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.

Abstract

A detailed mechanistic understanding of the direct functional consequences of DNA variation on gene regulatory mechanism is critical for a complete understanding of complex trait genetics and evolution. Here, we present a novel approach that integrates sequence information and DNase I footprinting data to predict the impact of a sequence change on transcription factor binding. Applying this approach to 653 DNase-seq samples, we identified 3,831,862 regulatory variants predicted to affect active regulatory elements for a panel of 1,372 transcription factor motifs. Using QuASAR, we validated the non-coding variants predicted to be functional by examining allele-specific binding (ASB). Combining the predictive model and the ASB signal, we identified 3,217 binding variants within footprints that are significantly imbalanced (20% FDR). Even though most variants in DNase I hypersensitive regions may not be functional, we estimate that 56% of our annotated functional variants show actual evidence of ASB. To assess the effect these variants may have on complex phenotypes, we examined their association with complex traits using GWAS and observed that ASB-SNPs are enriched 1.22-fold for complex traits variants. Furthermore, we show that integrating footprint annotations into GWAS meta-study results improves identification of likely causal SNPs and provides a putative mechanism by which the phenotype is affected.

1 Introduction

Genome-wide association studies (GWAS) have successfully identified large numbers of genetic variants associated with disease and normal trait variation (Mailman et al. 2007); yet a formidable challenge remains in determining the specific functional relevance of human DNA sequence variants. First, GWAS only identify large regions of association and in general, cannot directly pinpoint the true causative variants. Second, even after fine mapping, most of these genetic variants are located in non-coding regions that make it more difficult to infer the mechanism linking individual genetic variants with the disease trait. Third, we generally do not know in which cell-types/tissues, sequence variants identified by GWAS may have a functional impact.

Recent technological advances in molecular biology combined with high-throughput sequencing have made possible the analysis of many different types of molecular function across the entire genome. Functional genomics data collected by ENCODE (The ENCODE Project Consortium 2012), Roadmap Epigenomics (Bernstein et al. 2010), and other groups (e.g., Visel et al. 2009) have provided a great deal of information on the tissue-specific regulatory regions of the human genome. At the same time, several computational methods have been developed that can exploit the correlation structure of these types of data to identify tissue-specific regulatory regions. For example, different types of histone modifications have been integrated using hidden Markov models (HMMs) (Ernst and Kellis 2010), and dynamic Bayesian networks (DBNs) (Hoffman et al. 2013). Open chromatin and sequence motifs have been integrated in different types of mixture models (CENTIPEDE (Pique-Regi et al. 2011), PIQ (Sherwood et al. 2014) and others (Boyle et al. 2012; Neph et al. 2012)). A systematic comparison of CENTIPEDE predictions and ChIP-seq data from ENCODE on LCLs and K562 cells demonstrated a remarkable agreement in classifying motif instances as bound or unbound, and CENTIPEDE was used to create one of the most extensive map of transcription factor (TF) binding in LCLs. All these computationally and experimentally derived annotations for regulatory regions have been used to functionally characterize GWAS hits (Trynka et al. 2013; Schaub et al. 2012; Boyle et al. 2012; Dunham et al. 2012; Neph et al. 2012).

However, a simple positional overlap between a genetic variant and regulatory regions is not sufficient evidence to prove that the allelic status has a functional impact on binding. One of the major current limitations is that motif models for TF binding are generally not sufficiently well calibrated to make a prediction of the sequence impact on binding. As an additional challenge, the sequence elements (motifs) and the epigenetic state (e.g., open chromatin, DNA methylation state, histone marks) constitute a combinatorial grammar that is still poorly understood in a global genomic scale, making it difficult to understand the mechanistic link between sequence variation and complex traits. Quantitative trait loci (QTLs) for molecular cellular phenotypes (as defined in Dermitzakis (2012)), such as gene expression (eQTL) (Stranger et al. 2007), TF binding (Kasowski et al. 2010), and DNase I sensitivity (dsQTL) (Degner et al. 2012) have been crucial for providing stronger evidence and a better understanding of how genetic variants in regulatory sequences can affect gene regulation (Gibbs et al. 2010; Melzer et al. 2008; Gieger

et al. 2008). eQTL studies usually require a large collection of individuals which has limited the number of tissues and cellular conditions that have been explored to date (e.g., Dimas et al. 2009; GTEx Consortium 2013). Moreover, eQTL analysis cannot directly determine the true causative variant and the underlying molecular mechanism of regulation being affected. Catalogs such as the RegulomeDB (Boyle et al. 2012) and HaploReg (Ward and Kellis 2012) combine functional annotations, including those from eQTL studies, to build support for variants that may be truly functional. However, these annotations are derived by overlapping multiple studies with *ad hoc* methods and it is in general very challenging to combine the different sources of information with the appropriate weights, in the absence of a suitable joint model.

Allele-specific analysis of sequencing reads overlapping heterozygous loci represents an increasingly popular approach for exploring genetic effects on gene expression, chromatin state and TF binding using even one single individual (Pastinen 2010; McDaniell et al. 2010; Reddy et al. 2012; Cowper-Sallari et al. 2012; McVicker et al. 2013; Kasowski et al. 2013; Kilpinen et al. 2013). An advantage of allele specific approaches is that they effectively control for any trans-acting modifiers of gene expression, such as the genotype at other loci or environmental differences between samples. While it is possible to survey allele-specific binding (ASB) for a wide panel of TFs using DNase-seq footprinting (Degner et al. 2012) in a single experiment, few studies have characterized ASB, and only for a limited number of TFs in a limited number of cell-types or tissues (McDaniell et al. 2010; Reddy et al. 2012; Kilpinen et al. 2013). This is likely because ASB analysis requires considerable sequencing effort to infer ASB with high confidence. Alternatively, here we have developed an approach that combines computational predictions and empirical measurements for ASB to identify allele-specific effects on samples of low to intermediate sequencing depth data.

Specifically, we have extended the CENTIPEDE approach to generate a catalog of regulatory sites and binding variants for more than 600 experimental samples from the ENCODE and Roadmap Epigenomics projects using recalibrated sequence motif models for more than 800 TFs. This is the most comprehensive catalog of regulatory variation annotated through integration of sequence and functional data to date. We then incorporated ASB information to provide additional empirical evidence and to validate the accuracy of the computational predictions. We also examined genomic properties of the annotations, identifying characteristics that separate variants that disrupt binding from those that do not, and demonstrated the action of natural selection on TF binding sites. Finally, we used our catalog to annotate and interpret variants associated with complex traits. Our results show that this strategy provides a general framework for the identification of regulatory variants and the determination of their functional role in complex traits.

2 Results

2.1 Identification of regulatory sequences and factor activity

The CENTIPEDE approach allows one to predict TF activity from integrating sequence motifs together with functional genomics data, and gain the most information from high-resolution data such as DNase-seq or ATAC-seq (Buenrostro et al. 2013). In the original CENTIPEDE approach, the sequence models are pre-determined; e.g, k-mers or previously defined position weight matrix (PWM) models. Here we have extended CENTIPEDE by iteratively re-adjusting the sequence model for TF binding (Fig. 1A).

After scanning the genome for motif matches (using 1949 seed motifs), we extracted DNase-seq data at these sites using 653 samples publicly available from the ENCODE and Roadmap Epigenomics projects. The motifs and samples used are summarized in Tables S1 and S2. Using these samples, we applied the CENTIPEDE model to survey TF activity at the top-scoring motif matches (see Supplemental Methods) for each 1,272,697 tissue-TF pair. From this, we determined 1,891 TF motifs that have a footprint in at least one tissue. We then recalibrated the sequence models using the sequences of the top-scoring matches. Additionally, we included sequences with a low score in the original motif if a strong footprint is observed in the human data and the orthologous sequences in chimpanzee and rhesus macaque species have high PWM scores, as those sequences likely share an evolutionary history (Supplemental Methods). This novel strategy has the advantage of retaining the TF-motif identity better than considering a random sample of low scoring instances of the original motif in the genome.

In general, the improved sequence models have higher specificity than the original models, and have a higher overall information content (Fig. S2). Using this procedure, the probabilities of certain bases are readjusted, but the core part of the motif and its consensus sequence is largely maintained. To evaluate whether the updated sequence models derived from DNase-seq data are better at predicting TF binding than the original seed motifs, we compared to ChIP-seq data available for a small set of TFs from the ENCODE project (as these data are generated in independent experimental assays that should be highly TF-specific). Using precision recall operating characteristic (P-ROC) curve analysis, we determined that 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 (Fig. S3). This implies that the recalibrated motifs have a higher TF-specificity than the original motif models.

