Supplementary Material for: Are all genetic variants in DNase I sensitivity regions functional?

Gregory A. Moyerbrailean[1], Chris T. Harvey[1], Cynthia A. Kalita[1],
Xiaoquan Wen[2], Francesca Luca[1,*], Roger Pique-Regi[1*],

[1]Center for Molecular Medicine and Genetics, Wayne State University
[2]Department of Biostatistics, University of Michigan

[*]To whom correspondence should be addressed: rpique@wayne.edu,
fluca@wayne.edu.

1 Data sources

A summary of the data used in this paper can be found in Tables S1 and S2. Chromatin accessibility data used for the analysis presented in this study was obtained from two cohorts, the ENCODE Project and the Roadmap Epigenomics Project. The ENCODE Project data was downloaded from the main ENCODE data distribution center (EncodeDCC) at the University of California Santa Cruz (UCSC), publicly available at `ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/` (downloaded 07/2013). The Roadmap Epigenomics Project data was downloaded in the form of sequence read archives (SRAs) from the NCBI GEO repository, `http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/` (downloaded 07/2013).

Positional Weight Matrices (PWMs) for 1,949 transcription factors were obtained from the online databases TRANSFAC (Matys et al. 2006) (`http://www.gene-regulation.com/pub/databases.html`, downloaded11/01/11) and JASPAR (Sandelin et al. 2004) (`http://jaspar.genereg.net/`, downloaded 09/23/11).

Known genetic variants from the 1000 Genomes Project Phase 1 data were downloaded from `ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/`. Linkage disequilibrium (LD) data between variants was also obtained from the 1000 Genomes Project. The LD data used in this analysis comes from European individuals. Coding variants and their allele frequencies were obtained from dbSNP version 137, downloaded through the UCSC table browser (`http://genome.ucsc.edu/cgi-bin/hgTables`, downloaded 12/05/13).

Ensembl transcript positions used to annotate transcription start sites were obtained via the UCSC table browser (`http://genome.ucsc.edu/cgi-bin/hgTables`, downloaded 10/21/12).

Genetic variants identified via genome-wide association studies (GWAS) were extracted from the GWAS catalog (`https://www.genome.gov/26525384`, downloaded 7/16/13).

GWAS meta-analysis data and imputed statistics used to run fgwas via (Pickrell 2014) were obtained through personal communication. Annotations used in the model were downloaded from `https://github.com/joepickrell/1000-genomes` (downloaded 03/2014).

2 Data Preprocessing

2.1 Preprocessing for CENTIPEDE analysis

Pre-aligned DNase-seq reads from the Roadmap Epigenomics Project were not directly available, so raw reads were obtained in the form of Sequence Read Archives (SRA) files. We then converted the SRA files to FastQ format using the fastq-dump program from the NCBI SRA toolkit and aligned using a custom mapper previously described (Degner et al. 2012). To identify technical replicates, we extracted sample annotations from the SRA metadata database (downloaded 6/20/13) using the SRAdb R package from bioconductor(Zhu et al. 2013). Aligned reads from samples identified as technical replicates were then merged using samtools (Li et al. 2009).

Aligned DNase-seq data for ENCODE samples was obtained directly from the EncodeDCC as described in Section 1. The aligner used by ENCODE should not have an impact on our ability to run the CENTIPEDE algorithm, and thus we did not need to remap the reads as we did for the Roadmap Epigenomics samples or for the ASB analysis as in Section 2.2.

2.2 Preprocessing for allele-specific analysis

For allele-specific binding analysis of the Roadmap Epigenomics samples, we used the aligned samples described in Section 2.1. For ENCODE samples, we obtained raw sequence reads (fastq format) directly from the EncodeDCC and aligned them with the same custom mapper used to process the Roadmap Epigenomics samples. Reads sequenced on old Solexa machines were removed, as these reads were often of lower quality and more prone to base calling errors. Samples with fewer than 25 million reads were removed from analysis, as these typically displayed too low a coverage to be informative. To further minimize mapping errors and reference biases, we applied additional mappability filters (see Section 2.4) to all samples used for allele-specific analysis.

Aligned reads were then piled up on a filtered set of 1KG SNPs (see Section 2.3) using samtools mpileup and the hg19 reference genome, and reads were discarded if the SNP was either at the first or last base of the read to avoid the possibility of an experimental bias at these positions caused by the DNaseI cleavage preference (Degner et al. 2012). Finally, the following filters were applied to the read data: 1) the SNP must be covered by $>$4 reads, 2) $>$50% of the reads covering the SNP cannot start at the same position (i.e., PCR artifacts), and 3) the SNP cannot have a read coverage greater than 99.99% of the positions genome-wide, as such exaggerated coverage indicates potential mappability issues.

2.3 Selection of genetic variants

To create a core set of SNPs for ASB analysis, we started with all bi-allelic 1KG SNPs and first removed rare (MAF $<$ 5%) SNPs. To avoid the possibility of multiple SNPs in the same motif, we next removed SNPs within 25 bases up- or downstream of another SNP. Next we removed SNPs in regions prone to mapping biases, masking approximately 1% of the genome (Degner et al. 2012).

2.4 Mappability filtering

We created an array of hash tables containing all possible 20-mer reads (a 4-mer prefix indexes an array of 256 hash tables, the 16-mer suffix is used a hash key), the values of the hash tables record the locations a read can align to (up to a maximum of 128 locations). These arrays are then used for aligning the reads in the custom mapper (Degner et al. 2012). We can also use the hash tables to identify which locations of the genome can generate reads that can align to multiple locations. Two of these arrays have been created and are denoted as M0 and M1.

* M0 Hash Table - The 20-mers starting at each bp position of the genome are added into this table, as well as all 20-mers with alternate alleles (i.e., overlapping SNPs and InDels).

* M1 Hash Table - The same 20-mers as in M0 are generated, but for each genomic 20-mer (reference or variant) we consider all the possible single base pair error that could have occurred. Each 20-mer has 3x20=60 other 20-mers at Hamming distance of 1.

The M0 hash table is used to align reads with our mapper, so only reads that map without mismatches are used. Both hash tables are used to create the two following mappability tracks:

* M0 mappability track - For each base of the genome we record the number of locations that match the same exact 20-mer (considering all 1KG genetic variants). For allele specific measurements we only consider reads that originate at locations with mappability M0 value exactly equal to one (i.e., reads with unique mappability when there are no base-calling errors).

* M1 mappability track - For each base of the genome we record the number of locations that match the same exact 20-mer (considering all 1KG genetic variants) or any one base pair mismatch. For allele specific measurements we only consider reads that originate at locations with mappability M1 value ≤ 70. Using this value, a location can have up to 69 other loci with only one nucleotide different, reads from which could potentially map to the location of interest if a sequencing error occurs. However, considering a base-calling error rate of 0.01 and an average background coverage of $\sim 1X$, we expect $<< 0.5$ reads from other loci to incorrectly map at locations of interest.

The motivation behind the M1 mappability filter is that very repetitive regions of the genome, can generate reads that are very similar to other regions. Even when the base calling error is small, 20-mers with high similarity may generate reference calling biases when doing ASE analyses. We do not consider a more complex filter as the probability of a read with two base-calling errors at this read-length is very small.

3 Identification and Mapping of Active Transcription Factors

3.1 Identification of active transcription factors

For each of the 1,949 transcription factors for which we have positional weight matrices (PWMs), we scanned the genome for candidate motifs and calculated the PWM score according to the following formula (Stormo 2000):

$$\begin{aligned}\text{PWM score } (S_l) &= \sum\nolimits_{w=1}^{W} \log_2\left(\frac{\Pr\left(\text{seeing } S_{l+w}\text{on position } w|\text{PWM}\right)}{\Pr\left(\text{seeing } S_{l+w}|\text{ background}\right)}\right) = \\ &= \sum\nolimits_{w=1}^{W} \log_2\left(p\left[S_{l+w},w\right]\right) - \sum\nolimits_{w=1}^{W} \log_2\left(0.25\right) = \\ &= \sum\nolimits_{w=1}^{W} \log_2\left(p\left[S_{l+w},w\right]\right) - \log_2\left(0.25\right)W\end{aligned} \tag{1}$$

where S_l indicates the observed nucleotide at position l of sequence S, the PWM model is given by the probability $p[S_{l+w}, w]$ of observing the nucleotide S_{l+w}(A,C,G,T) at position w, and *W* is the motif length.

