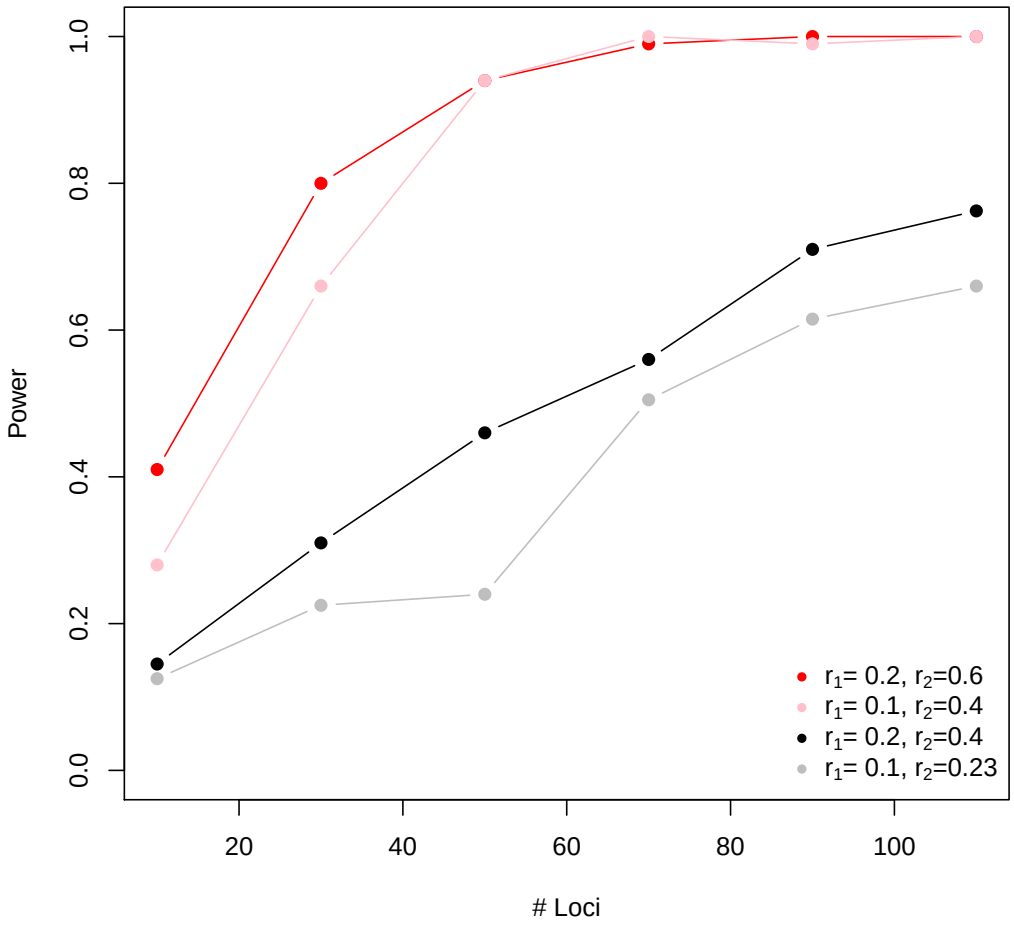

Figure 2. **Power to detect a significant annotation.** We simulated GWAS data under different levels of enrichment of causal SNPs in an annotation (see Supplementary Text), then evaluated the power of the method to detect the enrichment with different numbers of loci. In red and pink are $\log_2$-enrichments of 2.6, and in black and grey are $\log_2$-enrichments of 1.4.

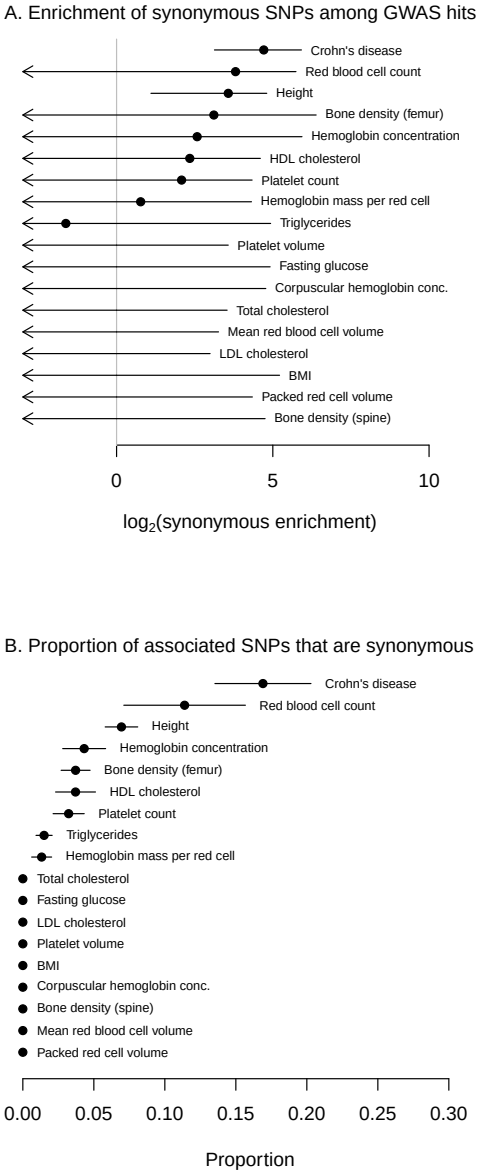

Figure 3. **Estimated role of synonymous polymorphisms in each trait. A. Estimated enrichment of synonymous SNPs.** For each trait, we fit a model including an effect of synonymous SNPs and an effect of SNPs within 5kb of a TSS. Shown are the estimated enrichments parameters and 95% confidence intervals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

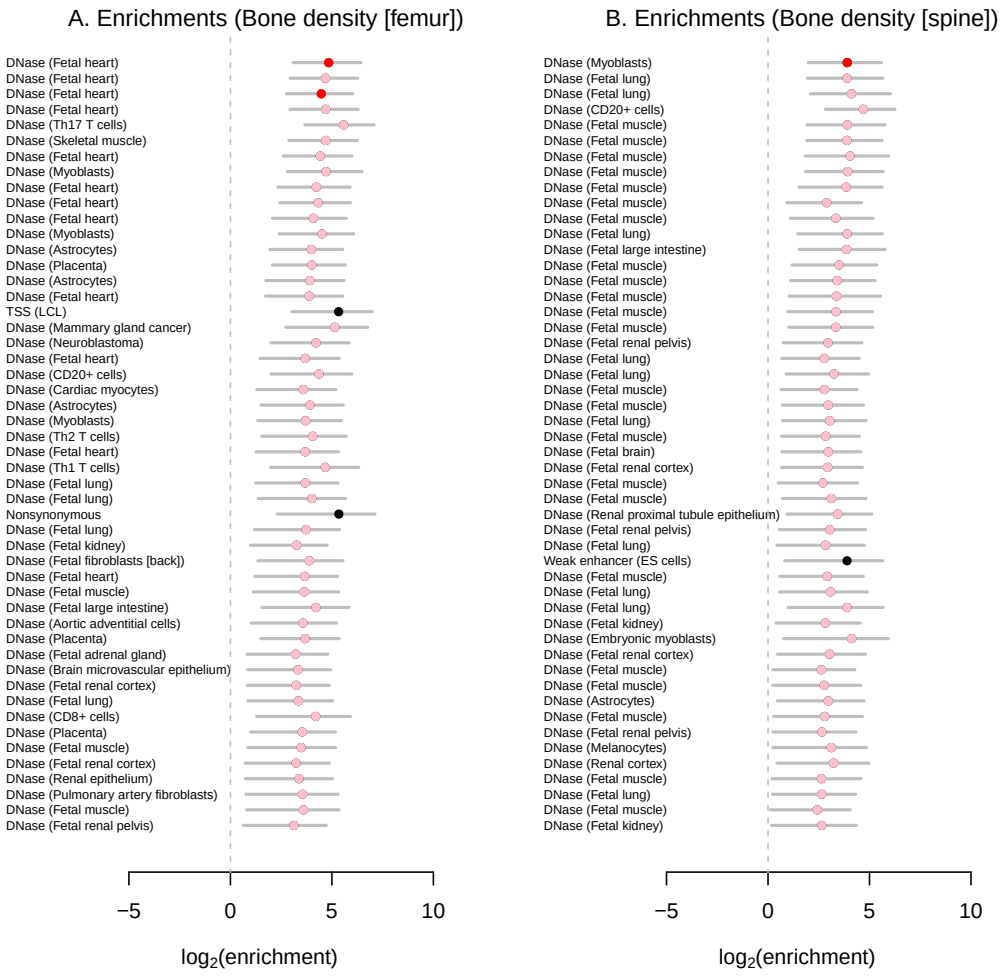

Figure 4. **Annotation effects in the bone mineral density data.** We estimated an enrichment parameter for each annotation individually in the GWAS for **A.** bone density in the femoral neck and **B.** bone density in the lumbar spine. Shown are the maximum likelihood estimates and 95% confidence intervals. Annotations are ranked according to how much each improves the fit of the model; shown are the 50 annotations that most improve the model (or if there were less than 50 significant annotations, all of the significant annotations). In red are the annotations included in the combined model, and in pink are annotations that are statistically equivalent to those in the combined model.

