12 Impact of functional information on understanding variation

ENCODE provides an initial interpretation of many human variants and plausible leads for the role of many variants identified in genome-wide association studies

In recent years, GWAS have greatly extended our knowledge of genetic loci associated with human disease risk and other phenotypes. The output of these studies is a series of SNPs ("GWAS SNPs") correlated with a phenotype, although not necessarily the functional variants. Strikingly, 88% of associated SNPs are either intronic or intergenic[75]. We examined 4,860 SNP-phenotype associations for 4,492 SNPs curated in the NHGRI GWAS catalogue[75]. We found that 12% of these SNPs overlap TF-occupied regions whereas 34% overlap DHSs (Figure 10A). Both figures reflect significant enrichments relative to the overall proportions of 1000 Genomes project SNPs (about 6% and 23%, respectively). Even after accounting for biases introduced by selection of SNPs for the standard genotyping arrays, GWAS SNPs show consistently higher overlap with ENCODE annotations (Figure 10A, see Supplementary Information). Furthermore, after partitioning the genome by density of different classes of functional elements, GWAS SNPs were consistently enriched beyond all the genotyping SNPs in function-rich partitions, and depleted in function-poor partitions (see Supplementary Figure M1). GWAS SNPs are particularly enriched in the segmentation classes associated with enhancers and TSSs across several cell types (see Supplementary Figure M2).

Examining the SOM of integrated ENCODE annotations (see above), we found 19 SOM map units showing significant enrichment for GWAS SNPs, including many SOM units previously associated with specific gene functions, such as the immune response regions. Thus, an appreciable proportion of SNPs identified in initial GWAS scans are either functional or lie within the length of an ENCODE annotation ($\sim$500 bp on average) and represent plausible candidates for the functional variant. Expanding the set of feasible functional SNPs to those in reasonable linkage disequilibrium, up to 71% of GWAS SNPs have a potential causative SNP overlapping a DNaseI site, and 31% of loci have a candidate SNP that overlaps a binding site occupied by a TF (see also refs 74,76).

The GWAS catalogue provides a rich functional categorization from the precise phenotypes being studied. These phenotypic categorizations are non-randomly associated with ENCODE annotations and there is striking correspondence between the phenotype and the identity of the cell type or TF used in the ENCODE assay (Figure 10B). For example, five SNPs associated with Crohn's disease overlap GATA2-binding sites (P-value 0.003 by random permutation or 0.01 by an empirical approach comparing to the GWAS-matched SNPs; see Supplementary information), and fourteen are located in DHSs found in immunologically relevant cell types. A notable example is a gene desert on chromosome 5p13.1 containing eight SNPs associated with inflammatory diseases. Several are close to or within DHSs in Th1 and Th2 cells as well as peaks of binding by TFs in HUVECs (Figure 10C). The latter cell line is not immunological, but factor occupancy detected there could be a proxy for binding of a more relevant factor, such as GATA3, in T-cells. Genetic variants in this region also affect expression levels of *PTGER477*, encoding the prostaglandin receptor EP4. Thus, the ENCODE data reinforce the hypothesis that genetic variants in 5p13.1 modulate the expression of flanking genes, and furthermore provide the specific hypothesis that the variants affect occupancy of a GATA factor in an allele-specific manner, thereby influencing susceptibility to Crohn's disease.

Non-random association of phenotypes with ENCODE cell types strengthens the argument that at least some of the GWAS lead SNPs are functional or extremely close to functional variants. Each of the associations between a lead SNP and an ENCODE annotation remains a credible hypothesis of a particular functional element class

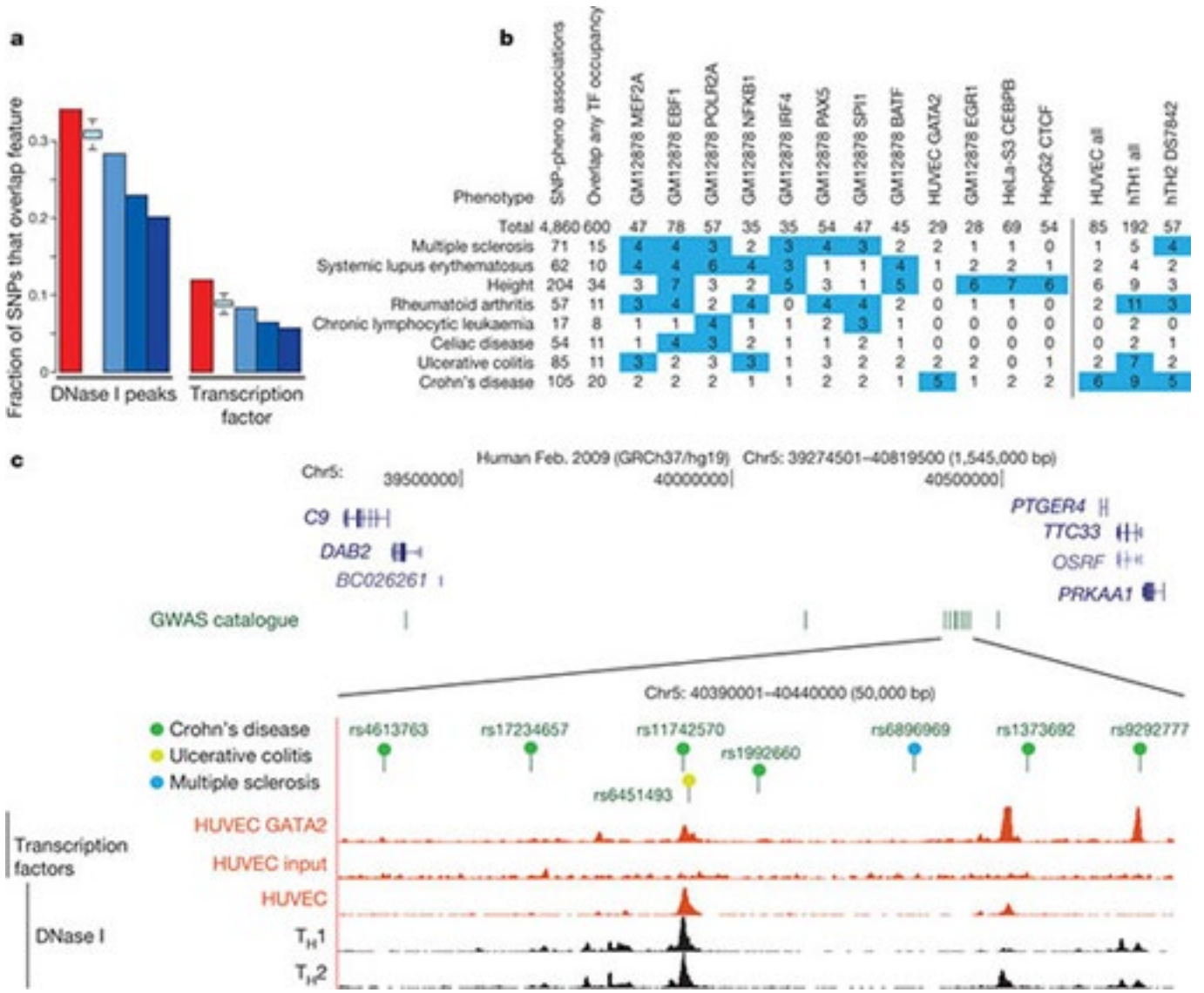


Figure 10 | **Comparison of genome-wide-association-study-identified loci with ENCODE data.** (**a**) Overlap of lead SNPs in the NHGRI GWAS SNP catalogue (June 2011) with DHSs (left) or transcription-factor-binding sites (right) as red bars compared with various control SNP sets in blue. The control SNP sets are (from left to right): SNPs on the Illumina 2.5M chip as an example of a widely used GWAS SNP typing panel; SNPs from the 1000 Genomes project; SNPs extracted from 24 personal genomes (see personal genome variants track at http://main.genome-browser.bx.psu.edu (ref. 80)), all shown as blue bars. In addition, a further control used 1,000 randomizations from the genotyping SNP panel, matching the SNPs with each NHGRI catalogue SNP for allele frequency and distance to the nearest TSS (light blue bars with bounds at 1.5 times the interquartile range). For both DHSs and transcription-factor-binding regions, a larger proportion of overlaps with GWAS-implicated SNPs is found compared to any of the controls sets. (**b**) Aggregate overlap of phenotypes to selected transcription-factor-binding sites (left matrix) or DHSs in selected cell lines (right matrix), with a count of overlaps between the phenotype and the cell line/factor. Values in blue squares pass an empirical *P*-value threshold ≤ 0.01 (based on the same analysis of overlaps between randomly chosen, GWAS-matched SNPs and these epigenetic features) and have at least a count of three overlaps. The *P* value for the total number of phenotype-transcription factor associations is <0.001. (**c**) Several SNPs associated with Crohn's disease and other inflammatory diseases that reside in a large gene desert on chromosome 5, along with some epigenetic features indicative of function. The SNP (rs11742570) strongly associated to Crohn's disease overlaps a GATA2 transcription-factor-binding signal determined in HUVECs. This region is also DNase I hypersensitive in HUVECs and T-helper T_H1 and T_H2 cells.