We then selected the top 5,000 scoring sequences and used this subset to train the CENTIPEDE model. We added orthologous locations for which the sequence was among the top 5,000 scoring motifs instances in chimp and in macaque and may not score as high in humans. This combined sequence set contained between 5,000 and 15,000 sequences per motif and were originally obtained from Pai *et al.* (in prep.). Using these locations and the DNase-seq data listed in Table S1, we applied the CENTIPEDE model for each sample/motif combination. For each pair we determined a Z-score using the following logistic model on the top 5,000 set:

$$\log\left(\frac{\pi_l}{(1-\pi_l)}\right) = \beta_0 + \beta_1 \times \text{PWM Score}_l \tag{2}$$

To determine an initial threshold for factor activity, we examined the footprint profiles at binding sites, stratified by Z-score (Figure S1). For most factors, a modest footprint was evident for sites with a Z-score of at least 5. Using this threshold, we detect 1,891 factors active in at least one sample. For these factors, we next generated a revised sequence model using motif sequences from the sample with the largest Z-score. Using sequences for which the posterior probability of being bound is >0.99, we calculated revised position weight matrices from the base frequencies of each position in the footprint (Figure S2).

3.2 Generation of CENTIPEDE binding predictions

Using a custom set of sequence models (see Section 3.1), we scanned the reference genome to identify all motif matches genome-wide. For each motif, we determined which positions to include for further analysis using a PWM match score calculated to include 95% of the sequences used to generate the model. Scanning was done in two stages. First, we identified every match above the threshold using Equation 1 as before. Next, we scanned the genome, this time only considering motifs that overlapped 1000 Genomes variants. For each of these matches we calculated two PWM scores, one for each allele.

To generate the binding predictions, first trained the CENTIPEDE model on motif locations that do not overlap a SNP for each sample/motif pair using the updated sequence models. As a check for how well calibrated the sequence models are for the data, we examined the correlation between the PWM scores and the observed DHS peaks. Using a Spearman correlation test and a nominal threshold of $p < 10^{-7}$, we discarded 519 sequence models, leaving us with data for 1,372 sequence models. We then applied these models to each sample/motif combination again, using motif instances overlapping known variants. Thus, for motif matches containing a variant, two binding predictions were made, one for each allele, using equation (2).

3.3 Validation of CENTIPEDE predictions

To compare the new sequence models to the originals, we performed precision recall operating characteristic (P-ROC) curve analysis using CENTIPEDE predictions and ChIP-seq peaks from GM12878 samples. We annotated a list of all binding sites identified by either CENTIPEDE or ChIP-seq, using the PWM score as the predictions and the presence or absence of a ChIP-seq signal as the labels. For sites with a ChIP-seq signal but no CENTIPEDE prediction, a PWM score of 0 was used. For each selected factor, we compared the precision-recall data for CENTIPEDE predictions from a genome-wide scan using the original PWM model and the updated PWM model (Figure S3).

4 Analysis of Allele-Specific Binding

4.1 Validation of genotype predictions

To verify the genotyping accuracy, we compared the genotype calls from applying QuASAR to DNase-seq data on LCLs for an individual fully resequenced by the 1KG Project (1KG individual NA12878). Of the 1,400 predicted heterozygous loci, all of them were confirmed to be true heterozygotes. Of the 11,278 predicted homozygous loci, only seven of them were actually heterozygotes, and all of the remaining true calls were homozygous for the predicted allele. The seven miscalled heterozygotes are likely cases of extreme allelic imbalance, and is to be expected, as QuASAR was designed to be conservative to avoid miscalling homozygous genotypes as heterozygotes with extreme allelic imbalance. The results of our comparisons are summarized in Table S3.

4.2 Postprocesing of allele-specific data

As a means to further filter out samples not well-suited for our method, including cancer tissues and samples from pooled individuals (e.g., Figure S4), we examined two parameters estimated by QuASAR, ρ and dispersion (D), as well as the allele frequencies of the variants. ρ is an estimate of the mean genotype frequency across heterozygous positions, and should center near 0.50. Deviation from this frequency in a sample can be an indication of genetic aberrations, such as in a cancer sample, where copy number variation can be extensive. We also examined the correlation (Pearson correlation coefficient) between the ρ estimate and the allele frequency for each heterozygous locus, as the two should be independent of each other. D represents the degree to which the read data follows a normal genomic distribution. Increasing D indicates a non-normal distribution of reads, for example, among samples with a high degree of copy number variation. After applying all filters, 317 samples remained for analysis. A summary of the post-processing results can be found in Table S4 and Figure S5.

5 Annotation of ASB with binding predictions

5.1 Combining predictions and ASB data

To determine which positions displaying ASB fall within a predicted footprint, we overlapped the allele ratios over heterozygous SNPs in DHS sites (DHS-SNPs) with CENTIPEDE footprint predictions in each sample. We then created a final set of annotated ASB-SNPs by aggregating the data across each sample and factor. For cases where a SNP is within multiple predicted binding sites, we selected the factor whose sequence model predicts the greatest log ratio between the prior log odds of binding for each allele. For SNPs predicted to have an effect on binding (effect-SNPs), we determined which ones were predicted to have an effect in the same direction as the observed allele ratio (e.g., the allele with a higher PWM score is observed more often in the DNase data). This generated a set of 204,757 SNPs across all samples. As the same SNP could affect multiple cell-types, this set of 204,757 SNPs reflects 961,297 observations of ASB. We then partitioned the data into three non-overlapping categories: 1) SNPs in predicted footprints whose binding effect is in the direction predicted, 2) all other SNPs in footprints, 3) all other DHS-SNPs. Because each annotation has a different prior expectation of being functional, we readjusted for multiple testing within each annotation separately by applying the Storey q-value method on the p-values obtained from the QuASAR test to estimate the false discovery rate (FDR).

5.2 Individual motif analysis of binding predictions

In order to evaluate the extent to which the newly defined sequence models accurately predict ASB, we compared CENTIPEDE predictions and ASB analysis for each motif individually. We examined motifs containing at least 10 heterozygous SNPs in footprints for which we can estimate the ASB allelic ratio $\hat{\rho}$. $\hat{\rho}$ is calculated from the sequencing data as,

$$\hat{\rho} = \frac{\text{\# reads w/ reference allele}}{\text{\# reads w/ reference or alternate allele}} \tag{3}$$

For each motif, we fit a logistic model,

$$\text{logit}(\hat{\rho}) \sim \beta_0 + \beta_1 * \Delta\text{logit}(p) \tag{4}$$

where $\Delta\text{logit}(p)$ is the change in log prior odds predicted by the sequence model in CENTIPEDE. We fit the model on SNPs displaying some allelic imbalance ($p < 0.1$) to focus on our predictions over true positives. Figure S6 shows the correlation between our prediction and the observed ASB for the most predictive sequence model, belonging to the factor AP-1.

6 Genomic Annotation

6.1 Allele frequency

Allele frequency for each 1KG SNP was obtained as described in Section 1. For each SNP, we calculated the minor allele frequency by taking the absolute value of the difference between the allele frequency and 0.5. To calculate the minor allele frequency for coding SNPs, we obtained coding SNP annotations from dbSNP (version 137). We classified coding SNPs into two categories, synonymous and non-synonomyous, the latter category encompassing missense, nonsense, and early stop variants. For the analysis, we only considered bi-alleleic SNPs, and those unambiguously categorized as either synonymous or non-synonymous. Of the 784,003 SNPs we analyzed, 298,986 (38%) are synonymous.

6.2 Distance to transcription start sites

For a given locus, distance to the nearest TSS was calculated as absolute value distances to the nearest annotated TSS. Using Ensembl gene annotations (see Section 1), we determined the distance for each SNP in our set. For TF binding sites, we determined the median distance for all binding sites genome-wide.

6.3 Identification of TF binding sites enriched for ASB

To identify which TF binding sites are enriched or depleted for ASB-SNPs, we calculated, for each factor, the proportion of binding sites containing ASB-SNPs to all binding sites containing a heterozygous SNP. As the proportions can be skewed at lower total numbers, we included only factors with at least 100 heterozygous SNPs across all binding sites genome-wide. For the 368 factors that met this criteria, we estimated the enrichment or depletion of ASB by calculating the fold-change between the ASB enrichment ratio and the average ASB enrichment ratio across all binding sites (with >100 heterozygous SNPs), using a binomial test to assess significant difference between the ratios. Factors whose binding sites are enriched or depleted for ASB-SNPs (at a nominal p-value cutoff of $p < 0.01$) are displayed in Table S7.

