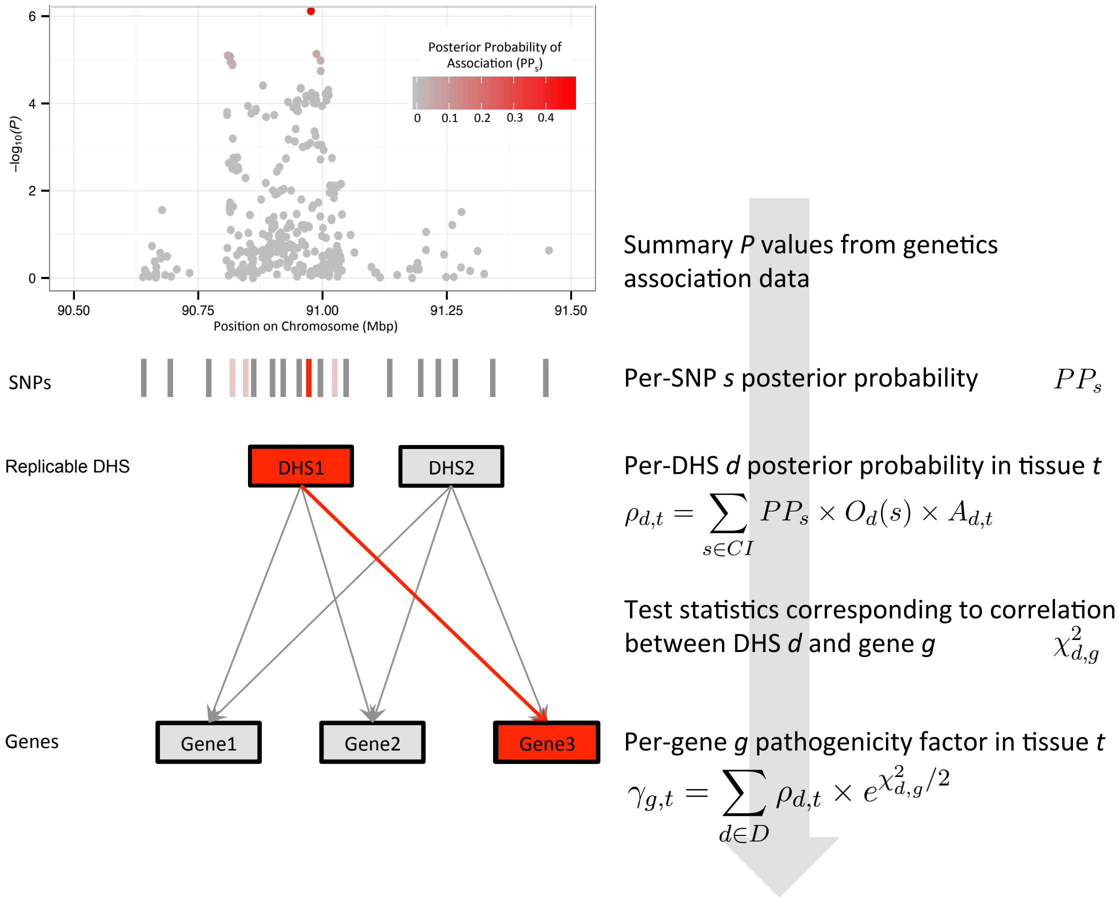
The American Journal of Human Genetics, Volume *101*

Supplemental Data

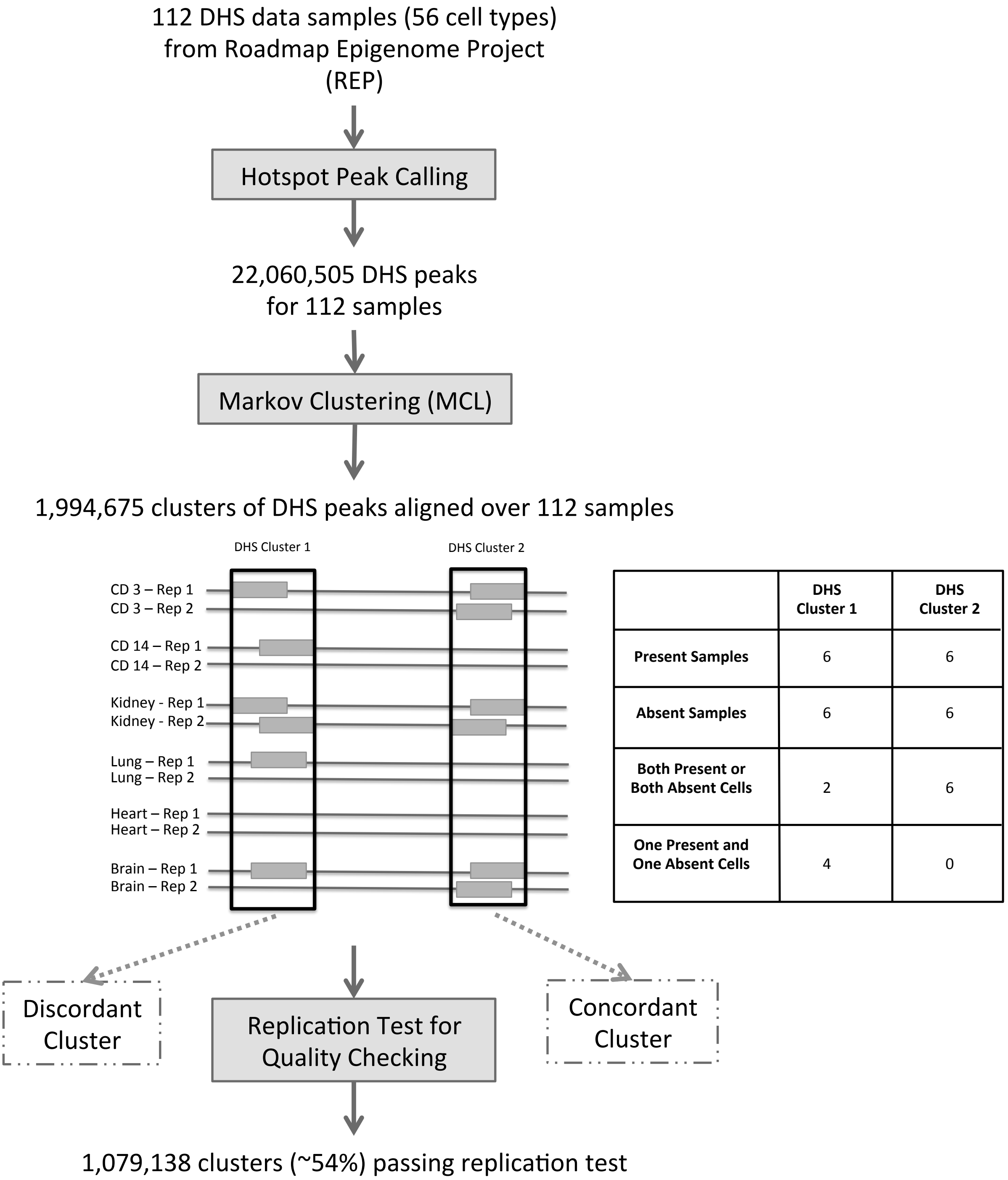
Integrative Genetic and Epigenetic Analysis Uncovers Regulatory Mechanisms of Autoimmune Disease

Parisa Shooshtari, Hailiang Huang, and Chris Cotsapas



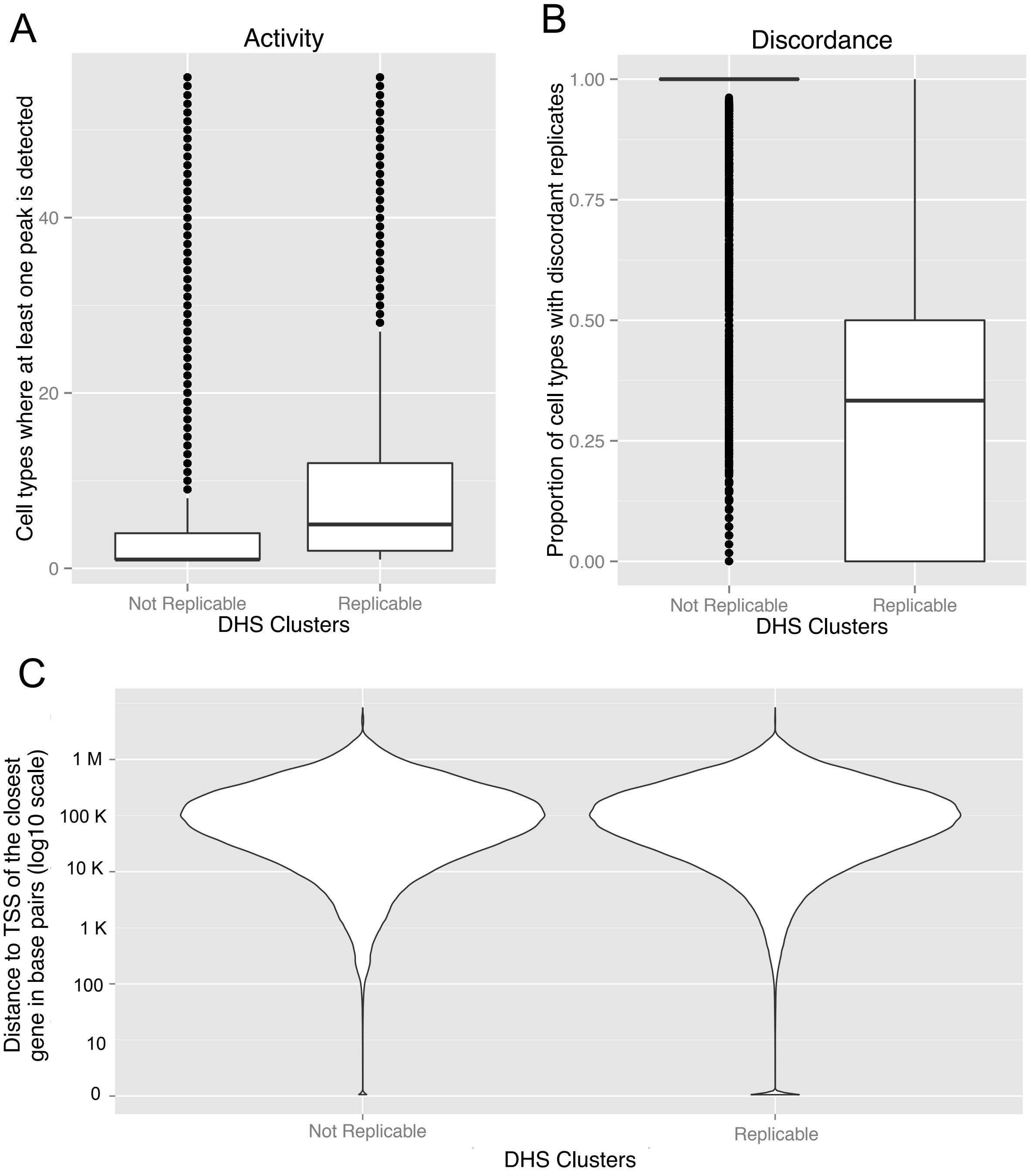
Posterior Probability of Association (PP_s)
0
0.1
0.2
0.3
0.4
6
4
2
0
$-\log_{10}(P)$
90.50
90.75
91.00
91.25
91.50
Position on Chromosome (Mbp)
SNPs
Replicable DHS
DHS1
DHS2
Genes
Gene1
Gene2
Gene3
Summary *P* values from genetics association data
Per-SNP *s* posterior probability PP_s
Per-DHS *d* posterior probability in tissue *t*
$\rho_{d,t} = \sum_{s \in CI} PP_s \times O_d(s) \times A_{d,t}$
Test statistics corresponding to correlation between DHS *d* and gene *g* $\chi^2_{d,g}$
Per-gene *g* pathogenicity factor in tissue *t*
$\gamma_{g,t} = \sum_{d \in D} \rho_{d,t} \times e^{\chi^2_{d,g}/2}$

