# 13 Impact of evolutionary selection on functional regions

**The imprint of evolutionary selection on ENCODE regulatory elements is manifested between species and within human populations.**

We examined negative selection using two measures that highlight different periods of selection in the human genome. The first measure, inter-species, pan-mammalian constraint (GERP-based scores; 24 mammals[8]) addresses selection during mammalian evolution. The second measure is intra-species constraint estimated from the numbers of variants discovered in human populations using data from the 1000 Genomes project[55] and covers selection over human evolution. In Figure 1, we plot both these measures of constraint for different classes of identified functional elements, excluding features overlapping exons and promoters that are known to be constrained. Each graph also shows genomic background levels and measures of coding-gene constraint for comparison. Since we plot human population diversity on an inverted scale, elements that are more constrained by negative selection will tend to lie in the upper and right hand regions of the plot.

For DNaseI elements (Figure 1B) and bound motifs (Figure 1C) most sets of elements show enrichment in pan mammalian constraint and decreased human population diversity, though for some cell types the DNaseI sites do not appear overall to be subject to pan-mammalian constraint. Bound TF motifs have a natural control from the set of TF motif with equal sequence potential for binding but without binding evidence from ChIP-seq experiments; in all cases, the bound motifs show both more mammalian constraint and higher suppression of human diversity.

Consistent with previous findings, we do not observe genome-wide evidence for pan-mammalian selection of novel RNA sequences (Panel D). There are also a large number of elements without mammalian constraint, between 17-90% for TF-binding regions as well as DHSs and FAIRE regions. Previous studies could not determine whether these sequences are either biochemically active, but with little overall impact on the organism, or are under lineage specific selection. By isolating sequences preferentially inserted into the primate lineage, which is only feasible given the genome-wide scale of this data, we are able to specifically examine this issue. The majority of primate-specific sequence is due to retrotransposon activity, but an appreciable proportion is non-repetitive primate-specific sequence. Of 104,343,413 primate-specific bases (excluding repetitive elements), 67,769,372 (65%) are found within ENCODE-identified elements. Examination of 227,688 variants segregating in these primate specific regions revealed that all classes of elements (RNA and regulatory) show depressed derived allele frequencies, consistent with recent negative selection occurring in at least some of these regions (Figure 1E). An alternative approach examining sequences that are not clearly under pan-mammalian constraint showed a similar result (Luke Ward and Manolis Kellis, personal communication). This suggests that an appreciable proportion of the unconstrained elements are lineage specific elements required for organismal function, consistent with long standing views of recent evolution[56], and the remainder are likely to be "neutral" elements[2] which are not currently under selection, but may still affect cellular or larger scale phenotypes without an effect on fitness.

The binding patterns of TFs are not uniform, and we can correlate both inter-and intra-species measures of negative selection with the overall information content of motif positions. The selection on some motif positions is as high as protein coding exons (Figure 1F, Luke Ward and Manolis Kellis, personal communication). These aggregate measures across motifs show that the binding preferences found in the population of sites are also relevant to the per-site behavior. By developing a per-site metric of population effect on bound motifs, we found that highly constrained bound instances across mammals are able to buffer the impact of individual variation[57].