6.4 Selection on transcription factor binding sites

Using the footprint annotations from CENTIPEDE, we identified all footprints that do not contain known human polymorphisms. We used the UCSC liftOver tool to obtain orthologous regions in the Chimpanzee genome (panTro3 assembly), using a minimum remap threshold of 10% using liftOver. At these loci in the chimp genome, we calculated PWM scores as in Section 3.1. Next, using the model obtained from CENTIPEDE on the human sites, we calculated the sequence-based probabilities of binding for the chimpanzee sites. Sites where the prior probability of binding differ from the humans sites were classified as divergent, and were further categorized by the difference in binding affinity: functional (analogous to non-synonymous) for those that differ by ≥20-fold, and silent for those that do not. For the polymorphic sites, we used binding sites with effect-SNPs as functional, and those with footprint-SNPs that are not effect-SNPs as silent. For each factor motif, we calculated the number of binding sites belonging to each category to build a contingency table similar to the McDonald-Kreitman test:

	Divergent	Polymorphic
Functional	D_f	P_f
Silent	D_s	P_s

Finally, we calculated a selection score using the following formula:

$$\text{Selection score} = \frac{D_f/D_s}{P_f/P_s} \tag{5}$$

To test for enrichment, we used a fisher exact test on the contingency table, and used the Storey q-value method to adjust for multiple testing. A full list of motif scores and the data used to calculate them can be found in Table S10.

7 Overlap with Genome-Wide Association Studies

7.1 Analysis of SNPs in the GWAS catalog

To account for the possibility that not all SNPs submitted to the GWAS catalog are indeed causal, we created a custom GWAS catalog by adding SNPs in linkage disequilibrium (LD) with each GWAS hit. Using 1000 Genomes LD data for European populations, we created a custom catalog mapping each GWAS hit to all SNPs on the same LD block at an r^2 cutoff of 0.8. The final file is a tabix-indexed bed file where each SNP entry has fields for its corresponding GWAS SNP, the r^2 between them, and associated GWAS traits. To identify overlap between our annotations and those associated with a GWAS trait, we intersected our results with this custom catalog. To test for enrichment, we used a fisher exact test to determine if the proportion of SNPs with a given annotation were enriched in the catalog.

7.2 Adding annotations to SNPs associated with complex traits

We integrated our CENTIPEDE footprint annotations into the top fitting models described in Pickrell (2014) for two GWAS meta-studies, lipid levels and height using the fgwas command line program. The lipids study is subdivided into four traits, high-density lipoprotein (HDL) levels, low-density lipoprotein (LDL) levels, triglyceride (TG) levels, and total cholesterol (TC) levels. For each of the five traits, we incorporated CENTIPEDE footprint annotations for associated SNPs and reapplied the top-fitting model. This was done individually for each factor to determine which factor binding sites were enriched for SNPs associated with each trait. A summary of the factors enriched for associations with each trait can be found in Table S6.

We also incorporated the following annotations as binary indicators, as per (Pickrell 2014): for HDL, gene rich and gene poor regions, repressed in HepG2, transcription start site in HepG2, coding exons, and repressed in K562 in HDL; for LDL and TC, gene rich and gene poor regions, 0 - 5,000bp from a transcription start site, 5,000 - 10,000bp from a transcription start site, repressed in HepG2, and nonsynonymous; for TG, gene rich and gene poor regions, repressed in HepG2, and 3UTR; for height, gene rich and gene poor regions, repressed in HeLa, transcribed in HUVEC, DNase of fetal lung, DNase of fetal muscle, 3UTR, and nonsynonymous. After applying the models, we assessed enrichment or depletion using the log_2(enrichment) values, excluding any motifs whose 95% confidence interval (CI) spanned zero. For each factor whose binding sites are either significantly enriched or depleted for trait-associated SNPs (Figure S8), we examined the SNPs whose posterior probability of association (PPA) with a trait had been increased by the addition of our annotation. Overall we found 23 SNPs for lipid levels and 15 SNPs for height whose associations were strengthened by our footprint annotations (Figure S9).

References

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G. E., *et al.*, 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, **482**(7385):390–4.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**:2078–2079.

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., *et al.*, 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, **34**:D108–D110.

Pickrell, J. K., 2014. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. *The American Journal of Human Genetics*, **94**(4):559–573.

Sandelin, A., Alkema, W., Engström, P., Wasserman, W. W., and Lenhard, B., 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic acids research*, **32**:D91–D94.

Stormo, G. D., 2000. DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)*, **16**(1):16–23.

Zhu, Y., Stephens, R., Meltzer, P., and Davis, S., 2013. SRAdb: query and use public next-generation sequencing data from within R. *BMC Bioinformatics*, **14**:19.

Table S1: **DNase samples and sources.** Listed for each sample is the cohort that produced it, the sample, and the number of reads.

See attached file, also available at `http://genome.grid.wayne.edu/centisnps/supp/`

Table S2: **Sources of additional data used in analyses.** Download dates and, where applicable, specific cell-types/tissues are also listed.

Type	Data Source	Cell-type/Tissue	Date Downloaded
PWM matrices	TRANSFAC	–	11/1/11
PWM matrices	JASPAR	–	9/23/11
GWAS Catalog	NIH	–	7/16/13
LD data	1KG	–	3/29/12
Gene annotations	Ensembl	–	10/21/12
ChIP-seq	ENCODE	GM12878	10/28/13
Genotypes	1KG	GM12878	10/30/13
Coding SNPs	dbSNP	–	12/05/13
SNP Annotations	`https://github.com/joepickrell/1000-genomes`	–	03/2014

Table S3: **Validation of genotype predictions.** A comparison of 1KG genotypes and those called by QuASAR for the 12,650 loci examined in the LCL cell line GM12878.

	1KG Hom	1KG Het
QuASAR Hom	11,271	7
QuASAR Het	0	1,372

Table S4: **Summary of post-processing filters.** The first three rows show the threshold and number of samples filtered for each parameter independently. After applying the three filters, the remaining samples for manually examined and known cancer samples were removed.

Parameter	Threshold	# Removed
ρ	<0.54	78
$\lvert\text{cor}(\rho, \phi)\rvert$	<0.15	31
$\frac{1}{D}$	>12	58
Cancer†	–	13

†without evident chromosomal abnormalities

Table S5: **Predictiveness of genomic characteristics on functional effects.** We considered the following characteristics in a regression analysis to determine their predictiveness as to whether a footprint-SNP is also an effect-SNP.

	Effect Size	p
Sequence Change	4.95×10^{-4}	$<10^{-16}$
TSS Distance	-7.22×10^{-8}	$<10^{-16}$
Number of Tissues	-2.08×10^{-3}	$<10^{-16}$
Minor Allele Frequency	-1.74×10^{-1}	$<10^{-16}$

Table S6: **Factor binding sites enriched for GWAS SNPs.** For each trait, factors whose binding sites are enriched for SNPs associated with the trait are listed. Shown also are the lower and upper limits of the 95% confidence interval.

See attached file, also available at *`http://genome.grid.wayne.edu/centisnps/supp/`*

Table S7: **Enrichment of ASB-SNPs within binding sites.** Factors with at least 100 heterozygotes in a predicted binding site are listed along with the counts, ratios, and enrichments of ASB-SNPs, footprint-SNPs, and switch-SNPs within them.

See attached file, also available at *`http://genome.grid.wayne.edu/centisnps/supp/`*

Table S8: **Comparison of multiple motifs for a single factor.** Motifs corresponding to the same transcription factor are similarly enriched or depleted for ASB-SNPs.