Figure S1: Overview of our regulatory fine-mapping approach. We begin with disease association summary statistics, then identify the credible interval (CI) set of SNPs and calculate posterior probabilities of association for each variant in the set. We then overlap these variants with DNase I Hypersensitive sites (DHS), and calculate the regulatory potential ρ as the proportion of posterior probability on each DHS. The sum of ρ across all DHS in a locus captures how likely the association is to be mediated by regulatory variation. We assess the significance of ρ in each tissue empirically, by randomizing the positions of each active DHS in the locus and recomputing. We use the χ^2 value derived from the correlation between DHS accessibility and nearby gene expression after correction for data structure, to calculate γ, the pathogenicity factor for each gene in the locus. We again calculate the significance of γ empirically, by randomly resampling values from the matrix of DHS–gene χ^2 correlation statistics.



	DHS Cluster 1	DHS Cluster 2
Present Samples	6	6
Absent Samples	6	6
Both Present or Both Absent Cells	2	6
One Present and One Absent Cells	4	0

Figure S2: Overview of our DHS clustering and QC pipeline. Our approach requires correlating gene expression to DHS accessibility across tissues. We therefore identify DHS peaks in each tissue corresponding to the same underlying regulatory region with Markov Clustering. To confirm the validity of these clusters, we perform a statistical replication test using replicate samples of the 56 Roadmap Epigenomics Project tissues as described in the Methods. We find that 54% of clusters show nominal evidence of replication (chi-squared $p < 0.05$).



A
Activity
Cell types where at least one peak is detected
0
20
40
Not Replicable
Replicable
DHS Clusters
B
Discordance
Proportion of cell types with discordant replicates
0.00
0.25
0.50
0.75
1.00
Not Replicable
Replicable
DHS Clusters
C
Distance to TSS of the closest gene in base pairs (log10 scale)
0
10
100
1 K
10 K
100 K
1 M
Not Replicable
Replicable
DHS Clusters

Figure S3: Characteristics of replicable and non-replicable DHS clusters. We clustered DHS peaks across two replicates in each of 56 REMC cell types, and applied a statistical test to determine which of these clusters show significant concordance across replicates (peaks seen in both or neither replicate, indicating an active or inactive state, respectively). We find that the 1,079,138/1,994,675 (54.1%) of clusters showing evidence of replication ($p < 0.05$) have different properties from the remaining clusters. (A) Replicable DHS are detected in more cell types. (B) Replicable DHS clusters show greater concordance between technical replicates, and we detect a DHS peak within the boundaries of the cluster in both replicates for 65% of active cell types. Non–replicable clusters only show evidence of activity in one technical replicate of a cell type the vast majority of the time. (C) Replicable and non–replicable DHS clusters are similarly positioned relative to the closest transcriptional start site (TSS). Promoter elements (those $<$ 100 bp from a TSS) tend to replicate, as shown by the second mode near zero in the right–hand distribution.

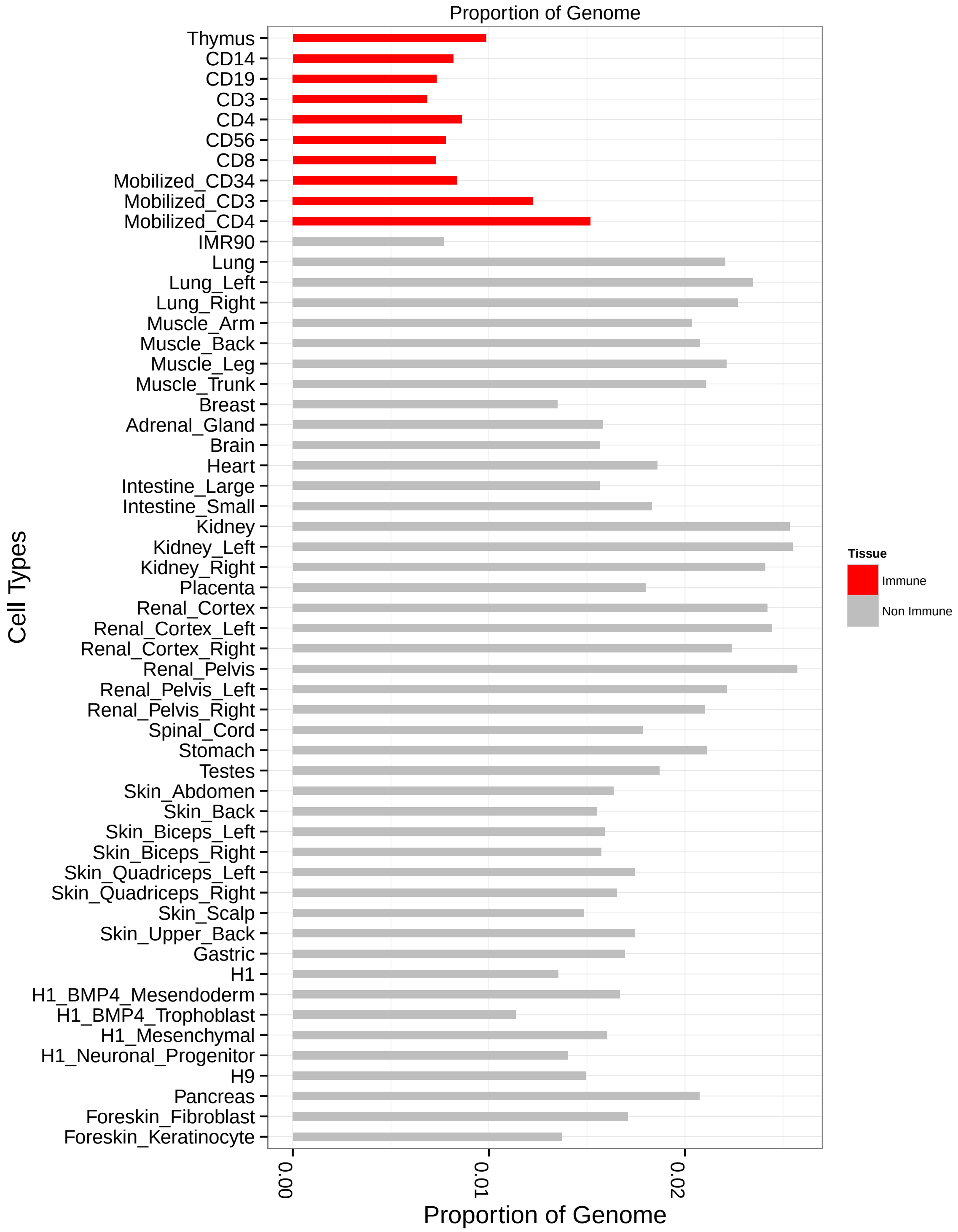


Figure S4: Proportion of the autosomal genome covered by replicable DHS clusters active in each Roadmap Epigenomics Project tissue. We define each of the 1,079,138 replicable DHS clusters as accessible in each tissue if we observe at least one DHS peak within its boundaries in one of the replicate samples in the underlying Roadmap data. We find that these clusters occupy 1.5–2.5% of the autosomal genome overall, and that clusters active in immune cell subpopulations (red) occupy a relatively small portion of the genome, around 1%.