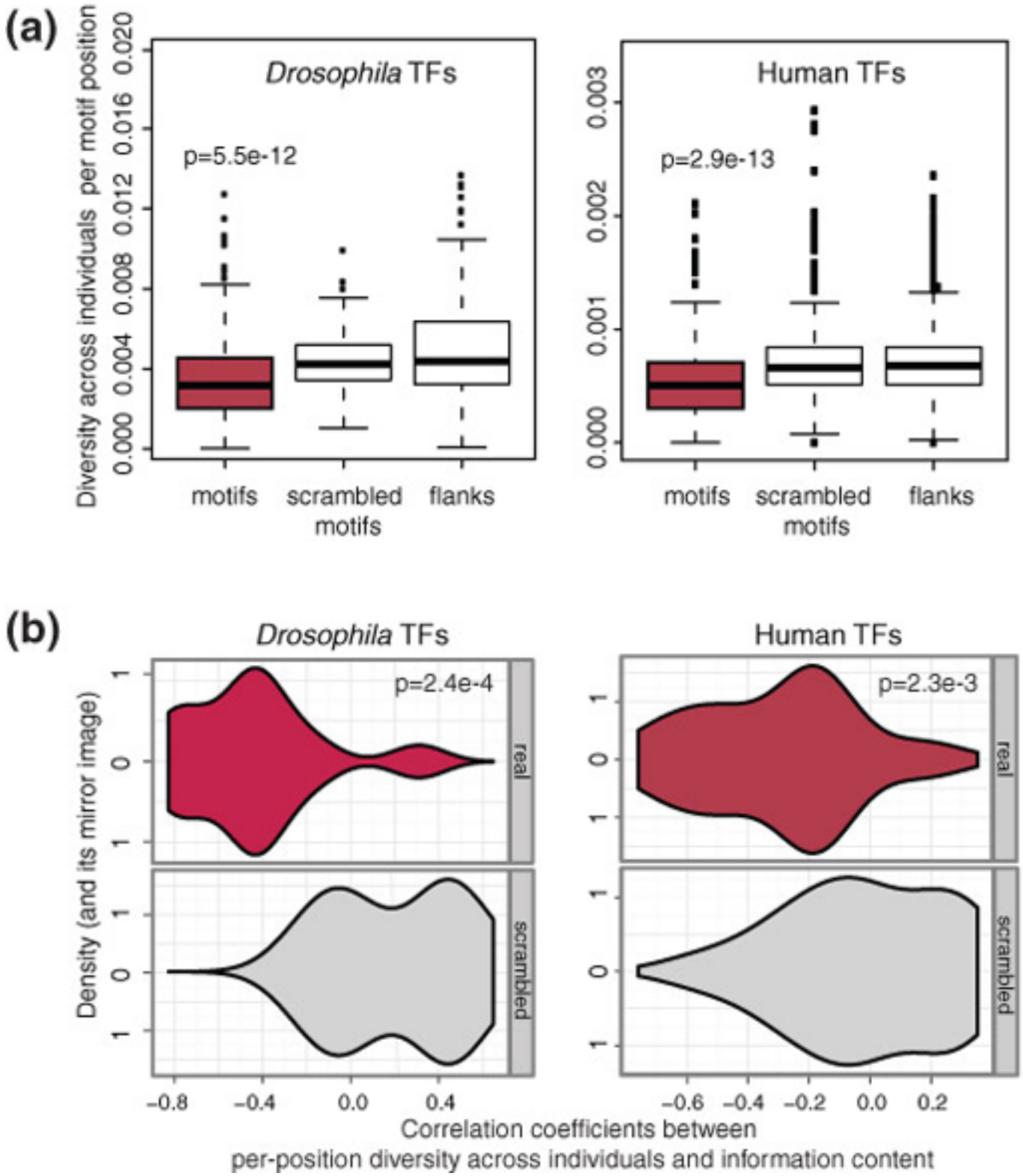


Figure 2 | Individual variation of the binding sites for 15 *Drosophila* and 36 human TFs selected for this study. (a) Distributions of position-wise diversity at motif positions (red), scrambled motifs and motif flanks at the TF-bound regions of *Drosophila* (left panel) and human (right) TFs; p-values are from Kruskal-Wallis non-parametric significance tests. (b) Violin plots (a combination of boxplots and two mirror-image kernel density plots) showing the correlation between individual variation and information content per motif position for the bound instances of *Drosophila* (left) and human (right) TFs included in this study (top, red) and their scrambled versions detected within the same bound regions (bottom, grey); p-values are from Wilcoxon two-sample non-parametric significance tests.

or cell type to explore with future experiments. Supplementary Tables M1, M2 and M3 list all 14,885 pairwise

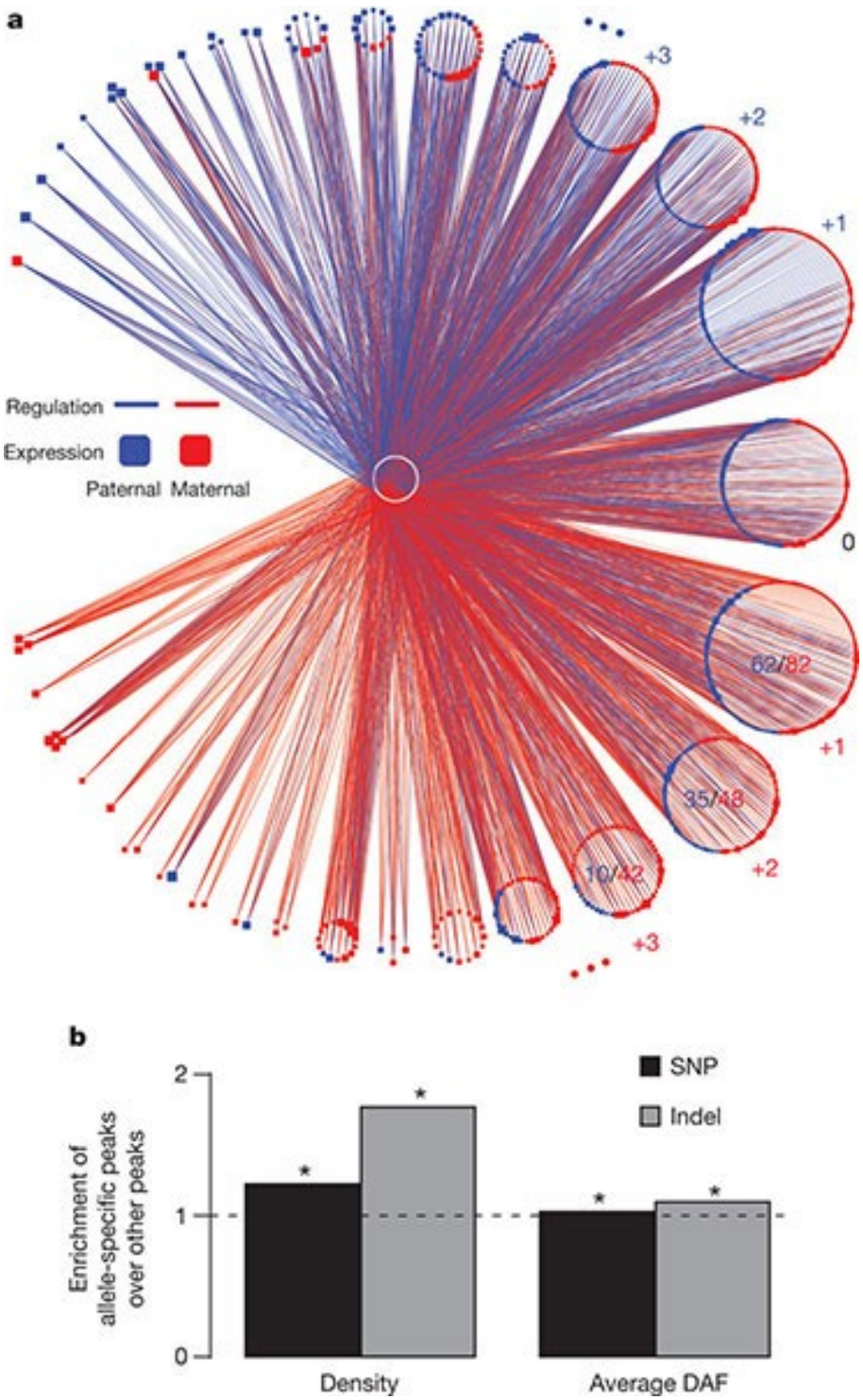
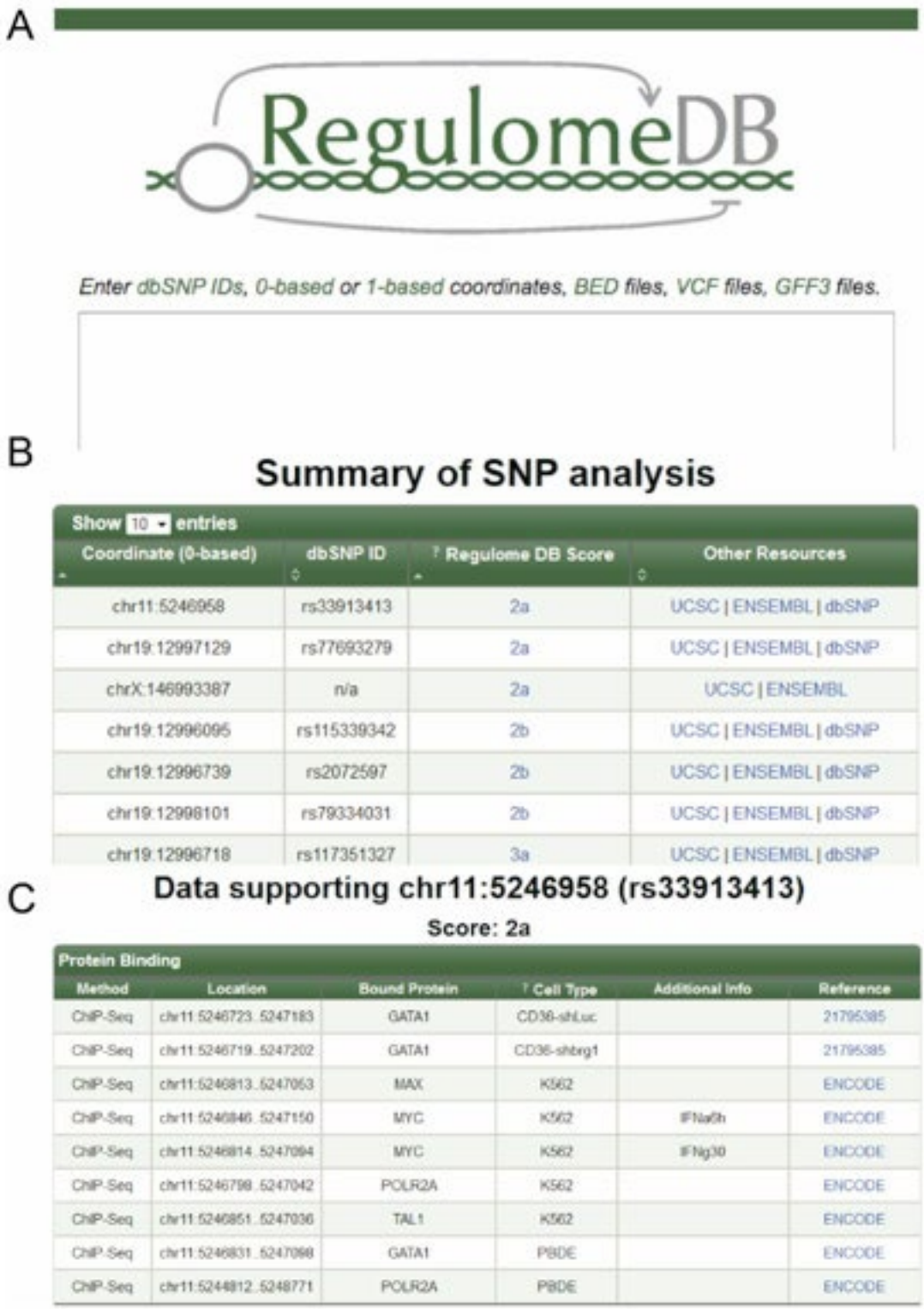