Factor	ID	Heterozygous SNPs	ASB SNPs	ASB/Het Ratio	Fold-enrichment	p-value
AP-1	M00517	461	10	0.021691974	1.036030117	0.869967821
	M00188	296	16	0.054054054	2.581675041	0.000608884
	M00199	111	22	0.198198198	9.466141817	1.81E-15
	M00924	181	13	0.071823204	3.430347206	0.000131074
	M00925	159	16	0.100628931	4.806137197	2.94E-07
	M00926	210	24	0.114285714	5.45839865	2.60E-11
	MA0099.2	1931	250	0.129466598	6.183452673	2.91E-114
CBF1	M01577	193	3	0.015544041	0.74239876	0.802620423
	M01699	403	8	0.019851117	0.948108967	1
	M01793	168	2	0.011904762	0.568583199	0.591560229
	M01911	135	5	0.037037037	1.768925491	0.211980416
CREB	M00113	451	14	0.031042129	1.48260276	0.136346848
	M00178	136	4	0.029411765	1.404734964	0.374540498
	M00916	950	32	0.033684211	1.608791208	0.011917593
	M00917	188	4	0.021276596	1.016191253	0.800445788
	M00801	675	59	0.087407407	4.174664144	1.08E-19
CTCF	M01259	3500	49	0.014	0.668653836	0.00309612
	MA0139.1	3426	43	0.01255108	0.599451985	0.000327385
	M01196	479	35	0.073068894	3.489842592	3.37E-10
E2F/E2F-1	M00024	1001	11	0.010989011	0.524846026	0.026362164
	M00425	949	16	0.016859852	0.805243194	0.49459686
	M00427	508	7	0.013779528	0.658123876	0.349487781
	M00918	1131	21	0.018567639	0.886808789	0.6773526
	M00920	1117	13	0.011638317	0.555857522	0.02754515
	M01114	577	5	0.008665511	0.41387337	0.039836283
	M00426	2017	7	0.003470501	0.165754558	3.17E-11
	M00516	1792	7	0.00390625	0.186566361	1.67E-09
	M00428	3085	49	0.015883306	0.758602392	0.050945353
	M00431	992	17	0.017137097	0.818484689	0.504382044
	M00940	2007	32	0.015944195	0.761510511	0.137608155
	M00939	1353	10	0.007390983	0.353000653	0.00012598
	M01251	224	2	0.008928571	0.426437375	0.342585393
	MA0024.1	158	2	0.012658228	0.60456948	0.77719828
Staf	M00262	1132	2	0.001766784	0.08438335	3.07E-08
	M00264	442	0	0	0	0.000167517
XBP1	M00251	257	5	0.019455253	0.929202111	1
	M01770	452	13	0.028761062	1.373656746	0.246263429
	M01970	674	6	0.008902077	0.425171996	0.029652731
	M01513	985	5	0.005076142	0.242441559	7.90E-05
	M01947	607	4	0.006589786	0.314734692	0.009734357

Table S9: **ASB effects for several immune-related factors.** For each factor listed, we calculated the aggregate ASB enrichment ratio across all sequence models corresponding to that factor.

Factor	Role in Innume Response	Average Number of Samples	Heterozygous SNPs	ASB SNPs	ASB/Het Ratio	Fold-enrichment	p-value
AP-1	Pro-inflammatory	39	3411	353	0.103	4.943	2.20E-16
c/EBP	Pro-inflammatory	5	125	7	0.056	2.675	0.01657
CREB	Anti-inflammatory	77	2947	127	0.043	2.058	1.53E-13
NF-kB	Pro-inflammatory	6	147	7	0.048	2.274	0.03591

Table S10: **Selection score for individual motifs.** For each factor motif, we used a modified MK test to calculated a selection score. Shown for each motif is the number of binding sites belonging to each category used in the MK test (divergent functional, divergent silent, polymorphic functional, and polymorphic silent) as well as the score.

See attached file, also available at `http://genome.grid.wayne.edu/centisnps/supp/`.

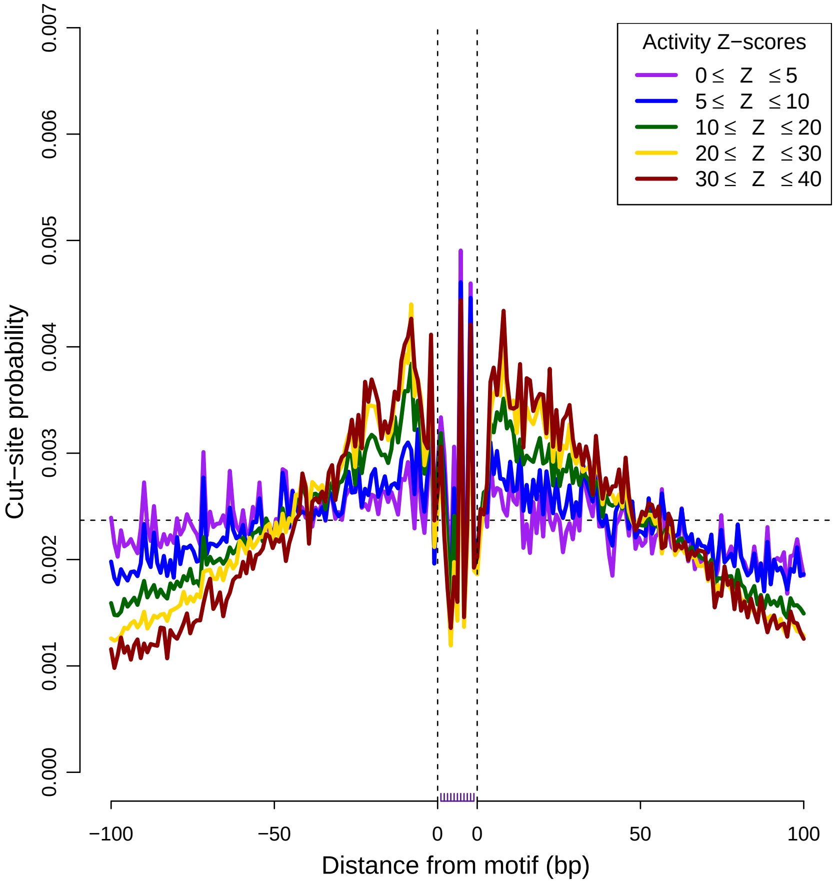

Figure S1: **Binding profiles of AP-1 motif M00172.** Each line is a composite footprint for AP-1 across all 653 tissues at the indicated Z-score intervals. The higher the Z-score, the more likely a factor is bound as predicted by the CENTIPEDE model.

