Supplementary Material for: Which genetic variants in DNase I sensitivity regions are functional?

Gregory A. Moyerbrailean[1], Chris T. Harvey[1], Cynthia A. Kalita[1],
Xiaoquan Wen[2], Francesca Luca[1,3,*], Roger Pique-Regi[1,4,*],

[1]Center for Molecular Medicine and Genetics, Wayne State University
[2]Department of Biostatistics, University of Michigan
[3]Department of Obstetrics and Gynecology, Wayne State University
[4]Department of Clinical and Translational Sciences, Wayne State University

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

Contents

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 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 (1KG) 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 1KG 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. Reads were then aligned using a custom mapper previously described in (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 S1. As the choice of aligner should have little impact on the ability to run the CENTIPEDE algorithm, we did not remap the reads as for the Roadmap Epigenomics samples.

2.2 Preprocessing for allele-specific analysis

Reads for allele specific analysis have to be carefully processed, as allele specific analysis is especially sensitive to biases in read data. To account for this, we aligned DNase-seq read data using a custom mapper with mappability filters (see Section S2.3). The Roadmap Epigenomics samples were previously aligned using our mapper (Section S2.1). To process the ENCODE samples in the same manner, we obtained raw sequence reads (FastQ format) directly from the EncodeDCC and reads were realigned with our custom mapper. 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 S2.3) to all samples used for allele-specific analysis.

2.3 Mappability filtering

We created an array of hash tables containing all possible 20-mer reads, where a 4-mer prefix indexes an array of 256 hash tables and the 16-mer suffix is used as 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 our 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 errors 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; which means that only reads that map without mismatches are used. Both hash tables are used to create the two following mappability tracks to assess alignment quality:

* 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). When aligning for allele specific analyses, 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 analyses 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. Using this threshold and 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 allele specific analyses. We do not consider a more complex filter as the probability of a read with two base-calling errors at 20bp read-length is very small.

2.4 Selection of genetic variants for ASB analysis

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).

Aligned reads were then piled up on this set of 1KG SNPs using samtools mpileup and the hg19 reference genome. 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 SNPs:

1. The SNP must be covered by >4 reads
2. 50% or more of the reads covering the SNP cannot start at the same position (i.e., PCR duplicates)
3. The coverage on the SNP cannot be in the top 0.01% sample-wise, as such exaggerated coverage usually indicates an unannotated copy number.

3 Identification and Mapping of Active Transcription Factors

3.1 Recalibration of position weight matrices

To recalibrate the PWMs in our two step approach we first created a reduced subset of motif matches. We obtained 1,949 seed PWMs from online databases as described in Section S1. Using these, 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.

Using these scores, we created for each motif a reduced set of locations (containing between 5,000 and 15,000 sites) that include both high and low scoring sequence matches to the original seed PWM. First we selected the top 5,000 best scoring sites in the human genome. Then, to expand the sequences included for a motif, we considered two different strategies. The first strategy was to randomly select additional sequences in which one randomly chosen base w of the PWM was not considered in the PWM score calculation. For the second strategy, which is the one we used for this paper, the additional sequences are selected using a heuristic that relies on sequence conservation due to evolutionary constraints across closely related species. In short, we conducted the following steps:

1. Scan the top 5,000 motif sequence matches in the chimp and macaque genomes

2. Lift over the coordinates to the human genome using the UCSC liftOver tool (excluding chains that are $<$10,000 bases or on very repetitive regions)

3. Calculate the PWM score again using the human sequence

Using this approach we add up to 10,000 new sites from the human genome that have low PWM scores. Compared to the sequences obtained using the first approach, these sequences are more likely to maintain the TF identity of the original PWM and to harbor true binding sites that will show a footprint.

Using these locations and the DNase-seq data listed in Table S1, we applied the CENTIPEDE model for each sample/motif combination. Then, we estimated the overall activity of each factor/sample combination by calculating a Z-score for the following logistic model that is used to calculate the prior probability of binding in CENTIPEDE:

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

where l represents each of the positions in our set of candidate sites, and π_l represents the probability of that position having a footprint. For most factors and experimental samples, a Z-score of at least 5 was the minimum for which a modest footprint was clearly evident (Figure S4). Using this Z-score value as a threshold, we detect 1,891 factors active (Z-score > 5) in at least one cell-type/tissue.

Finally, we used the CENTIPEDE binding predictions of this initial set to generate recalibrated sequence models. For each factor, we selected the best representative tissue (e.g., highest Z-score) and extracted the sequences predicted to have a factor bound (posterior probability > 0.99). Using these sequences, we calculated a new PWM from the base frequencies of each position in the footprint. A side-by-side comparison for several PWMs can be seen in Figure S6, and an evaluation with ChIP-seq data and ASB is available in Section 6.3.

3.2 Generation of CENTIPEDE binding predictions

Using a custom set of sequence models (see Section S3.1), we scanned the reference genome to identify all motif matches genome-wide. As a threshold on the scan, we calculated a match score separately for each sequence model designed to retain all sequences with at least a 10% prior probability of binding as in eq. (2). Scanning was done in two stages. First, we identified every match above the threshold using eq. (1) as before. Next, we scanned the genome, this time only considering motifs that overlapped 1KG variants. For each of these matches we calculated two PWM scores, one for each allele.

To generate the binding predictions, we 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 CENTIPEDE log ratio. 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. At the end of this process (Fig. S1) we generate an annotation of Footprints with a CENTIPEDE posterior probability > 0.99 divided in two major sets depending on whether they overlap with 1KG sequence variants. Those footprints overlapping genetic variants, which we call footprint-SNPs, are further classified based on CENTIPEDE's prior probability of binding (eq. (2)) for each allele into effect-SNPs and switch-SNPs. Effect-SNPs are footprint-SNPs with a 20-fold change in the prior odds of binding from one allele to the other. Switch-SNPs are Effect-SNPs where the prior probability for each allele crosses 0.5.

4 Analysis of Allele-Specific Binding

4.1 Validation of genotype predictions

To verify the genotyping accuracy, we compared the genotype calls from QuASAR for an individual fully resequenced by the 1KG Project (1KG individual NA12878). Of the 1,400 QuASAR-predicted heterozygous loci, all of them were confirmed to be true heterozygotes. Of the 11,278 QuASAR-predicted homozygous loci, only seven of them were actually heterozygotes. Additionally, all of the true homozygous calls were correct for the predicted allele. The seven miscalled heterozygotes are likely cases of extreme allelic imbalance, 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 S4.

4.2 Postprocesing of allele-specific data

To further filter out samples not well-suited for allele specific analysis, including cancer tissues and samples from pooled individuals (e.g., Figure S9), we examined two parameters estimated by QuASAR, ρ and M, as well as the non-reference allele frequency ϕ obtained from from 1KG. The ρ parameter represents the proportion of reads overlapping a SNP that match the reference allele. For heterozygous SNPs under the null model (no allelic imbalance) we would expect that the average $\hat{\rho}$ should be centered near 0.5. Deviation from 0.5 in a sample can be an indication of genetic aberrations, such as in a cancer sample, where copy number variation can be extensive or very high base-calling error rates. We also examined the correlation (Pearson correlation coefficient) between the ρ estimates and ϕ for each heterozygous locus, as the two should be independent of each other. Otherwise, this is a strong indication of sample mix-up or cross-sample contamination as new modes appear at $\rho = 0.25, 0.75$ or other intermediate frequencies with probabilities that are correlated with the reference allele frequency ϕ. The M parameter in QuASAR controls the degree of overdispersion of the beta-binomial distribution in the QuASAR model. A very high value of M indicates that the beta-binomial is almost a binomial distribution. On the other hand, a low value of M indicates more and more dispersion and a very high uncertainty in the underlying ρ being centered around 0.5, as is the case of samples with chromosomal aberrations and copy number alterations, as in cancer (see Figure S9B). After applying all filters, 316 samples remained for ASB analysis. A summary of the post-processing results can be found in Table S5 and Figure S10.

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 for 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. 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. 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). 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), following the strategy of Benjamini and Bogomolov (2014). The results of this analysis are summarized in the main text, Table 1.

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 S11 shows the correlation between our prediction and the observed ASB for the most predictive sequence model, belonging to the factor AP-1.

6 Evaluation of recalibrated sequence models

6.1 Precision versus recall analysis with ChIP-seq

To compare the new sequence models to the originals, we first performed precision recall operating characteristic (P-ROC) curve analysis using PWM scores of motif matches and ChIP-seq peaks from GM12878 samples. We annotated a list of all binding sites identified as a PWM match or as having a ChIP-peak, 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 PWM match, a PWM score of 0 was used. For each selected factor, we compared the precision-recall curve using the original PWM models and the updated PWM models (Figure S7). The curves show that in general, for a given precision (precision = 1 - FDR, false discovery rate), the updated sequence models have higher recall (sensitivity) than the original PWM in detecting ChIP-seq peaks.

6.2 Predicted binding strength correlation analysis with ChIP-seq

We also examined the correlation between the PWM scores (for the original seed and the recalibrated motif models) and the number of ChIP-seq reads. For each PWM we identified all matching sites genome-wide and extracted the ChIP-seq read coverage. Compared to the seed PWMs, we find that the revised PWMs are better correlated with the ChIP-seq data. Data for the individual comparisons can be seen in Figure S8. The new recalibrated models, seem to better capture the relationship between the prior probability of binding derived from the PWM score and TF occupancy as measured by ChIP-seq reads. This indicates that we are also capturing a wide range of binding events including weak binding sites.

6.3 PWM recalibration step impact on ASB

We also wanted to examined whether the recalibration process preferentially selected sites with the strongest binding, and therefore most affected by variation. If so, this would potentially bias our downstream ASB analysis, as we partition the SNPs based on their predicted impact on binding (Section S5.1). To see if this was the case, we compared ASB results within footprints predicted by the two sets of sequence models. Using the original seed PWMs, we ran CENTIPEDE as in Section S3.2.

For each set of sequence models, we compared the proportion of SNPs within footprints predicted to have an effect. We find that the recalibrated sequence models discover more variation within the footprints overall. However, the proportion of SNPs with low p-values ($p < 0.05$) in footprints predicted to affect binding versus those that do not, remains extremely similar between the old models (OR 2.14, Fisher $p = 1.6 \times 10^{-289}$) and the new models (OR 2.08, Fisher $p = 1.5 \times 10^{-263}$). Table S3 shows the values used for this comparison.

7 Genomic Annotation and Selection Signals

7.1 Allele frequency

Allele frequency for each 1KG SNP was obtained as described in Section S1. For each SNP, we calculated the minor allele frequency by taking the absolute value of the difference between the allele frequency and 0.5. We obtained coding SNP annotations from dbSNP (version

137). We classified coding SNPs into two categories, synonymous and non-synonomous, the latter category encompassing missense, nonsense, and early stop variants. For the analysis, we only considered bi-allelic SNPs, and those unambiguously categorized as either synonymous or non-synonymous. Of the 784,003 SNPs we analyzed in coding regions, 298,986 (38%) are synonymous.

7.2 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 S1), we determined the distance for each SNP in our set. For motif-wide analysis, we determined the median distance to the nearest TSS across all binding sites genome-wide.

7.3 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 (ASB enrichment ratio). 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 S8.