Figure 5 | Allelic Effects (a) An "allelic effects network" depicting the increasing coordination between ASB and ASE as the number of TFs regulating a target increases. Central white nodes denote TFs, and peripheral nodes denote targets, which are blue (red) if they are expressed from the paternal (maternal) allele. Blue (red) edges denote ASB to the paternal (maternal) allele. This network represents the strongest differences between the paternal- and maternal-specific regulatory networks. As one goes around the larger circle counter-clockwise (clockwise), each of the small circular clusters represents targets with progressively more paternal (maternal) regulation, indicated by the small blue (red) numbers to the side of the clusters. Moreover, within each of the clusters the fraction of predominantly paternally (maternally) expressed targets increases as one goes around the larger circle. As an illustration, this fraction is explicitly indicated by the ratios within three of the larger clusters at bottom right. (b) Relationship between TF allelicity and selection. The bar height is the ratio of the degree of selection (as measured by SNP density or average DAF) in those TF-binding peaks showing allelic behavior to the degree of selection in all other TF-binding peaks. Asterisks represent significant differences ($P<0.05$, Wilcoxon-rank-sum test). (More details in SOM/I.2 and Fig S10b,c.)



Supplementary Figure S2 | **Web-based Interface** Users are able to interface with our database by entering lists of SNVs or regions to identify common SNVs at http://www.RegulomeDB.org/ (**a**). They are then presented with a sorted list of the most important SNVs (**b**). These SNVs can be examined for the evidence used to rank them as well as a citation for the evidence.

associations across the ENCODE annotations. The accompanying papers have a more detailed examination of common variants with other regulatory information[76].

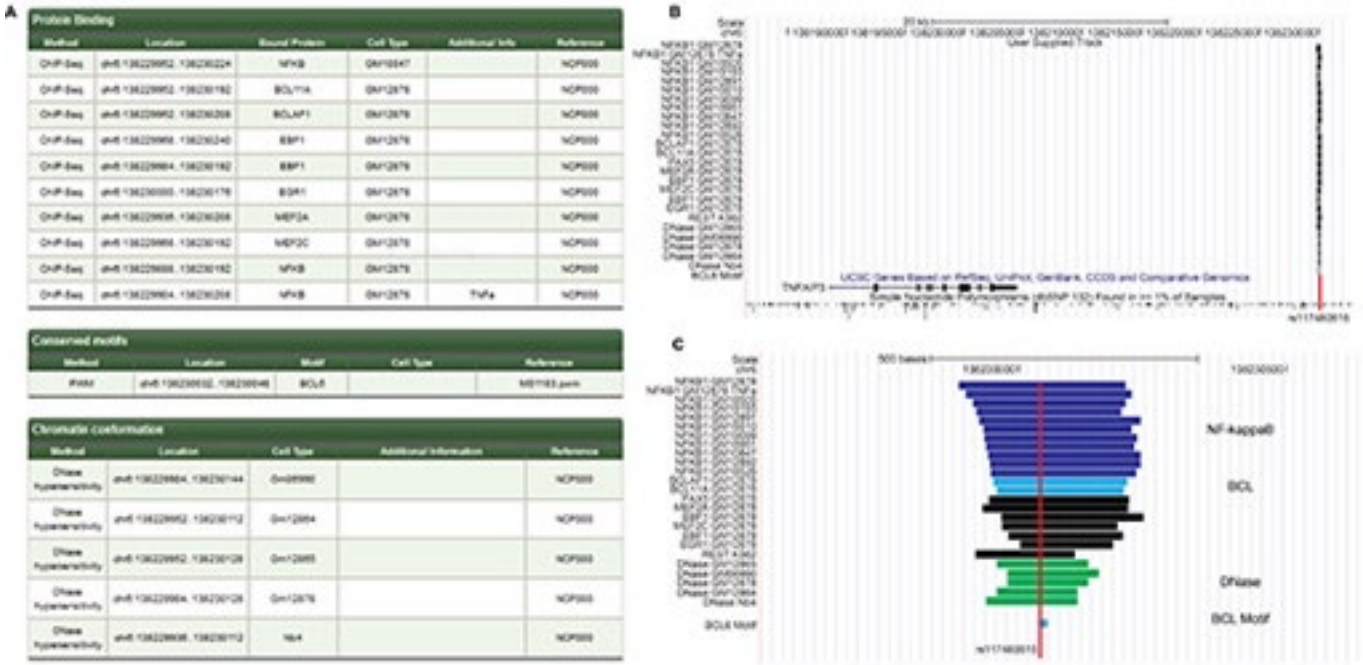


Figure 4 | ***TNFAIP3*-associated SNV** (**a**) RegulomeDB results for rs117480515 which is likely a functional variant associated with systemic lupus erythematosus. (**b**) This SNV was the most likely to be functional in the associated region but might be missed in a standard study because it lies >20 kb downstream from its target. (**c**) An enlargement of the region around rs117480515 (red line) shows the overlap with a large number of functional elements (NFKB, purple; BCL, light blue; and DNase, green) as well as the motif for BCL.

We observed reduced levels of individual variation at functional binding sites compared to reshuffled motif matches and flanking regions for other *Drosophila* factors as well as human TFs (Figure 2A). Notably, the significance of this effect was similarly high in *Drosophila* and humans, despite the fact that the SNP frequency differed approximately 11-fold (2.9% vs 0.25%, respectively), as closely reflected by the 7.5-fold difference in the number of varying TFBS. This is consistent with the overall differences in the total number of SNPs detected in these two species, likely resulting from their different ancestral effective population sizes[39]. We also observed a significant anti-correlation between variation frequency at motif positions and their information content in both species (Figure 2B).

Allelic Behavior in a Network Framework

We examined the relationship between sequence variation and TF regulation. In particular, we investigated the coordination between allele-specific binding and expression (ASB and ASE)[43,44]. We used the sequenced datasets for GM12878, which has a deeply sequenced diploid genome (SOM/I.1). We extended pairwise analysis of allele-specific behavior[20] to study higher-order coordination of multiple TFs regulating a common target. We first generated the unfiltered, promoter-regulation network for GM12878 and then identified a sub-network within it with 4,798 TF-target edges showing allele-specific regulation (SOM/I.2). This subnetwork is shown in Fig. 5a, where edges are colored red or blue to represent predominantly maternally or paternally regulated targets; the targets are similarly colored to indicate predominantly maternal or paternal expression. We find that of the 4,798 ASB cases of a single TF regulating its associated target, 57% show coordinated allelic binding and expression. We then find that for the cases in which two TFs regulate a common target, 63% are consistent (i.e., both TFs bind to the same allele that is expressed). For those cases in which triplets of TFs regulate a common target, the consistency increases to 65%. This trend continues, demonstrating that, as one increases the degree of combinatorial regulation, there is a progressively stronger relationship between expressed and regulated alleles.