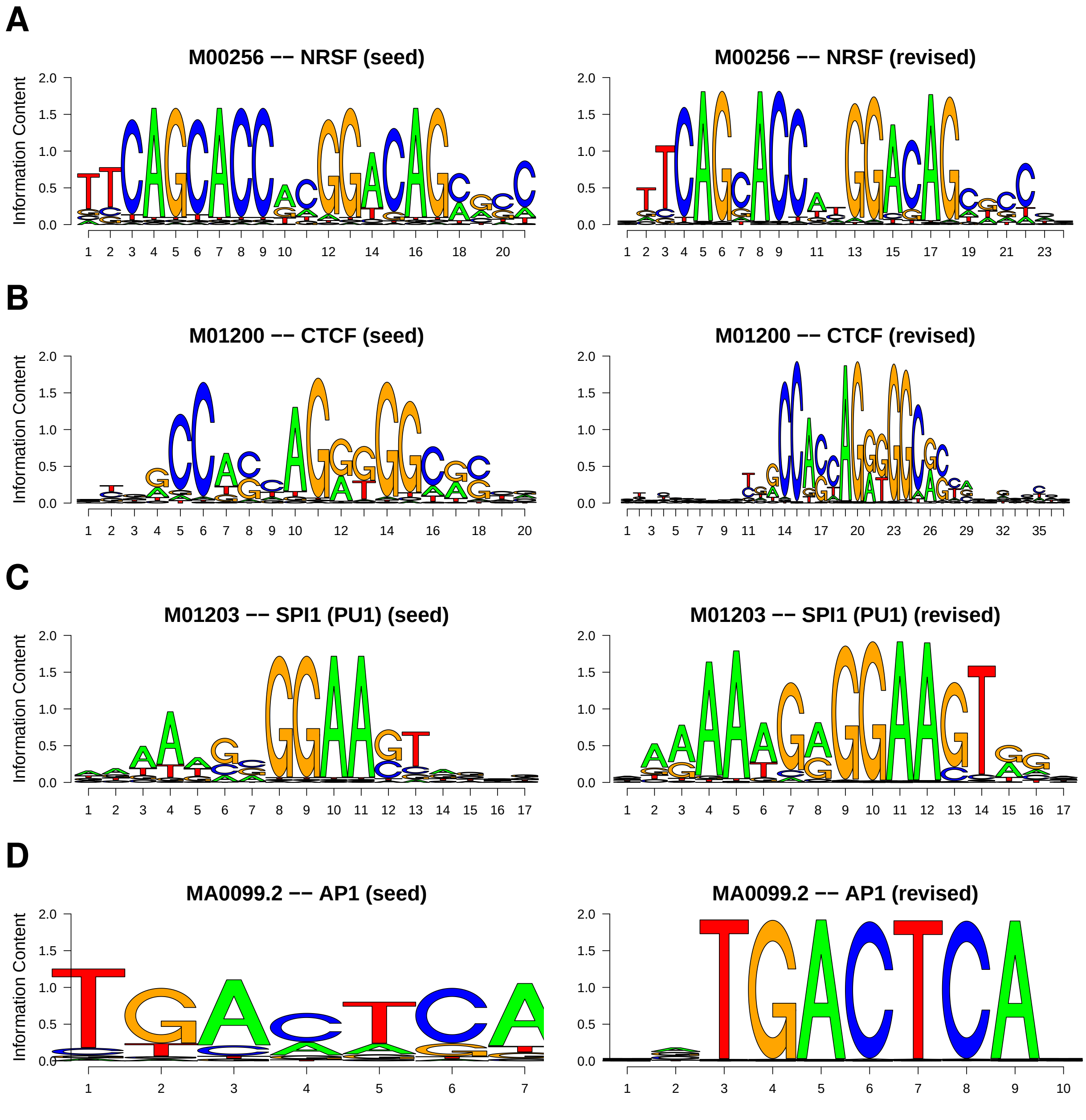

Figure S2: **Comparison between seed and revised sequence model.** For each factor motif, shown is the original seed sequence model (left) and the revised model (right). x-axis: position within motif, y-axis: information content. (A) NRSF (B) CTCF (C) PU.1 (D) AP-1

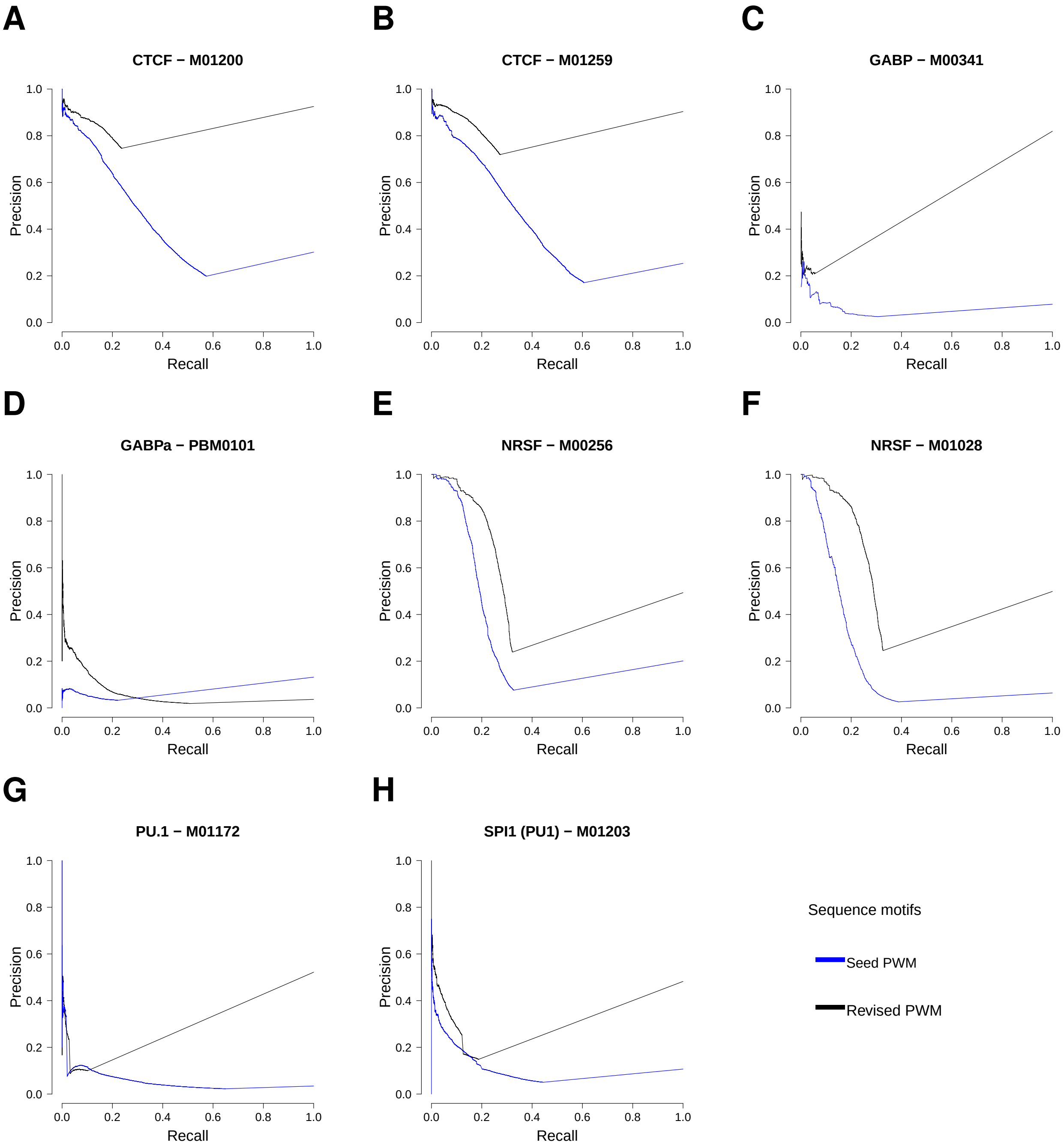

Figure S3: **Precision-recall curves for seed (blue) and revised (black) sequence models.** For each TF binding motif, CENTIPEDE-predicted footprints in GM12878 cells were compared using ENCODE ChIP-seq data as a gold standard. (A & B) CTCF (C & D) GABP (E & F) NRSF (G & H) PU.1