Using these newly updated sequence models we scanned the human genome for all possible matches both to the reference and to alternate alleles from genetic variants catalogued in the 1000 Genomes (1KG) Project (The 1000 Genomes Project Consortium 2012) and used the CENTIPEDE algorithm to assess the probability that each motif instance was bound by a TF. After scanning, we discarded sequence models that were still not well calibrated for a given sample (indicated by a lack of correlation between the PWM scores and the observed DHS peaks, see Supplemental Methods). Across all 653 tissues, we identified a total of 6,993,953 non-overlapping footprints corresponding to 1,372 active motifs and spanning 4.15% of the

genome. Each individual sample contained, on average, 280,000 non-overlapping footprints for 600 motifs and spanning 0.162% of the genome, indicating that footprints are highly tissue specific.

Considering all 1KG variants at any allele frequency (even singletons), we found 5,810,227 unique genetic variants in active footprints (footprint-SNPs), 3,831,862 of which are predicted to alter the prior odds of binding $\geq$20-fold (effect-SNPs) based on the logistic sequence model (Fig. 1A-3) hyperprior in the CENTIPEDE model. In some cases, effect-SNPs increase (or decrease) the binding affinity of a sequence with an already high (or low) affinity, and for these cases the functional impact is likely to be small. However, 264,965 of these SNPs are predicted to have a larger effect on the binding of a factor (switch-SNPs), switching from a high prior probability of binding ($>$0.5) to a low probability ($<$0.5).

2.2 Analysis of allele-specific binding using DNase-seq data

We next sought to validate the SNP functional predictions by examining ASB within DNase I hypersensitivity (DHS) regions. Generally, high-depth sequencing data from functional genomics experiments is needed to perform ASB analysis. However, most of the ENCODE and Roadmap Epigenomics DNase-seq samples were collected with the intention of mapping regulatory regions, and sequenced at a coverage that is generally insufficient for allele-specific analysis (Table S1). Additionally, ASB analysis requires prior identification of heterozygous SNPs, and for most of these samples, genotypes are not available. Despite these challenges, when SNPs have sufficient coverage and mapping biases are properly addressed, both ASB signal and genotype can be inferred from the functional genomics sequencing data itself (see Supplemental Methods). We used our recently developed approach, Quantitative Allele-Specific Analysis of Reads (QuASAR) (Harvey et al. 2014) to perform joint genotyping and ASB analysis using a filtered set of 1KG variants (Supplemental Methods). Briefly, for each sample, QuASAR first uses allele counts from all unambiguous reads overlapping a genetic variant to detect if the sample is heterozygous. QuASAR then uses the read counts to determine allelic imbalance, taking into account base calling error and genotype uncertainty.

Parameters of the QuASAR model also allow us to detect tissues with chromosomal abnormalities or samples from pooled individuals (Supplemental Methods), which we excluded from ASB analysis (Table S4; Fig. S4 and S5). Across the remaining 316 samples, we identified 204,757 heterozygous SNPs in DHS sites (DHS-SNPs) with coverage $>$10x. Across all motif instances, our predictions are overall highly concordant with the direction of ASB; 88% of the motif sequence models show positive correlation between the predicted and observed ASB (see Supplemental Methods). Based on our classification, among DHS-SNPs: 55,044 are footprint-SNPs, 26,773 are effect-SNPs, and 5,991 are switch-SNPs. Each of these nested SNP categories have marked differences in the distribution of p-values (Fig. 1C) for the QuASAR test of ASB. Compared to what would be expected from the null uniform distribution, effect-SNPs and switch-SNPs have 8x and 14x times more SNPs with $p < 0.001$ respectively, showing that our functional annotations can predict ASB. Furthermore, these enrichments for lower p-

values are much higher than those of DHS-SNPs and footprint-SNPs, indicating that identifying SNPs in DHS regions and/or footprints alone is not enough to predict functional effects.

In addition to the distribution of p-values, we examined the allelic ratio, $\hat{\rho}$, for SNPs in each of the categories. To see which categories, if any, had an observed allele ratio ($\hat{\rho}_{obs}$) markedly different than the expected value of 0.5 (50% of the reads coming from each allele), we calculated $\Delta\hat{\rho}$, defined as

$$\Delta\hat{\rho} = \frac{0.5 - \hat{\rho}_{obs}}{0.5} \quad (1)$$

We found that the average $\Delta\hat{\rho}$ for effect-SNPs and switch-SNPs ($\Delta\hat{\rho} = 0.119$) is significantly higher than that for DHS-SNPs and footprint-SNPs ($\Delta\hat{\rho} = 0.105$, Wilcoxon $p < 10^{-16}$, Fig. 1B). These results indicate that effect-SNPs and switch-SNPs have a higher impact on TF binding than DHS-SNPs or footprint-SNPs alone.

Our previous analyses indicate that the following three categories of SNPs, effect-SNP, footprint-SNPs that are not effect-SNPs, and the remaining DHS-SNPs, show characteristic differences in their impact on TF binding (Table 1). To improve the power of ASB detection and quantify the enrichment level of ASB signals in each category, we therefore followed the strategy of Benjamini and Bogomolov (2014) to perform multiple testing of ASB in each category separately, and use Storey's procedure (Storey 2003) to correct for multiple comparisons in each category. At an FDR threshold of 20%, we detect 3,217 unique loci displaying significant ASB (Table 1), hereafter referred to as ASB-SNPs. Several of the ASB-SNPs were significant in more than one cell-type, giving a total of 4,940 observations of ASB-SNPs. Based on the proportion of true null hypothesis ($\hat{\pi}_0$) estimated by the Storey FDR procedure, we estimate that 56% of the effect-SNPs show evidence of ASB. We should note that this estimate is conservative and can be considered a lower bound. When considering DHS-SNPs and Footprint-SNPs, $\hat{\pi}_0$ estimates (2.1 % and 3.1% respectively) indicate that most SNPs in DHS regions and even in the putative binding site do not affect binding.

2.3 Characterization of effect-SNPs

Using the extensive catalog of regulatory variants produced by CENTIPEDE across many cell-types/tissues and TFs, we next sought to identify characteristics that distinguish effect-SNPs from SNPs in footprints that do not impact binding.

Regions of the genome with demonstrated molecular function (e.g. genic regions) generally show reduced diversity (McVicker et al. 2009) and a site frequency spectrum skewed towards rare variants. This is due to negative (purifying) selection, which prevents deleterious alleles from reaching high frequencies in the population. We investigated whether a similar skew in the site frequency spectrum exists at functional non-coding variants (effect-SNPs). We observed that effect-SNPs are slightly more likely to be rare compared to those that do not affect binding (Fig. 2A), displaying a mild but significant 1.08-fold enrichment for having a MAF <1%. This observation is analogous to what has been seen among coding variation, where rare (< 0.5%) variants are 1-2 times more likely to be non-synonymous changes than synonymous (The

1000 Genomes Project Consortium 2012). Of effect-SNPs with MAF $< 1\%$, a slightly higher proportion are predicted to increase binding of a factor (54%), rather than decrease binding (46%) relative to the major allele.

We next examined the genomic location of footprint-SNPs relative to the nearest transcription start site (TSS). eQTL studies have found that variants associated with gene expression tend to occur close to TSS (Veyrieras et al. 2008; Gaffney et al. 2012; McVicker et al. 2013), with most signals occurring within 100kb of the TSS. We detect a similar trend among our annotations, with 83% of footprint-SNPs occurring within 100kb of the TSS. It is not surprising that we find a 1.12-fold depletion of effect-SNPs within 300 bases of a TSS (Fig. 2B), as this is typically considered to be the core promoter region (Cooper et al. 2006). Indeed, Degner et al. (2012) found that potential regulatory variants (variants associated with DNase I hypersensitivity, dsQTLs) that are also eQTLs are enriched within 1kb of the TSS. The observed depletion of effect-SNPs may reflect the fact that factors binding closer to the TSS may be housekeeping factors and those that recruit transcriptional machinery, and therefore changes in binding affinity are more likely detrimental to the cell.