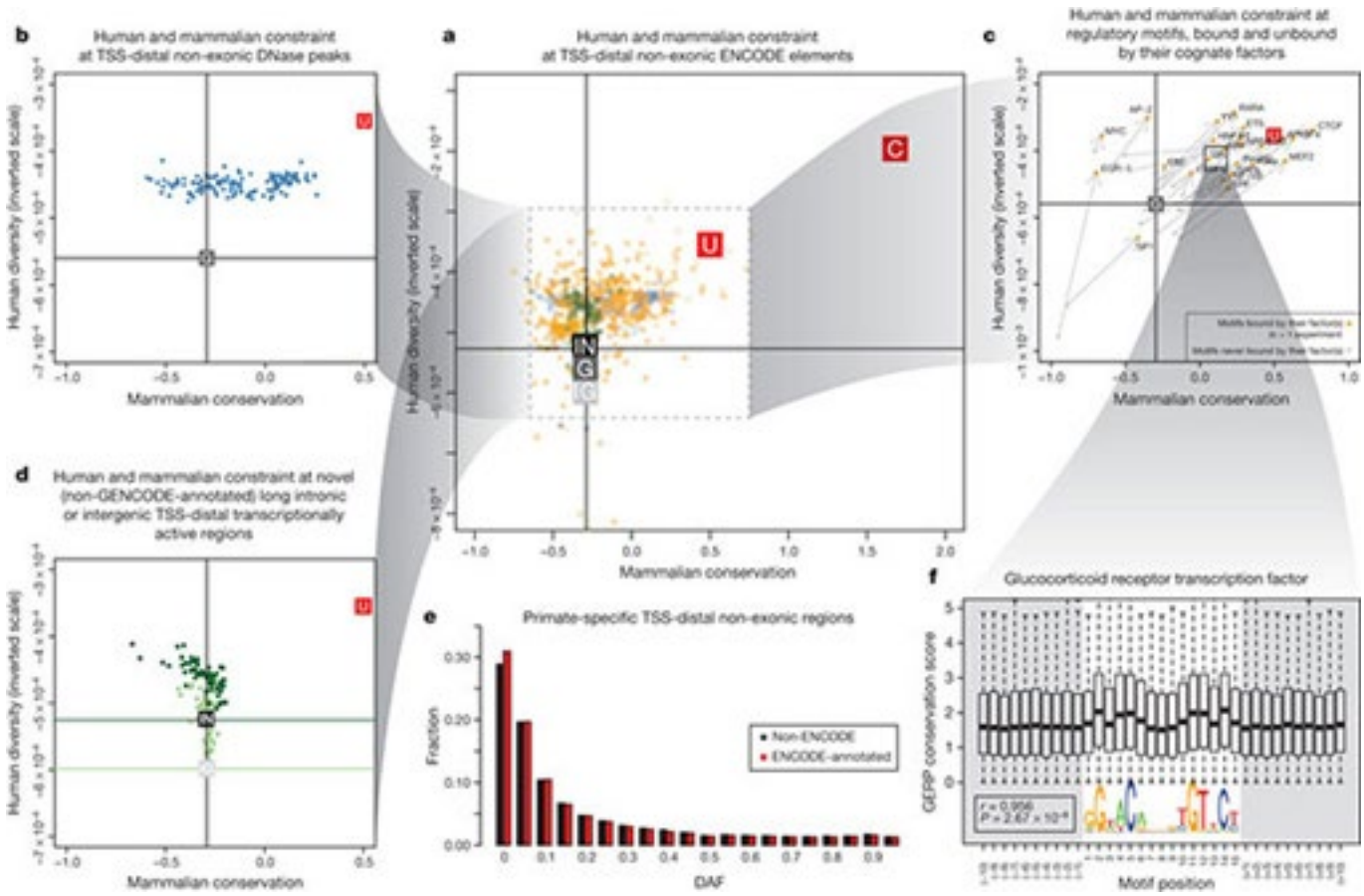


**Figure 1** | **Impact of selection on ENCODE functional elements in mammals and human populations.** (**a**) Levels of pan-mammalian constraint (mean GERP score; 24 mammals[8], *x* axis) compared to diversity, a measure of negative selection in the human population (mean expected heterozygosity, inverted scale, *y* axis) for ENCODE data sets. Each point is an average for a single data set. The top-right corners have the strongest evolutionary constraint and lowest diversity. Coding (C), UTR (U), genomic (G), intergenic (IG) and intronic (IN) averages are shown as filled squares. In each case the vertical and horizontal cross hairs show representative levels for the neutral expectation for mammalian conservation and human population diversity, respectively. The spread over all non-exonic ENCODE elements greater than 2.5 kb from TSSs is shown. The inner dashed box indicates that parts of the plot have been magnified for the surrounding outer panels, although the scales in the outer plots provide the exact regions and dimensions magnified. The spread for DHS sites (**b**) and RNA elements (**d**) is shown in the plots on the left. RNA elements are either long novel intronic (dark green) or long intergenic (light green) RNAs. The horizontal cross hairs are colour-coded to the relevant data set in **d**. (**c**), Spread of transcription factor motif instances either in regions bound by the transcription factor (orange points) or in the corresponding unbound motif matches in grey, with bound and unbound points connected with an arrow in each case showing that bound sites are generally more constrained and less diverse. (**e**) Derived allele frequency spectrum for primate-specific elements, with variations outside ENCODE elements in black and variations covered by ENCODE elements in red. The increase in low-frequency alleles compared to background is indicative of negative selection occurring in the set of variants annotated by the ENCODE data. (**f**) Aggregation of mammalian constraint scores over the glucocorticoid receptor (GR) transcription factor motif in bound sites, showing the expected correlation with the information content of bases in the motif. An interactive version of this figure is available in the online version of the paper.