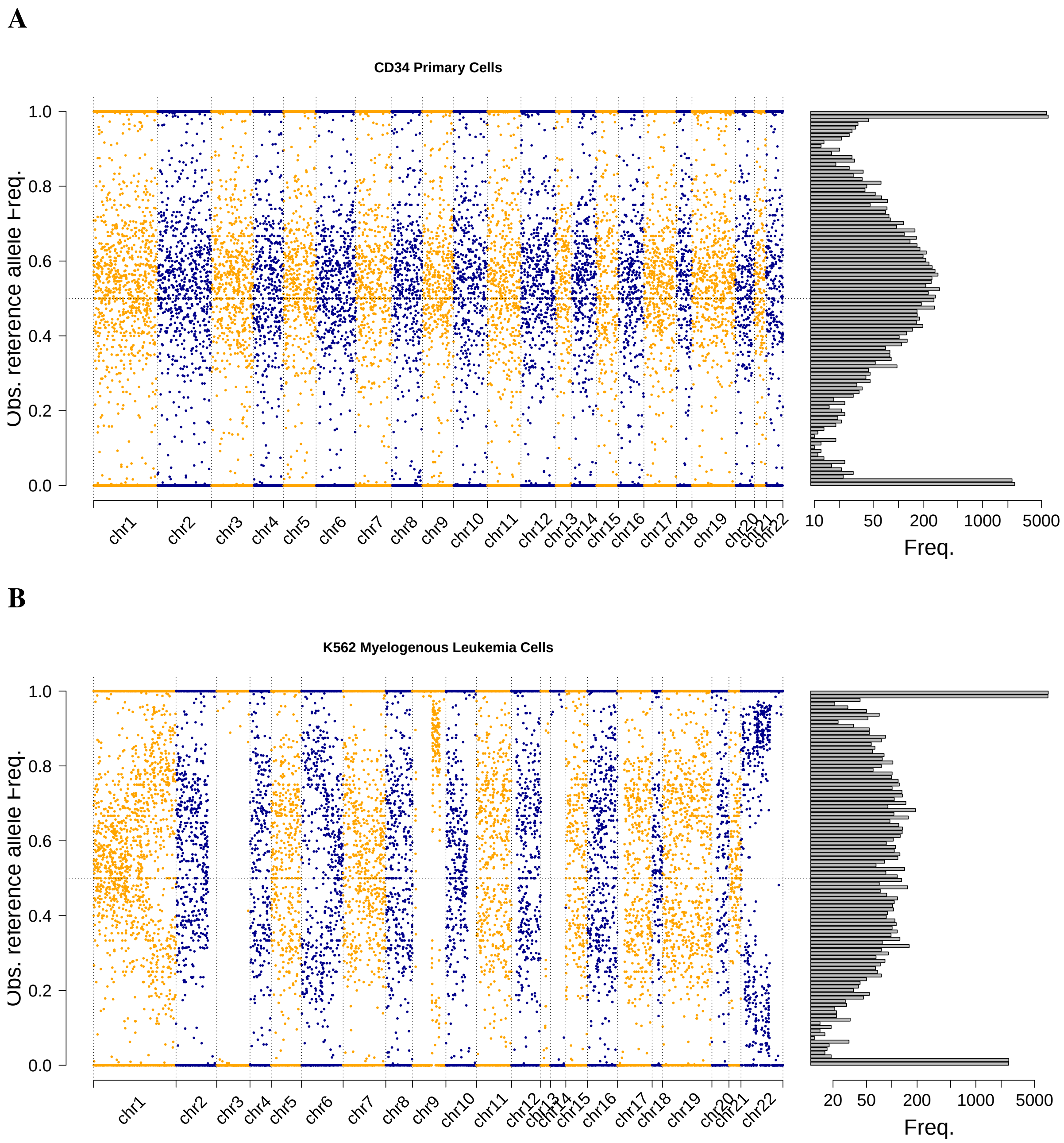

Figure S4: **Reference allele ratio at 1KG variants.** (A) Plot showing allele ratios for SNPs interrogated for CD34 primary cells. Three peaks on the histogram (right) correspond to homozygous reference (top), heterozygous (middle), and homozygous alternate (bottom) SNPs. (B) Plot showing allele ratios for SNPs interrogated for the cancer line K562. Signatures of chromosomal abnormalities are evident from the scatterplot, such as copy number variation and loss of heterozygosity.

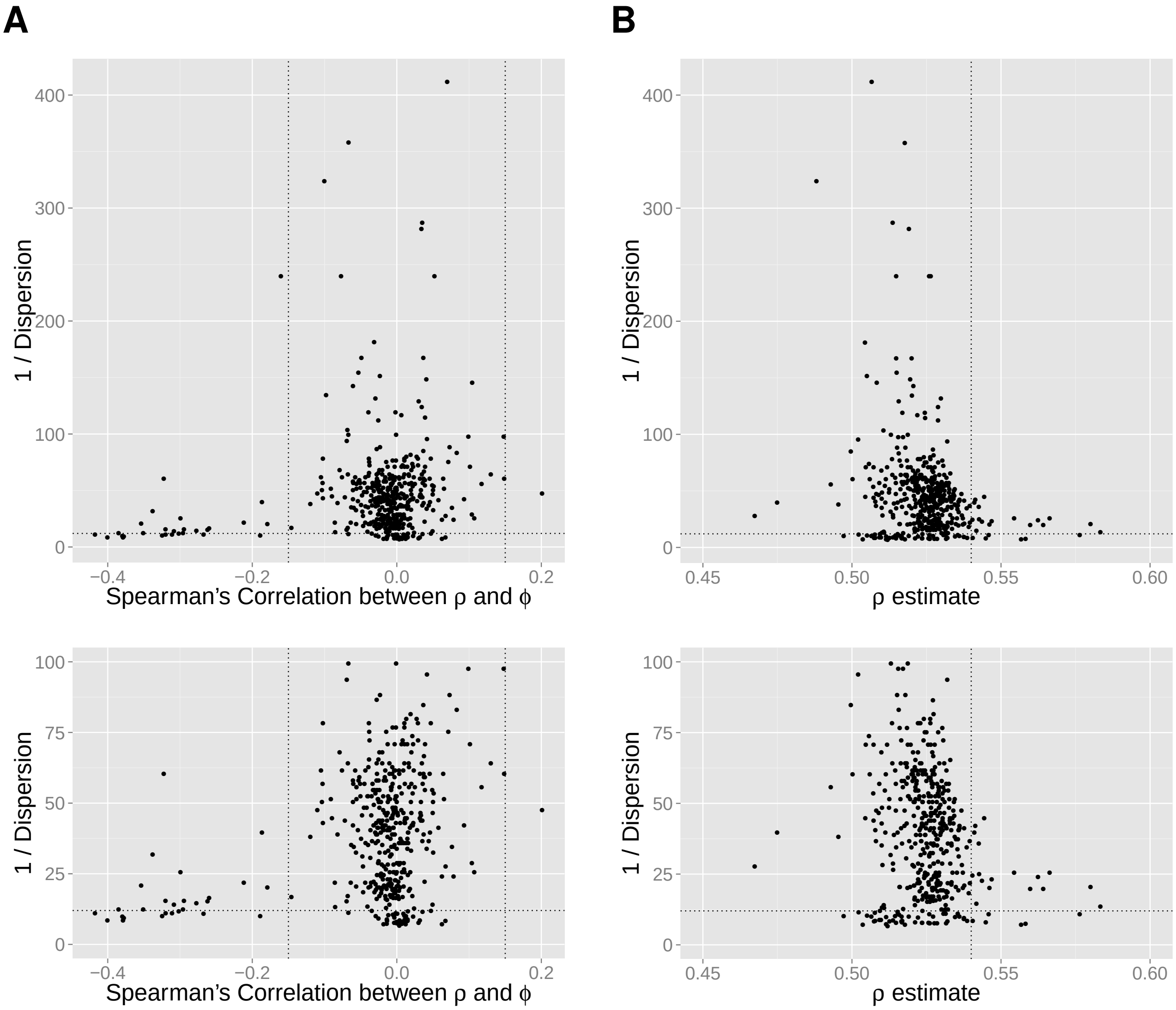

Figure S5: **Distribution of values used for post-ASB analysis filter criteria.** Dotted lines represent values used to filter samples. (A) Dispersion and correlation between ρ and ϕ (B) Dispersion and ρ estimation. Bottom plots show zoomed view of samples with $1/D < 100$.

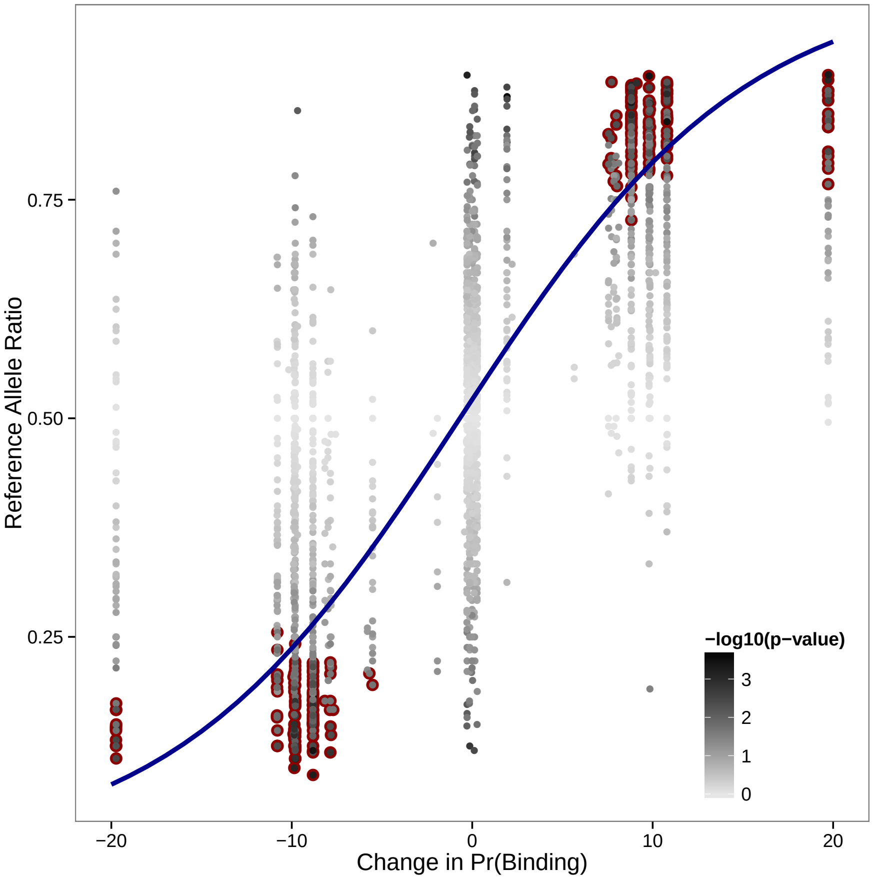

Figure S6: **Correlation between CENTIPEDE predictions and observed ASB.** SNPs identified in both the CENTIPEDE and ASB analysis are shown, shaded by p-value of allelic imbalance from QuASAR. Points circled in red display significant ASB at 20% FDR. The blue line is a logistic curve fit using points with a $p < 0.1$.