We also investigated whether effect-SNPs occur more frequently in tissue-specific footprints compared to footprints shared among tissues. We hypothesize that an effect-SNP in a footprint that is active in many tissues is more likely to have pleiotropic effects as each tissue where the motif is active will be affected. On the contrary, we hypothesize that an effect-SNP in a tissue-specific footprint will have functional consequences in a specific tissue/organ and possibly less severe phenotypic effects. We examined the number of samples for which a SNP was identified in an active regulatory footprint and discovered a 1.18-fold enrichment for effect-SNPs in motifs active in 5 or fewer samples (Fig. 2C).

Since allele frequency can be correlated with distance to the TSS, and shared footprints may also be more common at the promoter region, we tested all these features together in a joint model to control for one variable being a confounder of another. To this end, we employed a multiple regression logistic model to predict whether a footprint-SNP is more likely to affect TF binding based on distance to the TSS, tissue-specificity and allele frequency (Methods). All three factors are significant predictors when considered together, and the direction of the effect is the same as when considered separately (Table S5).

2.4 Motif-wise characteristics of regulatory variants

Next, we examined the distribution of ASB-SNPs across the different regulatory factors. Specifically, for each factor, we calculated the number of footprints in which we detected ASB-SNPs to the number of footprints in which we detected any heterozygous SNP (ASB enrichment ratio, Fig. S7). Using the average ratio across all factors, we determined which are enriched or depleted for ASB-SNPs, assessing significance with a binomial test. At a nominal p-value cutoff of $p < 0.01$, we detect 32 motifs enriched for ASB and 56 depleted for ASB (Fig. 3A; Table S7). In cases where multiple motifs correspond to a single factor, we asked if all binding sites were similarly enriched or depleted for ASB-SNPs. We generally found this to be the case

(Table S8), most notably for the factor AP-1. Six out of seven AP-1 motifs have binding sites enriched, and three of them are the top three enriched motifs in our dataset. We see the same pattern for motifs significantly depleted of ASB-SNPs, such as CTCF and E2F. For CTCF, two of the three motifs have binding sites depleted, while for E2F, all 14 motifs have binding sites depleted for ASB-SNPs (Table S8).

To see if ASB enrichment ratios are consistent across factors with similar functions, we looked at several factors similar to AP-1. AP-1, among other roles, has been shown to regulate proinflammatory genes such as those encoding cytokines (Dendorfer et al. 1994; Mukaida et al. 1994; Bailly et al. 1996). We examined three other factors with roles in the immune response, CREB (Zhao and Brinton 2004), c/EBP (Hu et al. 2000), and NF-κB (Thomas et al. 1997). Each of these factors are over 2-fold enriched for ASB-SNPs within their binding sites, whether they exert a pro- or anti-inflammatory effect (Table S9). It is important, however, to note that each of these factors regulates a variety of non-immune related genes, and further investigation is necessary to test the impact of regulatory variation on the immune response.

We then examined the genomic characteristics at motif instances to identify features that distinguish motifs enriched for ASB versus those that are not. Similar to what we hypothesized for effect-SNPs, we expected motifs for which we observe a high number of ASB-SNPs to be active in fewer cell-types/tissues. Additionally, we expected motifs more affected by ASB-SNPs to have binding sites farther from the TSS, reflecting the tendency of effect-SNPs to be depleted near promoter regions. As expected, we found that motifs enriched for ASB were significantly farther from the TSS, having an average median distance to the TSS of 23,443 bases compared to 17,599 bases for those depleted (*t*-test $p = 2.786\times10^{-7}$; Fig. 3B). Furthermore, motifs enriched for ASB were active in significantly fewer samples, on average active in 20% vs 40% for those depleted (*t*-test $p = 5.195\times10^{-9}$; Fig. 3C), indicating that TF with motifs with a high degree of ASB effects tend to be active in fewer tissues.

An important question in evolutionary biology is the extent to which selection has acted on *cis*-regulatory elements in humans (Wray 2007). While methods are being developed to address this question (Arbiza et al. 2013; Smith et al. 2013), such methods have only been applied to a narrow subset of TFs, and, in the case of Smith et al. (2013), rely on RNA expression data to classify mutations as up- or downregulating transcription relative to the reference enhancer sequence. Given our categorization of footprint-SNPs relative to their effect on factor binding, we performed an initial survey of selection across factor binding sites using a test similar to the McDonald-Kreitman (MK) test (McDonald and Kreitman 1991) (Supplemental Methods). The MK test, developed for protein-coding regions, is a nucleotide-based test that categorizes divergent and polymorphic variants into two classes, non-synonymous (functional change), and synonymous (silent change).

For our test, we consider the whole set of TF binding sites for a given motif as a functional unit, rather than the individual nucleotides or motif instances. For each motif, we examined all SNPs in 1KG and all fixed differences from orthologous sites in chimpanzee. Using the same definition as effect-SNPs, we defined which of the divergent and polymorphic changes are functional (as opposed to silent) and compared the proportion of fixed sites that are func-

tional to the proportion of polymorphic sites that are functional. Applying our modified MK test, we obtained a selection score for factor motifs with sufficient sites in each category (Fig. 4, Table S10). At an FDR of 1%, we observe 84 factors whose binding sites are enriched for fixed functional differences (higher selection scores), suggestive of positive selection acting on those sites. Among the top scoring motifs are several factors that regulate neural and developmental processes, including POU1F1, PHOX2B, DBX2, UNCX, and YY1. Additionally, 994 factors have binding sites depleted for fixed functional divergent sites (lower selection scores), suggestive of purifying selection. Of these, the top depleted factors include ARNT, RBPJ, CREB1, POU2F2, and MYC.

Because binding sites could show signs of background selection due to physical proximity to genes, rather than selection on the site itself, we compared the selection score for each factor to its median distance from the nearest TSS. We find that there is a mild but significant correlation (Spearman ρ = 0.16, $p = 5.592 \times 10^{-9}$), where factors that bind closer to the TSS tend to have lower selection scores (Fig. S10). However, without further investigation it is not clear how much of this signal is due to background selection from genes under negative selection, and how much is due to promoter regions themselves under purifying selection. Finally, we compared the tissue-specificity for each factor to its selection score. We found that factors active in many cell-types/tissues have lower selection scores (Spearman ρ = -0.20, $p = 1.194 \times 10^{-13}$, Fig. S10), suggesting a pattern of selective constraint on binding sites of broadly active factors.

2.5 Enrichment of binding variants in GWAS data

Trying to pinpoint causal variation within GWAS-identified regions of association is a significant challenge, particularly for non-coding regions. These identified regions can be broad, and the few SNPs reaching genome-wide significance are not necessarily functional themselves. Furthermore, GWAS variants may only have a functional impact in a subset of tissues and conditions, making it difficult to uncover the molecular mechanisms connecting sequence change and phenotypic effect. Given that our annotations comprise predicted and observed functional effects across multiple cell-types/tissues, we asked if they could help interpret genomic hits reported in the GWAS catalog.

To account for the possibility that the reported SNP is not itself causal, we defined the GWAS hit regions to include SNPs in linkage disequilibrium using an r^2 cutoff of 0.8 (from 1KG Project). In the GWAS hit regions, we examined effect-SNPs and ASB-SNPs compared to footprint SNPs and we found a 1.11-fold enrichment ($p < 2.2 \times 10^{-16}$, 95% CI: 1.099 - 1.136) for effect-SNPs and a 1.22-fold enrichment (95% CI: 1.06 - 1.39, $p < 5 \times 10^{-3}$) for ASB-SNPs.

We next asked whether ASB sites may be used to improve existing functional annotations. To do so we focused on category 2 SNPs from the RegulomeDB (Boyle et al. 2012) (SNPs with multiple regulatory annotations suggesting they affect TF binding, but not yet shown to be functional). Overlapping our ASB annotations with this set of variants, we detect a 1.6-fold enrichment ($p = 6.105 \times 10^{-5}$, 95% CI: 1.27 - 1.99) for GWAS catalog SNPs (compared to category 2 SNPs alone). This further confirms the importance of identifying functional sites at

a base-pair resolution as a large number of genetic variants in regulatory sites are not functional.