We proposed to express the deleterious effect of TFBS mutations in terms of *mutational load*, a known population genetics metric that combines the frequency of mutation with predicted phenotypic consequences that it causes[31, 32] (see Materials and methods for details). We adapted this metric to use the reduction in PWM score associated with a mutation as a crude but computable measure of such phenotypic consequences […]

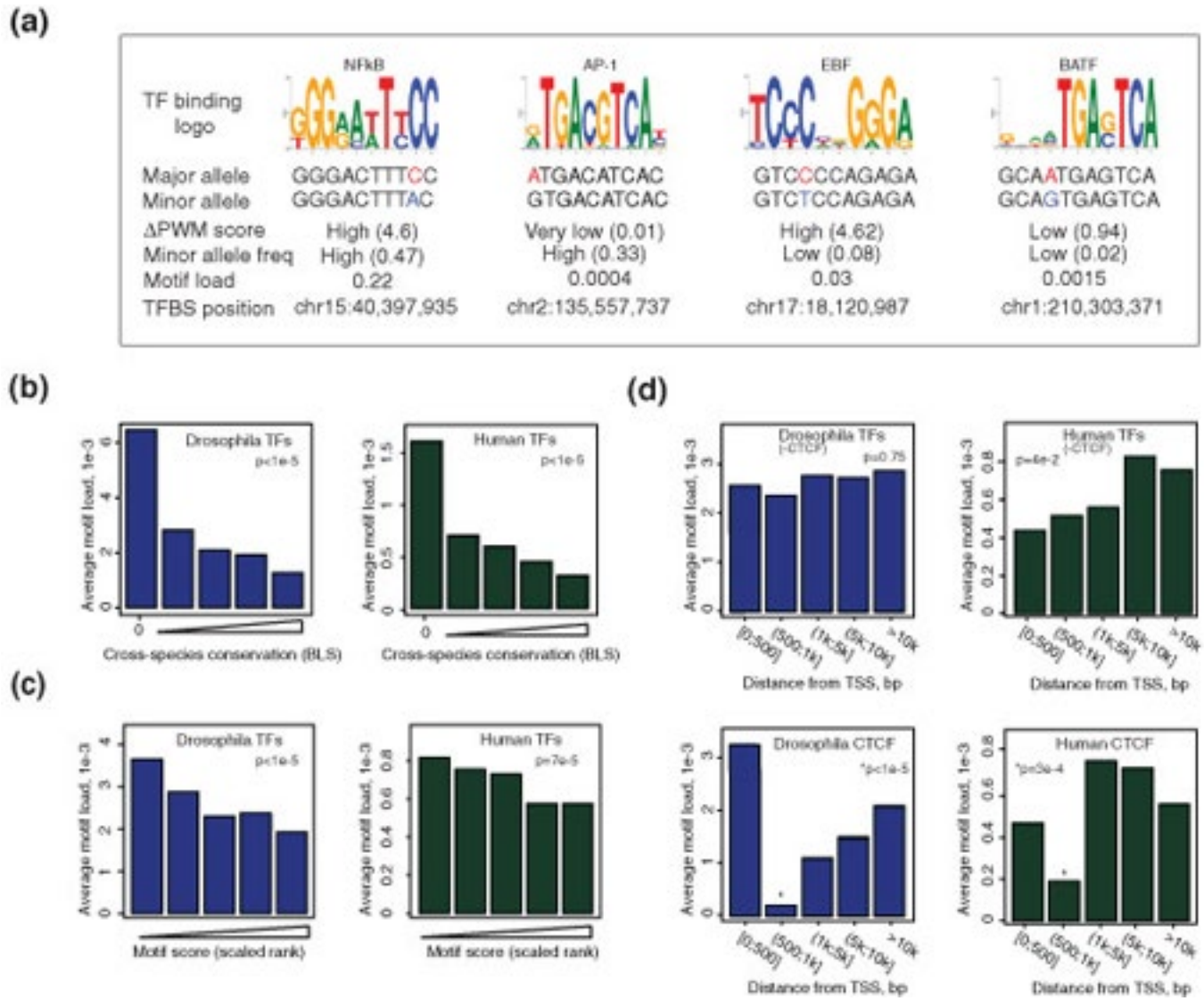


**Figure 3** | **Motif mutational load of *Drosophila* and human TFBSs located within different genomic contexts** (**a**) Examples of mutational load values for individual instances of four human TFs, (ranging from high to very low) showing different combinations of parameters that are combined in this metric: the reduction of Position Weight Matrix match scores at the minor allele ('ΔPWM score') and the number of genotypes within the mutation in the population (MAF; minor allele frequency). (**b**) Relationship between phylogenetic conservation and motif mutational load for *Drosophila melanogaster* (left) and human (right) TFs included in this study. Conservation is expressed as per-instance Branch Length Scores (BLS) for each instance computed against the phylogenetic tree of 12 *Drosophila* species. The average load for *D. melanogaster*-specific sites (BLS=0) is shown separately as these have an exceptionally high motif load. (**c**) Relationship between motif stringency and motif load in *Drosophila* (left) and humans (right). Motif stringency is expressed as scaled ranked PWM scores grouped into five incremental ranges of equal size (left to right), with average motif load shown for each range. (**d**) Relationship between distance from TSS and motif load in *Drosophila* (left) and humans (right) for all analysed TFs excluding CTCF (top) and for CTCF alone (bottom), with average motif load shown for each distance range. **(b-d)** Average motif load is computed excluding a single maximum value to reduce the impact of outliers. The p-values are from permutation tests, in which permutations are performed separately for each TF and combined into a single statistic as described in Materials and methods.

We do not assume that TFBS load at a given site reduces an individual's biological fitness. Rather, we argue that binding sites that tolerate a higher load are less functionally constrained. This approach, although undoubtedly