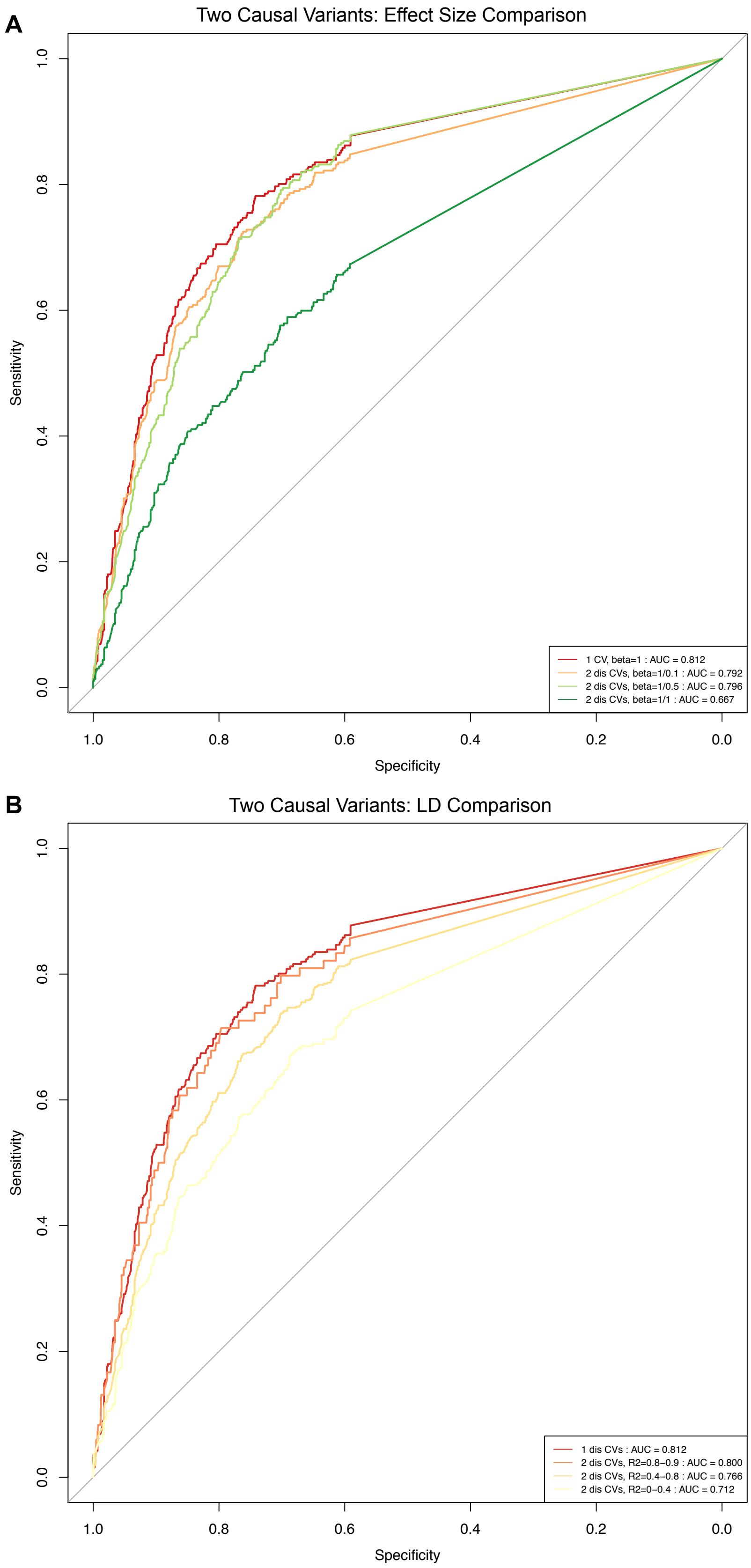

A
Two Causal Variants: Effect Size Comparison
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
1 CV, beta=1 : AUC = 0.812
2 dis CVs, beta=1/0.1 : AUC = 0.792
2 dis CVs, beta=1/0.5 : AUC = 0.796
2 dis CVs, beta=1/1 : AUC = 0.667
B
Two Causal Variants: LD Comparison
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
1 dis CVs : AUC = 0.812
2 dis CVs, R2=0.8–0.9 : AUC = 0.800
2 dis CVs, R2=0.4–0.8 : AUC = 0.766
2 dis CVs, R2=0–0.4 : AUC = 0.712

Figure S5: Statistical power and multiple independent associations in a locus. To assess the statistical power of our framework, we performed a series of simulations where we specified either one or two causal variants in a locus (as previously described in Chun *et. al.*[36]). We randomly selected one REMC cell type, fetal kidney, from which to draw DHS cluster data for these simulations. We performed positive simulations where the causal variant is on a DHS cluster, and negative simulations where it is not. For two variants, we performed positive simulations where the first causal variant is on a DHS cluster and the second is not, and negative simulations where neither is on a DHS cluster. We also varied the strength of the second variant's effect size relative to the first (shown in panel A), and the linkage disequilibrium between the two variants (shown in panel B), to establish how these parameters affect our ability to assess regulatory potential. (A) The single–variant simulations show our method is well–powered (AUC = 0.812, red line). It remains so when a second, weaker causal variant that is not on a DHS is introduced (yellow, light green), until the effect size of the second effect becomes equal to that of the primary causal variant (dark green). (B) When the two independent variants are in high LD, our power to detect regulatory potential remains high (orange and dark yellow) relative to the single–variant scenario. This is because the credible intervals due to the two causal variants would largely overlap and thus the combined data signal identifies the majority of the same variants. As LD reduces, our power diminishes as the combined association test statistics begin to become noisier, making the credible interval identification less accurate.

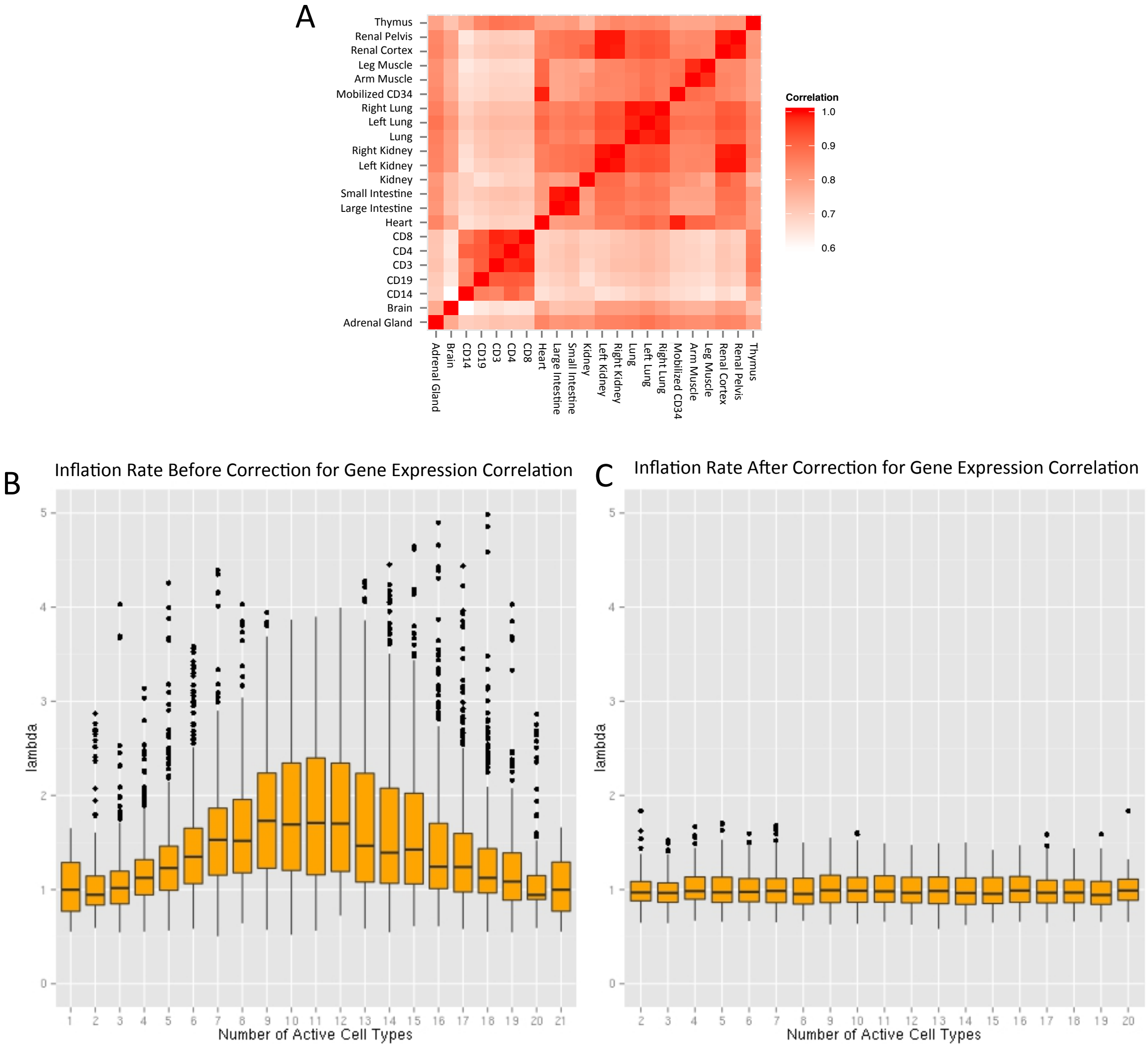


Figure S6: Adjusting for the correlation structure in gene expression data. (A) There is correlation between gene expressions of different cell types from relevant tissues. (B) This causes inflation in the P value of correlation between a DHS and the genes. (C) By adjusting for the correlation structure of gene expression data, we substantially reduced this inflation.

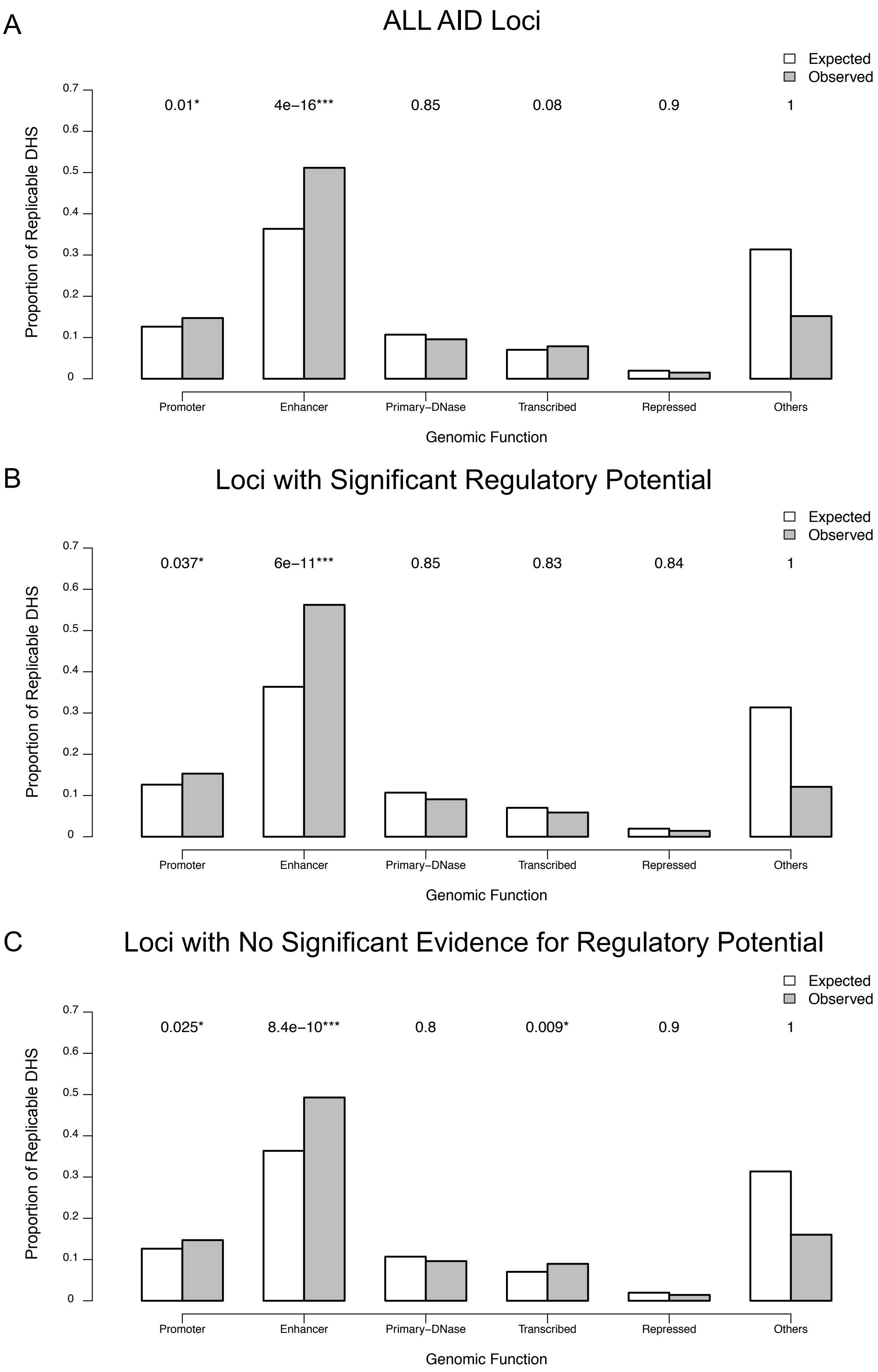

A
ALL AID Loci
□ Expected
▨ Observed
0.01*
4e−16***
0.85
0.08
0.9
1
Proportion of Replicable DHS
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Promoter
Enhancer
Primary−DNase
Transcribed
Repressed
Others
Genomic Function
B
Loci with Significant Regulatory Potential
□ Expected
▨ Observed
0.037*
6e−11***
0.85
0.83
0.84
1
Proportion of Replicable DHS
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Promoter
Enhancer
Primary−DNase
Transcribed
Repressed
Others
Genomic Function
C
Loci with No Significant Evidence for Regulatory Potential
□ Expected
▨ Observed
0.025*
8.4e−10***
0.8
0.009*
0.9
1
Proportion of Replicable DHS
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Promoter
Enhancer
Primary−DNase
Transcribed
Repressed
Others
Genomic Function