To test if our footprint annotations added support for associations of SNPs in meta-analysis studies of GWAS traits, we integrated our footprint annotations into a hierarchical model (Pickrell 2014) and applied the model to GWAS for two traits, lipid levels (Global Lipids Genetics Consortium 2013) and height (Lango Allen et al. 2010). For each trait examined, we identified factors whose binding sites were enriched for associated SNPs (Fig. 5, Fig. S8). Factors enriched for association with lipid levels include the liver-specific factor HNF4A, as well as several regulators of immune function, including SPIB, CREB1, IRF1, IRF2, and NR3C1 (GR). Factors enriched for association with height include the embryonic stem cell factor POU5F1 and the developmental regulators TBX15, FOXD3, NKX2-5.

We next looked at individual regions of association to explore whether our annotations improved fine mapping. Examining SNPs with a posterior probability of association (PPA) $>$0.2, we identified 25 SNPs for lipid levels and 14 SNPs for height within footprints (Fig. S9, Table 2). One SNP associated with LDL levels, rs532436, is within a footprint for USF, an E-box motif (Fig. 5A, right). Adding our annotation increased the PPA of the SNP from 39.7% to 94.7% (Fig. S9B). We found that the A allele, associated with a 0.0785 mg/dL increase of LDL in the blood, is predicted to have a lower binding probability. Additionally, the SNP is an eQTL identified in whole blood for SLC2A6 (GLUT6), a class III glucose transport protein, and we predict transcription factor activity at this site in several blood cell-types.

One of the SNPs associated with height, rs4519508, is in a binding site for the cell-cycle regulator family E2F (Fig. 5B, right). Our annotation increased the PPA from 10.5% to 44.4%, and it is the highest associated SNP in the association block (Fig. S9A). This example shows that while some signals may not be strong enough to reach genome-wide significance, our annotations may help rescue such variants by providing additional support for their association. This SNP occurs in a broadly active factor, E2F, with a footprint at this site in 368 samples, including ten samples (8 of them fetal) for which the predicted binding affinity of E2F is greatly reduced. Additionally, we detect ASB at this SNP in lung fibroblasts, further supporting a functional role for this variant. However, as this SNP was imputed and not part of the original GWAS, we cannot directly compare its effect on binding to its effect on height.

These results show that our integrated analysis provides support for some likely mechanisms linking regulatory sequence changes to complex organismal phenotypes.

3 Discussion

We have developed an approach for assessing functional significance of non-coding genetic variants. Our strategy integrates sequence information with functional genomics data to precisely annotate TF binding, and predict the impact of single nucleotide changes on these regions. By borrowing data from ENCODE and Roadmap Epigenomics, we generated one of the most comprehensive catalogs available to date annotating regulatory regions and functional genetic variants across the genome. We found that genetic variants that impact TF binding

are depleted in the core promoter regions, tend to have low allele frequency and are enriched in tissue-specific footprints. These properties largely reflect the family-wise characteristics of motifs, which are further reflected in signals of selection. Finally, we showed how regulatory annotations improve the identification of potential causal SNPs in GWAS, and we provide examples of putative molecular mechanisms behind the association signals for height and blood lipid levels.

Thus far the most common approach to make use of functional genomics data to identify regulatory variants, assumes that each SNP in a regulatory region is equally likely to be functional. A key finding in this study is that genetic variants in regulatory active sequences, as defined by DNase I sensitivity and footprinting, are in very large proportion silent; only 2.1% of SNPs in DHS regions and 3.1% of SNPs in CENTIPEDE footprints are estimated to have ASB. This is analogous to SNPs in coding regions, where most genetic changes are synonymous and do not result in an amino acid change. The sequence model developed in this study provides a very useful filter for non-coding genetic variants that are not functional, resulting in a tissue-specific and motif-specific annotation of effect-SNPs (56.5% of which are estimated to have an impact on ASB). This is crucial information to take into account when we attempt to understand the molecular mechanism behind GWAS hits and evolutionary signals of selection. As additional functional genomics studies are performed, across larger sample sizes, tissue types and cellular conditions, it will be important to further determine the functional subset of regulatory variants within binding sites to achieve greater power in functionally annotating genetic variants associated with complex traits.

A key feature of our annotation is that it spans a large collection of tissues and transcription factor motifs. This allowed us to trace some of the evolutionary history of TF binding and identify evolutionary constraints on specific molecular functions, which may reflect selective pressures during human history. For example, we observed that immune TFs are enriched for ASB sites, which supports the hypothesis that this may be a consequence of human adaptations to pathogen exposures. On the other hand, we identified neural development TFs that may have undergone positive selection in humans. The large number of regulatory variants predicted in our study, together with previously reported eQTL signals, and the overall relevance that they have in explaining complex traits provides further support for polygenic models of complex traits in humans. By taking advantage of the factor-specific annotations in our study, we identified motifs that are enriched for regulatory variants associated with relevant GWAS traits; e.g., immune TFs in the lipids study, and developmental TFs for height. Overall, the GWAS metaanalysis and selection signals in our study support the concept that variation in binding sites has been a major target of evolutionary forces and contribute to disease risk and complex phenotypes in human populations.

Methods

Identification of active regulatory sites. We used 1,949 PWM sequence models from the TRANSFAC (Matys et al. 2006) and JASPAR (Sandelin et al. 2004) databases to scan the genome for a set of locations with the best motif matches. Additionally, we included other regions with homologous sequences, obtained by scanning the best motif matches in two other primate genomes (rhesus and chimp) and finding the orthologous region in humans. For each sequence model, we used the sequences to calculate a new model (Supplemental Methods), which we then used to scan the genome and identify all genome-wide motif matches. Scanning was done in two stages. First, we identified every match and kept those above a motif-specific threshold match score (Supplemental Methods). Next, we scanned the genome again considering only the motifs that overlapped 1KG variants. For each of these matches we calculated two PWM scores, one for each allele. Using motif locations and DNase-seq data, we trained the CENTIPEDE model for each sample/motif combination using motifs devoid of variants. We then applied the model 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.

Identification of allele-specific binding. Starting from raw sequencing reads, we used a custom mapper (Degner et al. 2012) to align the reads to the hg19 reference genome. As allele-specific analysis is extremely sensitive to mapping errors and PCR duplicates, we employed several methods to reduce these sources of potential bias (Supplemental Methods). To detect allele-specific binding, we applied QuASAR (Harvey et al, in prep.) to the processed read data. QuASAR first genotypes SNPs in the dataset using the read counts, then determines the likelihood of alleleic imbalance at each heterozygous site. To adjust for multiple testing within each sample, we used the q-value method (Storey 2003) on p-values produced by QuASAR.

Annotation of ASB with binding predictions. To determine which positions displaying ASB fall within a predicted footprint, we overlapped DHS-SNPs identified by QuASAR with CENTIPEDE footprints for each sample. We classified a SNP as having an effect on binding if the difference in the prior log odds ratio (from the logistic sequence model in CENTIPEDE) between the two alleles was $\geq$3, indicating a $\geq$20-fold change in the prior odds of TF binding. To generate a final set of annotated SNPs, we aggregated the data from each sample and motif into one table. 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. SNPs were then partitioned based on their predicted effect on binding 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 using the q-value method (Storey 2003) on p-values produced by the QuASAR model.

Regression model for binding effect. To see which features of a SNP were predictors of functional effect, we performed multiple regression analysis using a logistic model considering the dependent binary variable E_l, indicating whether the footprint-SNP, l, is also an effect-SNP.

$$\text{logit}(E_l) \sim C_l + F_l + T_l + N_l \tag{2}$$

We considered the following variables related to the probability of a footprint-SNP being an effect-SNP: the fold-change in factor affinity predicted by the sequence model (C_l); the minor allele frequency (F_l); the absolute distance to the nearest transcription start site (T_l); the number of tissues for which the motif containing the footprint-SNP was predicted to be bound (N_l). The model was fit using the GLM function in R.

Identification of selection signals on TF motifs To identify divergent TF binding sites, we used the UCSC liftOver tool on binding sites without a polymorphism to obtain orthologous regions in the chimpanzee genome. Using the PWM model, we calculated PWM scores on the chimpanzee sequences. 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 for sites that change $\geq$20-fold between species (analogous to effect-SNPs), and silent for those that do not. For the binding sites containing a polymorphism, we separated them into similar groupings where the change is functional (i.e., effect-SNP) and those where it is not. For each factor motif, we then calculated the number of binding sites belonging to each of the four categories (divergent functional, divergent silent, polymorphic functional, and polymorphic silent) and calculated a selection score similar to the McDonald-Kreitman test (Supplemental Methods).