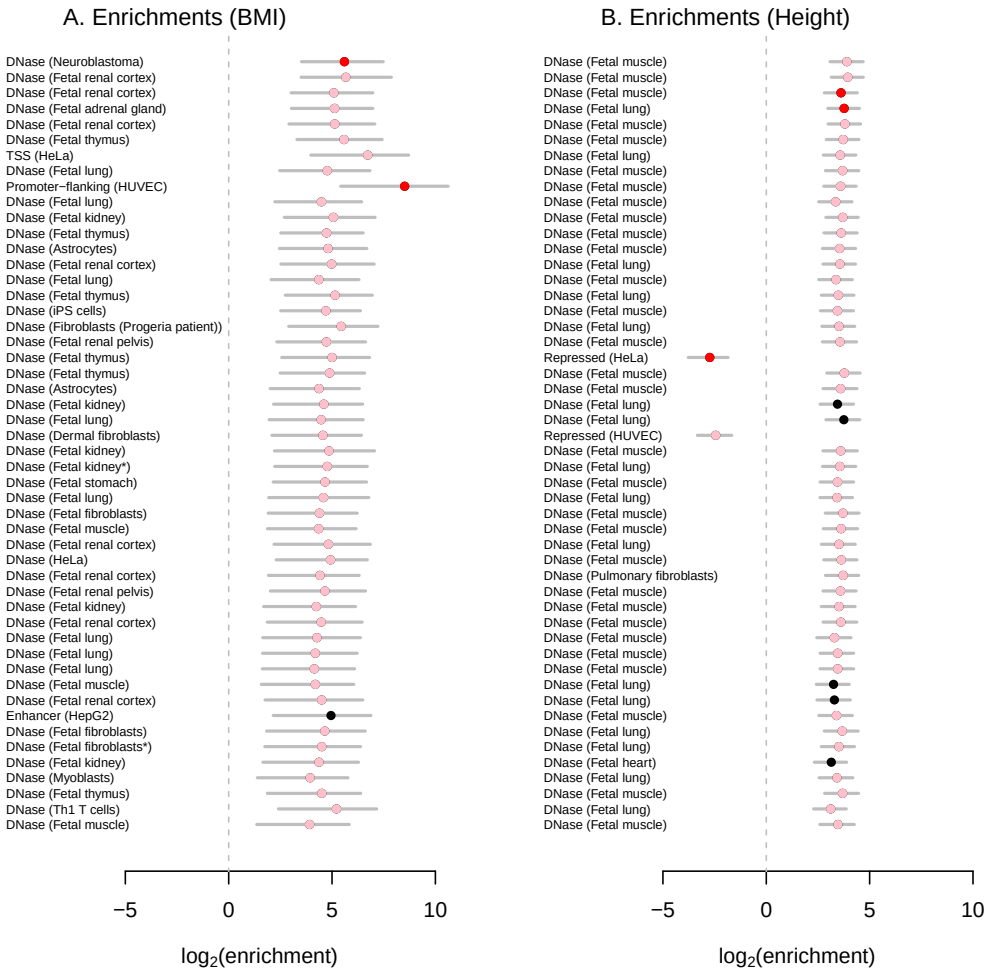

Figure 5. **Annotation effects in the GIANT data.** We estimated an enrichment parameter for each annotation individually in the GWAS for **A.** BMI and **B.** height. Shown are the maximum likelihood estimates and 95% confidence intervals. Annotations are ranked according to how much each improves the fit of the model; shown are the 50 annotations that most improve the model (or if there were less than 50 significant annotations, all of the significant annotations). In red are the annotations included in the combined model, and in pink are annotations that are statistically equivalent to those in the combined model.

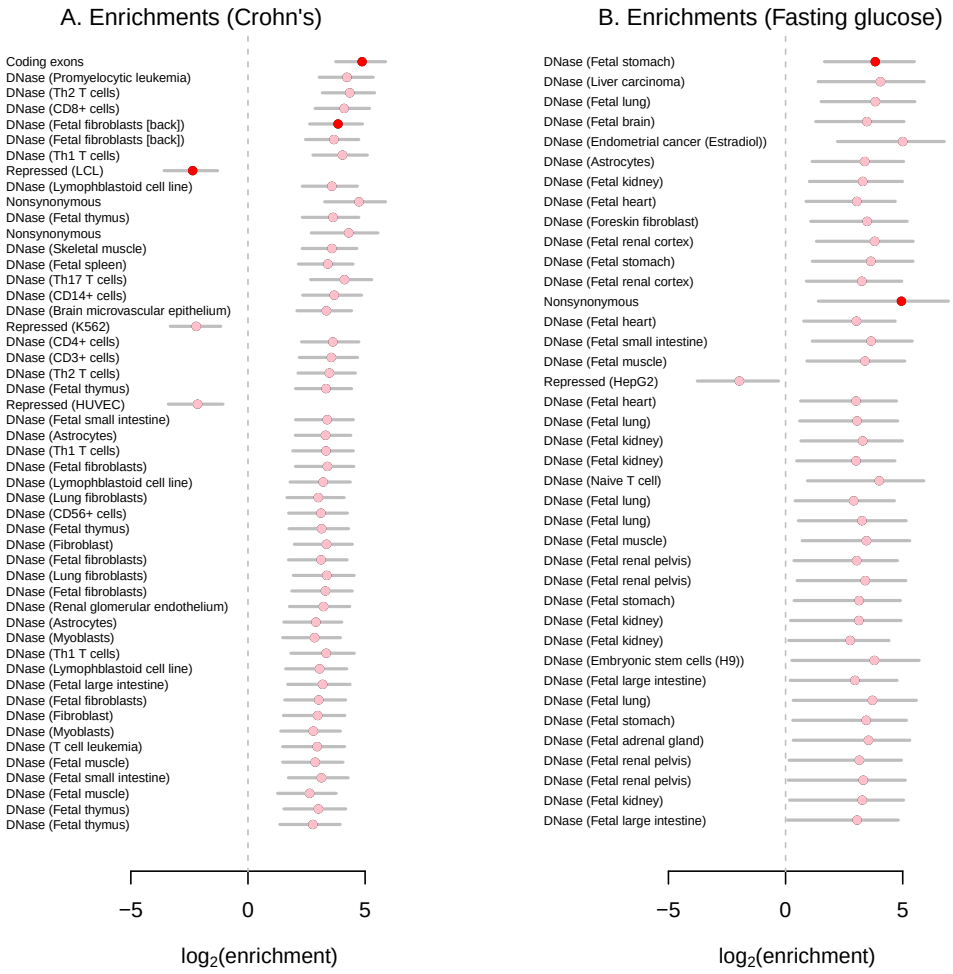

Figure 6. **Annotation effects in the Crohn's disease and fasting glucose data.** We estimated an enrichment parameter for each annotation individually in the GWAS for **A.** Crohn's disease and **B.** fasting glucose. Shown are the maximum likelihood estimates and 95% confidence intervals. Annotations are ranked according to how much each improves the fit of the model; shown are the 50 annotations that most improve the model (or if there were less than 50 significant annotations, all of the significant annotations). In red are the annotations included in the combined model, and in pink are annotations that are statistically equivalent to those in the combined model.

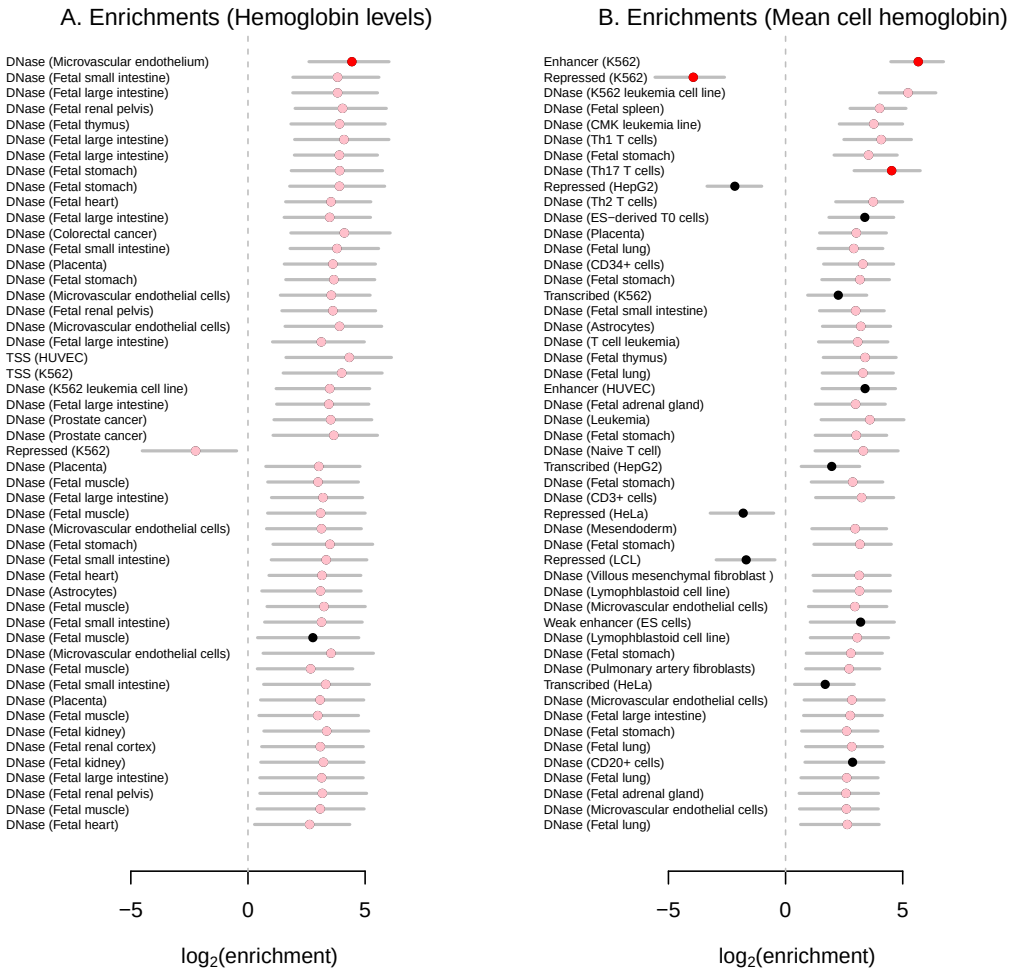

Figure 7. **Annotation effects in the red blood cell data.** We estimated an enrichment parameter for each annotation individually in the GWAS for **A.** hemoglobin levels and **B.** mean cellular hemoglobin. Shown are the maximum likelihood estimates and 95% confidence intervals. Annotations are ranked according to how much each improves the fit of the model; shown are the 50 annotations that most improve the model (or if there were less than 50 significant annotations, all of the significant annotations). In red are the annotations included in the combined model, and in pink are annotations that are statistically equivalent to those in the combined model.