7.4 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%. At these loci in the chimp genome, we calculated PWM scores as in Section S3.1. 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 $\geq$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 S11.

8 Overlap with Genome-Wide Association Studies

8.1 Analysis of SNPs in the GWAS catalog

We created an expanded GWAS catalog by adding SNPs in linkage disequilibrium (LD) with each GWAS hit, using 1KG LD data for European populations $r^2 > 0.8$. To identify overlap between our annotations and those associated with a GWAS trait, we intersected our results with this expanded catalog, but counting only one hit per GWAS loci. We used a Fisher exact test to determine if the proportion of effect-SNPs for a given annotation were enriched in the catalog.

8.2 Adding annotations to SNPs associated with complex traits

We integrated our CENTIPEDE footprint annotations into the combined models learned in Pickrell (2014) for GWAS meta-studies corresponding to 18 traits (Table S14), using the fgwas command line program. We assessed enrichment or depletion for footprint annotations using the log_2(enrichment) values, excluding any motifs whose 95% confidence interval (CI) spanned zero. For each TF motif whose binding sites are either significantly enriched or depleted for trait-associated SNPs (Figure S13), 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 88 unique SNPs whose associations were strengthened by our footprint annotations (Table S13).

8.3 Validating putative causal SNPs by reporter gene assays

To validate the predicted allelic effects on gene expression for 6 of the 88 SNPs identified by the fgwas analysis, we first constructed inserts containing the reference or alternate allele for each SNP of interest. Each region was amplified from genomic DNA extracted from LCLs (Coriell).

Primers were designed using the Infusion Clontech online primer design tool for inserts containing the SNP of interest ±75bp. Primers were ordered from IDT technologies. Inserts were amplified by PCR and pGL4.23 plasmid was linearized (inverse PCR) using Clontech Hi-fi PCR premix and following the manufacturer's instructions. PCR products and linearized plasmid were resolved on agarose gel, excised and purified using Nucleospin gel extraction and PCR cleanup kit (Clontech). Inserts were cloned into linearized pGL4.23 using the Infusion Cloning HD kit (Clontech). Transformation was done using Stellar Competent cells (Clontech) and DNA was extracted from selected colonies using the PureYield kit (Promega). The allelic status and the absence of artifactual mutations of each clone was validated by Sanger sequencing performed by Genewiz. Transfections were performed into GM18507 using the standard protocol for the Nucleofector electroporation (Lonza). After 10 hours we measured Firefly and Renilla (transfection control) luciferase activity using the Dual-Glo Luciferase Assay Kit (Promega) on the GloMax instrument (Promega). Luciferase activity was measured for four replicate experiments. We then used a t-test to identify significant differences in the expression of the reporter gene, calculated by the ratio of the firefly to the renilla activity, normalized to the ratio of the activity in the untransfected cells. We contrasted the activity of each construct to the pGL4.23 vector, to assess enhancer/repressor activity of each region. To evaluate allele-specific effects, we contrasted the activity of the reference allele to the alternate allele for each region. These results are summarized in Table S15.

References

Benjamini, Y. and Bogomolov, M., 2014. Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **76**:297–318.

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 source, 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: **Comparison of ASB within footprints between PWM models.** Shown is the number of ASB-SNPs within footprints identified by the two sets of PWM sequence models. The counts are stratified by p-value from the QuASAR test of ASB. Note that the old models, by default, only select sites with a PWM score $>$ 12; for comparison, the same constraint has been placed on the sequences used from the new models.

		QuASAR ASB Test	
		p $<$0.05	P $>$0.05
Seed sequence models	Effect-SNPs	4582	30834
	Footprint-SNPs (no effect)	6565	94325
Recalibrated sequence models	Effect-SNPs	4657	39675
	Footprint-SNPs (no effect)	5484	97272

Table S4: **Validation of genotype predictions.** A comparison of 1KG genotypes and those called by QuASAR for the 12,650 loci examined in the LCL GM12878.

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

Table S5: **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 were manually examined and known cancer samples were removed.

Parameter	Threshold	# Removed
Avg. ρ	$<$0.54	78
$\lvert\text{cor}(\rho, \phi)\rvert$	$<$0.15	31
1/Disp.=M	$>$12	58
Cancer†	–	13

†without evident chromosomal abnormalities

Table S6: **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 (bp)	-7.22×10^{-8}	$<10^{-16}$
Number of Tissues	-2.08×10^{-3}	$<10^{-16}$
Minor Allele Frequency (0-1)	-1.74×10^{-1}	$<10^{-16}$

[†] units are log-fold change in factor affinity between alleles.

Table S7: **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 S8: **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 S9: **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 S10: **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 Immune 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 S11: **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/`*.*

Table S12: **Active motifs in each sample.** For each sample, motifs were determined active if the Z-score, obtained from eq. 2, was > 5, and if the motif instances showed correlation with DHS peaks (Section S3.2).

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

Table S13: **SNPs associated with GWAS traits that fall in CENTIPEDE-predicted TF binding sites.** PPA, Posterior probability of association estimated by fgwas for each SNP. ”Before” indicates the PPA from the base model, ”after” indicates the PPA after adding footprint annotations to the model. The p-values listed are derived from the z-scores that are used as input for fgwas.

Trait	Motif	Factor	rsID	PPA before	PPA after	p-value
BMI	M00287	NF-Y	rs12641981	0.2133	0.9486	3.67×10^{-17}
BMI	M00287	NF-Y	rs13098327	0.3105	0.9709	5.70×10^{-08}
BMI	M01608	DAL82	rs4704230	0.2867	0.6469	8.95×10^{-08}
CD	M00197	ABF1	rs1052248	0.0518	0.2537	5.59×10^{-11}
CD	PBM0124	Elf4	rs2476601	0.7872	0.9985	2.18×10^{-09}
CD	M00433	Hmx3 (Nkx5-1)	rs3810936	0.0681	0.9740	4.87×10^{-16}
CD	M00664	STE12	rs3828917	0.0297	0.7407	2.10×10^{-07}
CD	M02032	SWI4	rs7746082	0.1133	0.6320	2.25×10^{-08}
FG	M00698	HEB	rs13266634	0.9479	0.9908	3.93×10^{-17}
FG	MA0019.1	Ddit3::Cebpa	rs2191348	0.2111	0.4478	4.19×10^{-21}
FNBMD	M01644	EDS1	rs10205005	0.1831	0.3384	2.20×10^{-07}
FNBMD	M01550	Mbp1	rs383911	0.4307	0.8084	1.12×10^{-16}
FNBMD	M00357	bZIP910	rs6426749	0.2642	0.5711	1.03×10^{-23}
FNBMD	M00241	Nkx2-5	rs7466269	0.7019	0.9613	1.91×10^{-08}
HB	M01538	Aro80	rs13219787	0.8098	0.9741	5.23×10^{-09}
HB	M00986	Churchill	rs198846	0.9989	0.9998	7.11×10^{-31}
HDL	M00332	Whn	rs1044973	0.3681	0.7553	7.65×10^{-08}
HDL	M01641	RFX1	rs12740374	0.2660	0.6177	3.10×10^{-08}
HDL	M01946	LYS14	rs676210	0.2959	0.6569	1.33×10^{-30}
HDL	M01461	EMX2	rs6907508	0.2188	0.4655	3.65×10^{-10}
Height	MA0041.1	Foxd3	rs10171985	0.0359	0.2022	2.99×10^{-07}
Height	M01114	E2F	rs11752007	0.7399	0.9523	4.84×10^{-08}
Height	M01264	TBX15	rs11752007	0.7399	0.9553	4.84×10^{-08}
Height	M01641	RFX1	rs12740374	0.0837	0.5101	3.81×10^{-05}
Height	MA0142.1	Pou5f1	rs17511102	0.9989	0.9999	6.32×10^{-13}
Height	M01641	RFX1	rs314263	0.2666	0.8066	4.16×10^{-18}
Height	M00792	SMAD	rs34529769	0.3947	0.8177	1.13×10^{-10}
Height	M00776	SREBP	rs3828559	0.1843	0.3429	2.74×10^{-06}
Height	M01641	RFX1	rs4073154	0.0628	0.4411	2.19×10^{-08}
Height	M01114	E2F	rs4519508	0.1050	0.4438	8.12×10^{-06}
Height	M01641	RFX1	rs4725984	0.1155	0.6012	2.00×10^{-08}
Height	M00104	CDP CR1	rs4973431	0.1841	0.6903	8.51×10^{-07}
Height	M00241	Nkx2-5	rs7466269	0.4258	0.8376	5.86×10^{-15}
Height	M00104	CDP CR1	rs894344	0.0866	0.5023	7.79×10^{-11}
Height	MA0041.1	Foxd3	rs9849338	0.3562	0.7910	2.53×10^{-14}
LDL	M00359	bZIP911	rs2075375	0.2384	0.8109	1.64×10^{-07}
LDL	M01863	ATF-3	rs217381	0.2705	0.6777	7.41×10^{-11}
LDL	PBM0176	HLH-29	rs217386	0.2596	0.5864	2.13×10^{-11}
LDL	M01812	TGA2	rs2479409	0.9314	0.9979	9.70×10^{-29}
LDL	M01812	TGA2	rs267733	0.9777	0.9992	3.26×10^{-08}