Expansion of GWAS Catalog. We created an expanded GWAS catalog by adding SNPs in linkage disequilibrium (LD) with each GWAS hit. Using 1000 Genomes data for European populations, we mapped 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.

Integrating functional annotations with GWAS To integrate functional annotations and GWAS results, we used the fgwas command line tool (Pickrell 2014). fgwas computes association statistics genome wide using all common SNPs from European populations in the 1KG Project, splitting the genome into blocks larger than LD. Summary statistics were imputed with ImpG using Z-scores from meta-analysis data. Using an empirical Bayesian framework implemented in the fgwas software, GWAS data were then combined with functional annotations. We then compared the informativeness of these annotations from each of the 1891 motifs with CENTIPEDE predicted regulatory sites to previously used genomic annotations (same features as the baseline model, see Supplemental Methods).

Data Access

Our footprint and footprint-SNP annotations are available as a custom UCSC Genome Browser hub. Please see `http://genome.grid.wayne.edu/centisnps/` for information on how to view and download the data.

Acknowledgements

Funding to support this research was provided by NIH 1 R01GM109215-01 (RPR and FL) and AHA14SDG20450118 (FL). All the computation was performed at the Wayne State University High Performance Computing Grid, and in the NSF supported Stampede Texas Advanced Computer Center (TACC) using an XSEDE start-up account (TG-MCB130080). We would like to thank Joe Pickrell for his assistance in running fgwas and in providing the lipids and height meta-analysis datasets, Jacob Degner for reviewing an earlier version of this manuscript, and the members of the Luca/Pique group for helpful discussions.

Competing Interests

The authors declare no competing interests in this study.

References

Arbiza, L., Gronau, I., Aksoy, B. a., Hubisz, M. J., Gulko, B., Keinan, A., and Siepel, A., 2013. Genome-wide inference of natural selection on human transcription factor binding sites. *Nature genetics*, **45**:723–9.

Bailly, S., Fay, M., Israël, N., and Gougerot-Pocidalo, M. A., 1996. The transcription factor AP-1 binds to the human interleukin 1 alpha promoter. *European cytokine network*, **7**(2):125–128.

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.

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M. A., Beaudet, A. L., Ecker, J. R., *et al.*, 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology*, **28**:1045–1048.

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. a., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., *et al.*, 2012. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, **22**(9):1790–1797.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, **10**:1213–8.

Cooper, S. J., Trinklein, N. D., Anton, E. D., Nguyen, L., and Myers, R. M., 2006. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. *Genome research*, **16**:1–10.

Cowper-Sallari, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., Moore, J. H., and Lupien, M., 2012. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature genetics*, **44**(11):1191–8.

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.

Dendorfer, U., Oettgen, P., and Libermann, T. A., 1994. Multiple regulatory elements in the interleukin-6 gene mediate induction by prostaglandins, cyclic AMP, and lipopolysaccharide. *Molecular and cellular biology*, **14**:4443–4454.

Dermitzakis, E. T., 2012. Cellular genomics for complex traits. *Nature reviews. Genetics*, **13**(3):215–20.

Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., *et al.*, 2009. Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science (New York, N.Y.)*, **325**:1246–1250.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. a., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., *et al.*, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414):57–74.

Ernst, J. and Kellis, M., 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, **28**(8):817–825.

Gaffney, D. J., McVicker, G., Pai, A. a., Fondufe-Mittendorf, Y. N., Lewellen, N., Michelini, K., Widom, J., Gilad, Y., and Pritchard, J. K., 2012. Controls of nucleosome positioning in the human genome. *PLoS genetics*, **8**(11):e1003036.

Gibbs, J. R., van der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. A., Lai, S. L., Arepalli, S., Dillman, A., Rafferty, I. P., Troncoso, J., *et al.*, 2010. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet*, **6**:e1000952.

Gieger, C., Geistlinger, L., Altmaier, E., Hrabé de Angelis, M., Kronenberg, F., Meitinger, T., Mewes, H.-W., Wichmann, H.-E., Weinberger, K. M., Adamski, J., *et al.*, 2008. Genetics meets metabolomics: a genome-wide association study of metabolite profiles in human serum. *PLoS genetics*, **4**:e1000282.

Global Lipids Genetics Consortium, 2013. Discovery and refinement of loci associated with lipid levels. *Nat Genet*, **45**(11):1274–1283.

GTEx Consortium, 2013. The Genotype-Tissue Expression (GTEx) project. *Nature genetics*, **45**:580–5.

Harvey, C. T., Moyerbrailean, G. A., Davis, G. O., and Wen, X., 2014. QuASAR: Quantitative allele specific analysis of reads. *bioRxiv*, **doi**:10.1101/007492.

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

Hu, H. M., Tian, Q., Baer, M., Spooner, C. J., Williams, S. C., Johnson, P. F., and Schwartz, R. C., 2000. The C/EBP bZIP domain can mediate lipopolysaccharide induction of the proinflammatory cytokines interleukin-6 and monocyte chemoattractant protein-1. *The Journal of biological chemistry*, **275**:16373–16381.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., *et al.*, 2010. Variation in transcription factor binding among humans. *Science (New York, N.Y.)*, **328**:232–235.

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J. B., Kundaje, A., Liu, Y., Boyle, A. P., Zhang, Q. C., Zakharia, F., Spacek, D. V., *et al.*, 2013. Extensive variation in chromatin states across humans. *Science (New York, N.Y.)*, **342**(6159):750–2.

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N. I., *et al.*, 2013. Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science (New York, N.Y.)*, **342**:744–7.

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

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., *et al.*, 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*, **39**(10):1181–1186.

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.

McDaniell, R., Lee, B.-K., Song, L., Liu, Z., Boyle, A. P., Erdos, M. R., Scott, L. J., Morken, M. A., Kucera, K. S., Battenhouse, A., *et al.*, 2010. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science (New York, N.Y.)*, **328**(5975):235–9.

McDonald, J. H. and Kreitman, M., 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, **351**:652–654.

McVicker, G., Gordon, D., Davis, C., and Green, P., 2009. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genetics*, **5**.

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J. K., *et al.*, 2013. Identification of Genetic Variants That Affect Histone Modifications in Human Cells. *Science (New York, N.Y.)*, **747**.

Melzer, D., Perry, J. R., Hernandez, D., Corsi, A. M., Stevens, K., Rafferty, I., Lauretani, F., Murray, A., Gibbs, J. R., Paolisso, G., *et al.*, 2008. A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet*, **4**:e1000072.

Mukaida, N., Morita, M., Ishikawa, Y., Rice, N., Okamoto, S., Kasahara, T., and Matsushima, K., 1994. Novel mechanism of glucocorticoid-mediated gene repression. Nuclear factor-kappa B is target for glucocorticoid-mediated interleukin 8 gene repression. *The Journal of biological chemistry*, **269**:13289–13295.

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., and Borenstein, E., 2012. Resource Circuitry and Dynamics of Human Transcription Factor Regulatory Networks. *Cell*, :1–13.

Pastinen, T., 2010. Genome-wide allele-specific analysis: insights into regulatory variation. *Nature reviews. Genetics*, **11**:533–538.

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.

Pique-Regi, R., Degner, J. F., Pai, A. a., Gaffney, D. J., Gilad, Y., and Pritchard, J. K., 2011. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome research*, **21**(3):447–55.

Reddy, T. E., Gertz, J., Pauli, F., Kucera, K. S., Varley, K. E., Newberry, K. M., Marinov, G. K., Mortazavi, A., Williams, B. A., Song, L., *et al.*, 2012. Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. *Genome research*, **22**(5):860–9.

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.

Schaub, M. a., Boyle, a. P., Kundaje, a., Batzoglou, S., and Snyder, M., 2012. Linking disease associations with regulatory information in the human genome. *Genome Research*, **22**(9):1748–1759.

Sherwood, R. I., Hashimoto, T., O'Donnell, C. W., Lewis, S., Barkal, A. a., van Hoff, J. P., Karun, V., Jaakkola, T., and Gifford, D. K., 2014. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nature Biotechnology*, **32**:171–178.

Smith, J. D., McManus, K. F., and Fraser, H. B., 2013. A novel test for selection on cis-regulatory elements reveals positive and negative selection acting on mammalian transcriptional enhancers. *Molecular Biology and Evolution*, **30**:2509–2518.