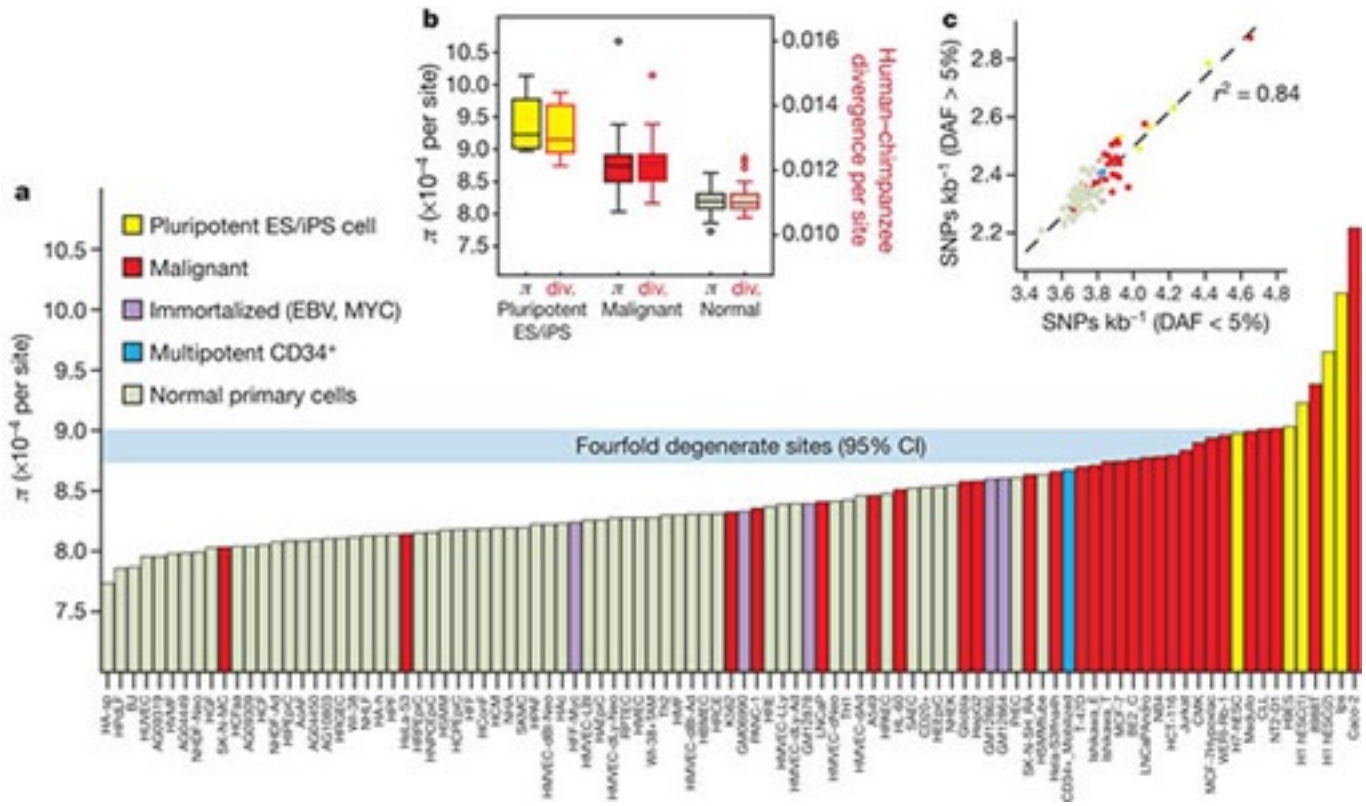


Figure 7 | **Genetic variation in regulatory DNA linked to mutation rate.** (**a**) Mean nucleotide diversity (π, *y* axis) in DHSs of 97 diverse cell types (*x* axis) estimated using whole-genome sequencing data from 53 unrelated individuals. Cell types are ordered left-to-right by increasing mean π. Horizontal blue bar shows 95% confidence intervals on mean π in a background model of fourfold degenerate coding sites. Note the enrichment of immortal cells at right. ES, embryonic stem; iPS, induced pluripotent stem. (**b**) Mean π (left *y* axis) for pluripotent (yellow) versus malignancy-derived (red) versus normal cells (light green), plotted side-by-side with human-chimpanzee divergence (right *y* axis) computed on the same groups. Boxes indicate 25-75 percentiles, with medians highlighted. c, Both low- and high-frequency derived alleles show the same effect. Density of SNPs in DHSs with derived allele frequency (DAF) $<5\%$ (*x* axis) is tightly correlated ($r^2 = 0.84$) with the same measure computed for higher-frequency derived alleles (*y* axis). Colour-coding is the same as in panel **a**.

The degree of allele-specific behavior of each TF can be quantified by a statistic we call "allelicity". The allelicity of a TF is defined as the fraction of SNPs that exhibit ASB out of all the SNPs that may potentially exhibit it (SOM/I.3). Thus, qualitatively, allelicity may be thought of as the sensitivity of a TF's binding to maternal-vs-paternal variants. Using our network described here, we find that TFs with higher degrees of allelicity tend to have more target genes, suggesting that less specific TFs tend to vary more in their binding with sequence (Table 1). Finally, and somewhat intriguingly, we find that small insertions and deletions (indels) tend to cause disproportionally more of these allelic events than do SNPs (Table S6g).

Using the AlleleSeq pipeline[32] on the SNPs in the GM12878 genome, we found that approximately 18% of both Gencode annotated protein coding and long non-coding genes exhibit allele-specific expression (ASE). The proportion of genes with ASE was similar in the three investigated RNA fractions (whole-cell, cytoplasm and nucleus, Table S9 and Supplementary Material).

We have developed a novel approach and database, RegulomeDB, which guides interpretation of regulatory variants in the human genome. RegulomeDB includes high-throughput, experimental data sets from ENCODE and other sources, as well as computational predictions and manual annotations to identify putative regulatory potential and identify functional variants. These data sources are combined into a powerful tool that scores

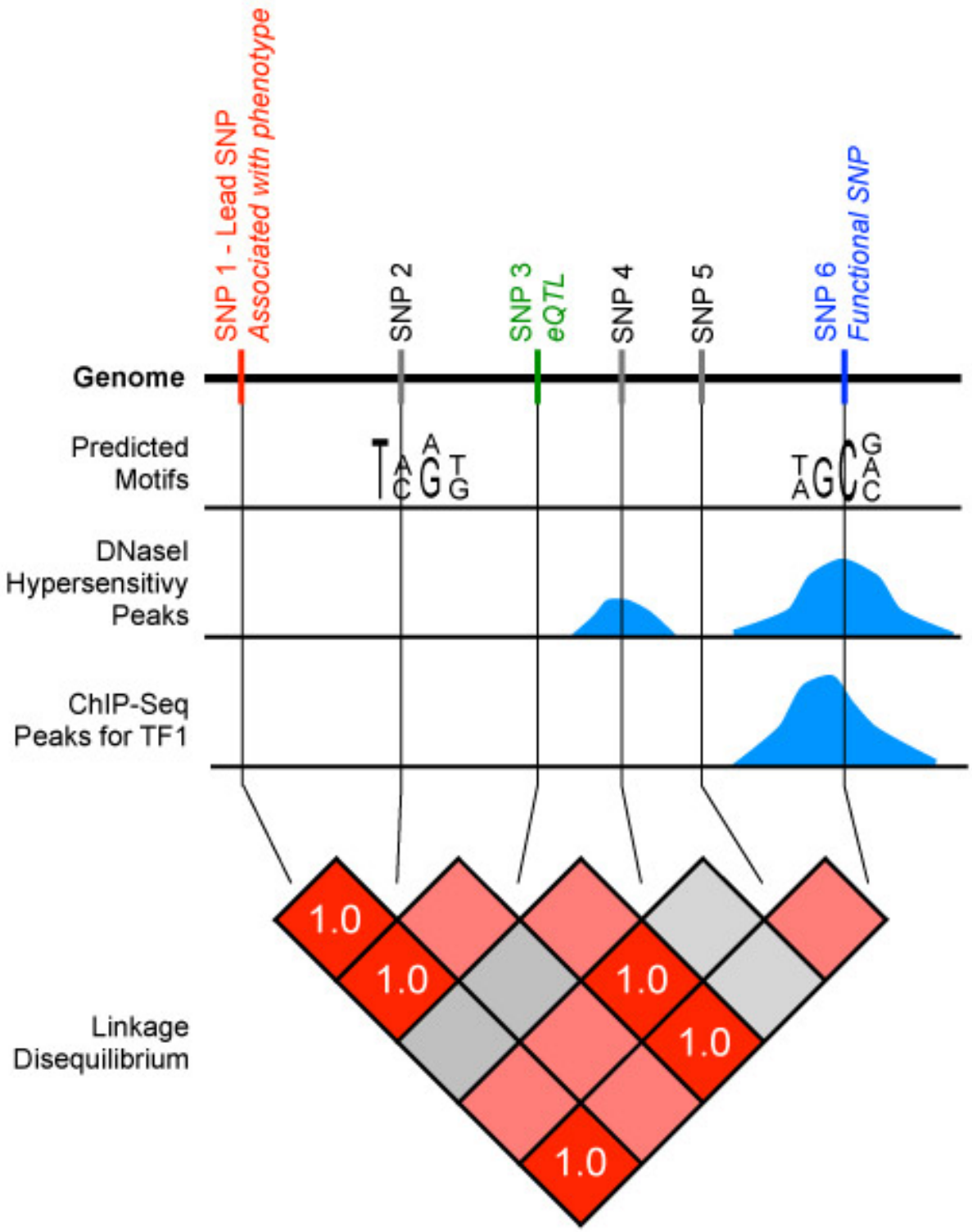


Figure 1 | **Schematic overview of the functional SNP approach.** This figure illustrates the approach we use to identify functional SNPs. Three different types of regulatory data are represented for an area of the genome: motif-based predictions, DNase I hypersensitivity peaks, and ChIP-seq peaks. This region contains six SNPs. SNP1 is associated with a phenotype in a genome-wide association study. SNP3 is an eQTL associated with changes in gene expression in a different study. SNP6 overlaps a predicted motif, a DNase I hypersensitivity peak, and a ChIP-seq peak. There are, therefore, multiple sources of evidence that SNP6 is in a regulatory region. Furthermore, SNP6 is in perfect linkage disequilibrium ($r^2 = 1.0$) with SNP1 and SNP3, meaning that there is transitive evidence due to the LD that SNP6 is also associated with the phenotype and is also an eQTL. SNP6 is therefore the most likely functional SNP in this associated region.

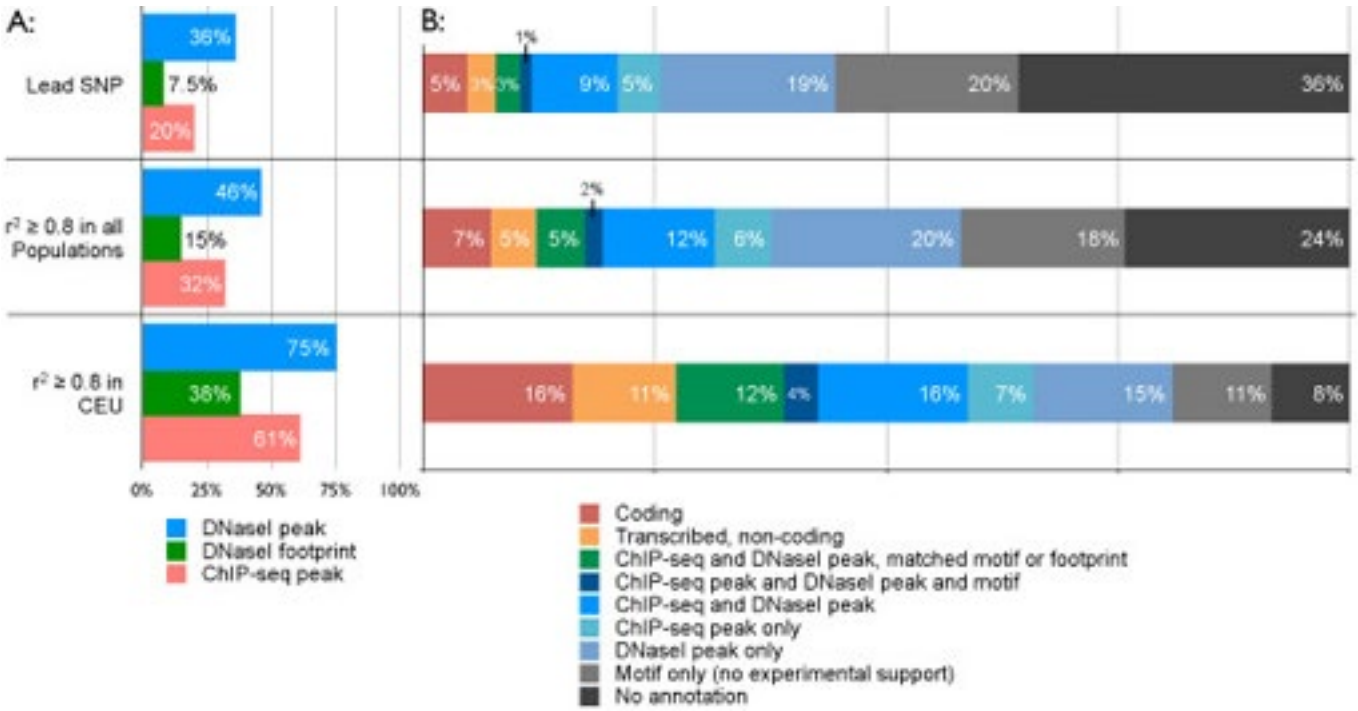


Figure 2 | **Proportions of associations for different types of functional data.** Proportions are shown for individual assays (**a**) and for all sources of evidence combined (**b**). Proportions are presented separately for lead SNPs and SNPs in strong linkage disequilibrium ($r^2 \geq 0.8$) with a lead SNP. For each association, we determine which SNP in the LD region is most strongly supported by functional data in order to generate the proportions in panel **b**. We separately consider SNPs in strong linkage disequilibrium with a lead SNP in all HapMap 2 populations, and SNPs in strong linkage disequilibrium with a lead SNP in the CEU population. For the latter case, we use only associations identified in populations of European descent, and show that we can map 80% of these associations to a functional SNP supported by experimental ENCODE data.