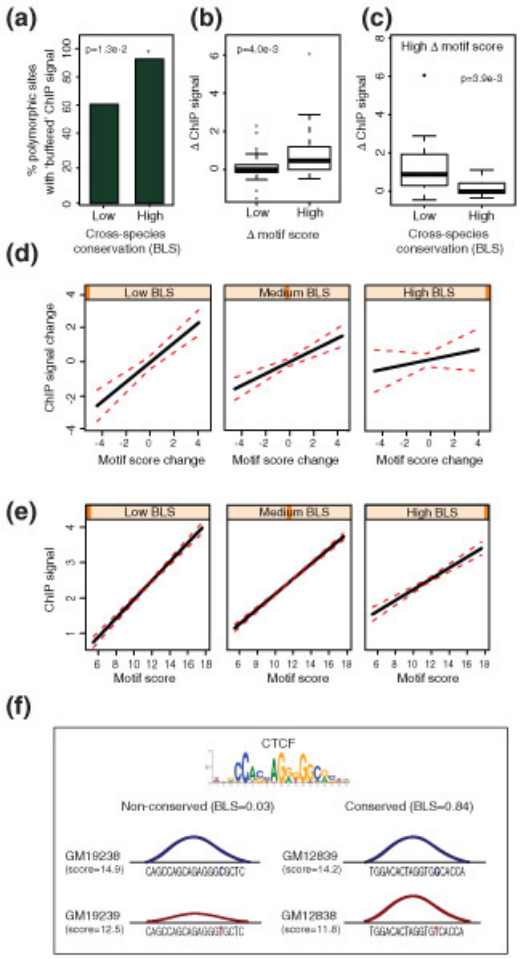


**Figure 5 | Evidence for the 'buffering' of variation at conserved CTCF binding sites. (a) Proportion of homozygous polymorphic CTCF binding sites with 'buffered' levels of ChIP signal depending on the sites' evolutionary conservation (less conserved: Branch Length Score $<$ 0.5, more conserved: Branch Length Score $\geq$ 0.5). Sites at which the minor variant retained at least two thirds of the major variant's signal were considered as 'buffered'. The p-value is from the Fisher test. Major and minor variants were defined on the basis of the global allele frequency data from [75,76]. (b) Differences in the CTCF binding signal ($\Delta$ ChIP signal) at homozygous polymorphic sites that show either 'low' (left) or 'high' (right) disparity in absolute motif match scores ($\Delta$ motif score) between the variants ($<$ 1 or $>$ 1, respectively). The ChIP signals are sign-adjusted relative to the direction of PWM score change. Site-specific signals from multiple individuals with the same genotype, where available, were summarised by mean. The p-value is from the Wilcoxon test. (c) Genotype-specific differences in the CTCF ChIP signal across individuals between homozygous polymorphic sites with appreciable differences in absolute PWM match scores ($\Delta$ motif score $>$ 1) at less conserved (Branch Length Score $<$ 0.5, left) and more conserved (Branch Length Score $>$ 0.5, right) CTCF motifs. The ChIP signals are sign-adjusted relative to the direction of PWM score change. Site-specific signals from multiple individuals with the same variant, where available, were summarised by mean. The p-value is from the Wilcoxon test. (d) An interaction linear model showing that interspecies motif conservation (expressed by Branch Length Scores) reduces the effect of motif mutations on CTCF binding. Shown are the effect plots predicting the relationship