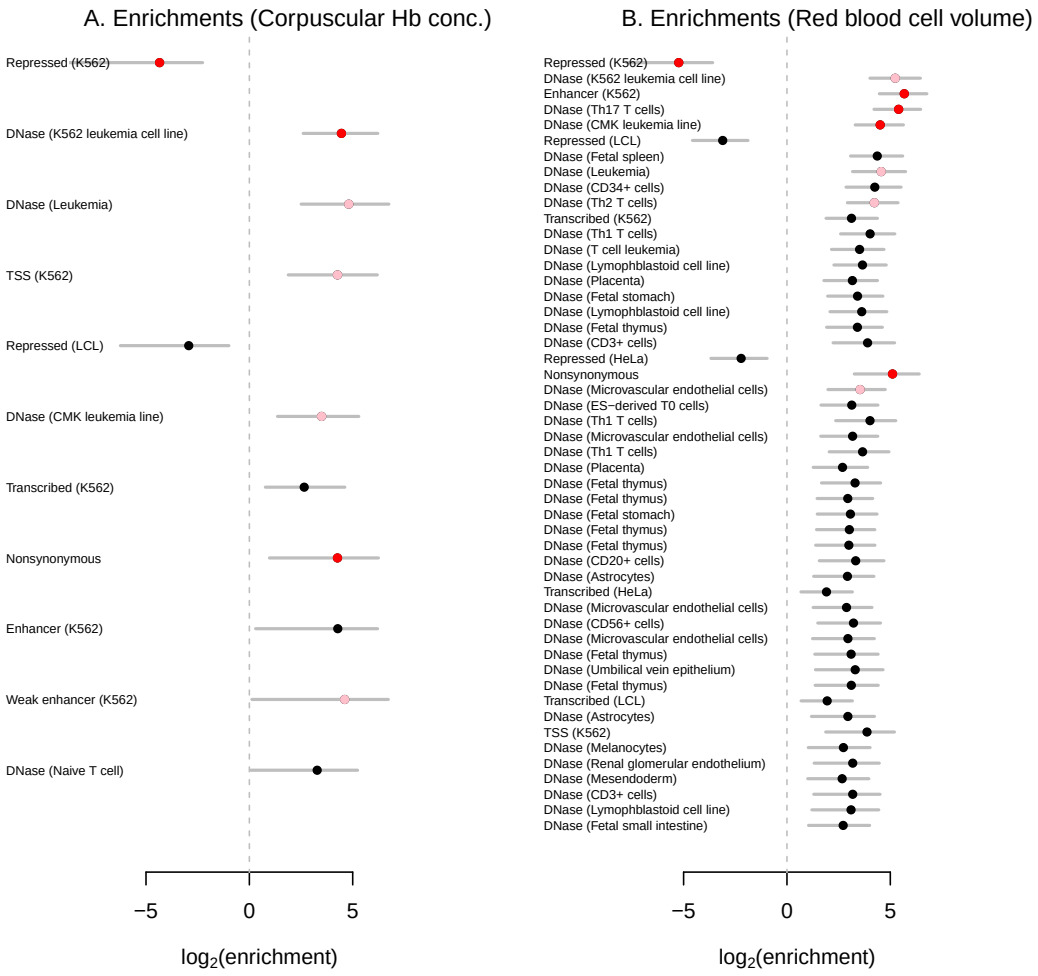

Figure 8. **Annotation effects in the red blood cell data.** We estimated an enrichment parameter for each annotation individually in the GWAS for **A.** mean corpuscular hemoglobin concentration and **B.** mean red cell volume. Shown are the maximum likelihood estimates and 95% confidence intervals. Annotations are ranked according to how much each improves the fit of the model; shown are the 50 annotations that most improve the model (or if there were less than 50 significant annotations, all of the significant annotations). In red are the annotations included in the combined model, and in pink are annotations that are statistically equivalent to those in the combined model.

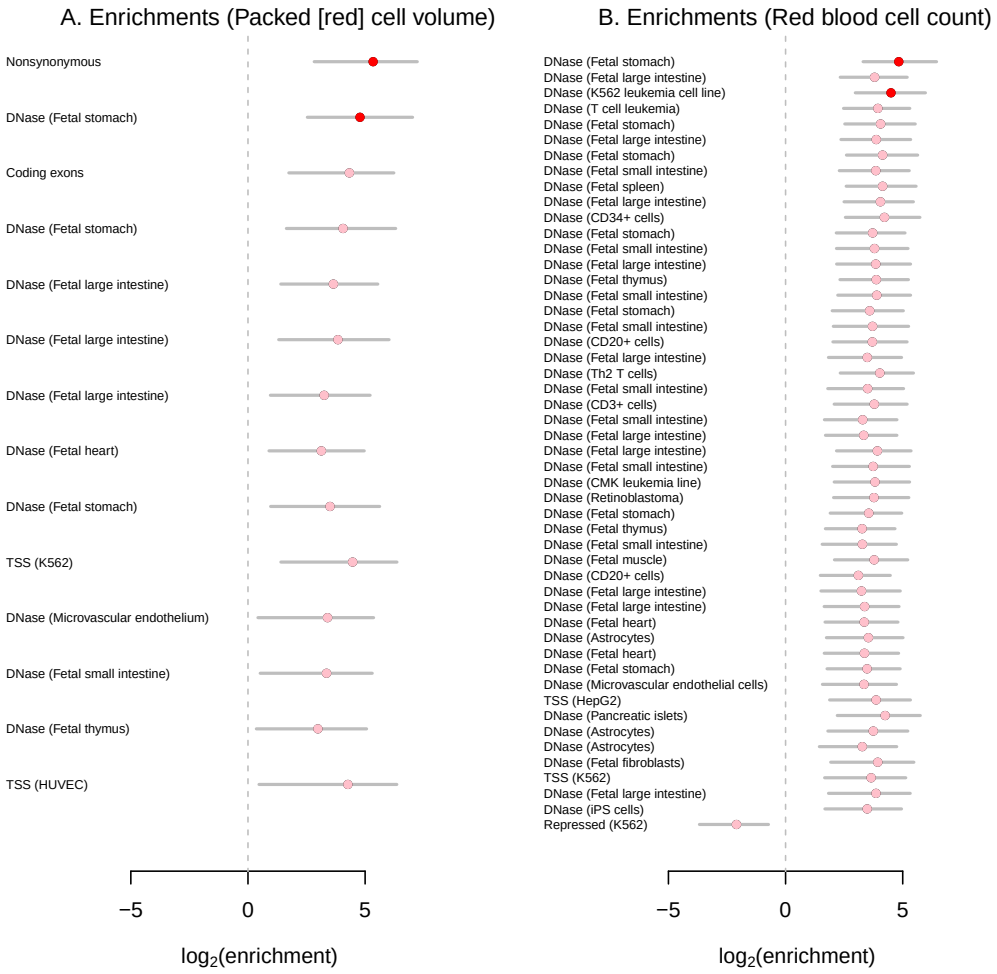

Figure 9. **Annotation effects in the red blood cell data.** We estimated an enrichment parameter for each annotation individually in the GWAS for **A.** packed cell volume and **B.** mean red cell count. Shown are the maximum likelihood estimates and 95% confidence intervals. Annotations are ranked according to how much each improves the fit of the model; shown are the 50 annotations that most improve the model (or if there were less than 50 significant annotations, all of the significant annotations). In red are the annotations included in the combined model, and in pink are annotations that are statistically equivalent to those in the combined model.