Figure S7: Functional annotation of replicable DHS harboring credible interval SNPs across 301 autoimmune and inflammatory risk loci. By overlapping these replicable DHS with ChromHMM data, we find they are significantly enriched for active enhancer ($P = 4e - 16$) and promoter annotation ($P = 0.01$, not significant after Bonferroni correction for number of annotations tested, panel A). We observe this enrichment both in the loci with significant regulatory potential and those with no significant evidence of risk on replicable DHS (panels B and C respectively).

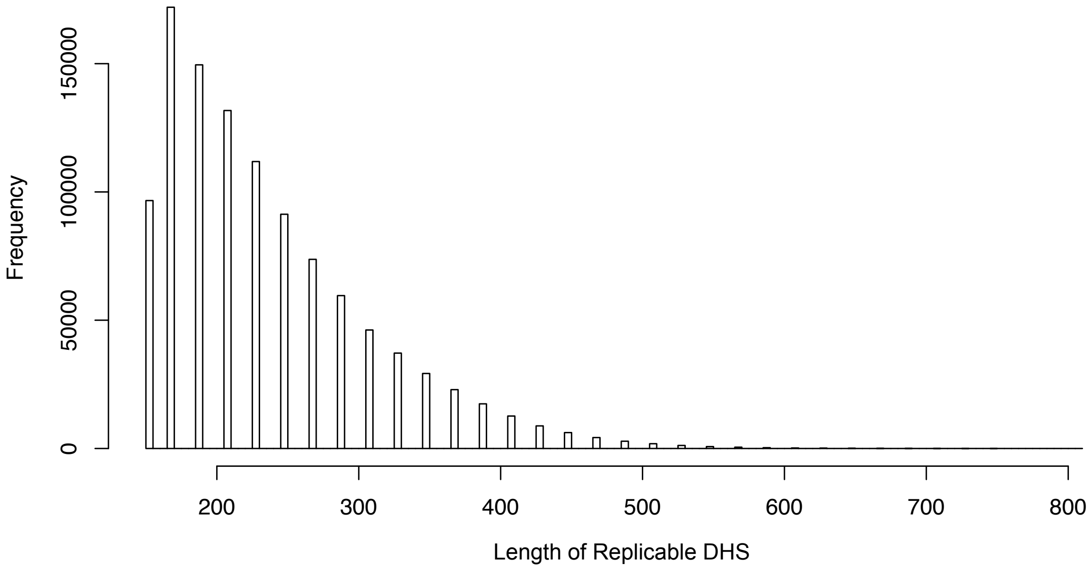


Figure S8: Replicable DHS size distribution. In 56 REP tissues with at least two replicate DHS sequencing runs, we called 22,060,505 narrow–sense 150bp peaks at a false discovery rate FDR $< 1\%$. We assembled these peaks into 1,994,675 DHS clusters using Markov clustering. Of these, 1,079,138 (54.1%) covering 8.5% of the genome passed nominal significance in a statistical replication test across the 56 pairs of samples (χ^2_1 test, $p < 0.05$). Shown here is the size distribution of the DHS clusters passing the replication test.

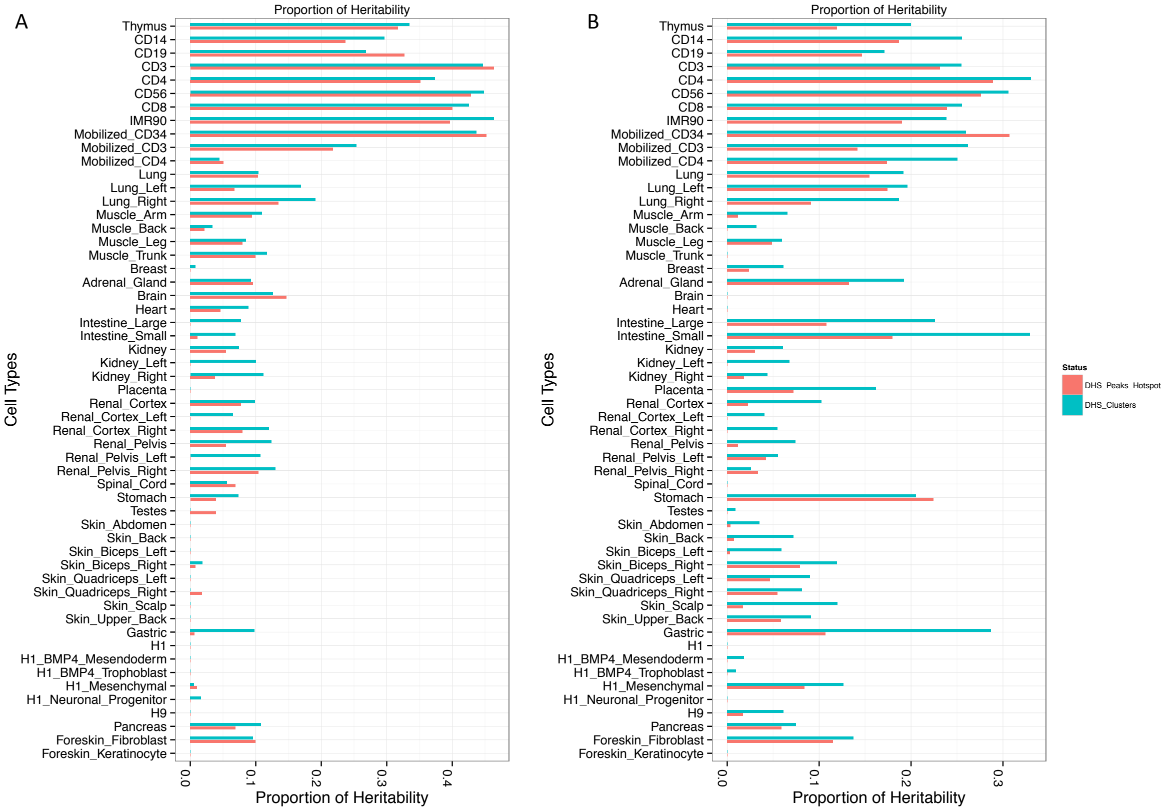


Figure S9: Proportion of heritability explained by replicable DHS. Of our DHS clusters, only 1,079,138/1,994,675 (54.1%) pass our statistical replication threshold. We reasoned that if our clustering was poorly calibrated or our statistical replication statistic underpowered, we would only capture about half the useful information across called DHS peaks. To evaluate this hypothesis, we asked if the 1,079,138 DHS clusters passing these filters explain a substantial amount of disease heritability ($h2g$). We used the partitioning heritability approach[6] to compute the proportion of disease heritability attributable in each REP tissue to (i) all DHS peaks called by Hotspot across the two sample replicates; and (ii) the subset of the 1,079,138 replicable DHS active in that tissue. As we called replicable DHS genome–wide and heritability estimates are also made genome–wide[6], we used publicly–available summary statistics from multiple sclerosis[22] (left) and inflammatory bowel disease[23] (right) for these analyses, rather than analyze Immunochip loci alone. We found that the 14.8% of the autosomal genome covered by all DHS peaks and the 8.5% covered by our replicable DHS explained effectively the same heritability, indicating our approach identifies most DHS relevant to disease pathogenesis. We note that differences between the heritability estimates are not significant and lie within the standard error of the estimates.