Continued on next page

Table S13 – continued from previous page

Trait	Motif	Factor	rsID	PPA before	PPA after	p-value
LDL	M00513	ATF3	rs267733	0.9777	0.9983	3.26×10^{-08}
LDL	M00359	bZIP911	rs2954021	0.0748	0.5249	6.13×10^{-29}
LDL	M00187	USF/E-box	rs532436	0.3971	0.7711	1.80×10^{-27}
LDL	M00513	ATF3	rs6920309	0.0595	0.4940	3.55×10^{-13}
LDL	MA0069.1	Pax6	rs9293637	0.1861	0.3661	2.62×10^{-11}
LDL	M00359	bZIP911	rs9438904	0.0327	0.3180	6.21×10^{-10}
LSBMD	M01733	MZF1	rs11898505	0.3838	0.8157	4.86×10^{-12}
LSBMD	M01525	Put3	rs1524068	0.2181	0.8539	2.41×10^{-15}
LSBMD	M01525	Put3	rs4869741	0.0866	0.6430	1.63×10^{-19}
MCH	M00235	AhR:Arnt	rs11968166	0.9267	0.9923	4.67×10^{-34}
MCH	M01770	XBP1	rs12718598	0.7329	0.9667	9.29×10^{-09}
MCH	M00942	CPRF-1	rs1800562	0.4999	0.9280	1.73×10^{-66}
MCH	M01234	IPF1	rs2236496	0.9794	0.9795	1.17×10^{-16}
MCH	M00496	STAT1	rs3851296	0.0255	0.2805	1.39×10^{-08}
MCH	M01175	CKROX	rs4729597	0.9522	0.9927	4.21×10^{-17}
MCH	M00942	CPRF-1	rs56050898	0.0773	0.5227	1.23×10^{-07}
MCH	M00513	ATF3	rs6568571	0.2151	0.4233	5.52×10^{-26}
MCH	M00235	AhR:Arnt	rs7664687	0.0299	0.2360	1.42×10^{-06}
MCH	M00496	STAT1	rs869785	0.3464	0.8891	1.87×10^{-14}
MCH	MA0093.1	USF1	rs911910	0.1820	0.5454	6.42×10^{-07}
MCH	M01770	XBP1	rs9660992	0.0685	0.4665	3.54×10^{-10}
MCHC	M01641	RFX1	rs11240734	0.2396	0.5485	5.91×10^{-12}
MCHC	M01658	AML1	rs12733102	0.3623	0.8985	3.22×10^{-06}
MCHC	M00986	Churchill	rs198846	0.9994	0.9999	3.73×10^{-21}
MCHC	M01030	Rim101p	rs4737009	0.3690	0.8040	2.43×10^{-11}
MCHC	M00345	GAMYB	rs4737010	0.4851	0.9013	2.14×10^{-11}
MCHC	M01658	AML1	rs9389268	0.2416	0.8496	9.60×10^{-15}
MCV	M01641	RFX1	rs10901252	0.1105	0.5953	4.06×10^{-08}
MCV	M01641	RFX1	rs11240734	0.5864	0.9445	2.22×10^{-09}
MCV	M00017	ATF	rs12718597	0.0883	0.5035	2.38×10^{-13}
MCV	M01770	XBP1	rs12718598	0.5803	0.9573	3.35×10^{-14}
MCV	M00942	CPRF-1	rs1800562	0.8678	0.9841	7.75×10^{-47}
MCV	M00496	STAT1	rs3851296	0.0945	0.7763	1.59×10^{-08}
MCV	M01175	CKROX	rs4729597	0.8192	0.9681	6.97×10^{-15}
MCV	M00942	CPRF-1	rs56050898	0.0558	0.3465	1.75×10^{-06}
MCV	M01863	ATF-3	rs6568571	0.5065	0.8577	8.07×10^{-23}
MCV	M01055	NAC69-1	rs6656196	0.4925	0.8656	1.46×10^{-08}
MCV	M01223	P50:P50	rs6730558	0.4072	0.8344	1.06×10^{-06}
MCV	M01197	ELF5	rs7022455	0.3546	0.8136	1.93×10^{-07}
MCV	M00041	ATF2:c-Jun	rs72667750	0.0494	0.2991	2.71×10^{-06}
MCV	M00496	STAT1	rs869785	0.3501	0.9456	4.83×10^{-15}
MCV	M01065	ABZ1	rs911910	0.2084	0.7617	1.07×10^{-08}
MCV	M01770	XBP1	rs9660992	0.2968	0.8784	5.18×10^{-10}
PCV	M00262	Staf	rs1934661	0.2199	0.5262	3.95×10^{-07}
PCV	M00187	USF/E-box	rs532436	0.4485	0.8706	7.48×10^{-18}
PLT	M00171	Adf-1	rs149290349	0.4414	0.8479	5.33×10^{-09}

Continued on next page

Table S13 – continued from previous page

Trait	Motif	Factor	rsID	PPA before	PPA after	p-value
PLT	M01814	AML2	rs17030845	0.2770	0.6496	6.34×10^{-11}
PLT	M00178	CREB	rs2336384	0.1062	0.5901	6.24×10^{-09}
PLT	M01492	Ynr063w	rs34592828	0.1358	0.2124	1.12×10^{-08}
PLT	M00739	E2F-4:DP-2	rs4731120	0.9853	0.9980	1.38×10^{-12}
PLT	M00178	CREB	rs540909	0.2168	0.7814	3.99×10^{-07}
PLT	M01784	UPC2	rs6141	0.6676	0.9428	3.09×10^{-08}
PLT	M01653	HMGIY	rs9399137	0.9999	1.0000	2.52×10^{-47}
RBC	M01797	SIRT6	rs10758656	0.7906	0.9607	3.79×10^{-10}
RBC	M00694	E4F1	rs12718598	0.4699	0.8635	3.04×10^{-09}
RBC	M01709	MAFA	rs13027161	0.6889	0.9273	3.53×10^{-13}
RBC	M00261	Olf-1	rs1434282	0.7906	0.9632	7.39×10^{-09}
RBC	M00262	Staf	rs1934661	0.4217	0.8423	4.51×10^{-07}
RBC	M00942	CPRF-1	rs532436	0.1135	0.2315	2.35×10^{-21}
RBC	M00694	E4F1	rs73019748	0.0875	0.4202	1.14×10^{-06}
TC	M01636	STB4	rs1556857	0.0666	0.4787	1.19×10^{-10}
TC	M00942	CPRF-1	rs1800562	0.9249	0.9946	1.24×10^{-08}
TC	M01636	STB4	rs2235215	0.6764	0.9655	2.79×10^{-10}
TC	M01812	TGA2	rs2479409	0.9696	0.9967	1.91×10^{-24}
TC	M00942	CPRF-1	rs532436	0.4694	0.9313	6.63×10^{-26}
TC	M00187	USF/E-box	rs532436	0.4694	0.8359	6.63×10^{-26}
TC	M01636	STB4	rs553427	0.0320	0.3031	2.64×10^{-12}
TG	M01617	ZMS1	rs13173241	0.1691	0.2924	2.58×10^{-10}
TG	M00486	Pax-2	rs2270924	0.2590	0.5997	2.51×10^{-11}
TG	M01848	TCP15	rs7789194	0.2586	0.6452	1.38×10^{-07}
TG	M01204	SPI-B	rs9686661	0.6413	0.9510	6.59×10^{-11}

Table S14: **Summary of GWAS meta analysis traits examined.** Shown for each trait is the trait abbreviation and the citation for the meta analysis study.

Abbreviation	Trait	Study
BMI	Body mass index	Speliotes, E.K., et al. (2010). Nat. Genet. 42, 937-948
CD	Chron disease	Jostins, L., et al. (2012). Nature 491, 119-124
FG	Fasting glucose levels	Manning, A.K., et al. (2012). Nat. Genet. 44, 659-669
FNBMD	Bone mineral density (femur)	Estrada, K., et al. (2012). Nat. Genet. 44, 491-501
HB	Hemoglobin levels	van der Harst, P., et al. (2012). Nature 492, 369-375
HDL	HDL cholesterol levels	Teslovich, T.M., et al. (2010). Nature 466, 707-713
Height	Height	Lango Allen, H., et al. (2010). Nature 467, 832-838
LDL	LDL cholesterol levels	Teslovich, T.M., et al. (2010). Nature 466, 707-713
LSBMD	Bone mineral density (lumbar spine)	Estrada, K., et al. (2012). Nat. Genet. 44, 491-501
MCH	Mean red blood cell hemoglobin	van der Harst, P., et al. (2012). Nature 492, 369-375
MCHC	Mean corpuscular hemoglobin concentration	van der Harst, P., et al. (2012). Nature 492, 369-375
MCV	Mean red blood cell volume	van der Harst, P., et al. (2012). Nature 492, 369-375
MPV	Mean platelet volume	Gieger, C., et al. (2011). Nature 480, 201-208
PCV	Packed red blood cell volume	van der Harst, P., et al. (2012). Nature 492, 369-375
PLT	Platelet counts	Gieger, C., et al. (2011). Nature 480, 201-208
RBC	Red blood cell count	van der Harst, P., et al. (2012). Nature 492, 369-375
TC	Total cholesterol levels	Teslovich, T.M., et al. (2010). Nature 466, 707-713
TG	Triglyceride levels	Teslovich, T.M., et al. (2010). Nature 466, 707-713

Table S15: **Reporter gene assay results.** For each of the 6 SNPs tested, listed are the results for the reference allele (top) and the alternate allele (bottom). Shown is the average and standard error (across four replicates) of the firefly luciferase activity normalized to the renilla luciferase activity, for each construct (Norm Expr) and for the pGL4.23 vector (Empty Vector). The last two columns are the t-test p-values comparing the activity of the reference allele to the alternate allele (vs ref), and of each allele to the pGL4.23 vector (vs empty).

rsID	Allele	Norm Expr	Std Err	Empty Vector	Std Err	p-value (vs empty)	p-value (vs ref)
rs532436	G	0.454	0.114	0.409	0.048	0.722	
	A	0.179	0.027	0.409	0.048	0.001	0.048
rs4519508	G	0.291	0.037	0.409	0.048	0.071	
	A	0.430	0.049	0.409	0.048	0.768	0.041
rs9686661	C	0.771	0.221	0.315	0.065	0.095	
	T	0.578	0.117	0.315	0.065	0.097	0.470
rs6730558	T	0.800	0.126	0.315	0.065	0.014	
	C	0.396	0.038	0.315	0.065	0.024	0.022
rs2336384	G	2.219	0.485	0.315	0.065	0.006	
	T	1.626	0.245	0.315	0.065	0.002	0.336
rs4973431	C	0.546	0.067	0.315	0.065	0.048	
	T	0.317	0.013	0.315	0.065	0.979	0.035

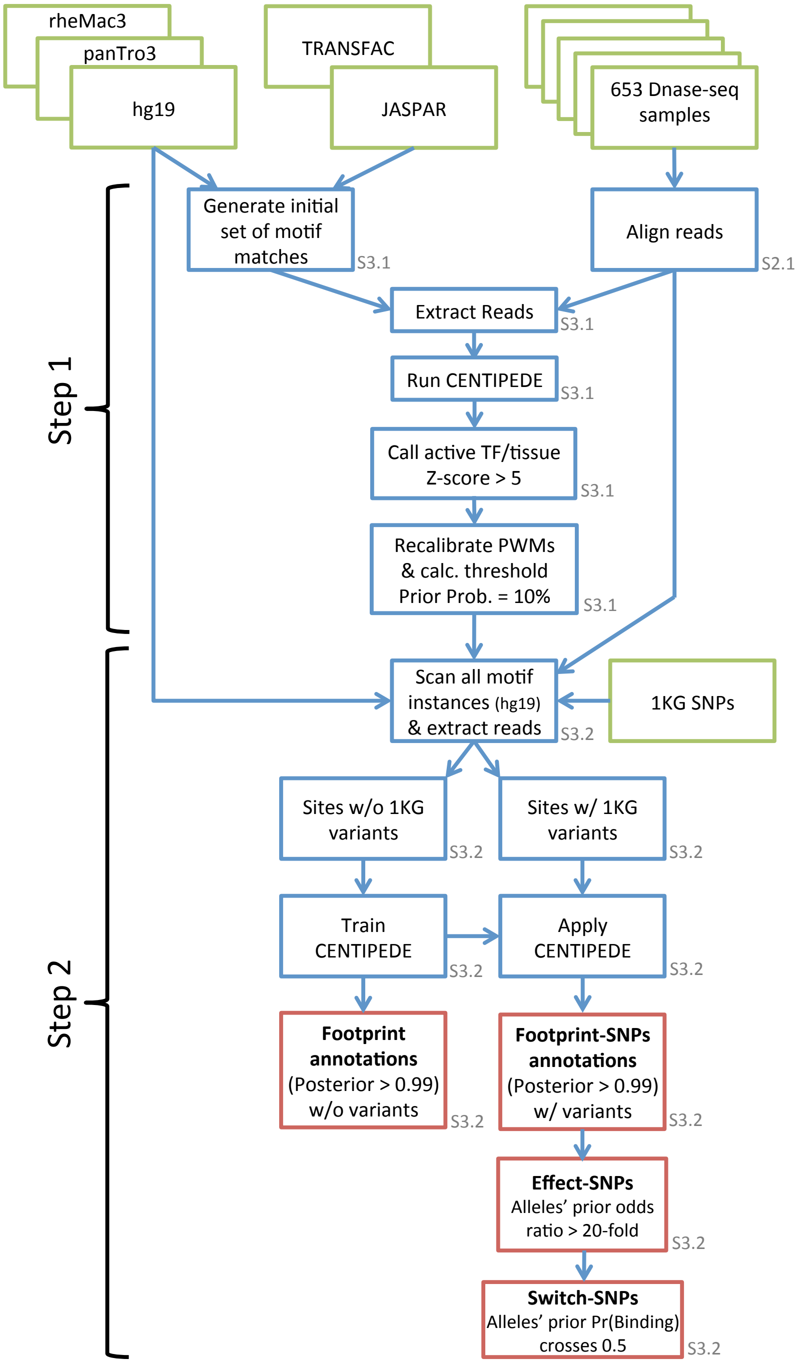
rheMac3
panTro3
hg19
TRANSFAC
JASPAR
653 Dnase-seq samples
Generate initial set of motif matches
S3.1
Align reads
S2.1
Extract Reads
S3.1
Run CENTIPEDE
S3.1
Call active TF/tissue Z-score > 5
S3.1
Recalibrate PWMs & calc. threshold Prior Prob. = 10%
S3.1
Step 1
Scan all motif instances (hg19) & extract reads
S3.2
1KG SNPs
Sites w/o 1KG variants
S3.2
Sites w/ 1KG variants
S3.2
Train CENTIPEDE
S3.2
Apply CENTIPEDE
S3.2
Footprint annotations (Posterior > 0.99) w/o variants
S3.2
Footprint-SNPs annotations (Posterior > 0.99) w/ variants
S3.2
Effect-SNPs Alleles' prior odds ratio > 20-fold
S3.2
Switch-SNPs Alleles' prior Pr(Binding) crosses 0.5
S3.2
Step 2