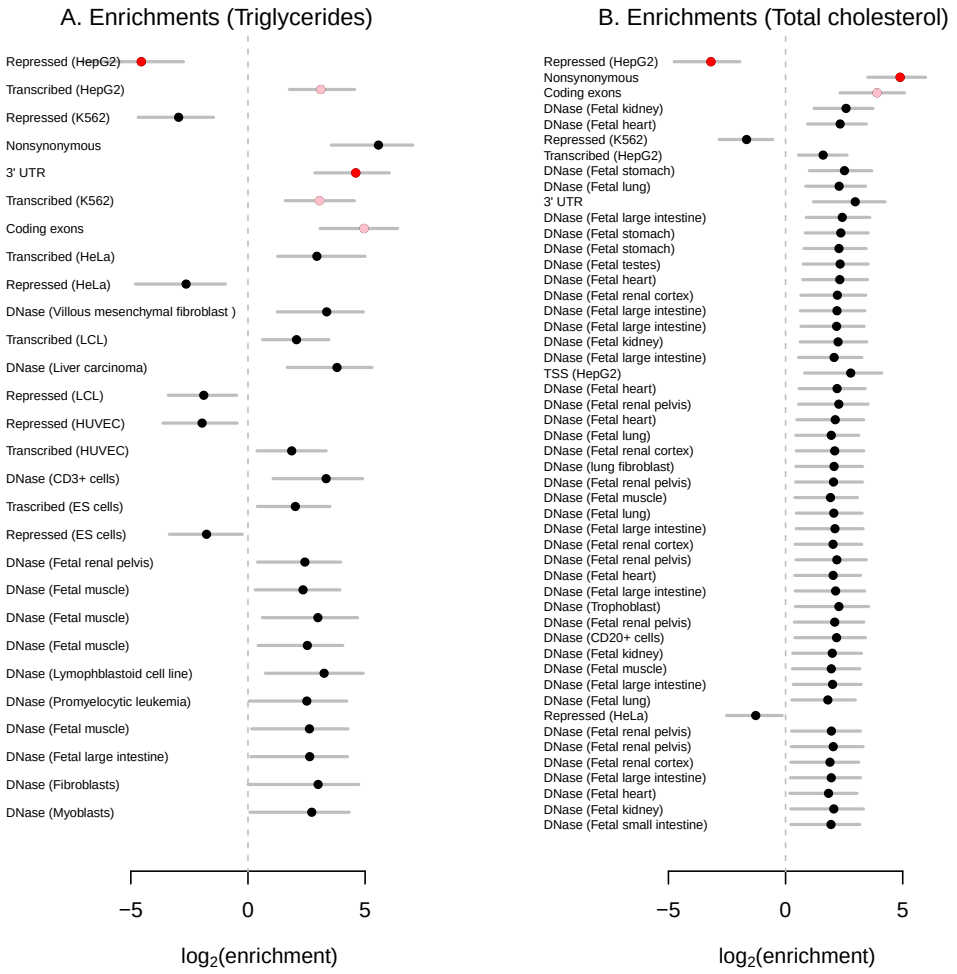

Figure 10. **Annotation effects in the lipids data.** We estimated an enrichment parameter for each annotation individually in the GWAS for **A.** triglyceride levels and **B.** total cholesterol. Shown are the maximum likelihood estimates and 95% confidence intervals. Annotations are ranked according to how much each improves the fit of the model; shown are the 50 annotations that most improve the model (or if there were less than 50 significant annotations, all of the significant annotations). In red are the annotations included in the combined model, and in pink are annotations that are statistically equivalent to those in the combined model.

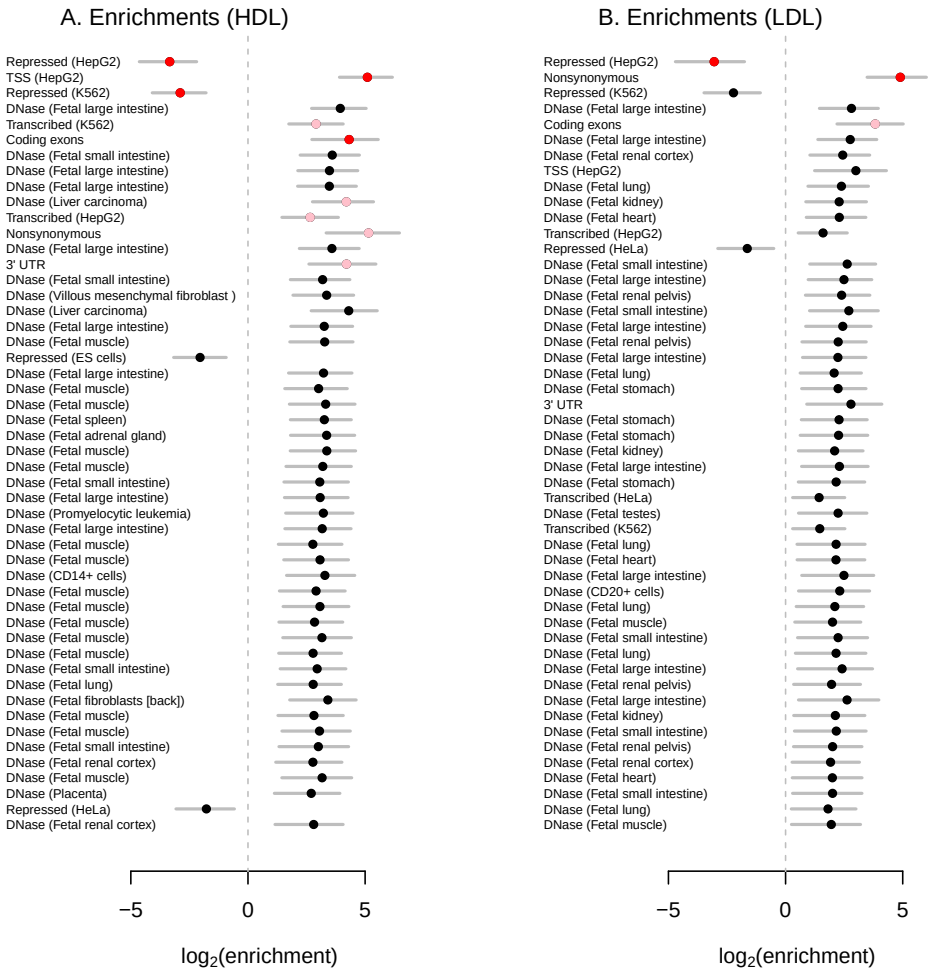

Figure 11. **Annotation effects in the lipids data.** We estimated an enrichment parameter for each annotation individually in the GWAS for **A.** HDL levels and **B.** LDL levels. Shown are the maximum likelihood estimates and 95% confidence intervals. Annotations are ranked according to how much each improves the fit of the model; shown are the 50 annotations that most improve the model (or if there were less than 50 significant annotations, all of the significant annotations). In red are the annotations included in the combined model, and in pink are annotations that are statistically equivalent to those in the combined model.

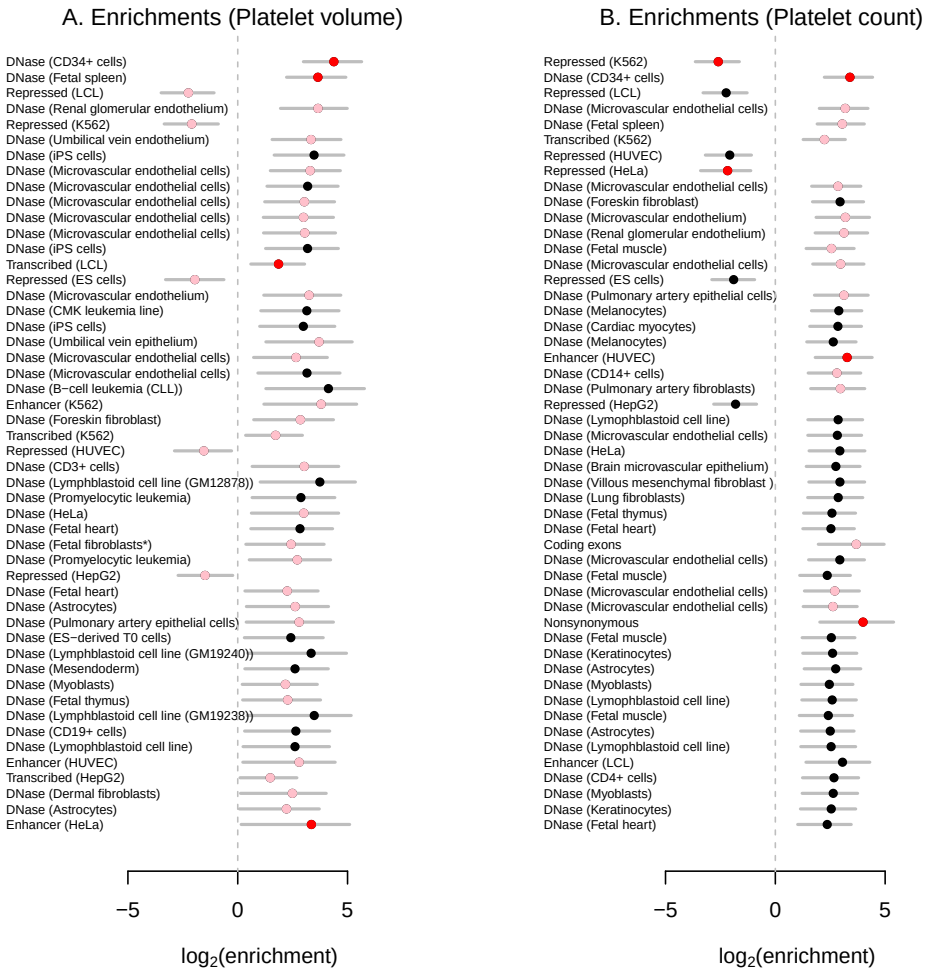

Figure 12. **Annotation effects in the platelet data.** We estimated an enrichment parameter for each annotation individually in the GWAS for **A.** mean platelet volume and **B.** platelet count. Shown are the maximum likelihood estimates and 95% confidence intervals. Annotations are ranked according to how much each improves the fit of the model; shown are the 50 annotations that most improve the model (or if there were less than 50 significant annotations, all of the significant annotations). In red are the annotations included in the combined model, and in pink are annotations that are statistically equivalent to those in the combined model.