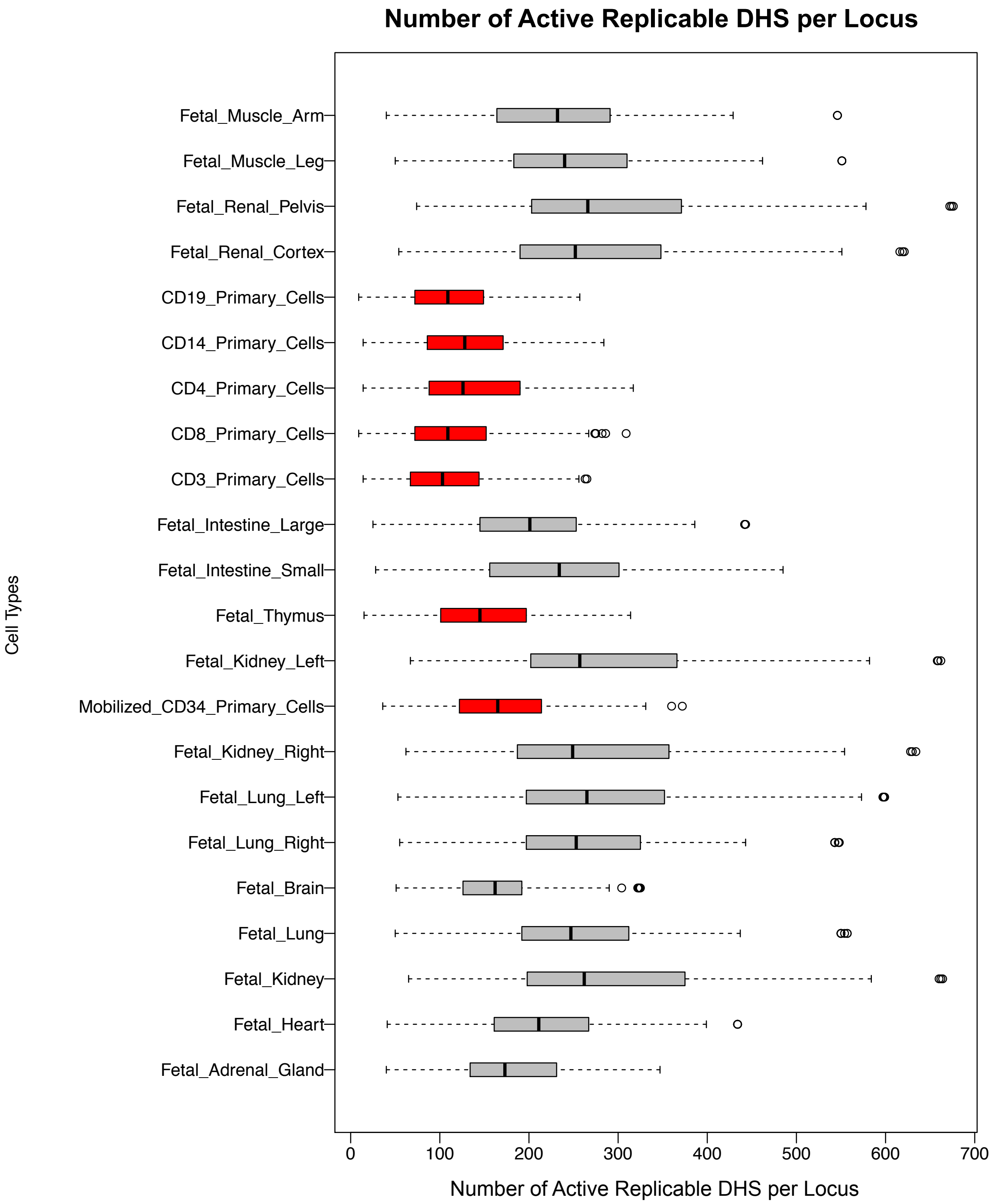


Figure S10: Number of replicable DHS in each risk locus accessible in Roadmap Epigenomics Project tissues. Different numbers of replicable DHS are accessible in each tissue. The immune cell subpopulations (red) have relatively low accessibility rates, indicating that the significant enrichment of risk signal that we observe on these DHS is not driven by sampling artifact.

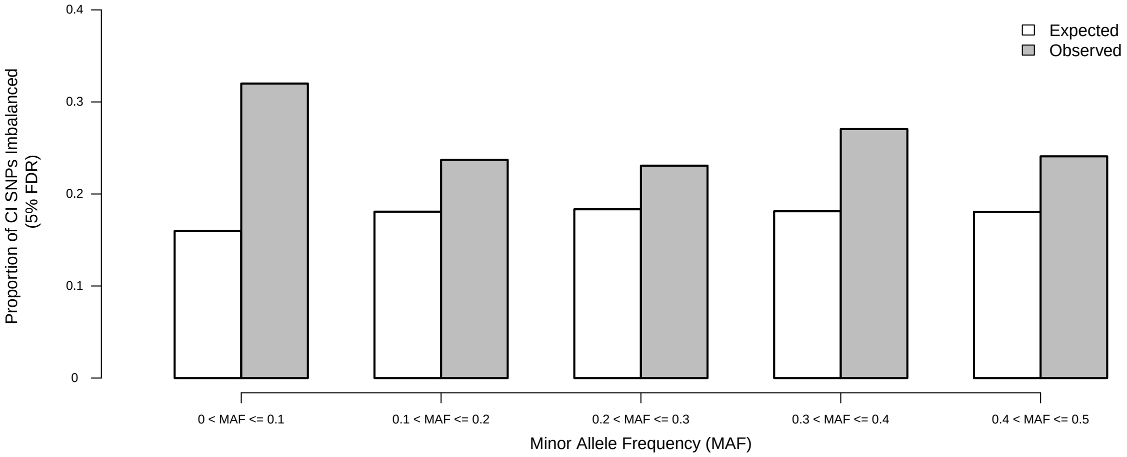


Figure S11: Replicable DHS harboring credible interval (CI) SNPs are more likely to show allele–specific accessibility across 301 loci with genome–wide association to one of nine autoimmune and inflammatory diseases. We observe that CI SNPs appear to alter the accessibility of replicable DHS in which they reside, and that this effect is independent of the minor allele frequency (MAF) of those SNPs.

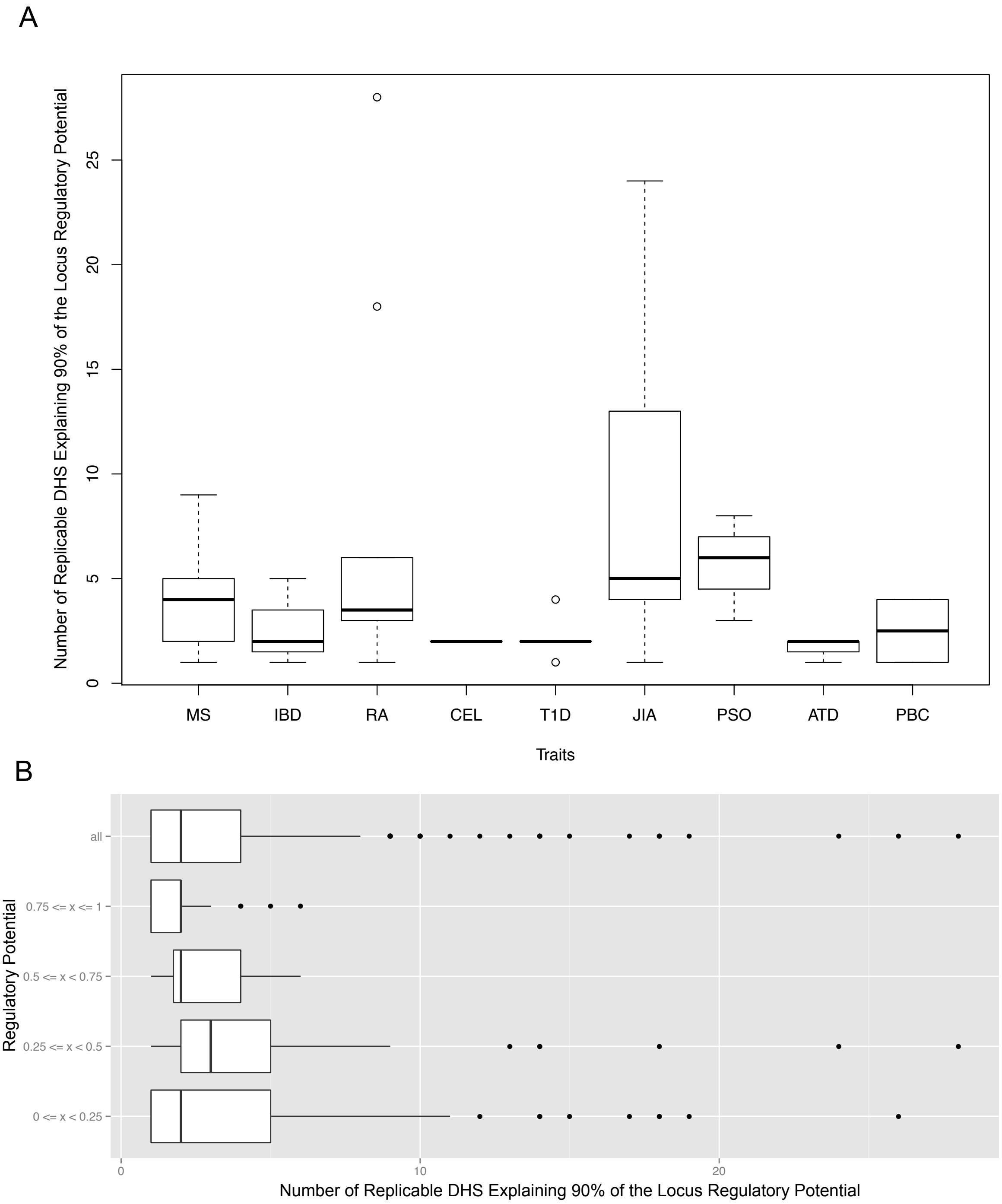
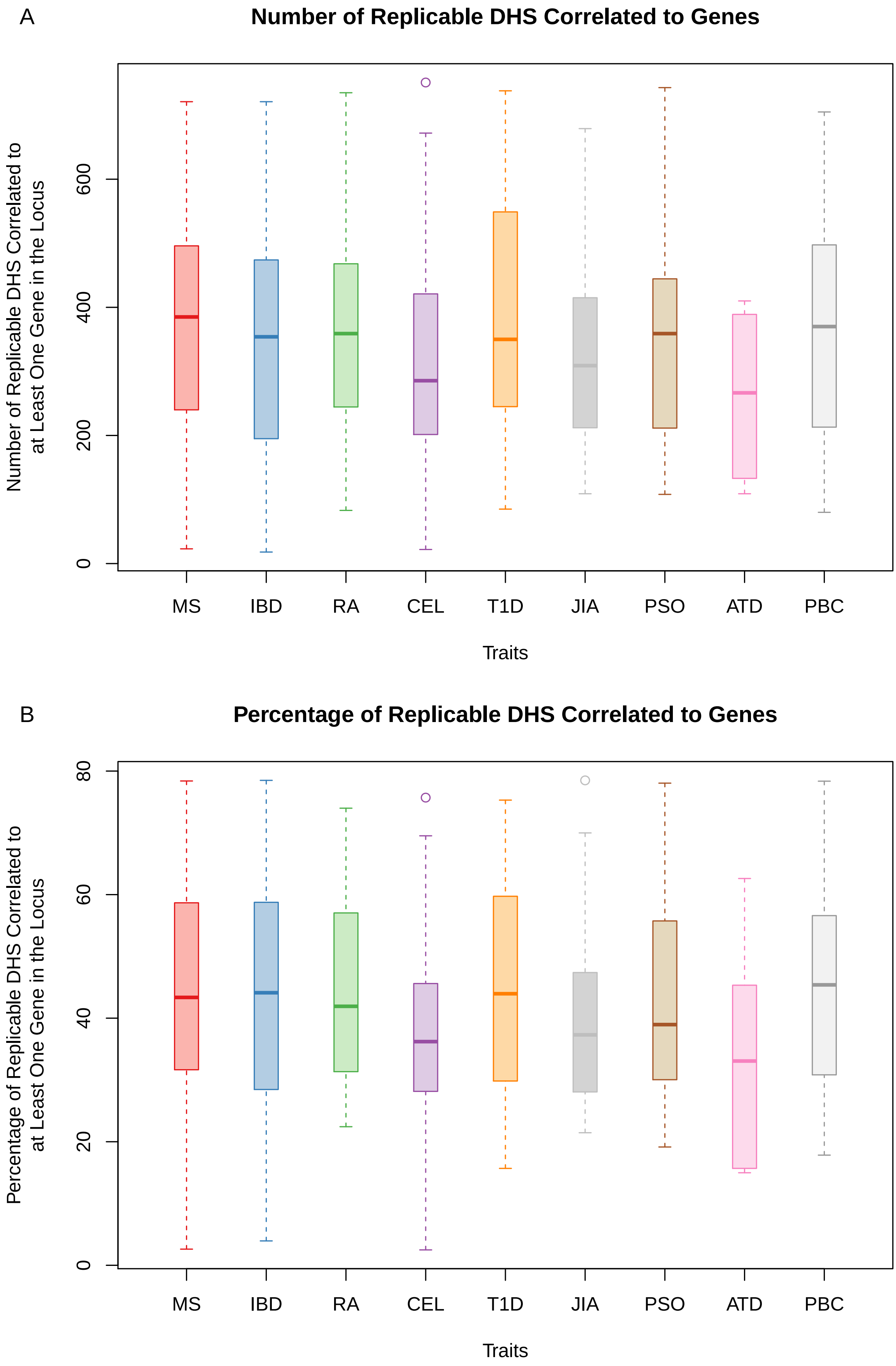


Figure S12: Small numbers of replicable DHS explain $>$ 90% of regulatory potential across 301 autoimmune and inflammatory disease risk loci. We find that a median of 3 replicable DHS explain 90% of regulatory potential (ρ, the sum of posterior probabilities of association) across the nine diseases, whereas the risk loci contain a median of 822 replicable DHS (panel A). This is not affected by the overall value of ρ (panel B). We can thus substantially reduce the number of prioritized replicable DHS per locus.