variants to help separate functional variants from a large pool and provides a small set of putative sites with testable hypotheses as to their function. We demonstrate the applicability of this tool to the annotation of noncoding variants from 69 full sequenced genomes as well as that of a personal genome, where thousands of functionally associated variants were identified. Moreover, we demonstrate a GWAS where the database is able to quickly identify the known associated functional variant and provide a hypothesis as to its function.

Classifying variants based on the above criteria is also highly informative to genome-wide association studies. We demonstrate this by repeating the search for a causative SNV for systemic lupus erythematosus in a 500MB region around the TNFAIP3 gene (Adrianto *et al.* 2011).

In the initial 500MB region, there are approximately 2,604 SNVs present at greater than 1% MAF (dbSNP132), of which 109 are classified by RegulomeDB as having a potentially functional consequence. Using an association test on 113 SNVs in the tested European and Asian populations we are able to identify 28 SNVs in association with the disease in common between Europeans and Asians. Of these SNVs, our approach classifies 3 as having potential functional consequence - each of which provides an easily testable hypothesis.

Furthermore, the study authors further reduced the size of the risk haplotype to a 16.3kb region through use of LD structure and conditional association analysis which resulted in 8 SNVs only one of which is assigned as putatively functional by RegulomeDB. This SNV is the same one that the study authors conclude to be the most likely functional polymorphism.

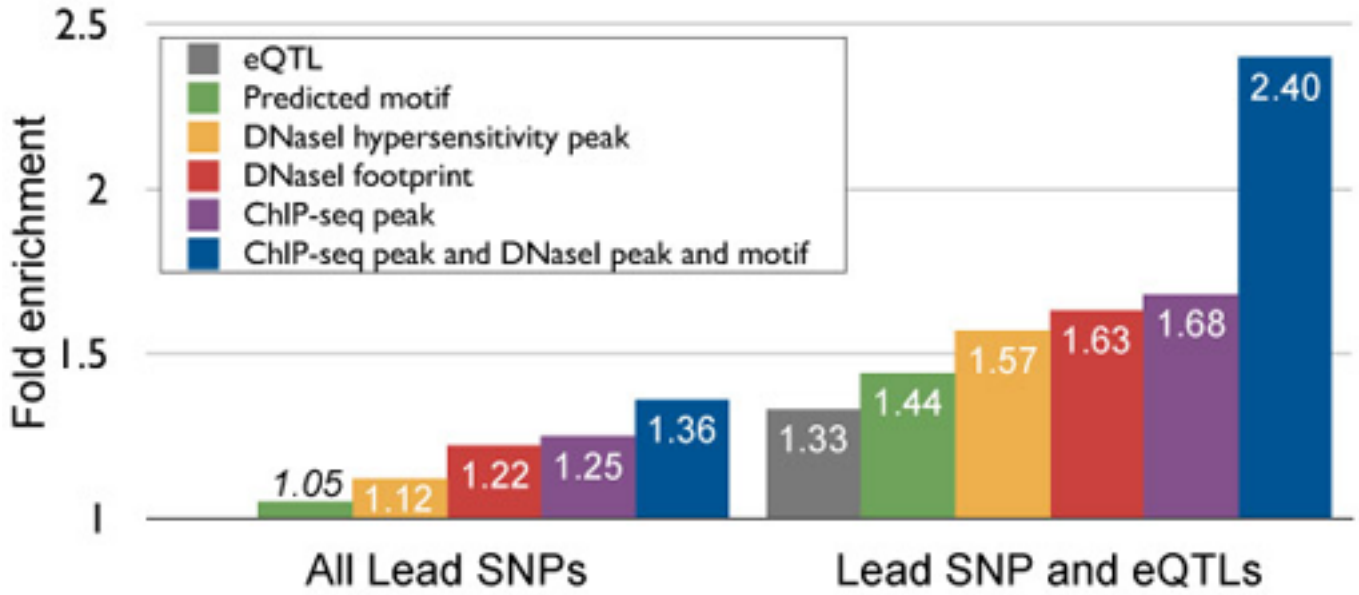


Figure 3 | **Overview of enrichment for different combinations of assays.** Enrichments are reported for all lead SNPs associated with a phenotype and separately for lead SNPs that are also eQTLs or in strong linkage disequilibrium with an eQTL. The enrichment for predicted motifs alone (italics) is not significant. These results show that combining multiple types of experimental evidence increases the observed enrichment.

The supporting evidence for this likely functional SNV (rs117480515) is detailed in Figure 4A. A set of immune associated proteins are shown by ChIP-seq to bind regions overlapping this SNV: NFKB, BCL11A, BCLAF1, EBF, MEF2A, and MEF2C (Figure 4B-C). However, there is only one putative binding site (based on PWMs) overlapping this SNV and that belongs to the BCL family indicating that BCL binding is disrupted by this polymorphism. In fact, the actual TT>A polymorphism decreases the information content match to the BCL consensus site by 3.24 bits and moves it below our PWM call threshold. The study authors demonstrate a decrease in NFKB binding with the polymorphism and conclude that this variant is likely to influence TNFAIP3 expression by decreasing factor binding in response to pro-inflammatory signals. However, in our analysis any NFKB binding sites are intact, and we find it likely that the actual cause of the binding disruption is due to a BCL motif disruption. It is possible that BCL binding assists NFKB binding at this genomic location.

The potential for single nucleotide variants within a transcription factor recognition sequence to abrogate binding of its cognate factor is well known[13]. The depth of sequencing performed in the context of our footprinting experiments provided hundreds- to thousands-fold coverage of most DHSs, enabling precise quantification of allelic imbalance within DHSs harbouring heterozygous variants. We scanned all DHSs for heterozygous single nucleotide variants identified by the 1000 Genomes Project[14] and measured, for each DHS containing a single heterozygous variant, the proportion of reads from each allele. We identified likely functional variants conferring significant allelic imbalance in chromatin accessibility and analysed their distribution relative to DNaseI footprints. This analysis revealed significant enrichment ($P < 2.2 \times 10\text{^}16$; Fisher's exact test) of such variants within DNaseI footprints (Supplementary Fig. 6). For example, rs4144593 is a common T-to-C (T/C) variant that lies within a DHS on chromosome 9. This variant falls on a high-information position within an NF1/CTF1 footprint and substantially disrupts footprinting of this motif, resulting in allelic imbalance in chromatin accessibility (Fig. 2a).

The DHS compartment as a whole is under evolutionary constraint, which varies between different classes and locations of elements[14], and may be heterogeneous within individual elements[35]. To understand the evolutionary forces shaping regulatory DNA sequences in humans, we estimated nucleotide diversity (p) in DHSs using publicly available whole-genome sequencing data from 53 unrelated individuals[36] (see Supplementary

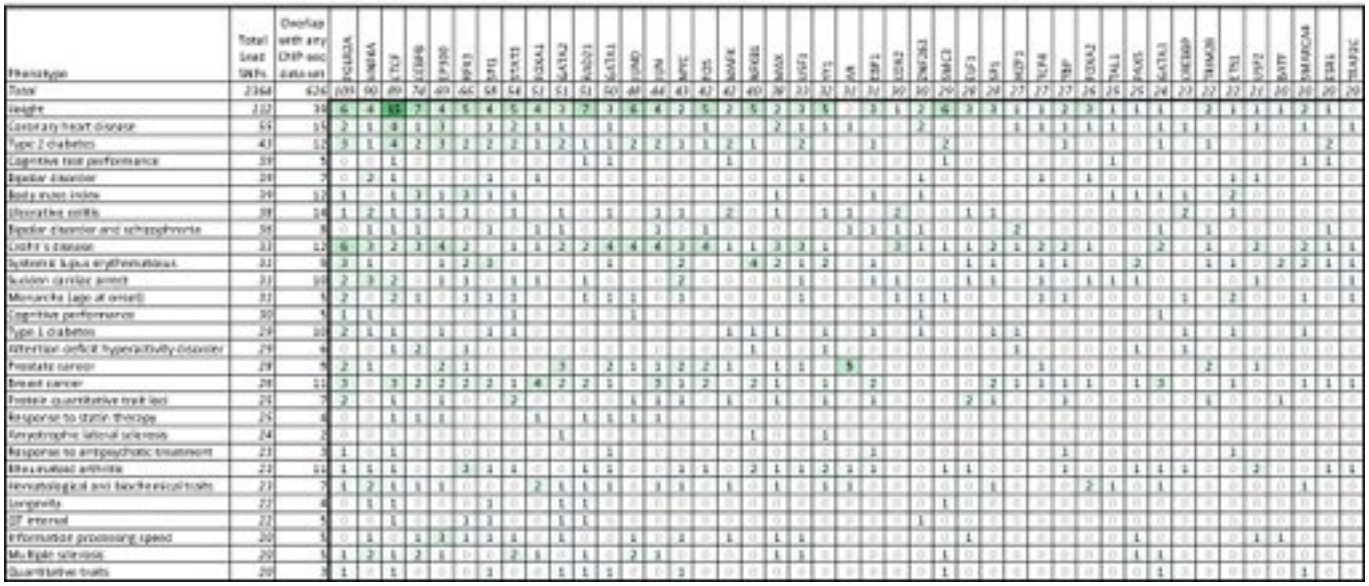


Figure 4 | **Phenotype level overview of the overlap between associations and ChIP-seq binding.** This matrix view shows phenotypes vertically and DNA binding proteins assessed using ChIP-seq horizontally. Each cell represents the number of lead SNPs for the respective phenotype that overlap with a ChIP-seq peak for the respective DNA binding protein or are in strong LD ($r^2 \geq 0.8$ in the CEU HapMap 2 population) with a SNP that overlaps such a peak. Only phenotypes with at least 20 lead SNPs and DNA binding proteins overlapping at least 20 functional SNPs are shown, but totals are computed over the entire data set. The significant interaction between height-associated functional SNPs and *CTCF*, as well as the association between prostate cancer-associated functional SNPs and androgen receptor (*AR*), are represented in bold font.