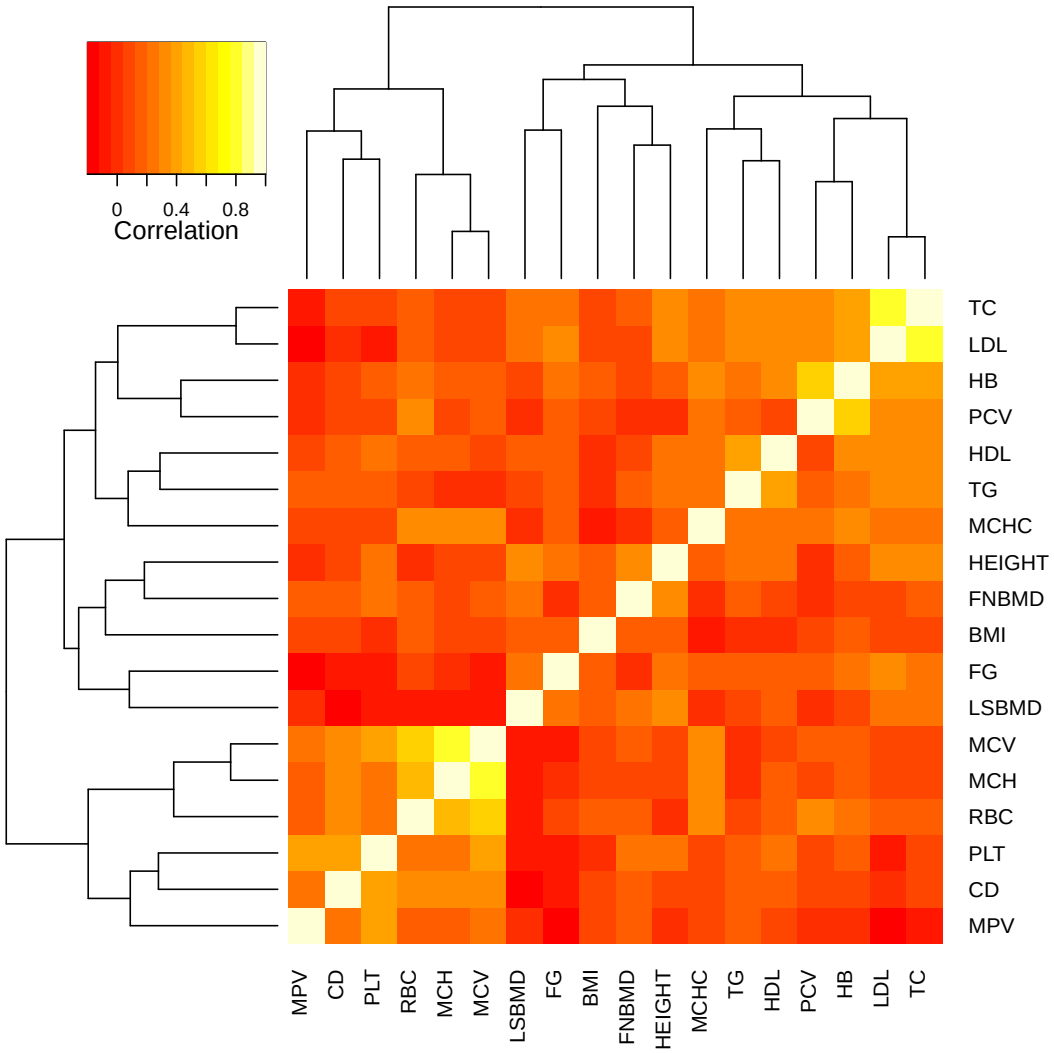

Figure 13. **Correlated patterns of enrichment across traits.** We estimated an enrichment parameter for each of 450 annotations for each of the 18 traits. For each pair of traits, we then estimated the Spearman correlation coefficient between the enrichment parameters. Plotted are these correlation coefficients. Orders of rows and columns were chosen by hierarchical clustering in R [R Core Team, 2013].

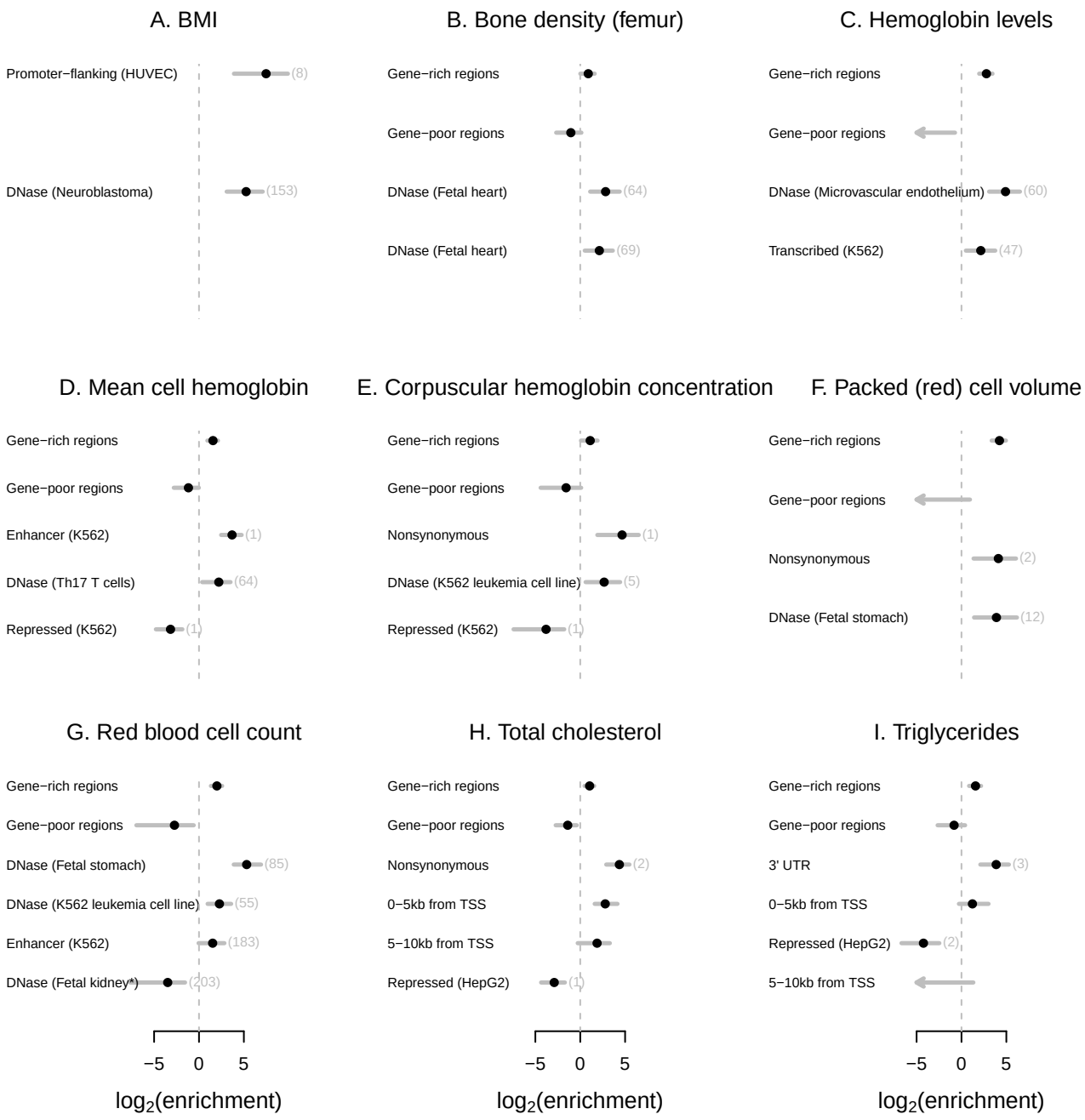

Figure 14. **Combined models for nine traits.** For each trait, we built a combined model of annotations using the algorithm presented in the Methods from the main text. Shown are the maximum likelihood estimates and 95% confidence intervals for all annotations included in each model. Note that though these are the maximum likelihood estimates, model choice was done using a penalized likelihood. In parentheses next to each annotation (expect for those relating to distance to transcription start sites), we show the total number of annotations that are statistically equivalent to the included annotation in a conditional analysis. For the other nine traits, see Figure 4 in the main text. *This annotation of DNase-I hypersensitive sites in fetal kidney (renal pelvis) has a positive effect when treated alone; see Supplementary Text for discussion.

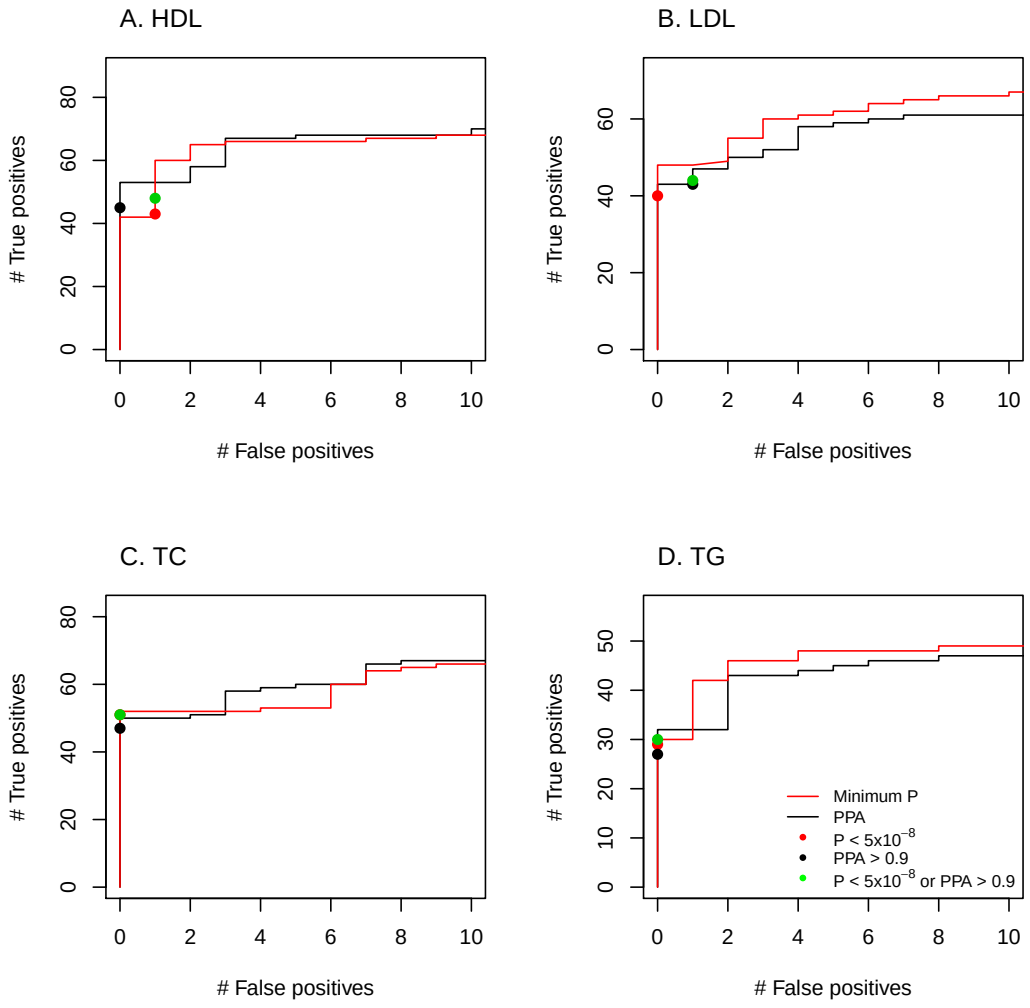

Figure 15. **Calibrating a PPA threshold similar to a P-value threshold.** For each of the four phenotypes in the lipids data, we plot the number of "true positives" and "false positives" obtained by different statistical thresholds; see Supplementary Text for details. Points show the positions of the thresholds used in the paper.