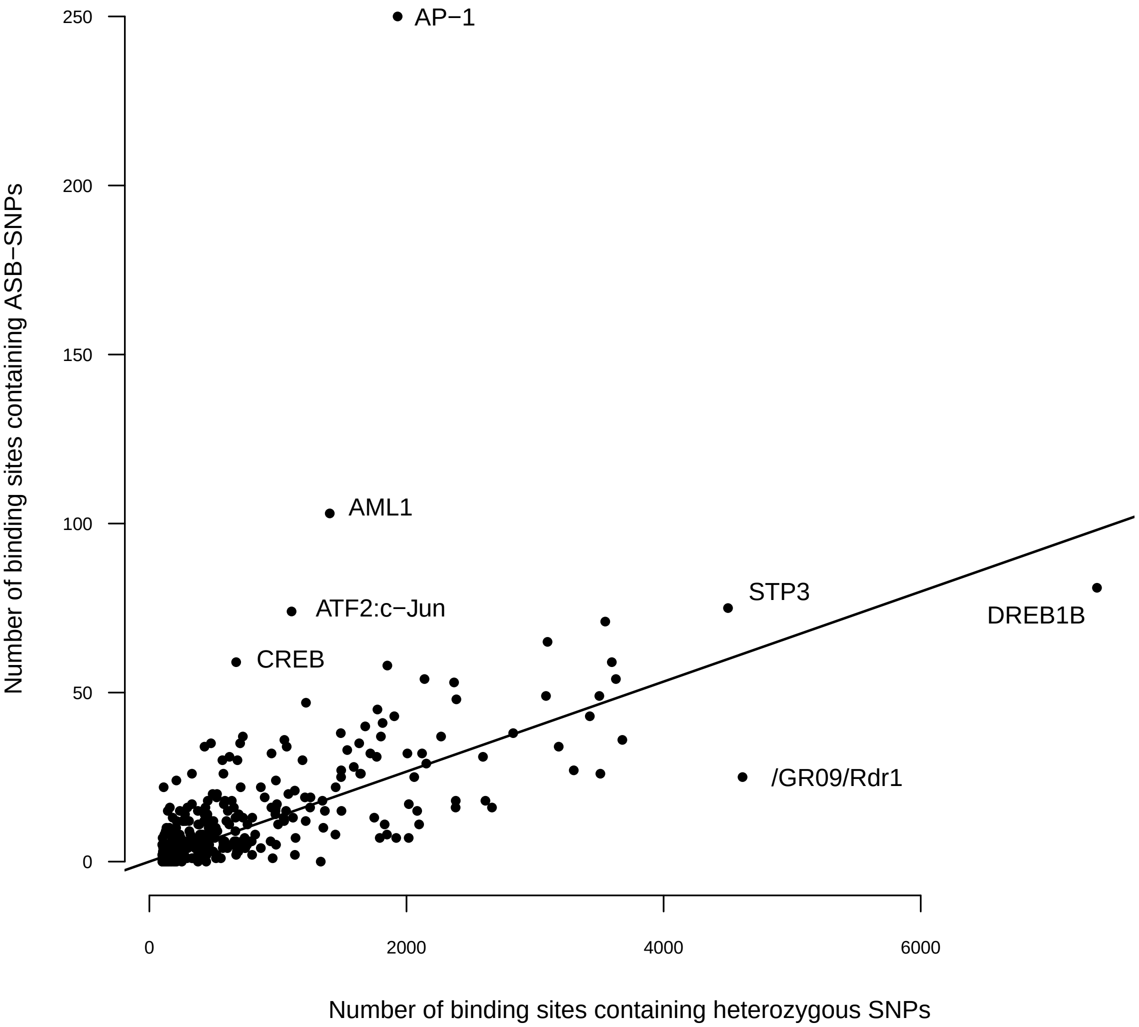

Figure S7: **Distribution of ASB enrichment ratios.** For all motifs with >100 heterozygous SNPs, an ASB enrichment ratio was calculated as # ASB-SNPs (20% FDR) / # heterozygous SNPs across all binding sites genome-wide. The black line shows the average ratio across all motifs. Several factors whose binding sites are highly enriched or depleted for ASB-SNPs are labeled.

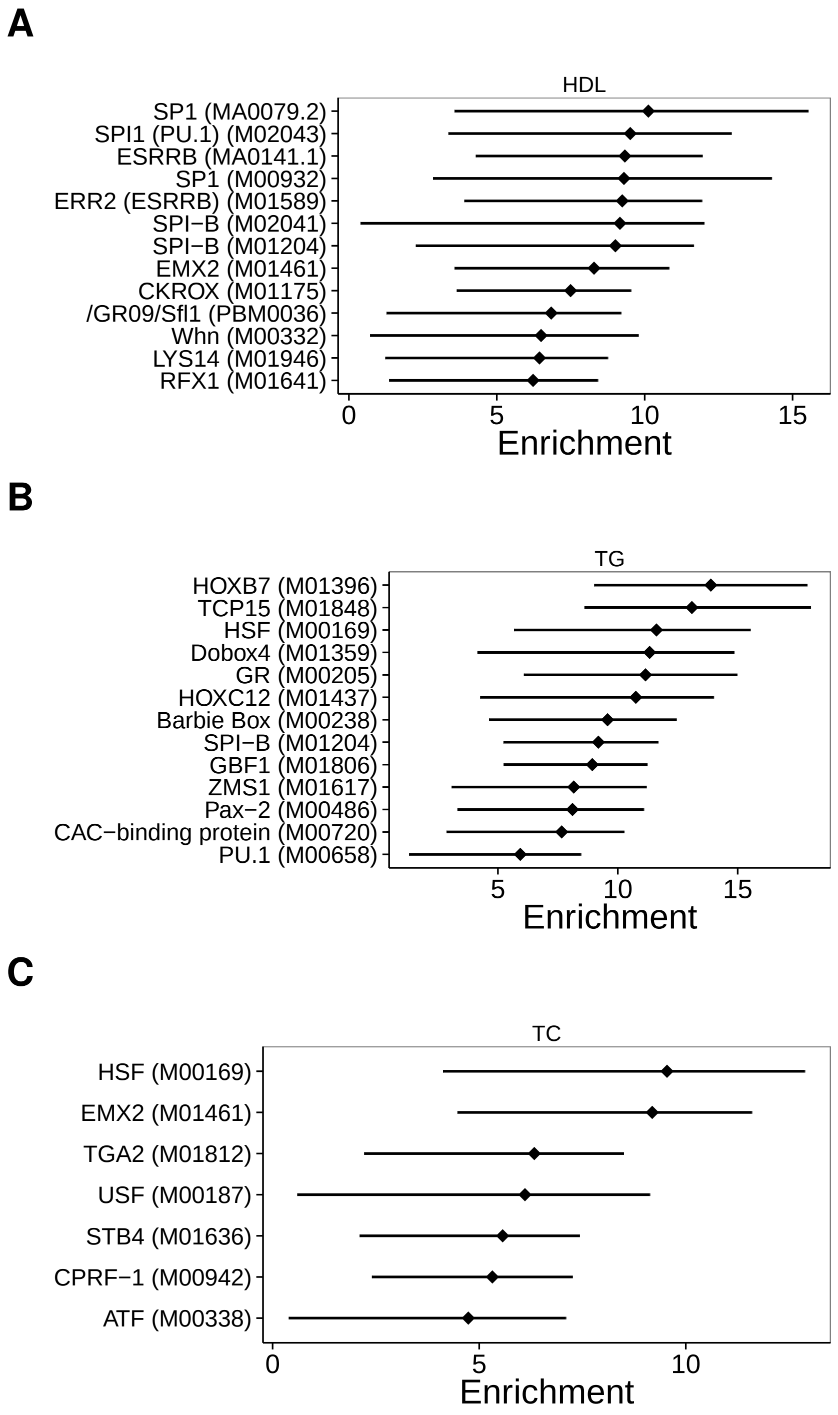

Figure S8: **Enrichment of factors for association with selected traits**. Shown are the log_2(enrichment) values with 95% confidence intervals for each factor whose binding sites are enriched for SNPs associated with (A) High-density lipoprotein (HDL) levels, (B) Triglyceride (TG) levels, and (C) Total cholesterol (TC) levels.