between the change of PWM score (at the minor versus the major variant) and the change of the associated ChIP signal at three hypothetical levels of evolutionary conservation: Branch Length Score (BLS)=0 (low; left), BLS=0.5 (medium; middle) and BLS=1 (high; right). Major and minor variants were defined on the basis of the global allele frequency data from [75,76]. (e) An interaction linear model showing that interspecies motif conservation (Branch Length Score) reduces the effect of motif stringency on the binding signal. Shown are the effect plots predicting the relationship between motif scores and ranked ChIP signal at three hypothetical conservation levels: BLS=0 (low; left), BLS=0.5 (medium; middle) and BLS=1 (high; right). (f) A schematic illustrating the observed effect of binding site mutations on CTCF binding signal at two polymorphic CTCF sites - one poorly conserved (BLS=0.03, left) and one highly conserved (BLS=0.84, right) - that have similar motif match scores (14.9 and 14.2, respectively). Sequences of higher- (top) and lower-scoring alleles (bottom) are shown on the figure. Mutations resulting in a similar loss of score (down to 12.5 and 11.8, respectively) resulted in a 53% loss of CTCF binding signal at the non-conserved site (left, compare the amplitudes of top [blue] to bottom [red] curves), in contrast to a mere 6% at the conserved site (right).**

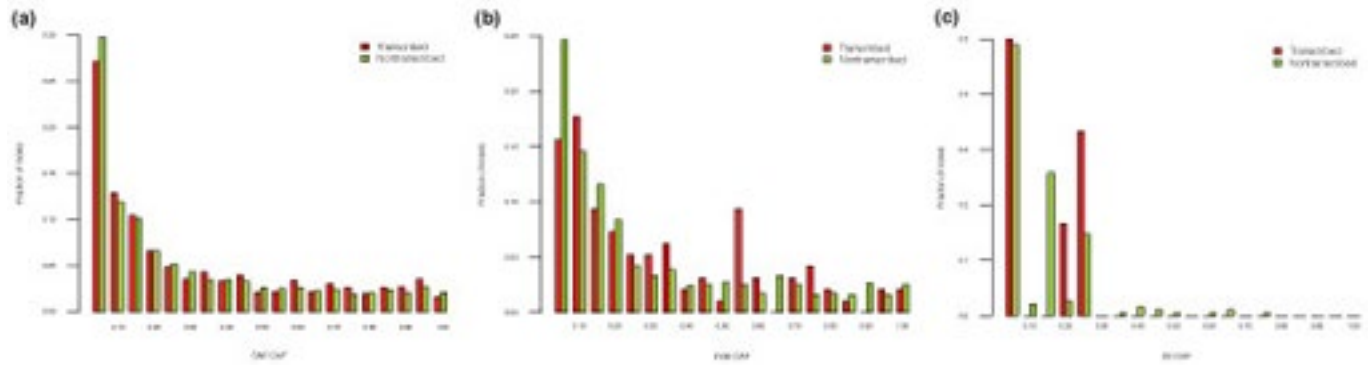


**Figure 7 | SNP and indel DAF spectra are shown for transcribed and non-transcribed pseudogenes.** Cumulative plots for variant DAFs in pseudogenes are also shown. The distribution of variant DAFs in transcribed and non-transcribed pseudogenes are not statistically different.

a crude one, makes it possible to consistently estimate TFBS constraints for different TFs and even different organisms and ask why TFBS mutations are tolerated differently in different contexts.