Storey, J. D., 2003. The positive false discovery rate: a Bayesian interpretation and the q -value. *The Annals of Statistics*, **31**(6):2013–2035.

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C., *et al.*, 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science (New York, N.Y.)*, **315**:848–853.

The 1000 Genomes Project Consortium, 2012. An integrated map of genetic variation. *Nature*, **135**:0–9.

The ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**:57–74.

Thomas, R. S., Tymms, M. J., McKinlay, L. H., Shannon, M. F., Seth, A., and Kola, I., 1997. ETS1, NFkappaB and AP1 synergistically transactivate the human GM-CSF promoter. *Oncogene*, **14**:2845–2855.

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S., and Raychaudhuri, S., 2013. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet*, **45**(2):124–130.

Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K., 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS genetics*, **4**(10):e1000214.

Visel, A., Rubin, E. M., and Pennacchio, L. A., 2009. Genomic views of distant-acting enhancers. *Nature*, **461**(7261):199–205.

Ward, L. D. and Kellis, M., 2012. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research*, **40**:D930–4.

Wray, G. A., 2007. The evolutionary significance of cis-regulatory mutations. *Nature reviews. Genetics*, **8**:206–216.

Zhao, L. and Brinton, R. D., 2004. Suppression of proinflammatory cytokines interleukin-1beta and tumor necrosis factor-alpha in astrocytes by a V1 vasopressin receptor agonist: a cAMP response element-binding protein-dependent mechanism. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **24**:2226–2235.

Table 1: **Summary of Allele-Specific Binding SNPs.** Each row represents a category that is a subset of the category in the previous row. Each column reports the number of heterozygous SNPs, SNPs displaying significant ASB (20% FDR), and the estimated proportion of non-null hypotheses using Storey's q-value approach. In parentheses are reported the numbers for SNPs that are not present in any of the subsequent subsets and are the basis for our partitioned q-value approach to detect ASB-SNPs.

	# Het SNPs	# ASB (20% FDR)	$1 - \hat{\pi}_0$
All DHS Hets	204,757 (179,137)	0 (0)	2.1 (1.7)%
Footprint-SNPs	55,044 (42,098)	0 (0)	3.1 (0.3)%
Effect-SNPs	26,773 (26,773)	3,217 (3,217)	56.5 (56.5)%

Table 2: **SNPs associated with GWAS traits that fall in CENTIPEDE-predicted TF binding sites.** PPA, Posterior probability of association. The PPAs for association with a trait are given each SNP. Before indicates the PPA from the base model, after indicates the PPA after adding footprint annotations to the model.

SNP	Trait	Factor	SNP PPA (before)	SNP PPA (after)
rs17511102	Height	Pou5f1	0.999	1.000
rs11752007	Height	TBX15	0.740	0.955
rs7466269	Height	Nkx2-5	0.426	0.838
rs34529769	Height	SMAD	0.395	0.818
rs314263	Height	RFX1	0.267	0.807
rs9849338	Height	Foxd3	0.356	0.791
rs4973431	Height	CDP CR1	0.184	0.690
rs4725984	Height	RFX1	0.116	0.601
rs12740374	Height	RFX1	0.084	0.510
rs894344	Height	CDP CR1	0.087	0.502
rs4519508	Height	E2F	0.105	0.444
rs4073154	Height	RFX1	0.063	0.441
rs3828559	Height	SREBP	0.184	0.343
rs10171985	Height	Foxd3	0.036	0.202
rs1044973	HDL	Whn	0.368	0.755
rs676210	HDL	LYS14	0.296	0.657
rs12740374	HDL	RFX1	0.266	0.618
rs6907508	HDL	EMX2	0.219	0.465
rs1800562	LDL	CPRF-1	0.981	0.999
rs267733	LDL	ATF3	0.978	0.998
rs2479409	LDL	TGA2	0.931	0.998
rs532436	LDL	USF/E-box	0.397	0.947
rs2075375	LDL	bZIP911	0.238	0.811
rs217381	LDL	ATF-3	0.271	0.678
rs217386	LDL	/Cell09/HLH-29	0.260	0.586
rs2162011	LDL	CPRF-1	0.044	0.548
rs2954021	LDL	bZIP911	0.075	0.525
rs6920309	LDL	ATF3	0.060	0.494
rs9293637	LDL	Pax6	0.186	0.366
rs2479409	TC	TGA2	0.970	0.997
rs1800562	TC	CPRF-1	0.925	0.995
rs2235215	TC	STB4	0.676	0.965
rs532436	TC	USF/E-box	0.469	0.836
rs1556857	TC	STB4	0.067	0.479
rs553427	TC	STB4	0.032	0.303
rs9686661	TG	SPI-B	0.641	0.951
rs2270924	TG	Pax-2	0.259	0.600
rs13173241	TG	ZMS1	0.169	0.292
rs7789194	TG	GBF1	0.259	0.259

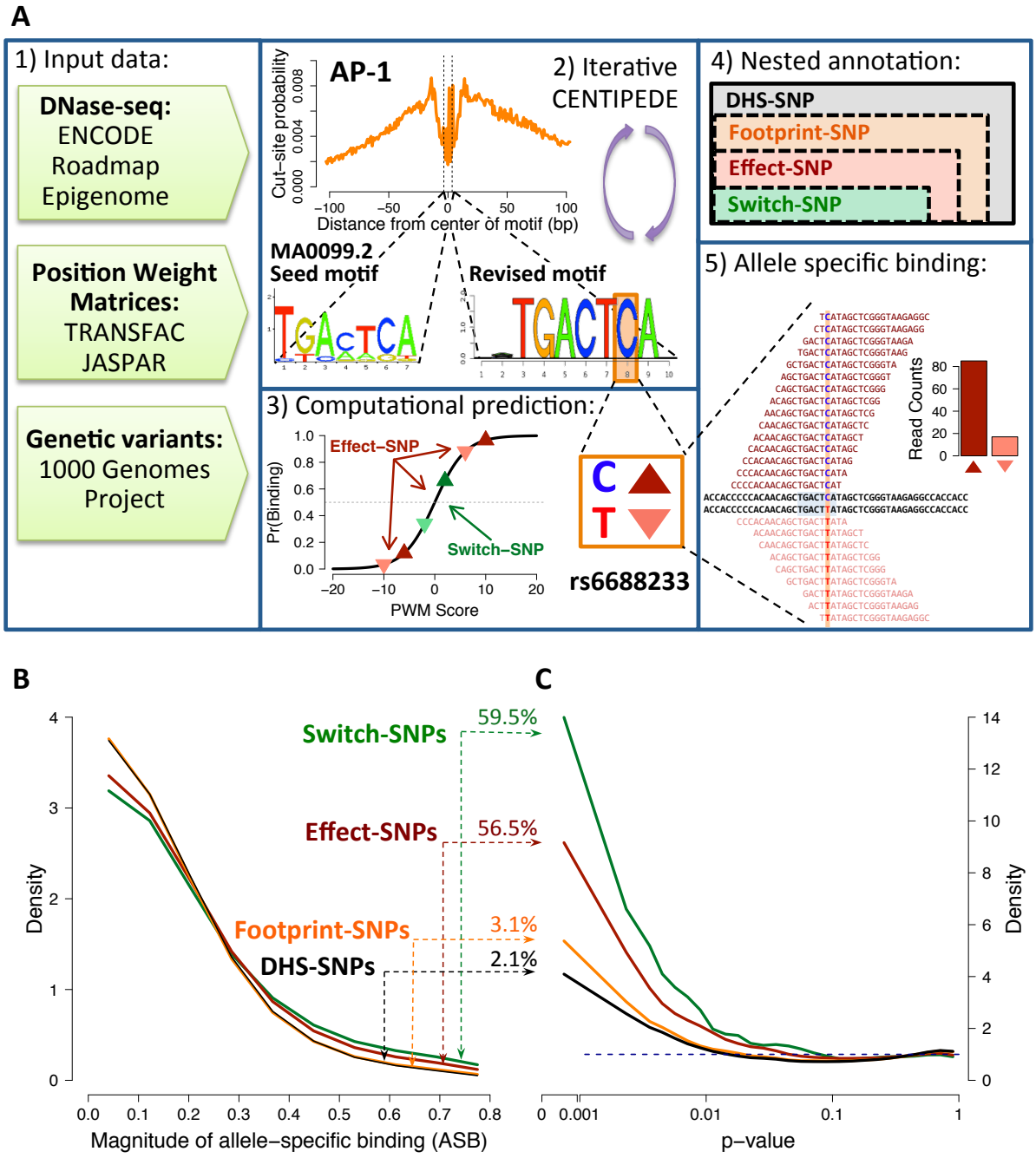

Figure 1: **Identification of allele-specific binding.** (A) A visual description of the methods. [(A1) Data sources. (A2) Iterative process of using CENTIPEDE to revise sequence models and call footprints. (A3) Computational predictions of factor binding. SNPs are categorized by the degree to which they are predicted to impact binding. (A4) The nested categories of functional annotations. (A5) Use of allele-specific analysis to detect differential binding.] (B) SNPs predicted to have effects on binding are enriched for a large magnitude of allelic imbalance [|(allele ratio - 0.5) / 0.5|]. Each line represents a density plot of the magnitude for SNPs within a functional category. Numbers on arrows represent the expected proportion of true signal $(1 - \hat{\pi}_0)$ for each category. (C) ASB p-value densities for SNP categories. The dotted blue line represents the null distribution.