Methods). We restricted our analysis to nucleotides outside of exons and RepeatMasked regions. To provide a comparison with putatively neutral sites, we computed p in fourfold degenerate synonymous positions (third positions) of coding exons. This analysis showed that, taken together, DHSs exhibit lower *p* than fourfold degenerate sites, compatible with the action of purifying selection.

Figure 7a shows *p* for the DHSs of all analysed cell types, with colour coding to indicate the origin of each cell type. Particularly striking is the distribution of diversity relative to proliferative potential. DHSs in cells with limited proliferative potential have uniformly lower average diversity than immortal cells, with the difference most pronounced in malignant and pluripotent lines. This ordering is identical when highly mutable CpG nucleotides are removed from the analysis.

If differences in *p* are due to mutation rate differences in different DHS compartments, the ratio of human polymorphism to human-chimpanzee divergence should remain constant across cell types. By contrast, differences in *p* due to selective constraint should result in pronounced differences. To distinguish between these alternatives, we first compared polymorphism and human-chimpanzee divergence for DHSs from normal, malignant and pluripotent cells (Fig. 7b). Differences in polymorphism and divergence between these three groups are nearly identical, compatible with a mutational cause. Second, raw mutation rate is expected to affect rare and common genetic variation equally, whereas selection is likely to have a larger impact on common variation. We consistently observe ~62% of single nucleotide polymorphisms (SNPs) in DHSs of each group to have derived-allele frequencies below 0.05. DHSs in different cell lines exhibit differences in SNP densities but not in allele frequency distribution (Fig. 7c). Collectively, these observations are consistent with increased relative mutation rates in the DHS compartment of immortal cells versus cell types with limited proliferative potential, exposing an unexpected link between chromatin accessibility, proliferative potential and patterns of human variation.

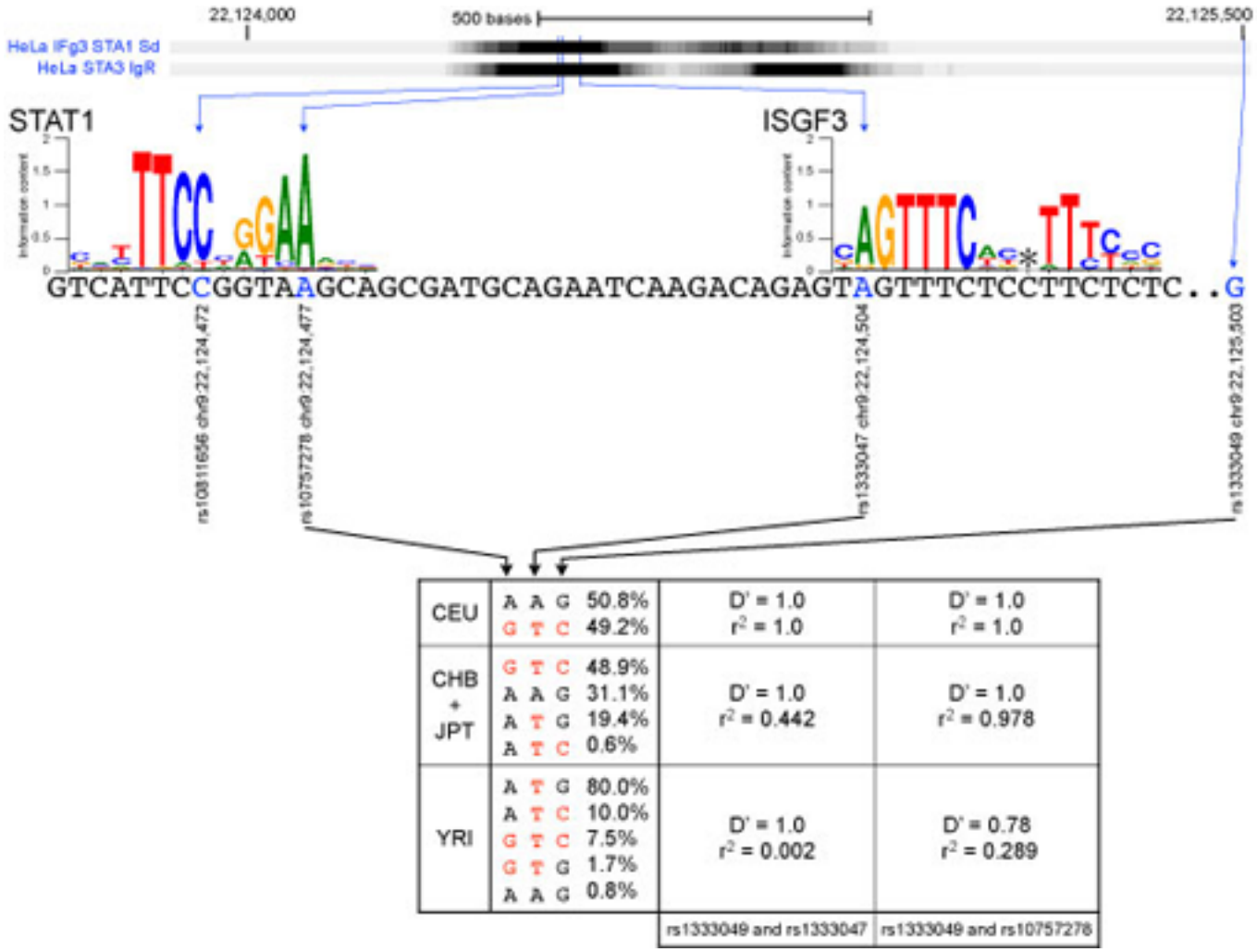


Figure 6 | Functional information and linkage disequilibrium patterns support the implication of rs1333047 in coronary artery disease. Functional data (ChIP-seq) generated by the ENCODE Consortium show evidence of *STAT1* binding in the 9p21 region associated with coronary artery disease. rs10757278 and rs1333047 are both located in the peak, whereas rs1333049 is a tag SNP that does not overlap any functional region in RegulomeDB. rs10757278 is part of a regulatory motif for *STAT1* binding, and rs1333049 is part of a regulatory motif for *ISGF3* binding. (*) The location at which a gap is inserted into the motif to handle variable linker length. Haplotype frequency and linkage disequilibrium data from the different HapMap2 populations show that all three SNPs are in perfect linkage disequilibrium in the CEU population, but not the CHB and JPT populations. In the YRI population, the frequency of the *A* allele at rs1333047 is only 0.8%. Risk alleles for all SNPs are determined using the haplotype associated with coronary artery disease in the CEU population (red). There is an absence of linkage disequilibrium between rs1333047 and rs1333049 in YRI, and the association between rs1333049 and rs10757278 and coronary artery disease has not been replicated in populations of African descent.

We call *functional SNP* any SNP that appears in a region identified as associated with a biochemical event in at least one ENCODE cell line. Functional SNPs can be further subdivided into SNPs that overlap coding or non-coding transcripts, and SNPs that appear in region identified as potentially regulatory, such as ChIP-seq peaks and DNaseI hypersensitive sites. We call the SNPs that are reported to be statistically associated with a phenotype *lead SNPs*. For each lead SNP we first determine whether the lead SNP itself is a functional SNP, then find all functional SNPs that are in strong linkage disequilibrium with the lead SNP.

We first annotated each lead SNP with transcription information from GENCODE v7 and regulatory information from RegulomeDB. Overall, 44.8% of all lead SNPs overlap with some ENCODE data, making them functional

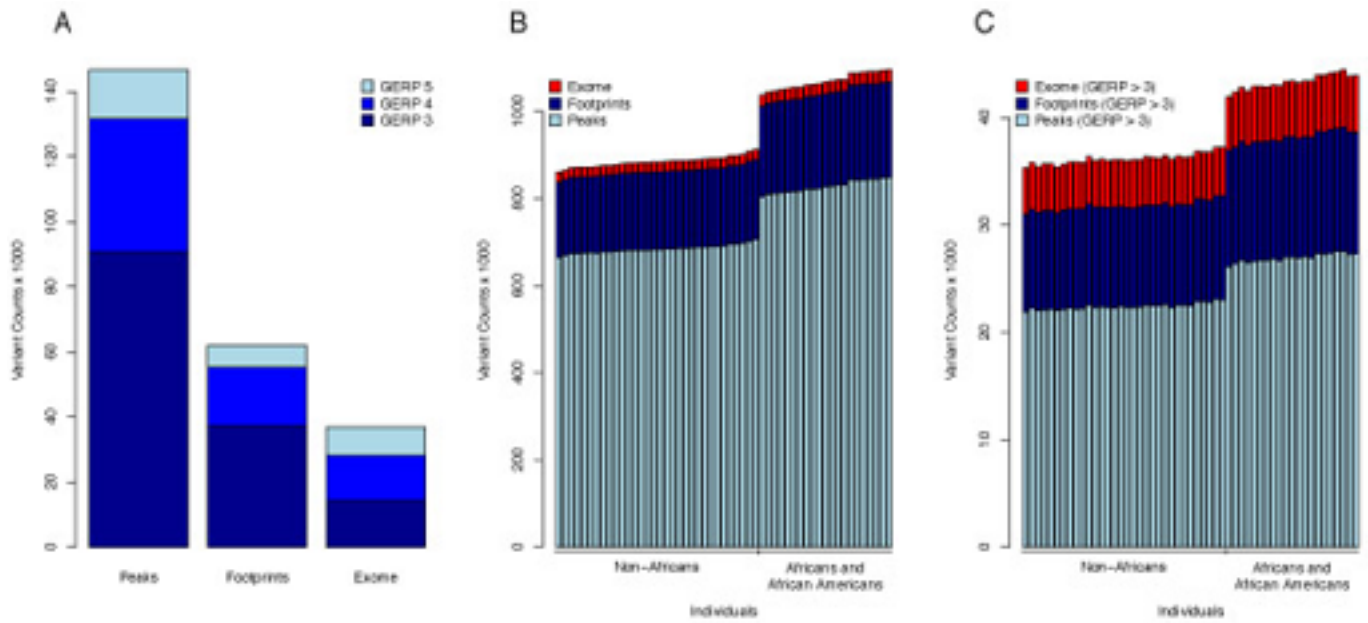


Figure 2 | **Characteristics of regulatory variation among individuals.** (**a**) Total number of variants in DNase I peaks, footprints, and the exome stratified by GERP score. (**b**) Distribution of the number of variants per individual in DNase I peaks, footprints, and the exome. (**c**) Distribution of the number of variants per individual with GERP ≥ 3 in DNase I peaks, footprints, and exomes.