We first asked whether motif load would be able to detect the expected link between evolutionary and individual variation. We used a published metric, Branch Length Score (BLS)[40], to characterise the evolutionary conservation of a motif instance. This metric utilises both a PWM based model of the conservation of bases and allows for motif movement. Reassuringly, mutational load correlated with BLS in both species, with evolutionary non-conserved motifs (BLS=0) showing by far the highest degree of variation in the population (Figure 3B). At the same time, ~40% of human and fly TFBSs with an appreciable load (L>5e-3) still mapped to reasonably conserved sites (BLS>0.2, ~50% percentile in both organisms), demonstrating that score-reducing mutations at evolutionary preserved sequences can be tolerated in these populations.

Using this metric, we confirmed our original findings, suggesting that TFBSs with higher PWM scores are generally more functionally constrained compared to 'weaker' sites (Figure 3C). The fraction of detected sites mapping to bound regions remained similar across the whole analysed score range, suggesting that this relationship is unlikely to be an artefact of higher false-positive rates at 'weaker' sites (Additional file 1: Figure S4A). This global observation, however, does not rule out the possibility that a weaker match at some sites is specifically preserved to ensure dose-specific TF binding. This may be the case, for example, for Drosophila Bric-à-brac motifs, which exhibited no correlation between motif load and PWM score (Additional file 1: Figure S4B), consistent with the known dosage-dependent function of Bric-à-brac in embryo patterning[41].

We then used motif load to address whether TFBSs proximal to transcription start sites (TSS) are more constrained compared to more distant regulatory regions. We found this to be the case in the human, but not Drosophila (Figure 3D; see Discussion). CTCF binding sites in both species were a notable exception, tolerating the lowest mutational load at locations 500bp-1kb from TSS, but not closer to the TSS (Figure 3D, bottom panel), suggesting that the putative role of CTCF in establishing chromatin domains[42] is particularly important in proximity of gene promoters.

To gain further insight into the functional effects of TFBS mutations, we used a dataset that mapped human CTCF binding sites across four individuals from[16] (see Materials and methods for more details). TFBS mutations detected in this dataset often did not result in a significant loss of binding, with ~75% mutated sites retaining at least two thirds of the binding signal. This was particularly prominent at conserved sites (BLS>0.5), 90% of which showed this 'buffering' effect (Figure 5A). To address whether buffering could be explained solely by the flexibility of CTCF sequence preferences, we analysed between-allele differences in the PWM score at polymorphic binding sites. As expected, globally CTCF binding signal correlated with the PWM score of the