Figure S1: **Flowchart detailing steps of the CENTIPEDE-based annotation of regulatory regions and variants.** Numbers next to boxes refer to the corresponding section in the Supplement.

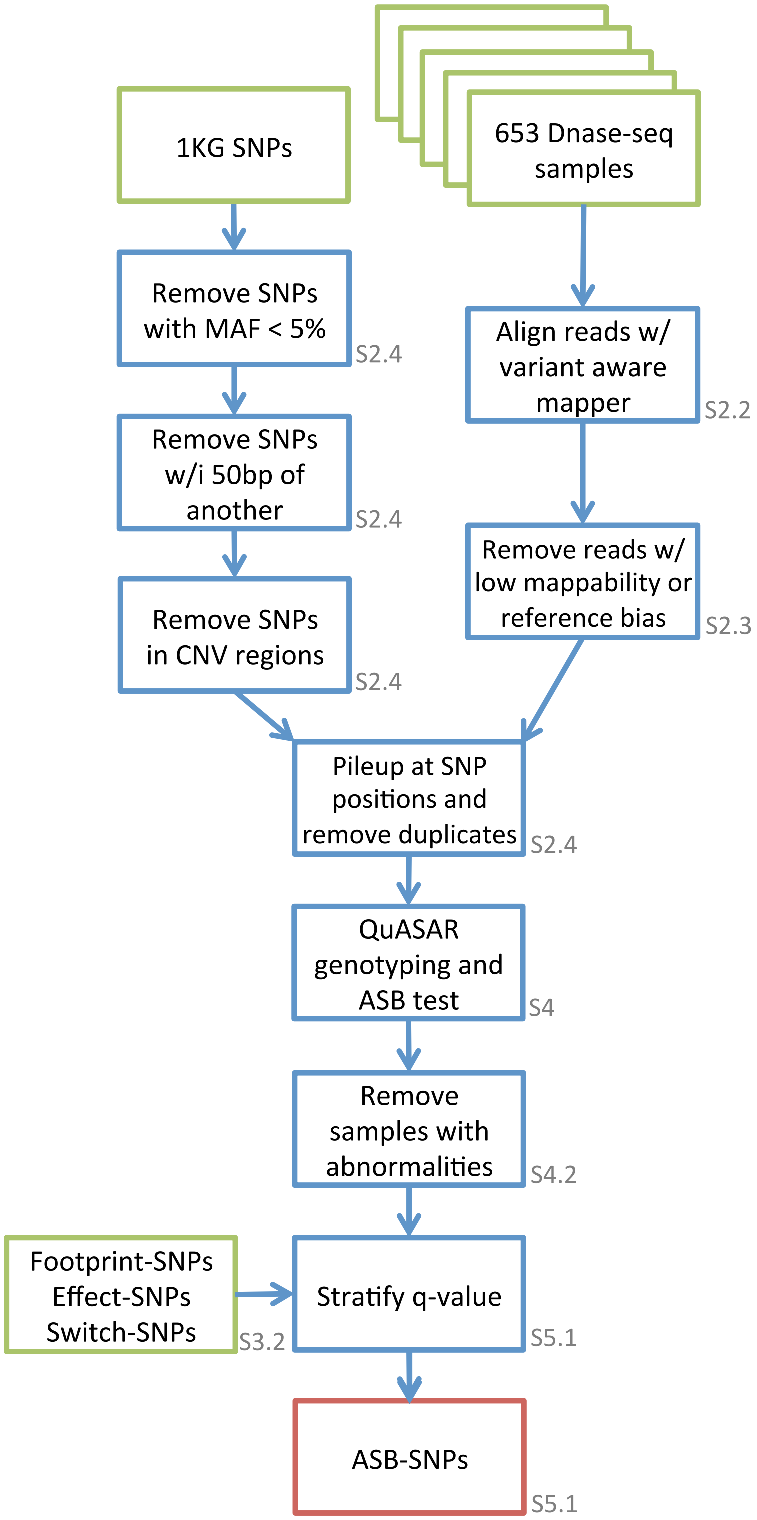
1KG SNPs
653 Dnase-seq samples
Remove SNPs with MAF < 5%
S2.4
Align reads w/ variant aware mapper
S2.2
Remove SNPs w/i 50bp of another
S2.4
Remove reads w/ low mappability or reference bias
S2.3
Remove SNPs in CNV regions
S2.4
Pileup at SNP positions and remove duplicates
S2.4
QuASAR genotyping and ASB test
S4
Remove samples with abnormalities
S4.2
Footprint-SNPs Effect-SNPs Switch-SNPs
S3.2
Stratify q-value
S5.1
ASB-SNPs
S5.1

Figure S2: **Flowchart detailing ASB analysis pipeline.** Numbers next to boxes refer to the corresponding section in the Supplement.

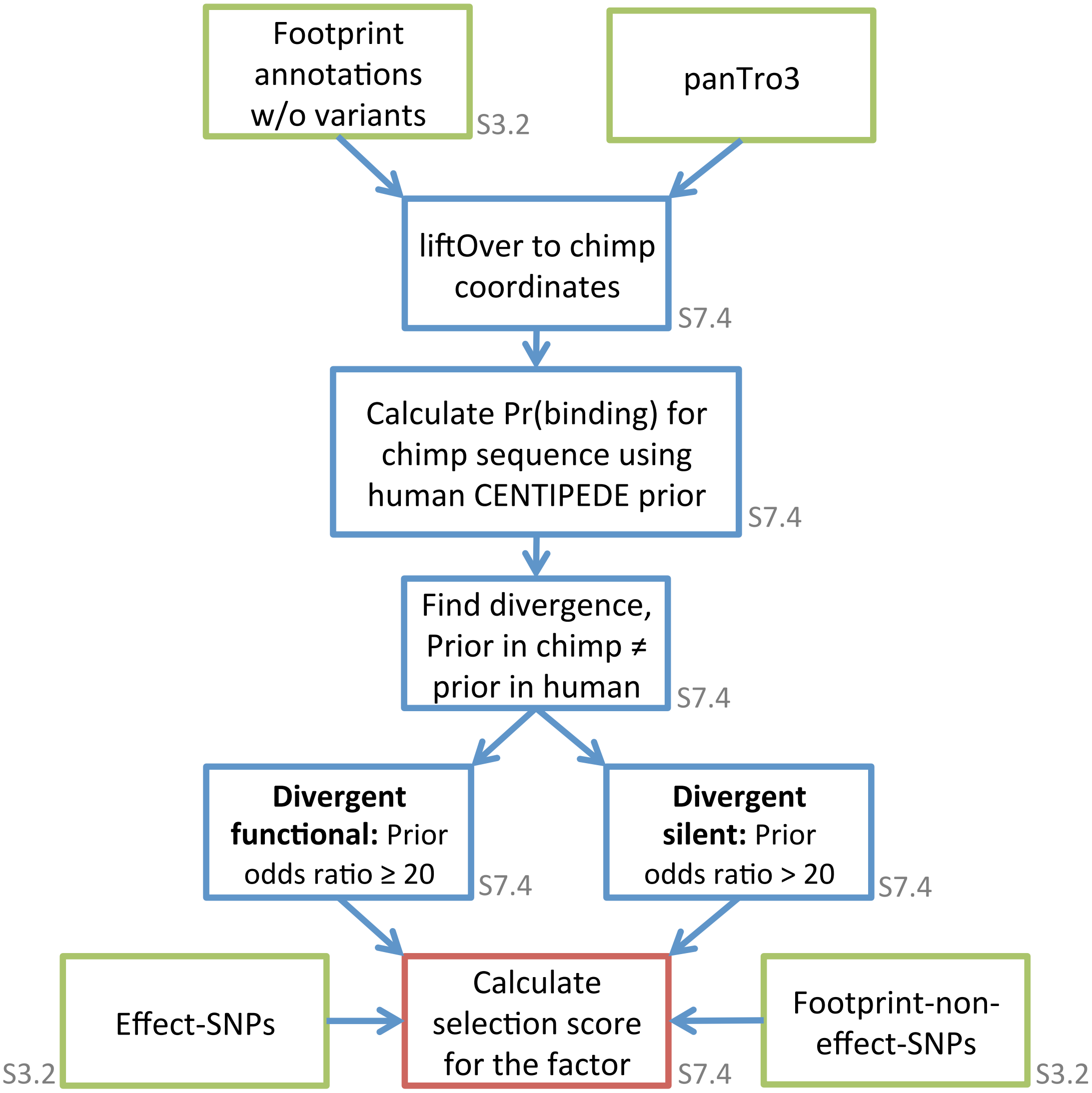
Footprint annotations w/o variants
S3.2
panTro3
liftOver to chimp coordinates
S7.4
Calculate Pr(binding) for chimp sequence using human CENTIPEDE prior
S7.4
Find divergence, Prior in chimp ≠ prior in human
S7.4
Divergent functional: Prior odds ratio ≥ 20
S7.4
Divergent silent: Prior odds ratio > 20
S7.4
Effect-SNPs
S3.2
Calculate selection score for the factor
S7.4
Footprint-non-effect-SNPs
S3.2

Figure S3: **Flowchart detailing analysis pipeline for identifying selection across TFBS.** "Prior Odds Ratio $>$ 20" is the same criteria as the one used to define effect-SNPs. Numbers next to boxes refer to the corresponding section in the Supplement.