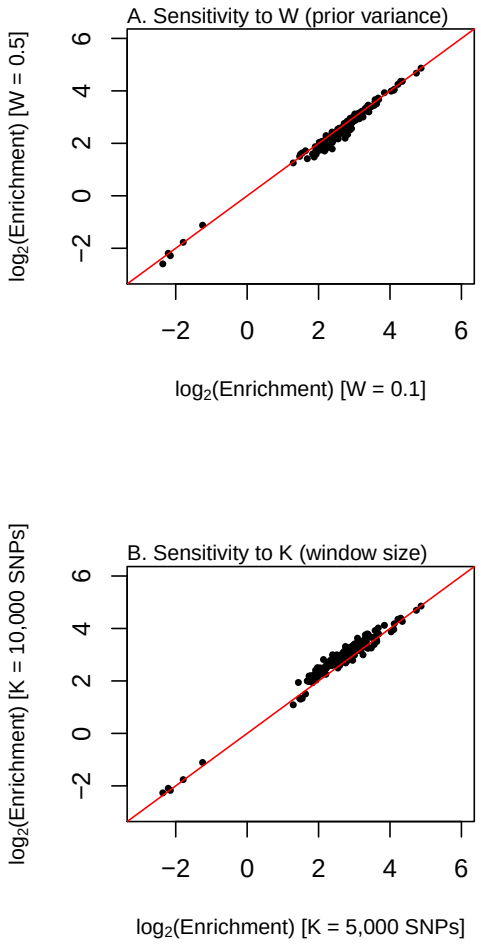

Figure 16. **Robustness of parameter estimates to preset parameters. A. Prior variance on effect size.** We estimated an enrichment parameter for each annotation in Crohn's disease using prior variances of 0.1 or 0.5. Shown are the estimates for all annotations with 95% confidence intervals that did not overlap 0 in at least one of the two runs. In red is the $y = x$ line. **B. Window size.** We estimated an enrichment parameter for each annotation in Crohn's disease using window sizes of 5,000 and 10,000 SNPs. Shown are the estimates for all annotations with 95% confidence intervals that did not overlap 0 in at least one of the two runs. In red is the $y = x$ line.

Phenotype	λ_{GC} (before imputation)	λ_{GC} (after imputation)
Height	1.04	0.99
BMI	1.04	0.97
BMD (femoral neck)	1.0	0.92
BMD (lumbar spine)	1.0	0.93
Crohn's	1.27	0.71
FG	1.08	0.97
HB	1.07	0.99
MCH	1.13	1.0
MCHC	1.07	0.85
MCV	1.13	1.0
PCV	1.09	0.97
RBC	1.14	1.01
TC	1.0	0.93
TG	1.0	0.92
HDL	1.0	0.94
LDL	1.0	0.93
PLT	1.08	1.01
MPV	1.04	0.96

Table 1: **Genomic control inflation factors before and after imputation.** We show λ_{GC} [Bacanu et al., 2002] before and after imputation for all 18 GWAS included in this study.

Phenotype	Proportion [95% CI]
BMI	0.022 [0.013, 0.032]
FNBMD	0.028 [0.019, 0.040]
LSBMD	0.028 [0.019, 0.041]
Crohn's	0.078 [0.059, 0.10]
FG	0.020 [0.012, 0.03]
HB	0.010 [0.006, 0.015]
HDL	0.034 [0.026, 0.044]
Height	0.131 [0.111, 0.153]
LDL	0.034 [0.026, 0.045]
MCH	0.035 [0.025, 0.047]
MCHC	0.018 [0.011, 0.027]
MCV	0.046 [0.034, 0.059]
MPV	0.025 [0.017, 0.035]
PCV	0.003 [0.002, 0.005]
PLT	0.036 [0.028, 0.047]
RBC	0.023 [0.016, 0.033]
TC	0.052 [0.040, 0.067]
TG	0.023 [0.015, 0.032]

Table 2: **Estimates of the fraction of regions containing an associated SNP for each phenotype.** We show the estimates of $\frac{1}{1+e^{-\kappa}}$, the proportion of regions from the middle third of the distribution of gene density that contain associated SNPs (see Equation 7 in the main text), along with the 95% confidence interval of this parameter.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
BE2_C-DS14625	DNAse-I in BE(2)-C neuroblastoma cell line	5.25 [3.09, 7.10]	5.15	5.60 [3.51, 7.47]
HUVEC PF	Genome segmentation in HUVEC cells: promoter-flanking	7.47 [3.90,9.91]	7.18	8.51 [5.41, 10.62]

Table 3: **Combined model learned for BMI.** Shown are the exact annotation names and parameters learned for BMI, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	0.89 [0.01, 1.62]	0.88	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-1.05 [-2.68, 0.12]	-0.95	NA
fHeart-DS12810	DNase-I in fetal heart	2.83 [1.11, 4.40]	2.45	4.83 [3.08, 6.43]
fHeart-DS16621	DNase-I in fetal heart	2.12 [0.50, 3.64]	2.21	4.47 [2.76, 6.03]

Table 4: **Combined model learned for bone mineral density (femur).** Shown are the exact annotation names and parameters learned for FNBMD, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	0.52 [-0.40, 1.27]	0.53	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-1.49 [-3.65, -0.13]	-1.33	NA
HSMM_D-DS15542	DNase-I in skeletal muscle myoblasts	4.23 [2.24, 5.97]	3.75	3.90 [1.98, 5.58]

Table 5: **Combined model learned for bone mineral density (spine).** Shown are the exact annotation names and parameters learned for LSBMD, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) Effect [95% CI]	Penalized effect	Marginal effect [95%
High gene density	Regional annotation: top 1/3 of gene density	1.18 [0.61, 1.72]	1.18	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-2.18 [-4.10, -0.92]	-2.03	NA
fSkin_fibro_upper_back-DS19696	DNase-I in fetal skin fibroblasts from the upper back	5.21 [4.08, 6.20]	4.78	3.84 [2.63, 4.89]
gm12878.combined.R	Genome segmentation of GM12878: repressed	-1.83 [-3.06, -0.78]	-1.79	-2.35 [-4.50, -1.05
fSkin_fibro_abdomen-DS19561	DNase-I in fetal skin fibroblasts from abdomen	-2.34 [-3.85, -1.18]	-1.86	2.77 [1.27, 3.94]
huvec.combined.T	Genome segmentation of HUVEC: transcribed	1.20 [0.25, 2.15]	1.17	1.63 [0.61, 2.65]
Distance to TSS [0-5 kb]	From 0-5 kb from a TSS	1.18 [0.17, 2.15]	1.17	NA
Distance to TSS [5-10 kb]	From 5-10 kb from a TSS	0.45 [-1.38, 1.75]	0.40	NA

Table 6: **Combined model learned for Crohn's disease.** Shown are the exact annotation names and parameters learned for Crohn's disease, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
fStomach-DS17878	DNase-I in fetal stomach	3.66 [1.66, 5.31]	3.62	3.82 [1.66, 5.50]
Nonsynonymous	nonsynonymous SNPs	4.28 [1.53, 6.10]	4.13	4.95 [1.40, 6.95]
Distance to TSS [0-5 kb]	From 0-5 kb from a TSS	1.83 [0.22, 3.40]	1.75	NA
Distance to TSS [5-10 kb]	From 5-10 kb from a TSS	2.68 [0.76, 4.28]	2.54	NA

Table 7: **Combined model learned for fasting glucose.** Shown are the exact annotation names and parameters learned for FG, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) 95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	2.78 [1.98, 3.46]	2.80	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-40.6 [$-\inf$, -0.72]	-5.44	NA
HMVEC_dAd-DS12957	DNase-I in microvascular endothelium	4.91 [3.09, 6.52]	4.86	4.43 [2.60, 6.02]
k562.combined.T	Genome segmentation of K562: transcribed	2.15 [0.49, 3.77]	2.12	1.82 [0.01, 3.55]

Table 8: **Combined model learned for hemoglobin levels.** Shown are the exact annotation names and parameters learned for hemoglobin levels, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	1.69 [1.13, 2.19]	1.56	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-1.17 [-0.13, 0.69]	-0.20	NA
hepg2.combined.R	Genome segmentation of HepG2: repressed	-1.83 [-3.12, -0.68]	-1.79	-3.35 [-4.63, -2.19]
hepg2.combined.TSS	Genome segmentation of HepG2: TSS	3.10 [1.79, 4.20]	2.84	5.09 [3.91, 6.16]
ens_coding_exons	Ensembl: coding exons	3.16 [1.51, 4.40]	2.73	4.31 [2.73, 5.55]
k562.combined.R	Genome segmentation of K562: repressed	-1.43 [-2.65, -0.30]	-1.43	-2.90 [-4.08, -1.79]