A Number of Replicable DHS Correlated to Genes

B Percentage of Replicable DHS Correlated to Genes

Figure S13: Number and percentage of replicable DHS correlated to at least one gene across 301 autoimmune disease risk loci. Across nine autoimmune and inflammatory diseases (each shown in a different color) we find that approximately half the replicable DHS in risk loci are correlated to the expression of at least one gene in the 2Mbp locus (DHS–gene correlation $P < 0.05$).

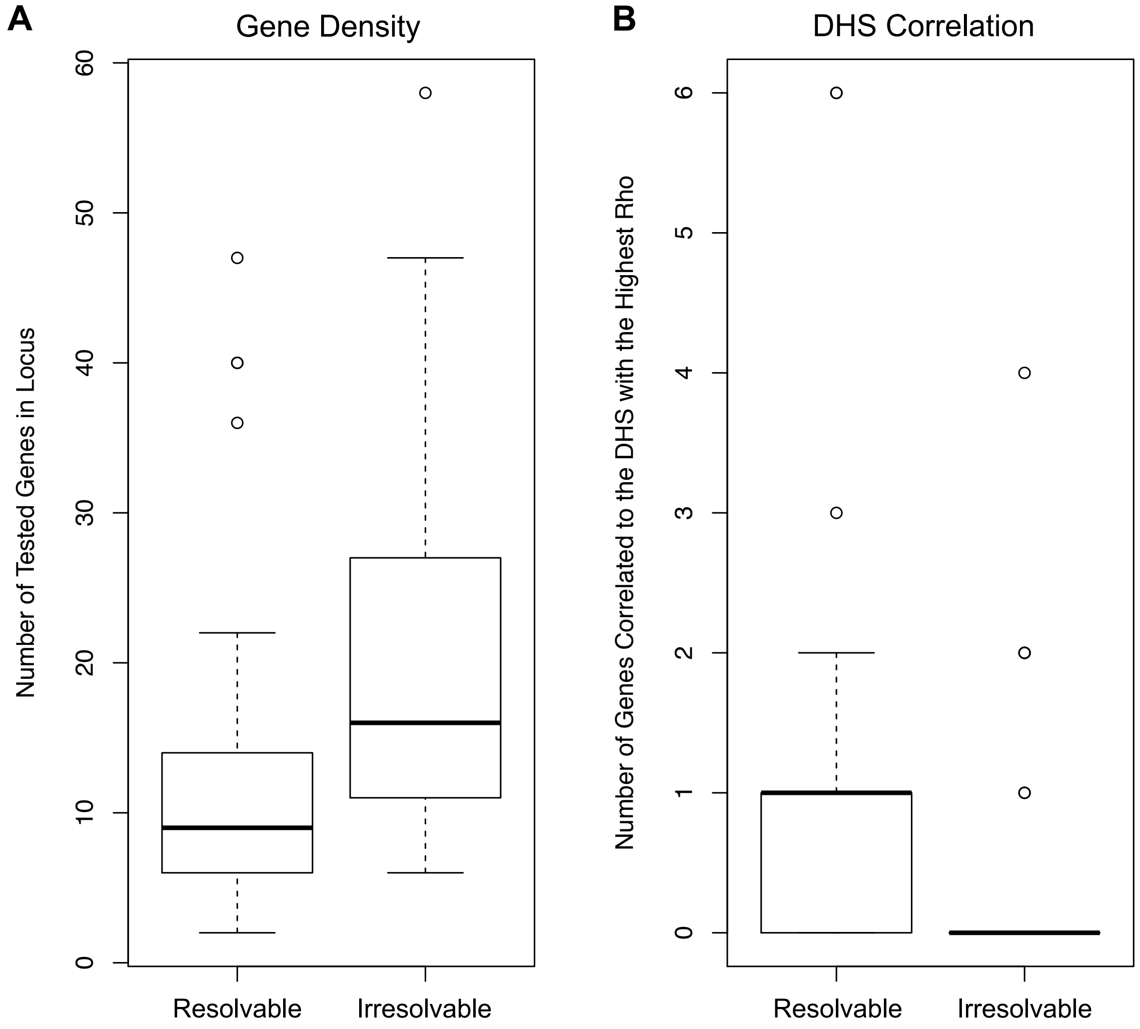


Figure S14: DHS harboring associated variants in irresolvable loci are not correlated to gene expression. We compared the 53/78 loci with significant ρ for which we could identify a single candidate gene to the 25 loci we could not resolve. (A) The 25 irresolvable loci harbor, on average, more genes than the 53 others. (B) We also find that the replicable DHS with the highest ρ in the 25 irresolvable loci are not correlated to the expression levels of genes in the region across the REMC, even before multiple testing correction for the number of genes encoded in each locus. This is in contrast to the 53 resolvable loci, where the replicable DHS with the highest ρ show strong correlation to at least one gene in the region. This lack of correlation for the 25 irresolvable loci suggests that these DHS do not influence any of the tested genes in that locus, rather than that we lack statistical power to resolve these 25 loci.

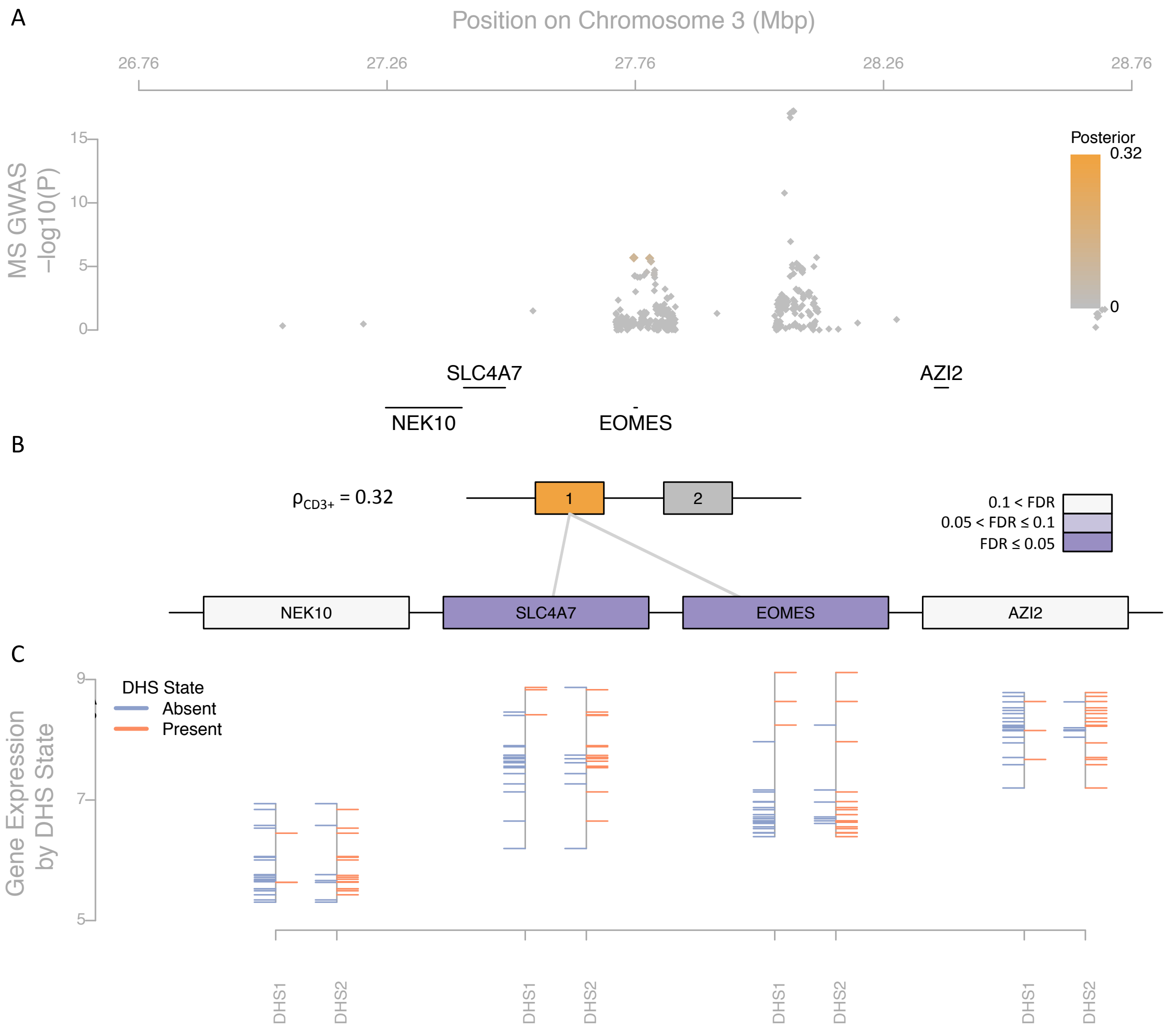


Figure S15: Regulatory fine–mapping identifies a replicable DHS and changes to *EOMES* and *SLC4A7* regulation in CD3$^+$ T cells as mediating multiple sclerosis risk on chromosome 3. A genome–wide significant association to MS risk on chromosome 3 localizes to the *EOMES* locus (A). Different combinations of replicable DHS are active in each of the Roadmap Epigenomics Project tissues we examined; the enrichment is most significant in replicable DHS active in CD3$^+$ T cells (32% of the overall posterior probability; FDR $<$ 0.1, panel B). By partitioning the posterior probability of association attributable to each replicable DHS by the strength of this correlation, we find that *EOMES* and *SLC4A7* show significant enrichment (purple, FDR $<$ 0.05). The expression levels of the two genes are markedly higher in tissues where the replicable DHS we identify is accessible (orange) than in tissues where it is inaccessible (blue; panel C).