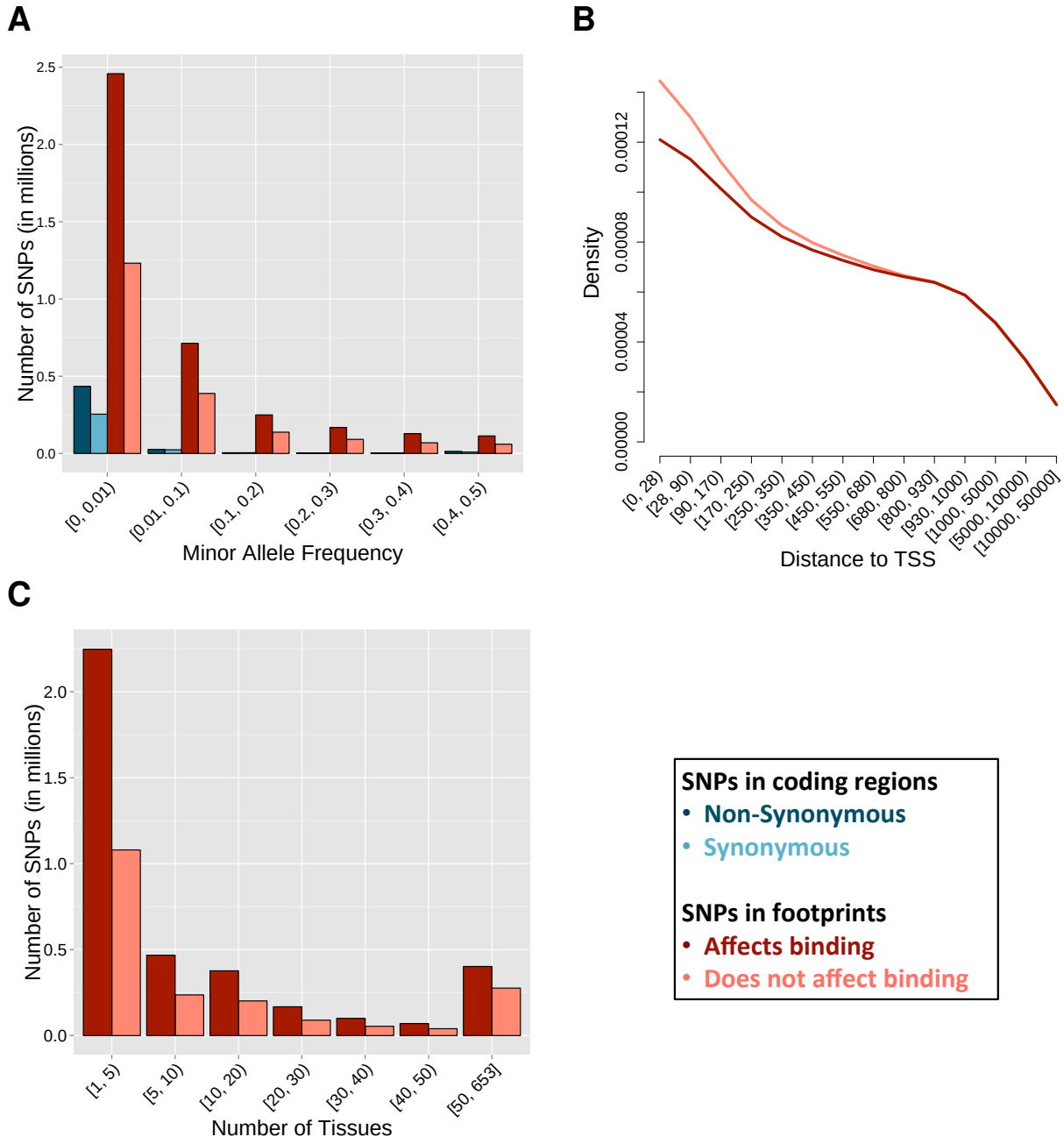

Figure 2: **Characterization of binding SNPs.** (A) Comparison of the minor allele frequency of SNPs predicted to affect binding (dark red) or not (light red). Minor allele frequency at coding SNPs, separated into non-synonymous (dark blue) and synonymous (blue), is shown for comparison. MAF is in bins of 10%, with the exception of rare (MAF $< 1\%$) SNPs. (B) Proportion of SNPs at increasing distance from the nearest transcription start site (TSS) up to 50Kb. Fewer SNPs affecting binding (dark red) are present near a TSS compared to those with no effect (light red). Distance is absolute distance, regardless of direction (up- or downstream) from TSS. (C) Stratification of footprint-SNPs by the number of tissues for which the footprint was predicted active, colored by binding effect (dark red) or not (light red). Number of tissues is binned by 5 or 10 until 50, where the remainder is binned.