SNPs according to our definition, and 13.1% of the lead SNPs are supported by more than one type of functional evidence. Specifically, 223 lead SNPs (4.7%) overlap coding regions, 146 (3.1%) overlap with the non-coding part of an exon, 1714 (36.3%) overlap with a DNaseI peak in at least one cell line, 355 (7.5%) overlap with a DNaseI footprint, and 938 (19.9%) overlap with a ChIP-seq peak for at least one of the assessed proteins in at least one cell line.

For each lead SNP we next located the set of SNPs that are in strong linkage disequilibrium ($r^2 \geq 0.8$) with the lead SNP in all four HapMap 2 populations, and annotate each SNP in this set. As expected, the fraction of lead SNPs in strong linkage disequilibrium with a SNP overlapping each type of functional evidence is larger than when considering lead SNPs alone (Figure 2), and 58% of all associations are in strong linkage disequilibrium with at least one functional SNP. A similar increase can be observed for functional SNPs supported by multiple sources of evidence. We repeated the same analysis for the 2464 lead SNPs that have been associated with a phenotype in a population of European descent, using SNPs in strong linkage disequilibrium ($r^2 \geq 0.8$) with the lead SNP in the European HapMap population only. A total of 81% of the lead SNPs are in strong LD with at least one functional SNP, and 59% of the associated SNPs are in strong linkage disequilibrium with a functional SNP supported by multiple sources of evidence (Figure 2C).

We performed randomizations in order to compare the fraction of lead SNPs that are functional SNPs or are in linkage disequilibrium with a functional SNP, to the expected fraction amongst all SNPs. We found that associated regions are significantly enriched for functional SNPs identified using DNase-seq and ChIP-seq. Furthermore, enrichments increased, both when integrating multiple ENCODE assays and when adding eQTL information. We used a subset of 2364 lead SNPs for which sufficient information is available, and built 100 random matched SNP sets in which each lead SNP is replaced by a similar SNP (see Methods for details). We compared the fraction of lead SNPs overlapping functional regions in the set of actual lead SNPs to the fractions observed in the random sets, and computed enrichment values in order to show that the fraction of associated SNPs that overlap functional regions is higher than expected.

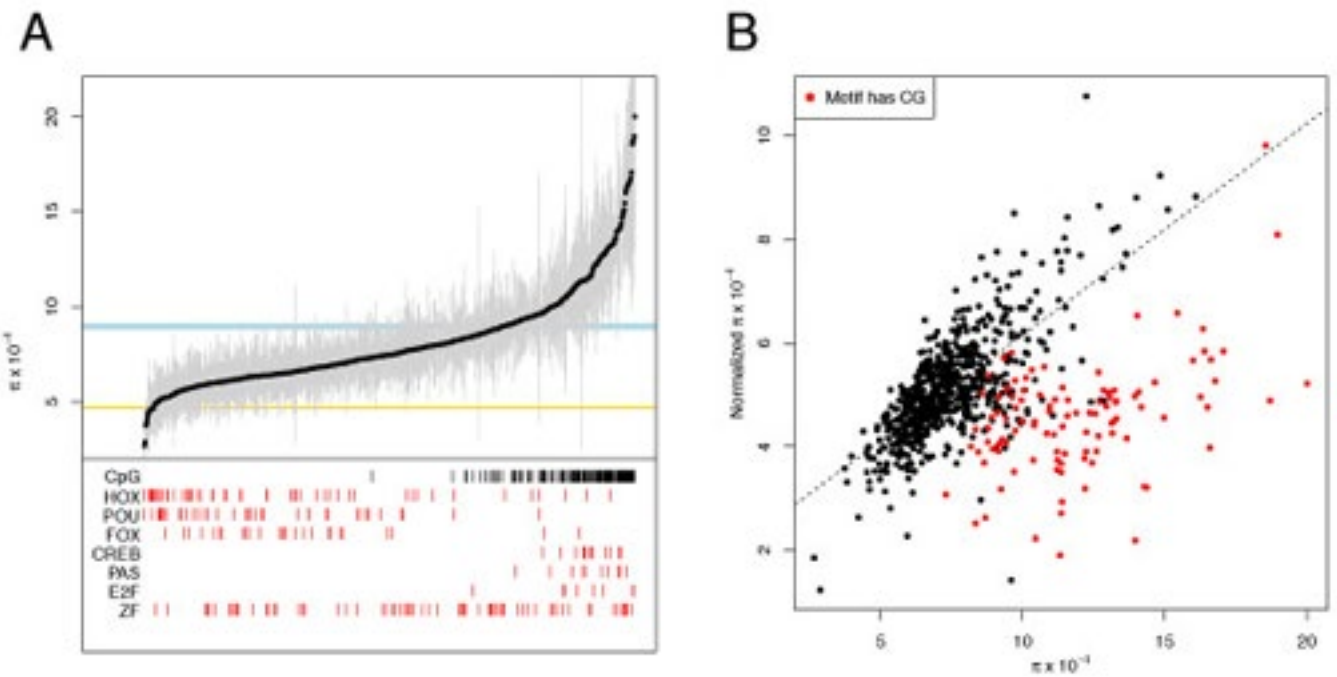


Figure 3 | **Significant variation of diversity between 732 *cis*-regulatory motifs.** (**a**) For each motif, average diversity is plotted as a black circle, and 95% confidence intervals obtained by bootstrapping are shown as gray lines. The light blue and yellow rectangles denote the 95% confidence intervals of diversity in fourfold synonymous sites (FFSs) and the exome, respectively. (Red vertical lines) Motifs that belong to the indicated class of transcription factor. (Black vertical lines) Motifs where at least 50% of all instances of that motif contain a CpG dinucleotide. (**b**) Normalized diversity in motifs versus non-normalized diversity. Motifs with a CpG (defined as above) are plotted in red. (Dashed line) Best fit for non-CpG motifs ($r = 0.70$, $P < 10^{-16}$).

ENCODE data can be used in order to compare multiple functional SNPs that are in LD with a given lead SNP. We used a two-step approach to compare the functional annotation of two SNPs. First, if one of the SNPs is in a coding region according to GENCODE v7 and the other one is not, the coding SNP is considered to be more likely to be functional. Similarly, a SNP in a non-coding part of an exon is considered to be more likely to be functional than a SNP in an intergenic region or an intron. Second, if both SNPs are not in exons, then we compared the amount of evidence across data sources supporting the functional role of the SNP using a scoring scheme integrated in RegulomeDB (see Supplementary Methods). We hypothesized that a SNP supported by multiple types of evidence (eg. a ChIP-seq peak and a DNaseI footprint) is more likely to be functional than a SNP supported by a single experimental modality. We find that most associations where the lead SNP is in LD with at least one other SNP, the SNP with the most strongly supported functional SNP is not the lead SNP itself, but another SNP in the LD region (22.4% compared to 13.6% when using LD in all populations, 56.8% compared to 13.6% percent when considering CEU only, Table 1). These results show that in most cases, the associated SNP reported in a GWAS is not the most likely to play a biological role in the phenotype according to ENCODE data.

This result is of particular importance for the interpretation of GWAS results, as LD patterns differ markedly between populations. If the functional SNP is in strong LD with the lead SNP in the population in which the GWAS was performed, but not in a different population, then the lead SNP will not be associated with the phenotype in this second population. An example of this situation is functional SNP rs1333047 (described in thread component 16), which lies in a region associated with coronary artery disease. This SNP is in perfect LD with two lead SNPs in populations of European descent in which the studies identifying the associations were performed, but not in populations of African descent, in which the associations could not be replicated (Assimes *et al.* 2008, Kral *et al.* 2011, Lettre *et al.* 2011); see Supplementary Information.