Table 9: **Combined model learned for HDL levels.** Shown are the exact annotation names and parameters learned for HDL, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	1.50 [1.13, 1.86]	1.49	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-0.95 [-1.62, -0.36]	-0.94	NA
helas3.combined.R	Genome segmentation of HeLa: repressed	-1.50 [-2.39, -0.71]	-1.50	-2.74 [-3.78, -1.85]
fMuscle_lower_limb-DS18174	DNase-I in fetal muscle from lower limb	2.27 [1.50, 3.02]	2.24	3.61 [2.81, 4.40]
Nonsynonymous	Nonsynonymous SNPs	3.74 [2.55, 4.65]	3.58	4.27 [2.77, 5.32]
fLung-DS15573	DNase-I in fetal lung	2.09 [1.30, 2.80]	2.05	3.77 [2.97, 4.50]
huvec.combined.T	Genome segmentation of HUVEC: transcribed	1.27 [0.52, 1.96]	1.24	1.63 [0.89, 2.34]
ens_utr3_exons	Ensembl: 3' UTRs	1.57 [0.00, 2.64]	1.54	2.93 [1.34, 3.98]

Table 10: **Combined model learned for height.** Shown are the exact annotation names and parameters learned for height, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	1.77 [1.21, 2.27]	1.72	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-0.72 [-1.98, 0.25]	-0.71	NA
hepg2.combined.R	Genome segmentation of HepG2: repressed	-2.78 [-4.36, -1.51]	-2.70	-3.04 [-4.70, -1.76]
Nonsynonymous	Nonsynonymous SNPs	4.24 [2.74, 5.40]	3.97	4.89 [3.48, 6.02]
Distance to TSS [0-5 kb]	From 0-5 kb from a TSS	3.13 [1.96, 4.56]	2.84	NA
Distance to TSS [5-10 kb]	From 5-10 kb from a TSS	1.63 [-0.65, 3.12]	1.17	NA

Table 11: **Combined model learned for LDL levels.** Shown are the exact annotation names and parameters learned for LDL, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	1.56 [0.94, 2.11]	1.51	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-1.17 [-2.80, -0.01]	-1.10	NA
k562.combined.E	Genome segmentation of K562: enhancers	3.68 [2.47, 4.75]	3.53	5.67 [4.49, 6.74]
k562.combined.R	Genome segmentation of K562: repressed	-3.17 [-4.80, -1.86]	-2.97	-3.94 [-5.57, -2.61]
hTH17-DS11039	DNase-I in Th17 T cells	2.21 [0.35, 3.51]	2.06	4.53 [2.93, 5.74]

Table 12: **Combined model learned for mean cell hemoglobin.** Shown are the exact annotation names and parameters learned for MCH, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	1.11 [0.09, 1.93]	1.17	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-1.57 [-4.41, 0.10]	-1.36	NA
k562.combined.R	Genome segmentation of K562: repressed	-3.81 [-7.43, -1.79]	-3.42	-4.34 [-8.94, -2.27]
K562-DS9767	DNase-I in K562 cells	2.67 [0.61, 4.44]	2.47	4.46 [2.60, 6.22]
Nonsynonymous	Nonsynonymous SNPs	4.66 [1.90, 6.52]	4.03	4.27 [0.97, 6.25]

Table 13: **Combined model learned for mean corpuscular hemoglobin concentration.** Shown are the exact annotation names and parameters learned for MCHC, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	1.36 [0.76, 1.86]	1.31	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-1.51 [-3.06, -0.39]	-1.46	NA
k562.combined.R	Genome segmentation of K562: repressed	-3.91 [-6.25, -2.38]	-3.69	-5.24 [-7.76, -3.59]
k562.combined.E	Genome segmentation of K562: enhancer	3.10 [1.86, 4.15]	2.96	5.67 [4.47, 6.77]
hTH17-DS11039	DNase-I in Th17 T cells	2.31 [0.81, 3.48]	2.25	5.40 [4.21, 6.46]
Nonsynonymous	Nonsynonymous SNPs	4.54 [2.34, 5.92]	4.13	5.11 [3.26, 6.39]
CMK-DS12393	DNase-I in CMK leukemia line	1.28 [0.04, 2.35]	1.34	4.52 [3.30, 5.64]
Distance to TSS [0-5 kb]	From 0-5 kb from a TSS	0.38 [-1.59, 0.65]	-0.33	NA
Distance to TSS [5-10 kb]	From 5-10 kb from a TSS	0.89 [-0.40, 1.83]	0.84	NA

Table 14: **Combined model learned for mean red cell volume.** Shown are the exact annotation names and parameters learned for MCV, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	1.95 [1.30, 2.52]	1.88	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-2.06 [-4.73, -0.40]	-1.63	NA
CD34-DS12274	DNase-I in CD34+ cells	3.02 [1.69, 4.26]	2.76	4.37 [2.99, 5.64]
gm12878.combined.T	Genome segmentation of GM12878: transcribed	2.35 [1.07, 3.53]	1.83	1.86 [0.59, 3.04]
helas3.combined.E	Genome segmentation of HeLa: enhancer	2.80 [0.75, 4.23]	2.27	3.35 [0.16, 5.09]
fSpleen-DS17448	DNase-I in fetal spleen	1.93 [0.59, 3.15]	1.88	3.65 [2.22, 4.92]

Table 15: **Combined model learned for mean platelet volume.** Shown are the exact annotation names and parameters learned for MPV, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	4.24 [3.36, 4.95]	3.72	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-40.60 [-inf, 0.94]	-1.92	NA
Nonsynonymous	Nonsynonymous SNPs	4.11 [1.34, 6.07]	3.61	5.34 [2.83, 7.23]
fStomach-DS17172	DNase-I in fetal stomach	3.90 [1.40, 6.17]	3.48	4.78 [2.54, 7.03]

Table 16: **Combined model learned for packed red cell volume.** Shown are the exact annotation names and parameters learned for PCV, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	1.67 [2.81, 2.64]	2.14	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-1.63 [-3.40, -0.38]	-1.51	NA
k562.combined.R	Genome segmentation in K562: repressed	-1.60 [-2.63, -0.66]	-1.60	-2.60 [-3.65, -1.64]
CD34-DS12274	DNase-I in CD34+ cells	1.82 [0.59, 2.86]	1.80	3.39 [2.24, 4.43]
Nonsynonymous	Nonsynonymous SNPs	3.38 [1.31, 4.79]	3.00	3.98 [2.02, 5.38]
huvec.combined.E	Genome segmentation in HUVEC: enhancers	1.67 [0.16, 2.84]	1.59	3.27 [1.82, 4.41]
helas3.combined.R	Genome segmentation in HeLa: repressed	-1.17 [-2.37, -0.13]	-1.14	-2.18 [-3.40, -1.11]

Table 17: **Combined model learned for platelet count.** Shown are the exact annotation names and parameters learned for PLT, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	1.99 [1.33, 2.58]	1.96	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-2.74 [-6.97, -0.59]	-2.18	NA
fStomach-DS17878	DNAse-I in fetal stomach	5.31 [3.87, 6.91]	4.83	4.83 [3.30, 6.45]
k562.combined.E	Genome segmentation of K562: enhancer	1.53 [-0.04, 2.83]	1.56	4.28 [1.41, 5.90]
fKidney_renal_pelvis_R-DS18663	DNase-I in fetal renal pelvis	-3.49 [-7.68, -1.56]	-2.80	2.48 [0.04, 4.17]
K562-DS9767	DNase-I in K562 leukemia line	2.28 [0.97, 3.58]	2.25	4.50 [2.97, 5.97]

Table 18: **Combined model learned for red blood cell count.** Shown are the exact annotation names and parameters learned for RBC, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	1.05 [0.48, 1.56]	1.04	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-1.40 [-2.74, -0.39]	-1.34	NA
hepg2.combined.R	Genome segmentation of HepG2: repressed	-2.90 [-4.36, -1.72]	-2.84	-3.19 [-4.76, -1.95]
Nonsynonymous	Nonsynonymous SNPs	4.36 [2.90, 5.48]	4.18	4.89 [3.51, 5.99]
Distance to TSS [0-5 kb]	From 0-5 kb from a TSS	2.76 [1.62, 4.15]	2.58	NA
Distance to TSS [5-10 kb]	From 5-10 kb from a TSS	1.88 [-0.27, 3.29]	1.56	NA

Table 19: **Combined model learned for total cholesterol.** Shown are the exact annotation names and parameters learned for total cholesterol, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

Annotation	Description	$\log_2$(Effect) [95% CI]	Penalized effect	Marginal effect [95% CI]
High gene density	Regional annotation: top 1/3 of gene density	1.56 [0.85, 2.18]	1.49	NA
Low gene density	Regional annotation: bottom 1/3 of gene density	-0.82 [-2.65, 0.40]	-0.78	NA
hepg2.combined.R	Genome segementation of HepG2: repressed	-4.24 [-6.68, -2.47]	-3.75	-4.56 [-7.11, -2.76]
ens_utr3_exons	Ensembl: 3' UTRs	3.87 [2.11, 5.28]	3.46	4.60 [2.86, 6.03]

Table 20: **Combined model learned for triglyceride levels.** Shown are the exact annotation names and parameters learned for triglycerides, along with the penalized effect sizes and the effect of each annotation in a single-annotation model.