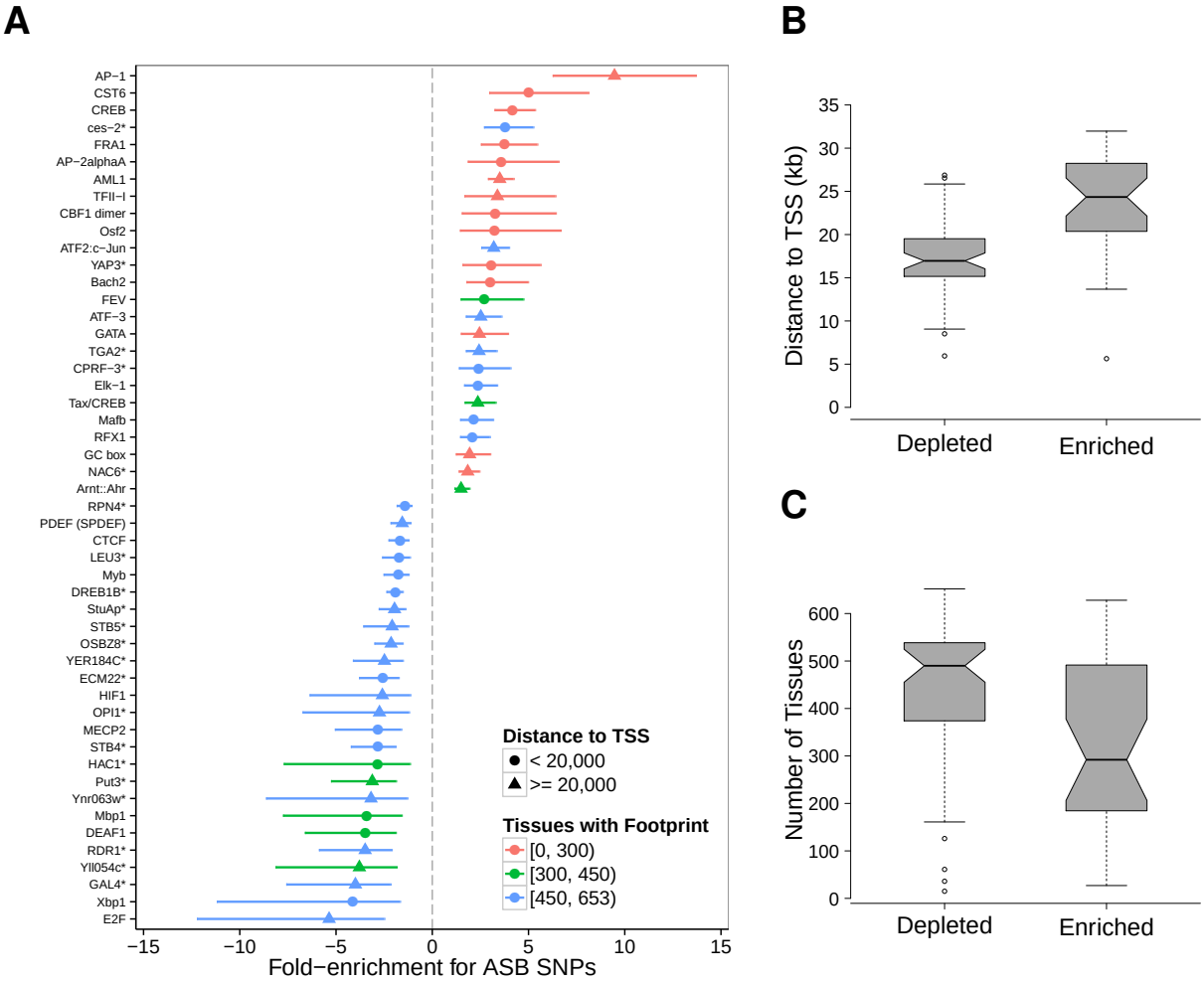

Figure 3: **Characterization of binding elements.** (A) Plot showing factors whose binding sites are significantly enriched or depleted for ASB variants (p-value <0.01), with indication of the number of tissues affected (color) and the median distance to the TSS (shape). Horizontal lines represent the 95% confidence interval of the ASB enrichment ratio. An asterisk denotes a possible human analog for the specified factor. Redundant motifs were excluded from the plot. (B) Barplot showing the distance to the nearest TSS between motifs either enriched or depleted for ASB-SNPs. (C) Barplot showing the number of tissues where a motif was predicted to be active for motifs either enriched or depleted for ASB-SNPs.

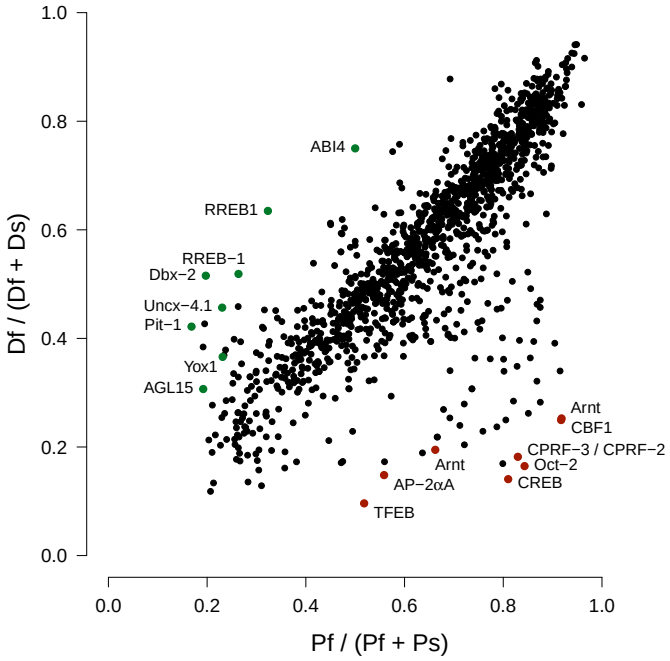

Figure 4: **Examining selection on TF binding sites.** Comparison of fixed functional (D_f) to fixed silent (D_s) (y-axis) versus polymorphic functional (P_f) to polymorphic silent (P_s) (x-axis) variants across all of the binding sites for each TF examined. Scores towards the top left are suggestive of positive selection (excess of fixed functional changes), with several of the top significant examples indicated by green points and the factor name. Scores towards the bottom right are suggestive of negative (purifying) selection, with several of the top significant examples indicated by red points and the factor name.

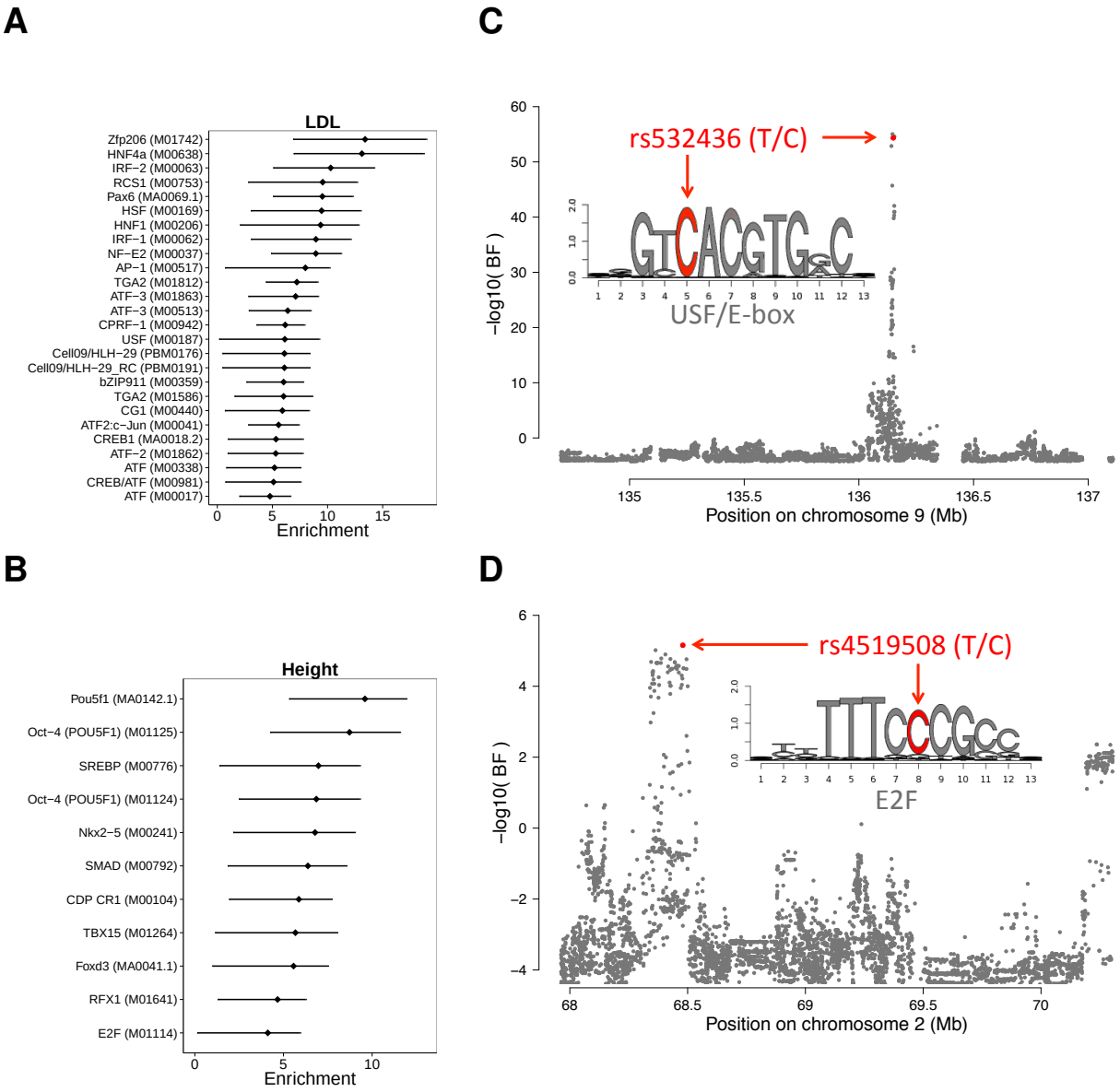

Figure 5: **Integration of annotations into GWAS results.** (A/B) Enrichment (log_2(change in prior odds w.r.t the baseline model)) of factors for association with (A) low-density lipoprotein levels and (B) height. Error bars are drawn for 95% confidence intervals. (C/D) Association plots showing the prior association (Bayes factor) of each SNP in the displayed region for (A) low-density lipoprotein levels and (B) height. Shown in red are SNPs with a posterior probability of association >0.4.