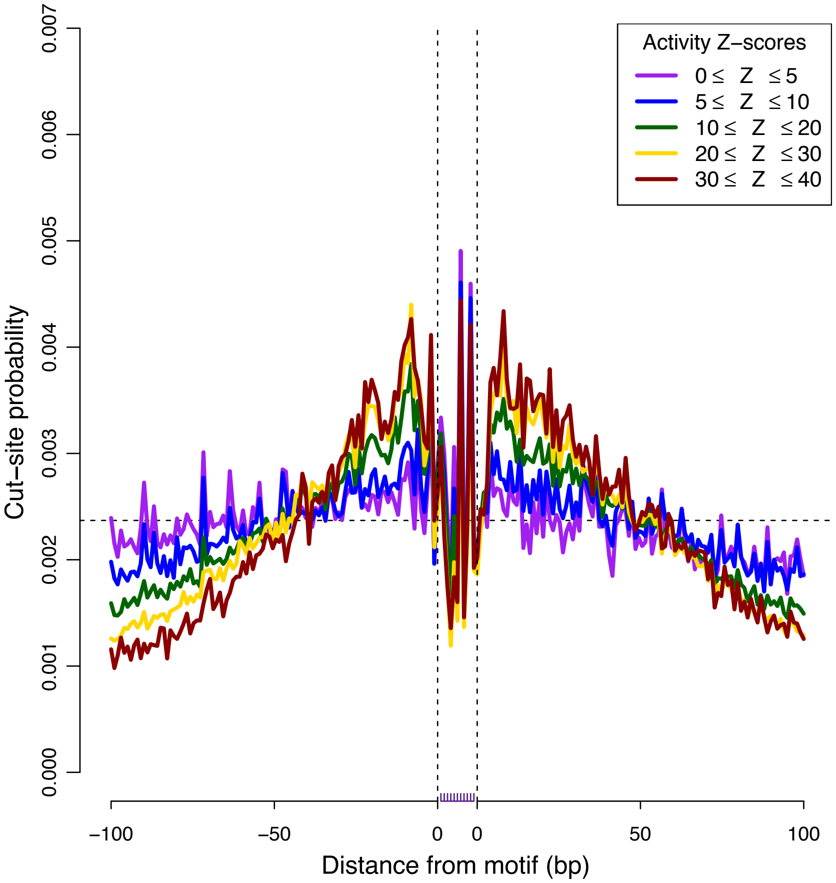

Figure S4: **Binding profiles of AP-1 motif M00172.** Footprint profiles are aggregated across all binding sites in all 653 samples, and stratified by Z-score (color). The higher the Z-score, the more likely a factor is bound as predicted by the CENTIPEDE model.

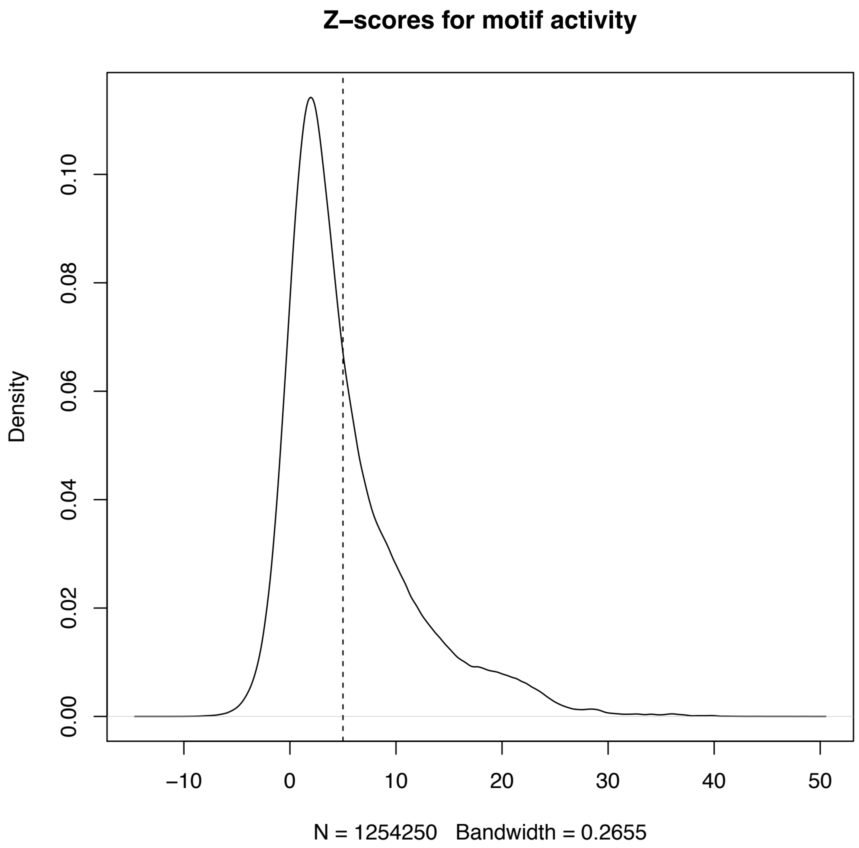

Figure S5: **Distribution of Z-scores across samples and motifs.** Shown is the full distribution of Z-scores (calculated with eq. 2) across every sample-motif pair. The dotted vertical line at Z = 5 shows the selected threshold for factor activity.

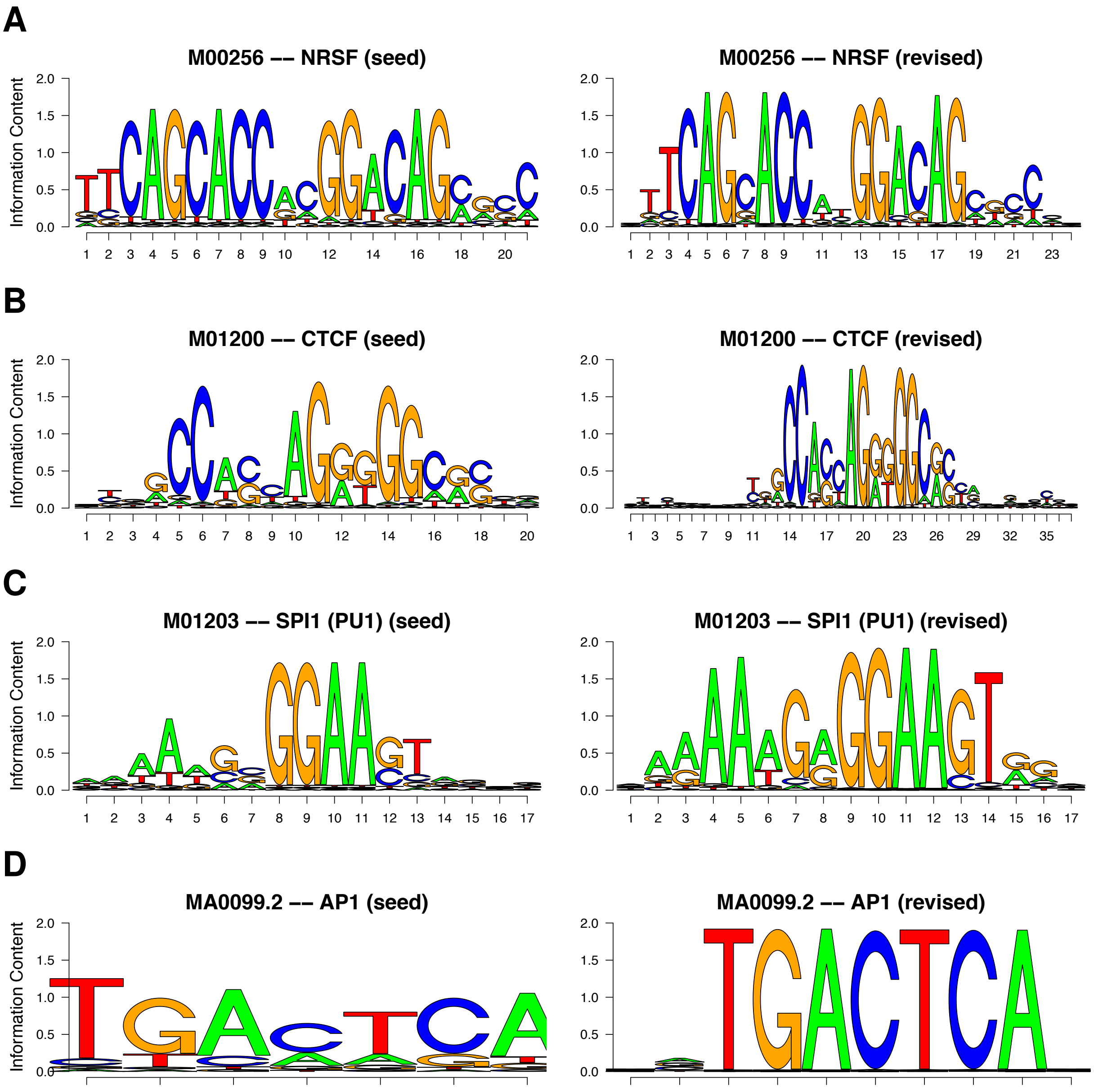

Figure S6: **Comparison between seed and revised sequence model.** For each factor motif, shown is the original seed sequence model (left) and the revised model (right). x-axis: position within motif, y-axis: information content. (A) NRSF (B) CTCF (C) PU.1 (D) AP-1