	PPA		P-value		combined	
Phenotype	True positives	False positives	True positives	False positives	True positives	False positives
HDL	45	0	43	1	48	1
LDL	43	1	40	0	44	1
TC	47	0	51	0	51	0
TG	27	0	29	0	30	0

Table 21: **Comparison of loci identified in the lipids data with different methods.** We ranked genomic regions in GWAS of four lipid traits according to their minimum P-value or posterior probability of association from Teslovich et al. [2010]. We then evaluated false positives and false negatives by comparison to a larger GWAS [Global Lipids Genetics Consortium et al., 2013]. See Supplementary Text for details.

trait	region (hg19)	Regional PPA	lead SNP (P-value)	Nearest gene	Successful replication (SNP, r^2 with lead)
BMI	chr13:27,75,5426-29,745,954	0.94	rs9512699 (6×10^{-8})	MTIF3	[Speliotes et al., 2010] (rs4771122, 0.73)
BMD (femur)	chr1:170,892,281-173,086,517	0.93	rs6701929 (2×10^{-7})	DNM3	[Estrada et al., 2012] (rs479336, 0.93)
HDL	chr1:25,427,217-29,426,896	0.96	rs6659176 (1.5×10^{-6})	NR0B2	[Global Lipids Genetics Consortium et al., 2013] (rs12748152, 0.85)
HDL	chr1:93,534,311-95,828,501	0.93	rs2297707 (1×10^{-6})	TMED5	[Global Lipids Genetics Consortium et al., 2013] (rs12133576, 0.79)
HDL	chr1:108,743,042-111,481,349	0.97	rs12740374 (6×10^{-8})	CELSR2	[Global Lipids Genetics Consortium et al., 2013] (rs12740374)
HDL	chr2:85,349,339-88,736,950	0.98	rs1044973 (1.5×10^{-7})	TGOLN2	No (sample size not increased in Global Lipids Genetics Consortium et al. [2013])
HDL	chr10:45,535,916-50,321,467	0.93	rs10900223 (1.4×10^{-7})	MARCH8	[Global Lipids Genetics Consortium et al., 2013] (rs970548, 0.99)
MCV	chr3:139,060,509-141,377,851	0.98	rs13059128 (3.8×10^{-7})	ZBTB38	[van der Harst et al., 2012] (rs6776003, 0.48)
MCV	chr9:134,164,493-136,620,584	0.90	rs8176662 (7.5×10^{-7})	ABO	NA
MCV	chr20:24,615,239-30,836,608	0.98	rs6088962 (7.5×10^{-7})	BCL2L1	NA
TG	chr16:31,050,033-49,644,030	0.95	rs1549293 (2.7×10^{-7})	KAT8	[Global Lipids Genetics Consortium et al., 2013] (rs749671, 0.80)
LDL	chr1:91,146,258-93,672,688	0.97	rs7542747 (2.3×10^{-7})	RPAP2	[Global Lipids Genetics Consortium et al., 2013] (rs4970712, 0.75)
LDL	chr1:146,751,272-152,014,485	0.98	rs267733 (7×10^{-8})	ANXA9	[Global Lipids Genetics Consortium et al., 2013] (rs267733)
LDL	chr2:116,901,934-119,001,466	0.98	rs1052639 (6.6×10^{-8})	DDX18	[Global Lipids Genetics Consortium et al., 2013] (rs10490626, 0.53)
LDL	chr13:31,693,235-34,119,073	0.93	rs4942505 (9.8×10^{-8})	BRCA2	[Global Lipids Genetics Consortium et al., 2013] (rs4942505)
LDL	chr17:7,456,344-9,908,665	0.92	rs4791641 (2.6×10^{-7})	PFAS	No ($P = 1.3 \times 10^{-7}$ in [Global Lipids Genetics Consortium et al., 2013])
MCHC	chr7:76,062,644-78,334,941	0.93	rs58176556 (5.4×10^{-8})	PHTF2	NA
Height	chr2:240,701,166-243,060,642	0.98	rs13006939 (3.9×10^{-7})	SEPT2	[Lango-Allen et al., 2010] (rs12694997, 0.99)
Height	chr3:11,167,568-13,294,698	0.98	rs2276749 (3.0×10^{-6})	VGLL2	NA
Height	chr3:13,294,698-15,353,840	0.93	rs2597513 (1.1×10^{-7})	HDAC11	[Lango-Allen et al., 2010] (rs2597513)
Height	chr3:55,068,506-57,000,141	0.94	rs7637449 (1.3×10^{-6})	CCDC66	[Lango-Allen et al., 2010] (rs9835332, 0.87)
Height	chr4:72,048-2,570,837	0.98	rs3958122 (6.0×10^{-8})	SLBP	[Lango-Allen et al., 2010] (rs2247341, 0.99)
Height	chr5:71,376,237-73,712,303	0.98	rs34651 (2.5×10^{-7})	TNPO1	NA
Height	chr6:108,017,102-110,694,347	0.95	rs1476387 (2.2×10^{-6})	SMPD2	[Lango-Allen et al., 2010] (rs1046943, 0.93)
Height	chr7:22,074,248-23,998,552	0.99	rs12534093 (5.6×10^{-8})	IGF2BP3	[Lango-Allen et al., 2010] (rs12534093)
Height	chr7:46,327,426-48,083,339	0.97	rs12538905 (2.6×10^{-7})	IGFBP3	NA
Height	chr9:87,279,007-89,667,667	0.90	rs405761 (1.3×10^{-7})	ZCCHC6	[Lango-Allen et al., 2010] (rs8181166, 0.82)
Height	chr11:12,559,691-14,685,886	1.0	rs7926971 (7.3×10^{-8})	TEAD1	[Lango-Allen et al., 2010] (rs7926971)
Height	chr11:14,685,886-17,491,336	0.93	rs757081 (2.2×10^{-6})	NUCB2	[Lango-Allen et al., 2010] (rs1330, 0.60)
Height	chr15:62,349,517-64,370,301	0.97	rs7178424 (2.2×10^{-7})	C2CD4A	[Lango-Allen et al., 2010] (rs7178424)
Height	chr17:19,924,256-26,838,292	0.96	rs9895199 (3.6×10^{-7})	KCNJ12	[Lango-Allen et al., 2010] (rs4640244, 0.79)
Height	chr17:45,331,502-47,944,460	0.99	rs9904645 (2.2×10^{-7})	ATP5G1	NA
Height	chr22:32,075,899-33,846,972	0.97	rs1012366 (6.9×10^{-8})	SYN3	[Lango-Allen et al., 2010] (rs4821083 [not in 1000 Genomes])
Crohn's	chr2:42,522,756-44,575,426	1.0	rs17031095 (2.6×10^{-7})	THADA	[Jostins et al., 2012] (rs10495903, 0.95)
Crohn's	chr10:59,615,595-61,881,674	1.0	rs1832556 (2.0×10^{-7})	IPMK	[Jostins et al., 2012] (rs2790216, 0.94)
Crohn's	chr11:61,269,649-64,734,682	0.98	rs174568 (2.8×10^{-7})	FADS2	[Jostins et al., 2012] (rs4246215, 0.86)
Crohn's	chr13:99,900,420-102,096,823	0.94	rs3742130 (2.3×10^{-5})	GPR18	[Jostins et al., 2012] (rs9557195, 0.91)
Crohn's	chr15:67,140,517-70,199,927	0.93	rs11639295 (6.4×10^{-7})	SMAD3	[Jostins et al., 2012] (rs17293632, 0.10)
Crohn's	chr17:17,986,955-26,038,545	0.92	rs2945406 (4.1×10^{-7})	KSR1	[Jostins et al., 2012] (rs2945412, 0.13)
PLT	chr1:44,022,121-47,087,366	0.99	rs4468203 (3.2×10^{-7})	GPBP1L1	NA
PLT	chr9:90,221,450-92,241,847	0.90	rs9410382 (1.9×10^{-6})	S1PR3	NA
PLT	chr11:32,343,164-34,501,064	0.93	rs7481878 (7.2×10^{-8})	QSER1	NA
MCH	chr4:86,147,717-88,340,969	0.98	rs6819155 (2.3×10^{-7})	APP1	NA
MCH	chr14:102,971,016-107,289,436	0.93	rs17616316 (1.5×10^{-7}	EIF5	[van der Harst et al., 2012] (rs17616316)
HB	chr15:75,349,145-78,654,148	0.90	rs1874953 (4.2×10^{-7})	NRG4	[van der Harst et al., 2012] (rs11072566, 0.93)
BMD (spine)	chr17:43,556,652-46,084,026	0.99	rs117504376 (3.1×10^{-7})	MAPT (chr17 inversion)	[Estrada et al., 2012] (rs1864325, 0.99)
RBC	chr20:54,899,828-57,013,873	0.96	rs737092 (4.5×10^{-7})	MIR5095	[van der Harst et al., 2012] (rs737092)
MPV	chr14:67,315,438-69,802,709	0.91	rs117823369 (3.9×10^{-6}	DCAF5	NA
FG	chr9:111,051,626 - 112,662,634	0.96	rs76817627 (3.4×10^{-7})	FAM206A	NA

Table 22: **Sub-threshold associations with high posterior probability.** In each GWAS, we identified regions of the genome with a posterior probability of association greater than 0.9 but with no P-values less than 5×10^{-8}. Shown are the positions of these regions for each trait. See Supplementary Text for details. LD between lead SNPs and replication SNPs was computed from the 1000 Genomes Project haplotypes in Europeans; the exact file versions are listed in Section 3.

References

Bacanu, S.-A., Devlin, B., and Roeder, K., 2002. Association studies for quantitative traits in structured populations. *Genetic epidemiology*, **22**(1):78–93.

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

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

Global Lipids Genetics Consortium, Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., *et al.*, 2013. Discovery and refinement of loci associated with lipid levels. *Nat Genet*, **45**(11):1274–83.

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

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

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

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

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

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

R Core Team, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

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

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

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

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., Burtt, N. P., Fuchsberger, C., Li, Y., Erdmann, J., *et al.*, 2012. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet*, **8**(8):e1002793.