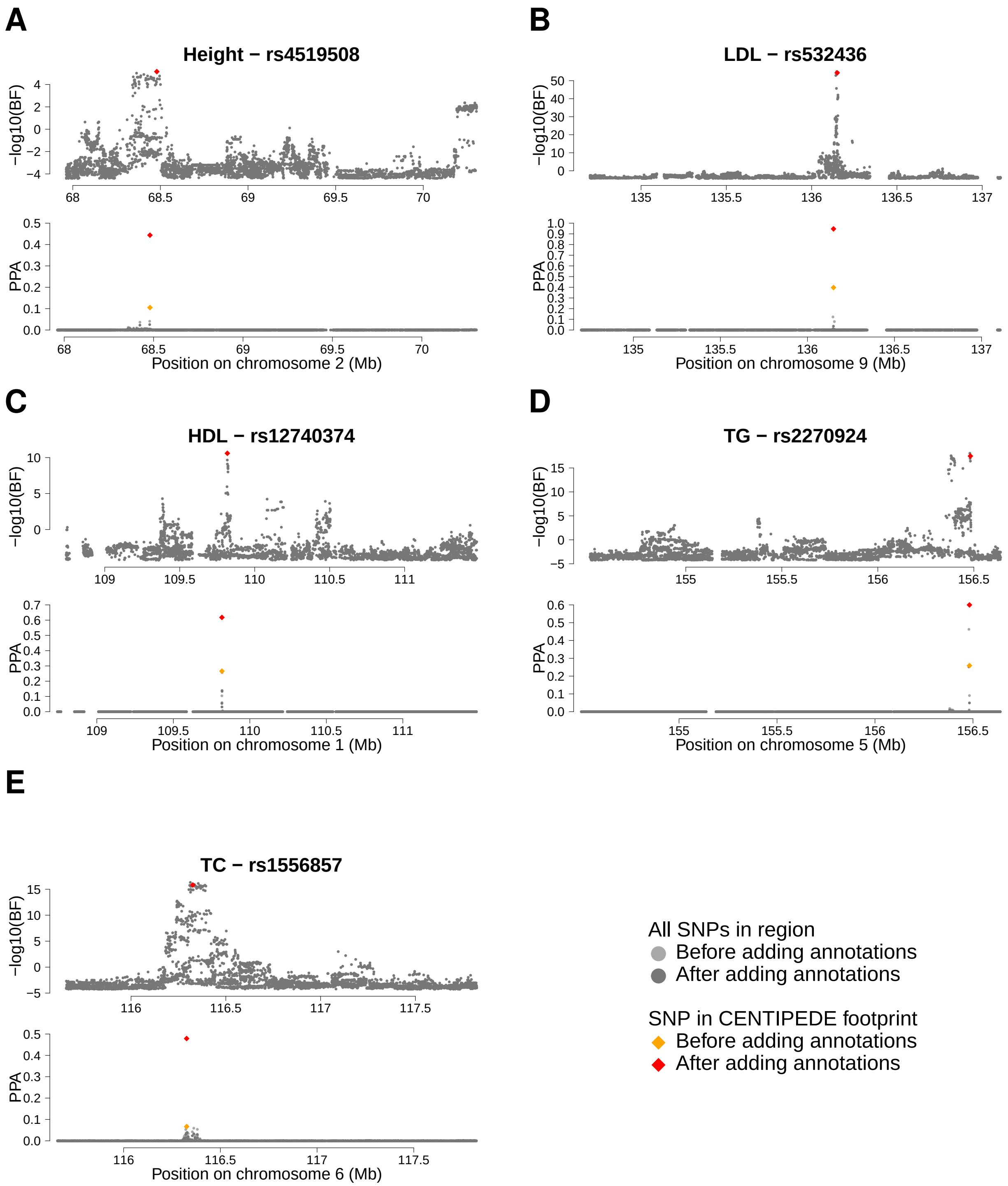

Figure S9: **Association plots identifying SNPs in footprints.** (A-E) Prior (top) and posterior (bottom) probabilities of association to the indicated trait for SNPs in the region.

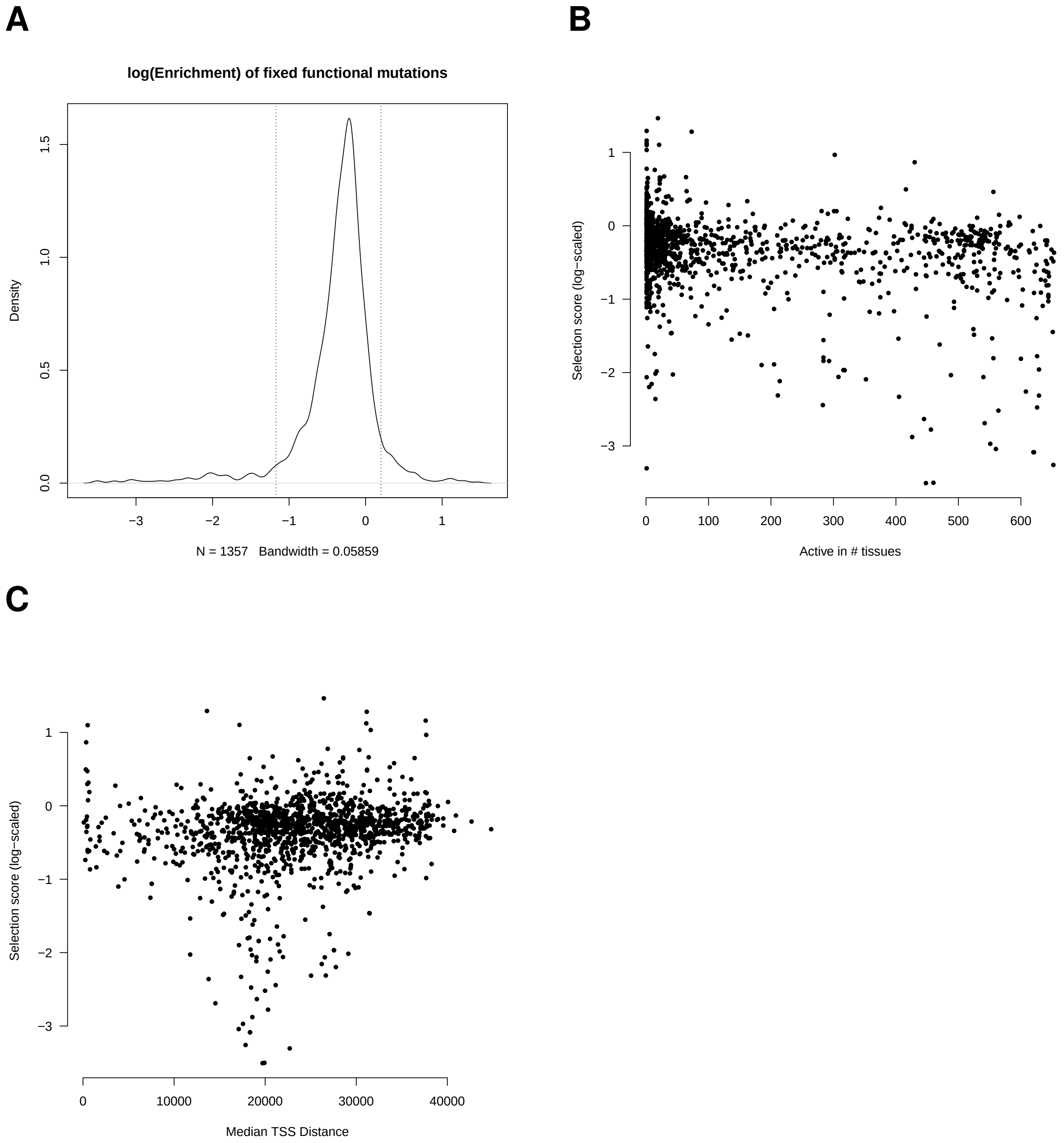

Figure S10: **Identifying selection of TF binding sites**. (A) Density plot showing the distribution of selection scores from the modified MK test. (B) Comparison of selection scores to the number of tissues each factor is predicted to be active in. (C) Comparison of selection scores to the median distance to the TSS across all sites for a given factor.