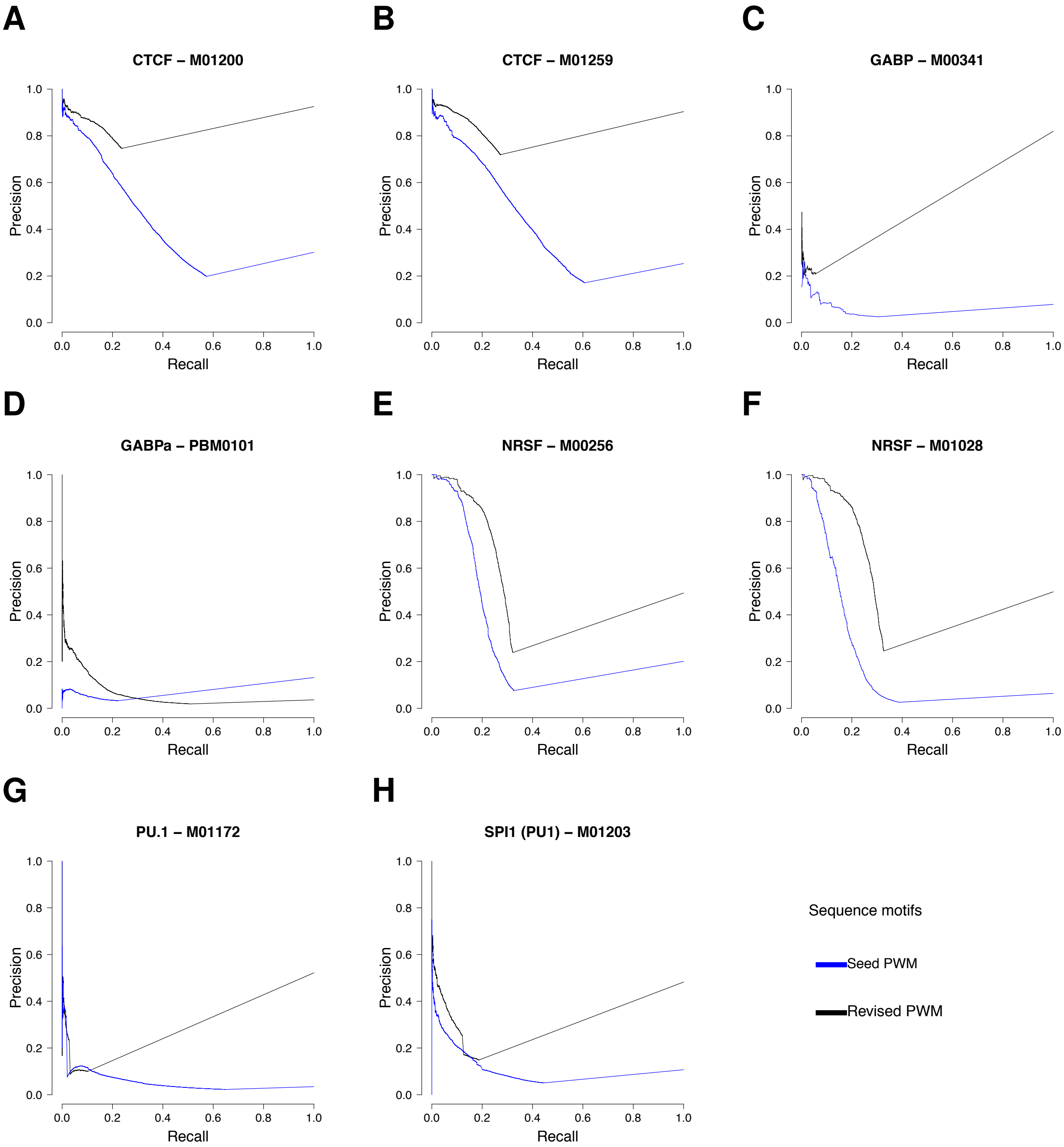

Figure S7: **Precision-recall curves for seed (blue) and revised (black) sequence models.** For each TF binding motif, CENTIPEDE-predicted footprints in GM12878 cells were compared using ENCODE ChIP-seq data as a gold standard. (A & B) CTCF (C & D) GABP (E & F) NRSF (G & H) PU.1

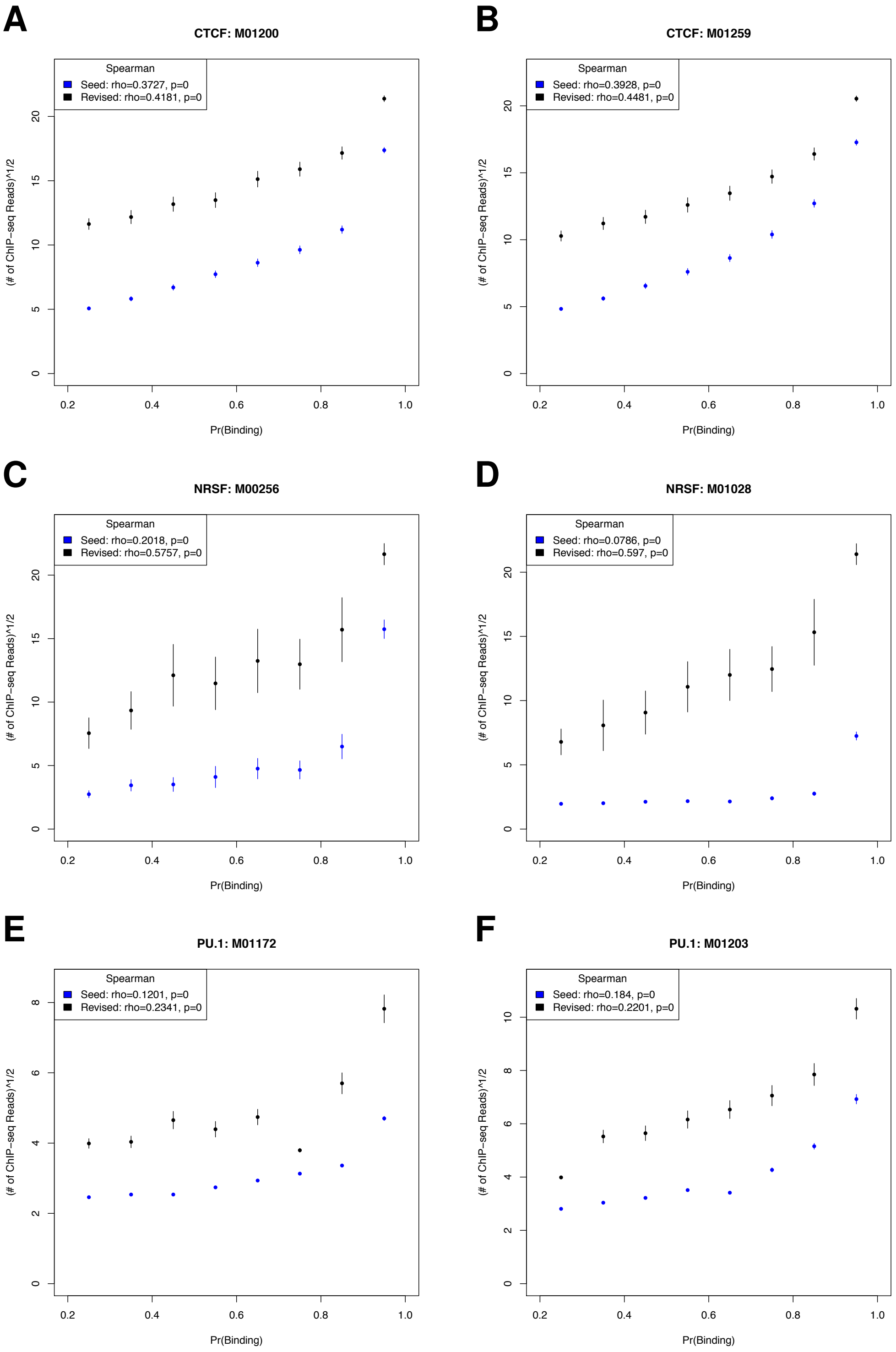

Figure S8: **Comparison of prior Pr(binding) derived from PWM scores to ChIP-seq read data across all motif matches using seed (blue) and revised (black) sequence models.** Due to thresholds on the match score (see Section S3.2), few models have data Pr(binding) < 0.2. For ease of display data is binned in 10% increments. Points represent the average number of ChIP-seq reads within that bin and vertical lines represent the 95% confidence interval. Spearman correlation (legend) is calculated using the full data set without binning. (A & B) CTCF, (C & D) NRSF, (E & F) PU.1

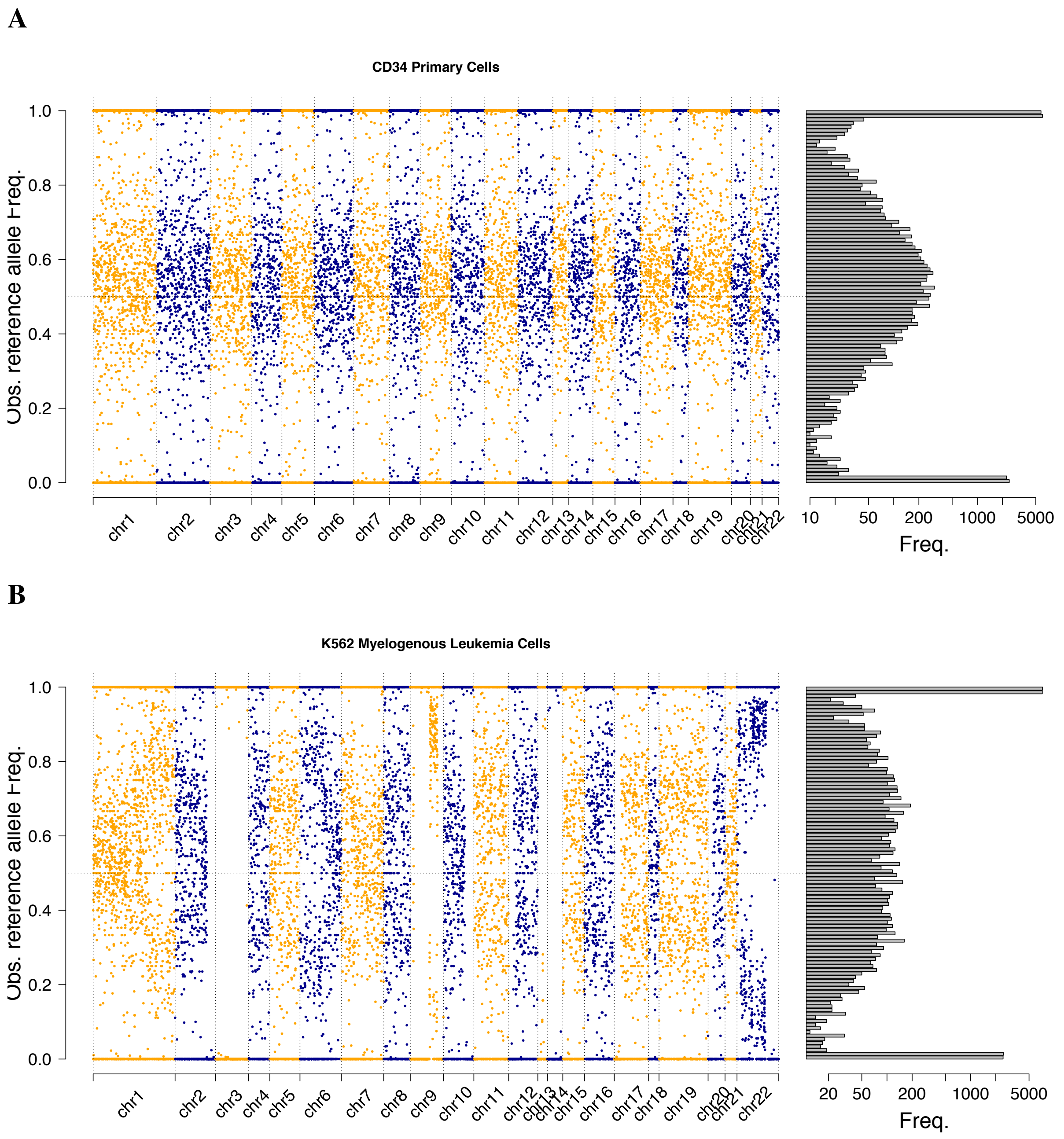

Figure S9: **Reference allele ratio ρ at 1KG variants.** (A) Plot showing ρ allele ratios for SNPs interrogated for CD34 primary cells (used for ASB analysis). Three peaks on the histogram (right) correspond to homozygous reference (top), heterozygous (middle), and homozygous alternate (bottom) SNPs. (B) Plot showing ρ allele ratios for SNPs interrogated for the cancer line K562 (discarded for ASB analysis). Signatures of chromosomal abnormalities are evident from the scatterplot, such as copy number variation and loss of heterozygosity.