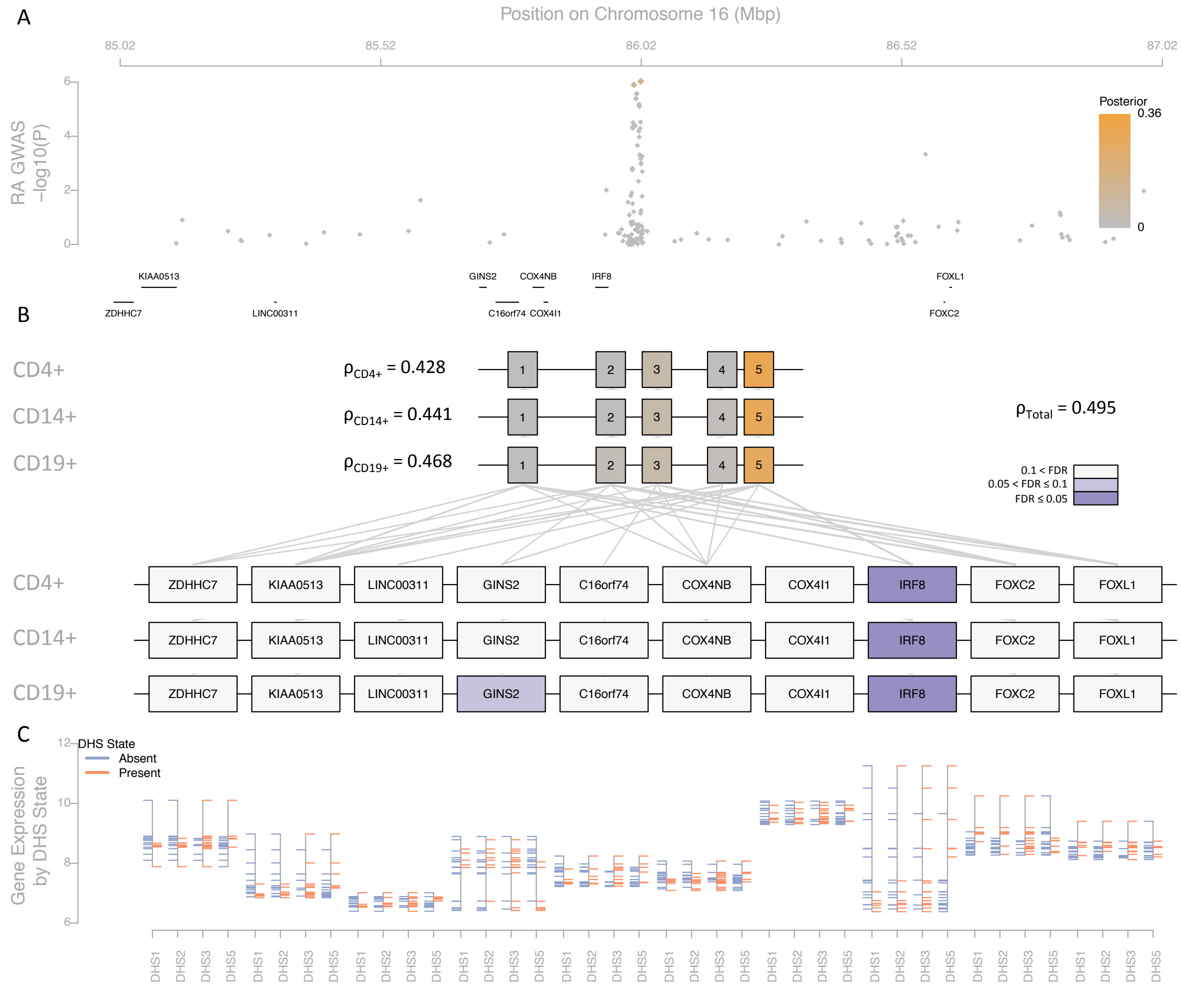


Figure S16: Regulatory fine–mapping identifies a replicable DHS and changes to *IRF8* regulation across immune subpopulations as mediating rheumatoid arthritis risk on chromosome 16. A genome–wide significant association to RA risk on chromosome 16 localizes near *IRF8* (panel A). Different combinations of replicable DHS are active in each of the Roadmap Epigenomics Project tissues we examined. The enrichment is most significant in replicable DHS active in CD4$^+$ T cells, CD14$^+$ monocytes, and CD19$^+$ B cells, each explaining the majority of the 49.5% of the overall posterior probability attributable to all DHS in the locus (FDR $<$ 0.1, panel B). By partitioning the posterior probability of association attributable to each replicable DHS by the strength of this correlation, we find that *IRF8* shows significant enrichment in all three cell types (purple, FDR $<$ 0.05), and its expression levels are markedly higher in tissues where the replicable DHS we identify (DHS5) is accessible (orange) than in tissues where it is inaccessible (blue; panel C).

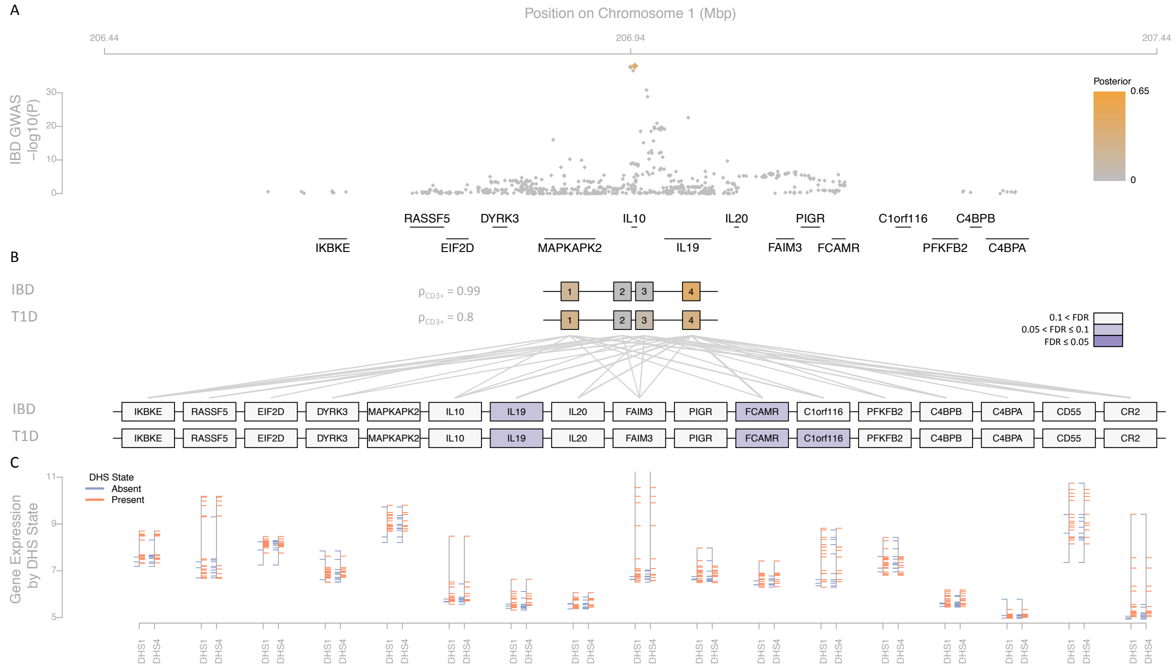


Figure S17: Regulatory fine–mapping on chromosome 1 identifies two replicable DHS and changes to *IL19* and *FCAMR* regulation in CD3$^+$ T cells, as driving risk to IBD and T1D. Association to IBD and T1D localizes to the coding region of *IL10* (IBD shown in panel A). We found significant regulatory potential in CD3$^+$ T cell subsets for both diseases, which independently localize to the same replicable DHS in the two diseases and explain 99% and 80% of IBD and RA posterior probability of association, respectively (panel B). In each case, we can independently prioritize *IL19* and *FCAMR* regulation in CD3$^+$ T cells. The expression levels of both genes are markedly higher in tissues where the two replicable DHS we identify are accessible (orange) than in tissues where they are inaccessible (blue; panel C).

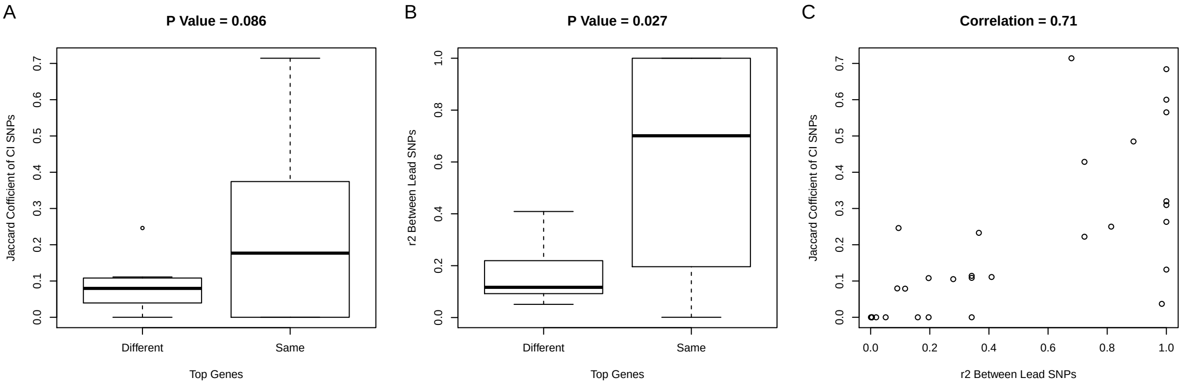


Figure S18: Comparing prioritized genes identified for pairs of traits at the shared loci. By comparing the prioritized genes at the shared loci, we found that pairs of traits with (A) higher similarity between their CI SNPs, and (B) higher LD between their lead SNPs tend to target the same genes more often. (C) LD between lead SNPs and the similarity between CI SNPs (measured by Jacquard coefficient) are correlated.