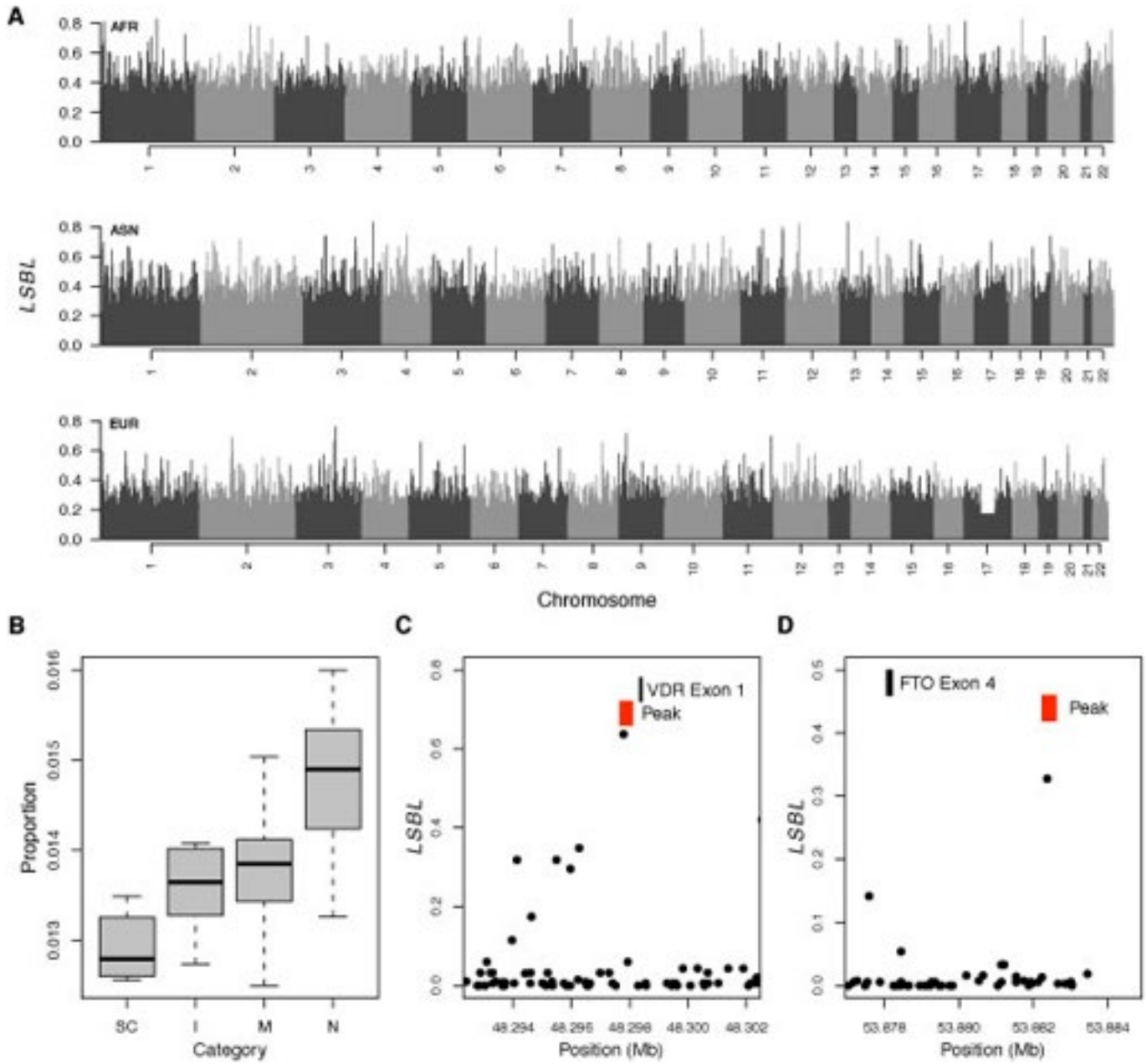


Figure 6 | Genome-wide distribution of population structure in regulatory DNA. (a) Genome-wide distribution of locus-specific branch lengths (LSBLs) for Africans, Asians, and Europeans, respectively. Note that the valley of uniform LSBL on chromosome 17 in Europeans corresponds to the *MAPT* region that is segregating a large chromosomal inversion (Zody et al. 2008). (b) Distribution of the proportion of highly differentiated DNase I peaks found for different categories of cell types. (SC) Stem cells (iPS/ES); (I) immortalized; (M) malignant; (N) normal/primary cell types. (c) Distribution of African LSBL across intron 1 of *VDR*. (d) Distribution of European LSBL across intron 4 of *FTO*. In panels c and d, peaks are shown as red rectangles and exons as black rectangles.

In addition to considering individual associations separately, we can group associated SNPs in order to search for patterns at the phenotype level. We first assessed whether there are specific sequence binding proteins that tend to overlap functional SNPs associated with certain phenotypes more often than expected, using only associations in populations of European descent (Figure 4). We found a strong association (P-value 9×10^{-5}) between height and *CTCF* ChIP-seq peaks. A total of 39 SNPs associated with height overlap a ChIP-seq peak or are strong linkage disequilibrium ($r^2 \geq 0.8$ in the CEU population) with a SNP that overlaps a ChIP-seq peak, and 15 of those (38%) overlap a peak for *CTCF* (Supplementary Table 5), compared to 89 out of 626 SNPs (14%) when considering all phenotypes.

A second novel functional SNP is in the 9p21 region, a gene desert that contains multiple SNPs that are strongly associated with several common diseases. Lead SNP rs1333049 has been associated with coronary artery disease in multiple studies in populations of European (WTCCC 2007, Samani *et al.* 2007, Broadbent *et al.* 2008, Wild *et al.* 2011) as well as Japanese and Korean descent (Hiura *et al.* 2008, Hinohara *et al.* 2008). In the HapMap 2 CEU population, this SNP is part of a haplotype block that includes rs10757278 and rs1333047, both of which are in perfect LD with rs1333049. There is no evidence in ENCODE supporting a functional role for rs1333049. However, both rs10757278 and rs1333047 overlap a DNase hypersensitivity peak as well as ChIP-seq peaks for *STAT1* and *STAT3* in HeLA-S3 cells. Furthermore, rs10757278 lies in a *STAT1* binding site, and rs1333047 lies in a binding site and a DNaseI footprint for Interferon-stimulated gene factor 3 (*ISGF3*). Figure 6 provides an overview of this region. Although the functional role of rs10757278 has been previously reported (Harismendy *et al.* 2011), evidence of the functional role of rs1333047 is novel. Interestingly, while only 27 base pairs separate the two SNPs, they are in perfect linkage disequilibrium in the CEU population only. The frequency of the 'A' allele at rs1333047 in the Yoruba in Ibadan, Nigeria (YRI) HapMap 2 population is only 0.8%, compared to 50.8% in the CEU population. This allele is part of the protective haplotype found in GWAS performed in populations of European descent. The 'A' allele is part of the motif for *ISGF3* binding, whereas the 'T' allele is not.

On average, individuals contain 24.2k $\pm$ 2.3k, 10.1k $\pm$ 0.92k, and 4.7k $\pm$ 0.40k high GERP variants in peaks, footprints, and the exome, respectively (Fig. 2C). Although evolutionary constraint is not a perfect proxy for function, these results suggest that individuals possess more regulatory versus protein-coding variants. Assuming the probability that a variant is functional is the same between coding and noncoding DNA for a given GERP value, we estimate that individuals contain up to seven times as many regulatory compared with protein-coding variants.

The unique scope of the data sets analyzed here allows us for the first time to systematically investigate genomic patterns of variation in DNA sequence motifs. To this end, we scanned DNaseI footprints for 732 known motifs (see Methods), and for each motif we calculated nucleotide diversity, π, averaged across all instances of the motif in these regions. To facilitate interpretation of motif diversity, we also calculated π for fourfold synonymous sites, a proxy for neutrally evolving DNA, and protein-coding sequences. As shown in Figure 3A, average diversity varies by over seven-fold across known regulatory motifs, ranging from 2.67×10^{-4} to 2.0×10^{-3}. Approximately 60% of motifs have average diversities significantly lower than fourfold synonymous sites (Figure 3A), indicative of purifying selection.

Figure 3A also highlights motif diversity for several important classes of transcriptional regulators. For example, HOX-, POU-, and FOX-domain factors are heavily enriched in developmental regulators and controllers of cellular differentiation. Motifs for transcription factors belonging to these classes are markedly shifted toward lower diversity, and motifs for several individual factors exhibit levels of diversity that are reduced beyond that of protein-coding sequences (Figure 3A). By contrast, diversity in motifs for tandem zinc finger transcription factors, which comprise the largest and most diverse class of human transcription factors, are distributed relatively evenly across the diversity spectrum (Figure 3A). Members of this group include core regulatory factors such as CTCF and YY1, developmental regulators such as BLIMP1 and ZIC3, and numerous chromatin repressors such as RREB1, NRSF, and the KRAB-ZNF family of proteins. Because many of the canonical motifs for these factors contain one or more CG dinucleotides, we hypothesized that the increased average diversity for these factors might be a consequence of higher mutation rates at CpG sites. To explore this hypothesis, we identified factors for which $>$50% of the motif instances in regulatory DNA contained CpGs, which revealed that the ubiquitous presence of CpG sites is a common characteristic of motifs with high levels of diversity (Figure 3A).

A large number of genome-wide scans for recent positive selection have been performed in humans (reviewed in Akey 2009). Typically, these studies focus only on patterns of DNA sequence variation and are not informed

by functional genomics data, although genome-wide analyses have been pursued on computationally predicted motifs (e.g., Sethupathy *et al.* 2008). The large compendium of experimentally characterized regulatory regions provides a unique data set to interrogate for signatures of recent positive selection.

The genome-wide distributions of population structure in DNaseI peaks in the African, Asian, and European populations are shown in Figure 6A. We pursued two distinct approaches to interpret this data. First, to obtain general insights into the characteristics of DNaseI peaks that exhibit large allele frequency differences between populations, we focused on peaks in the 1% tail of the empirical distribution of LSBLs in each population (Figure 6A). Next, we identified all genes within 50 kb of these peaks ($n = 3{,}372$, 3,224, and 3,099 such genes in Africans, Asians, and Europeans, respectively), and tested for enrichment of KEGG pathways. As shown in Table 1, this set of genes is significantly enriched for 15 KEGG pathways, seven of which are shared between two or more populations (including pathways related to cancer, axon guidance, and WNT signaling). Interestingly, the most significantly enriched pathway in Europeans is melanogenesis (Table 1), suggesting that in addition to protein-coding variants (Lamason *et al.* 2005), regulatory polymorphisms influencing pigmentation phenotypes have also been a target of recent positive selection. Moreover, our African sample is significantly enriched for chemokine and adipocytokine signaling pathways (Table 1), which is particularly interesting given the known differences in prevalence of insulin resistance and Type 2 diabetes in individuals of African ancestry (Reimann *et al.* 2007).

We also investigated the distribution of DNaseI peaks that exhibit unusually large levels of population structure across cell types. To this end, we classified the 138 types into normal, immortalized, malignant, and pluripotent (iPS/ES) categories. The proportion of DNaseI peaks that are in the 1% tail of the empirical distribution of LSBLs is significantly different across cell type categories (Kruskal-Wallis test, $p = 3.2 \times 10^{-12}$). Primary/normal cell lines had the highest proportion of differentiated peaks, whereas iPS/ES cell lines had the lowest proportion of differentiated peaks (Figure 6B). The higher proportion of differentiated DNaseI peaks in primary/normal cell lines is driven by a wide variety of cell types including astrocytes (spinal cord (HA-sp), cerebellar (HA-c), and cortical (HA-h)), renal glomeral endothelial cells (HRGEC), and cardiac fibroblasts (HCFaa). Although these results are intriguing and offer preliminary insights into the types of tissues that contribute to fitness differences among individuals, more definitive inferences will require an even broader sampling of cell types.