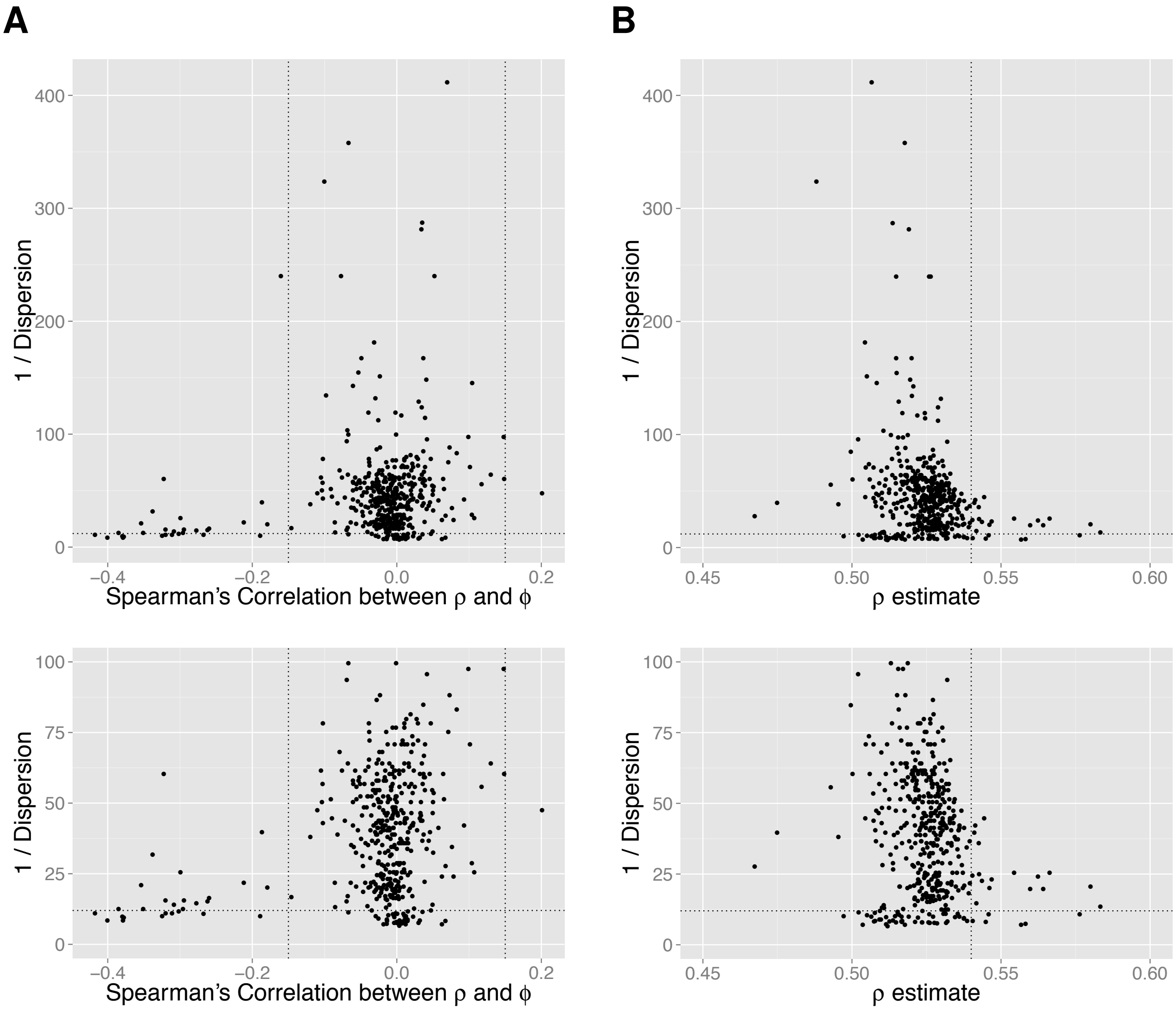

Figure S10: **Distribution of values used for post-ASB analysis filter criteria.** On all four panels y-axis represents the parameter M that is reciprocally related to the dispersion of rho in the QuASAR model. Dotted lines represent values used to filter samples. (A) Dispersion and correlation between ρ and ϕ (B) Dispersion and ρ estimation. Bottom plots show zoomed view of samples with $M < 100$.

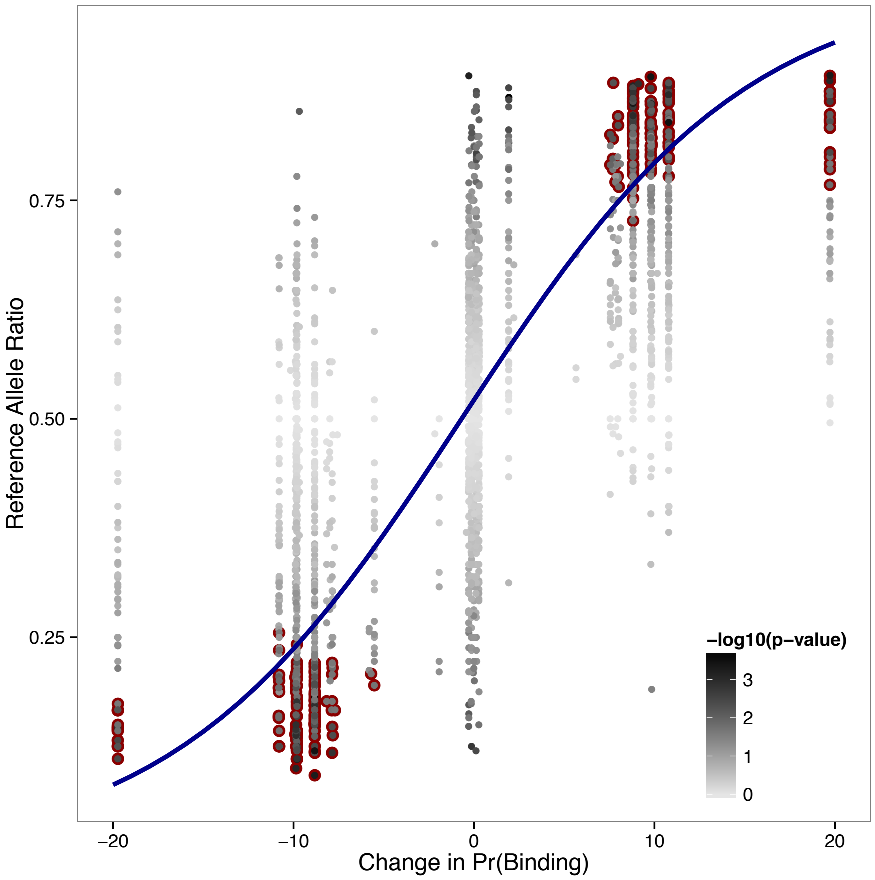

Figure S11: **Correlation between CENTIPEDE predictions and observed ASB.** SNPs identified in both the CENTIPEDE and ASB analysis are shown, shaded by p-value of allelic imbalance from QuASAR. Points circled in red display significant ASB at 20% FDR. The blue line is a logistic curve fit using points with a $p < 0.1$.

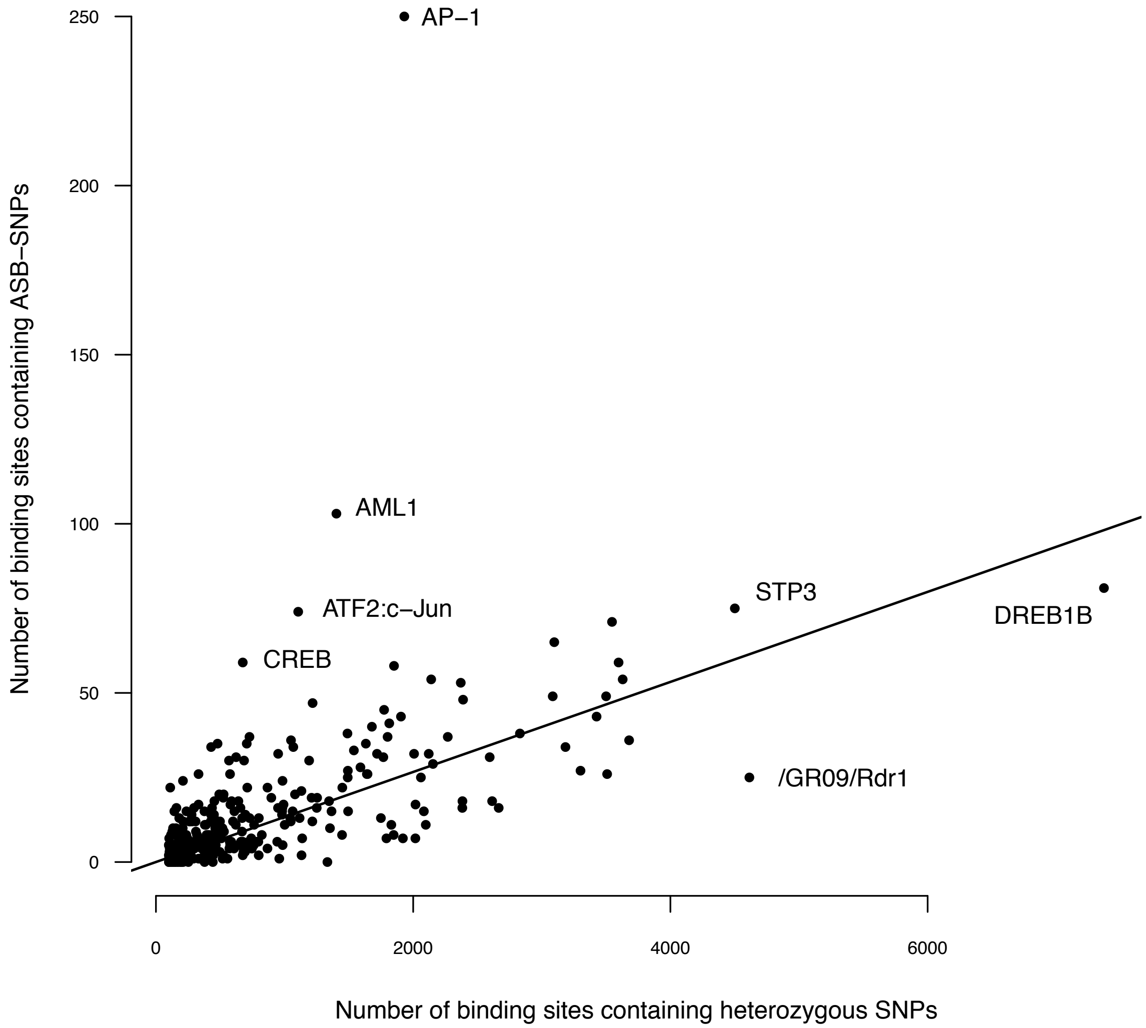

Figure S12: **Distribution of ASB enrichment ratios.** For all motifs with >100 heterozygous SNPs, an ASB enrichment ratio was calculated as # ASB-SNPs (20% FDR) / # heterozygous SNPs across all binding sites genome-wide. The black line shows the average ratio across all motifs. Several factors whose binding sites are highly enriched or depleted for ASB-SNPs are labeled.