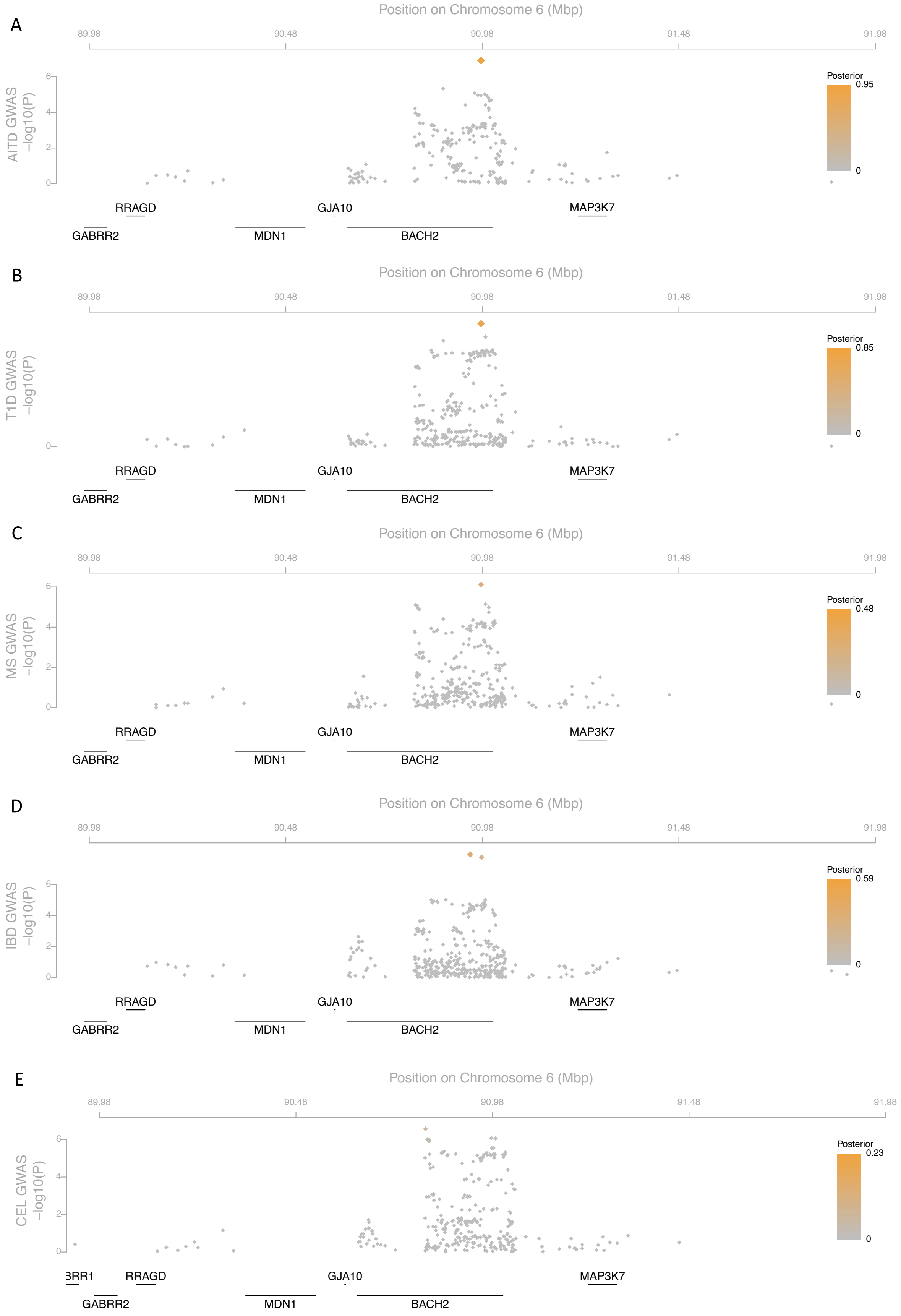

A
Position on Chromosome 6 (Mbp)
89.98
90.48
90.98
91.48
91.98
AITD GWAS
−log10(P)
0
2
4
6
Posterior
0.95
0
RRAGD
GJA10
MAP3K7
GABRR2
MDN1
BACH2
B
Position on Chromosome 6 (Mbp)
89.98
90.48
90.98
91.48
91.98
T1D GWAS
−log10(P)
0
Posterior
0.85
0
RRAGD
GJA10
MAP3K7
GABRR2
MDN1
BACH2
C
Position on Chromosome 6 (Mbp)
89.98
90.48
90.98
91.48
91.98
MS GWAS
−log10(P)
0
2
4
6
Posterior
0.48
0
RRAGD
GJA10
MAP3K7
GABRR2
MDN1
BACH2
D
Position on Chromosome 6 (Mbp)
89.98
90.48
90.98
91.48
91.98
IBD GWAS
−log10(P)
0
2
4
6
Posterior
0.59
0
RRAGD
GJA10
MAP3K7
GABRR2
MDN1
BACH2
E
Position on Chromosome 6 (Mbp)
89.98
90.48
90.98
91.48
91.98
CEL GWAS
−log10(P)
0
2
4
6
Posterior
0.23
0
3RR1
RRAGD
GJA10
MAP3K7
GABRR2
MDN1
BACH2

Figure S19: genetic association data in the BACH2 locus on chromosome 6 for five autoimmune and inflammatory diseases. We show genetic association data for AITD, T1D, MS, IBD and celiac disease, which we cannot conveniently display in figure 3 of the main paper. The first three traits show significant regulatory potential ρ in CD3$^+$ T cells in this locus, and each independently shows significant evidence for the role of *MDN1* in disease risk. The last two traits show no significant ρ or γ in any tissue.

Legends of Supplemental Tables

Table S1: Roadmap Epigenomics Project tissues used in our analyses. We used 56 tissues with at least two replicate samples, of which 22 had matched expression data (top panel). Samples from immune subpopulations are shown in bold font.

Table S2: Replicable DHS harboring credible interval SNPs across 301 loci associated to nine autoimmune and inflammatory disease loci. A total of 1954 replicable DHS (1470 unique) harbor at least one credible interval SNP. The last column gives the total posterior probability attributable to all SNPs on each replicable DHS (though the vast majority have exactly one SNP).

Table S3: Regulatory fine-mapping in 301 disease risk loci across nine autoimmune and inflammatory diseases. For each disease, we give the most associated SNP, the exact coordinates of each locus, the number of replicable DHS, credible interval (CI) SNPs, and genes in the locus, the total number of DHS harboring a CI SNP, and the total regulatory potential (ρ) attributable to all these replicable DHS. Then, for each locus, we show each Roadmap Epigenome Mapping Consortium tissue with significant excess of regulatory potential (FDR $<$ 0.1), indicating the total number of replicable DHS in the locus active in the tissue, the number of these harboring CI SNPs, the sum of ρ they account for, and the p-value of the enrichment test. Finally, within each of these tissues, we show each gene with significant pathogenicity factor γ (FDR < 0.1) with the accompanying p-value.

Table S4: Regulatory fine–mapping indicates risk variants for multiple diseases in the same loci affect the same genes. In 25 loci harboring associations to exactly two diseases each, we find that the most associated variants are often different (top row). However, the credible interval sets in these loci overlap significantly (hypergeometric $p < 0.001$, second row), and this overlap is greater than that of the most associated variants alone (increase in Jaccard coefficient). This overlap is also true when comparing the subset of CI SNPs on DHS and for the number of DHS harboring a CI SNP across diseases (third and fourth rows). When we compare prioritized genes, we see further increase in overlap relative to most associated variants *and* to prioritized DHS (bottom row). Thus, identifying risk–mediating genes partially overcomes the limited resolution of analyses only focusing on genetic association data.

Cell Types with Matched DHS and Gene Expression	**CD14 Primary Cells**
	CD19 Primary Cells
	CD3 Primary Cells
	CD4 Primary Cells
	CD8 Primary Cells
	Fetal Thymus
	Mobilized CD34 Primary Cells
	Fetal Adrenal Gland
	Fetal Brain
	Fetal Heart
	Fetal Large Intestine
	Fetal Small Intestine
	Fetal Kidney
	Fetal Left Kidney
	Fetal Right Kidney
	Fetal Lung
	Fetal Left Lung
	Fetal Right Lung
	Fetal Arm Muscle
	Fetal Leg Muscle
	Fetal Renal Cortex
	Fetal Renal Pelvis
Cell Types with DHS Data Only	**Mobilized CD3 Primary Cells**
	Mobilized CD4 Primary Cells
	CD56 Primary Cells
	Breast vHMEC
	Fetal Back Muscle
	Fetal Trunk Muscle
	Fetal Placenta
	Fetal Left Renal Cortex
	Fetal Right Renal Cortex
	Fetal Left Renal Pelvis
	Fetal Right Renal Pelvis
	Fetal Spinal Cord
	Fetal Stomach
	Fetal Testes
	Fibroblasts Fetal Skin (Abdomen)
	Fibroblasts Fetal Skin (Back)
	Fibroblasts Fetal Skin (Left Biceps)
	Fibroblasts Fetal Skin (Right Biceps)
	Fibroblasts Fetal Skin (Left Quadriceps)
	Fibroblasts Fetal Skin (Right Quadriceps)
	Fibroblasts Fetal Skin (Scalp)
	Fibroblasts Fetal Skin (Upper Back)
	Gastric
	H1
	H1 - BMP4 Derived Mesendoderm Cultured Cells
	H1 - BMP4 Derived Trophoblast Cultured Cells
	H1 - Derived Mesenchymal Stem Cells
	H1 - Derived Neuronal Progenitor Cultured Cells
	H9
	IMR90
	Pancreas
	Penis Foreskin Fibroblast Primary Cells
	Penis Foreskin Keratinocyte Primary Cells
	Penis Foreskin Melanocyte Primary Cells

Table S1: Roadmap Epigenomics Project tissues used in our analyses. We used 56 tissues with at least two replicate samples, of which 22 had matched expression data (top panel). Samples from immune subpopulations are shown in bold font.

	Concordance	**Discordance**	**Jaccard Coefficient**
Number of Most Associated SNPs	3	22	0.12
Number of CI SNPs (mean)	14.4	59.44	0.21
Number of Prioritized CI SNPs (mean)	4	13.36	0.26
Number of Prioritized Replicable DHS (mean)	3.52	11.8	0.28
Number of Prioritized Genes (mean)	1.88	2.75	0.39

Table S4: Regulatory fine–mapping indicates risk variants for multiple diseases in the same loci affect the same genes. In 25 loci harboring associations to exactly two diseases each, we find that the most associated variants are often different (top row). However, the credible interval sets in these loci overlap significantly (hypergeometric $p < 0.001$, second row), and this overlap is greater than that of the most associated variants alone (increase in Jaccard coefficient). This overlap is also true when comparing the subset of CI SNPs on DHS and for the number of DHS harboring a CI SNP across diseases (third and fourth rows). When we compare prioritized genes, we see further increase in overlap relative to most associated variants *and* to prioritized DHS (bottom row). Thus, identifying risk–mediating genes partially overcomes the limited resolution of analyses only focusing on genetic association data.