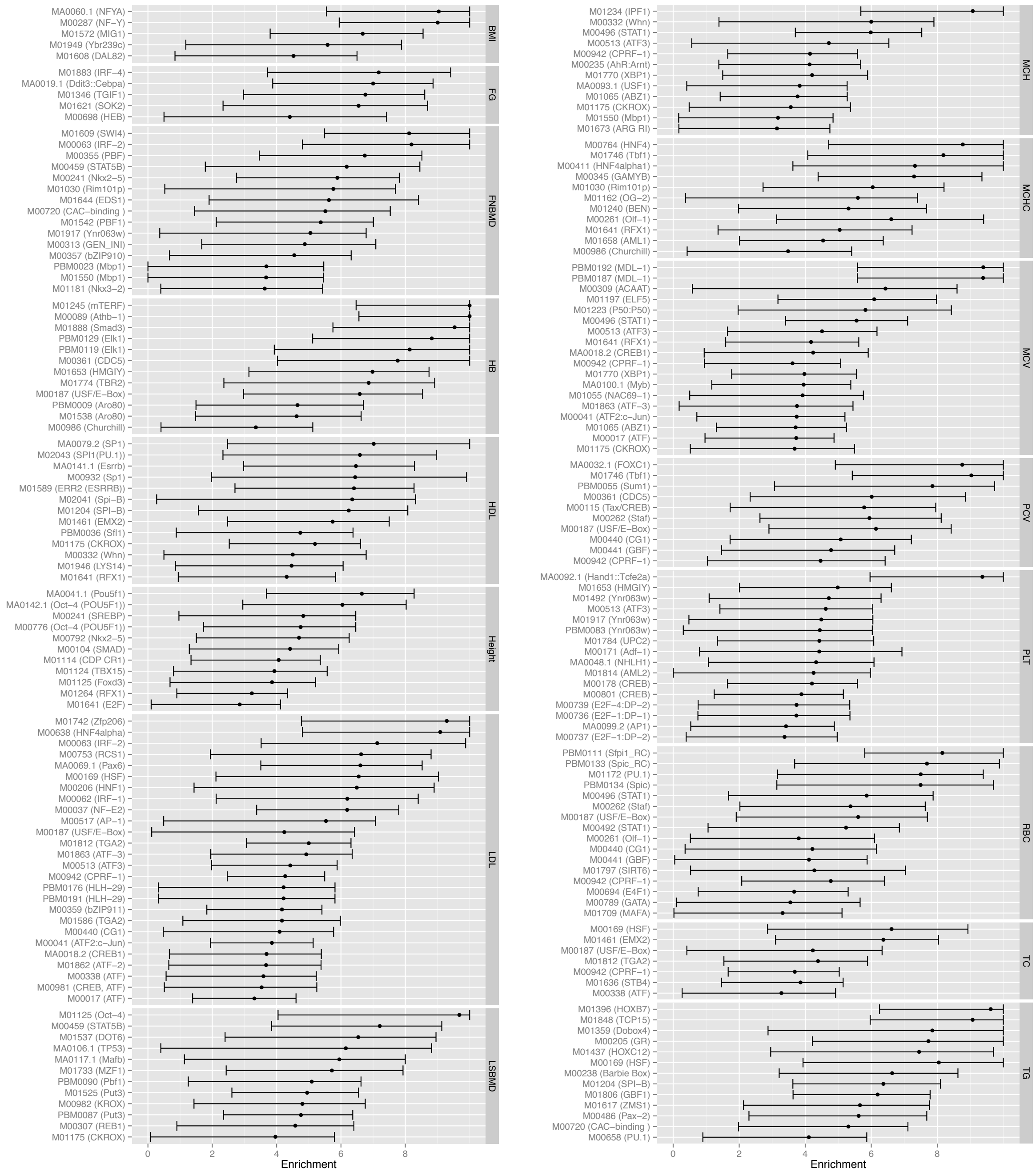

Figure S13: **Enrichment of factors for association with selected traits**. Shown are the log_2(enrichment) values with 95% confidence intervals for each factor whose binding sites are enriched for SNPs associated with the traits in Table S14. x-axis is truncated at 10 for ease of display.

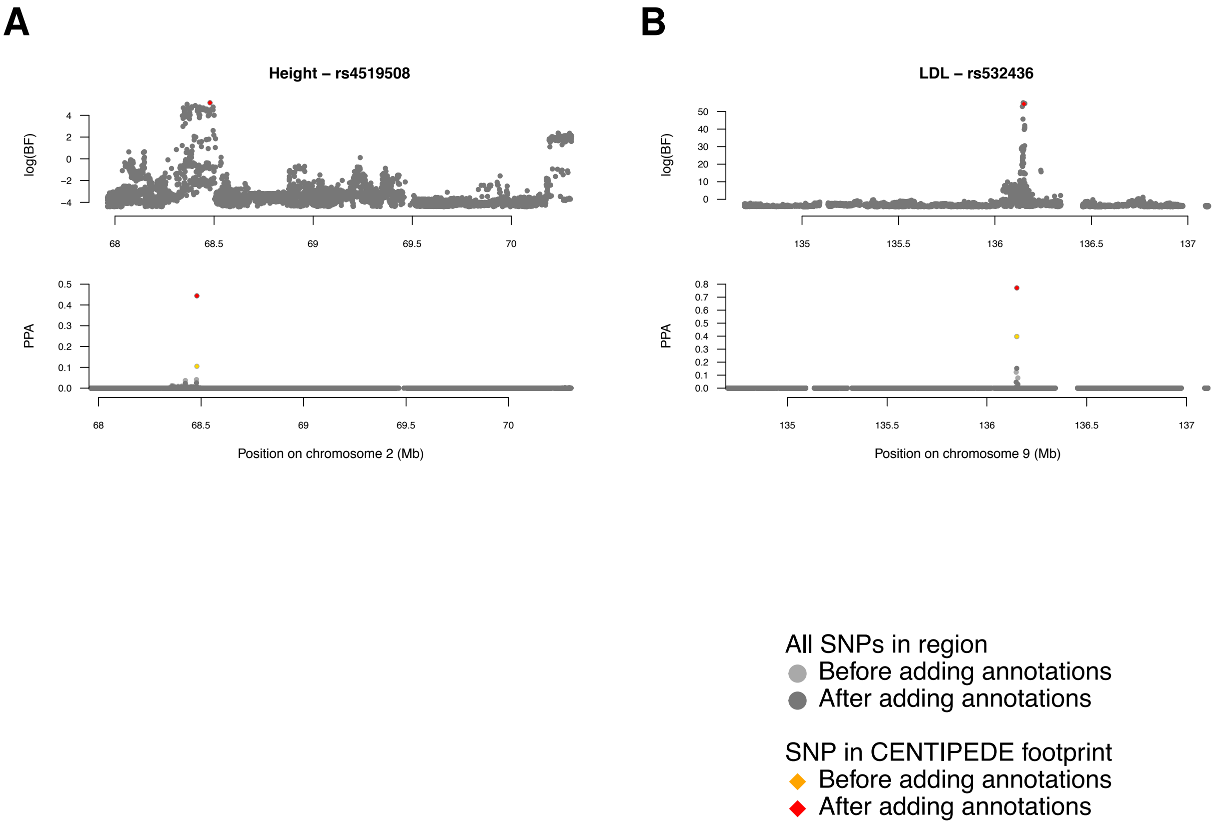

Figure S14: **Association plots identifying SNPs in footprints.** Log Bayes factor (top) and posterior probabilities (bottom) of association to the indicated trait for all genetic variants in the regions containing rs4519508 and rs532436.

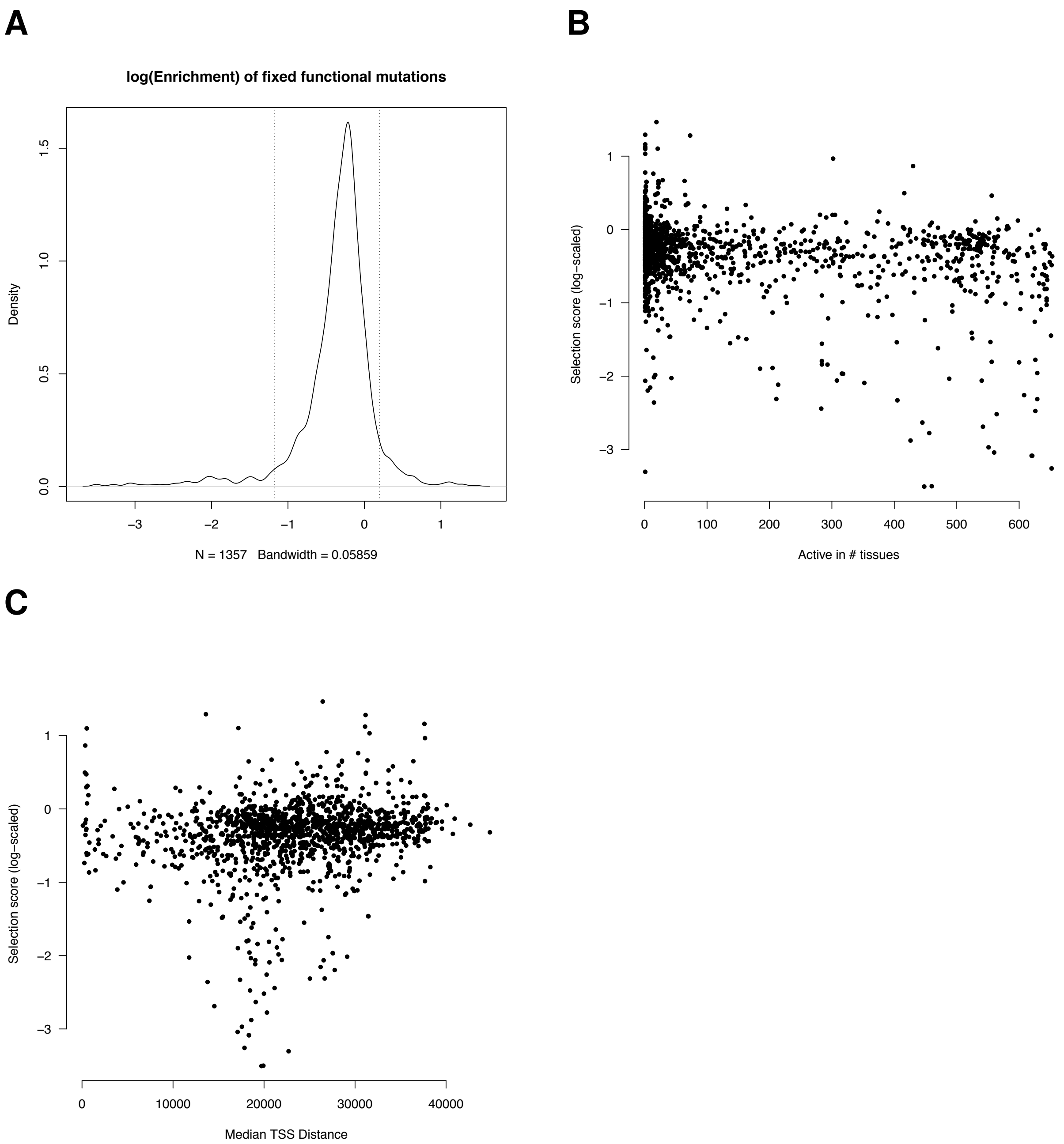

Figure S15: **Identifying selection signals in TF binding sites**. (A) Density plot showing the distribution of selection scores from the modified MK test. (B) Comparison of selection scores to the number of tissues each factor is predicted to be active in. (C) Comparison of selection scores to the median distance to the TSS across all sites for a given factor.