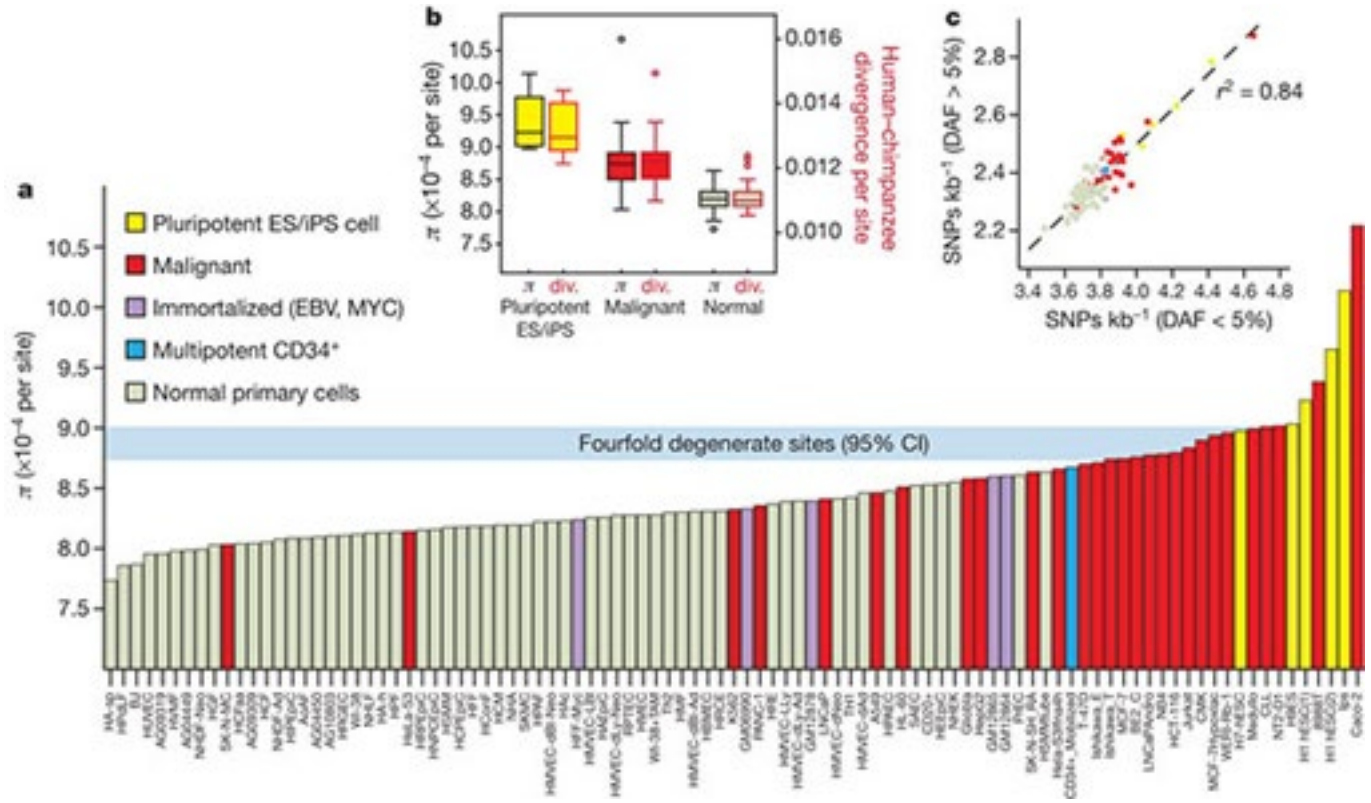


**Figure 7** | **Genetic variation in regulatory DNA linked to mutation rate.** (**a**) Mean nucleotide diversity ($\pi$, *y* axis) in DHSs of 97 diverse cell types (*x* axis) estimated using whole-genome sequencing data from 53 unrelated individuals. Cell types are ordered left-to-right by increasing mean $\pi$. Horizontal blue bar shows 95% confidence intervals on mean $\pi$ 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 $\pi$ (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**.

underlying motifs (Additional file 1: Figure S6A). Consistent with this, alleles with minor differences in PWM match generally had little effect on the binding signal compared to sites with larger PWM score changes (Figure 5B), suggesting that the PWM model adequately describes the functional constraints of CTCF binding sites. At the same time, we found that CTCF binding signals could be maintained even in those cases, where mutations resulted in significant changes of PWM score, particularly at evolutionary conserved sites (Figure 5C). A linear interaction model confirmed that the effect of motif mutations on CTCF binding was significantly reduced with increasing conservation (Figure 5D; interaction term p=2.9e-2). These effects were not due to the presence of additional CTCF motifs (as 96% of bound regions only contained a single motif), while differences between more and less conserved sites could not be explained away by differences in the PWM scores of their major alleles (not shown). A CTCF dataset from three additional individuals generated by a different laboratory[44] yielded consistent conclusions (Additional file 1: Figure S6BCD), suggesting that our observations were not due to overfitting.

Taken together, CTCF binding data for multiple individuals show that mutations can be buffered to maintain the levels of binding signal, particularly at highly conserved sites, and this effect cannot be explained solely by the flexibility of CTCF's sequence consensus. We asked whether mechanisms potentially accountable for such buffering would also affect the relationship between sequence and binding in the absence of mutations. Training

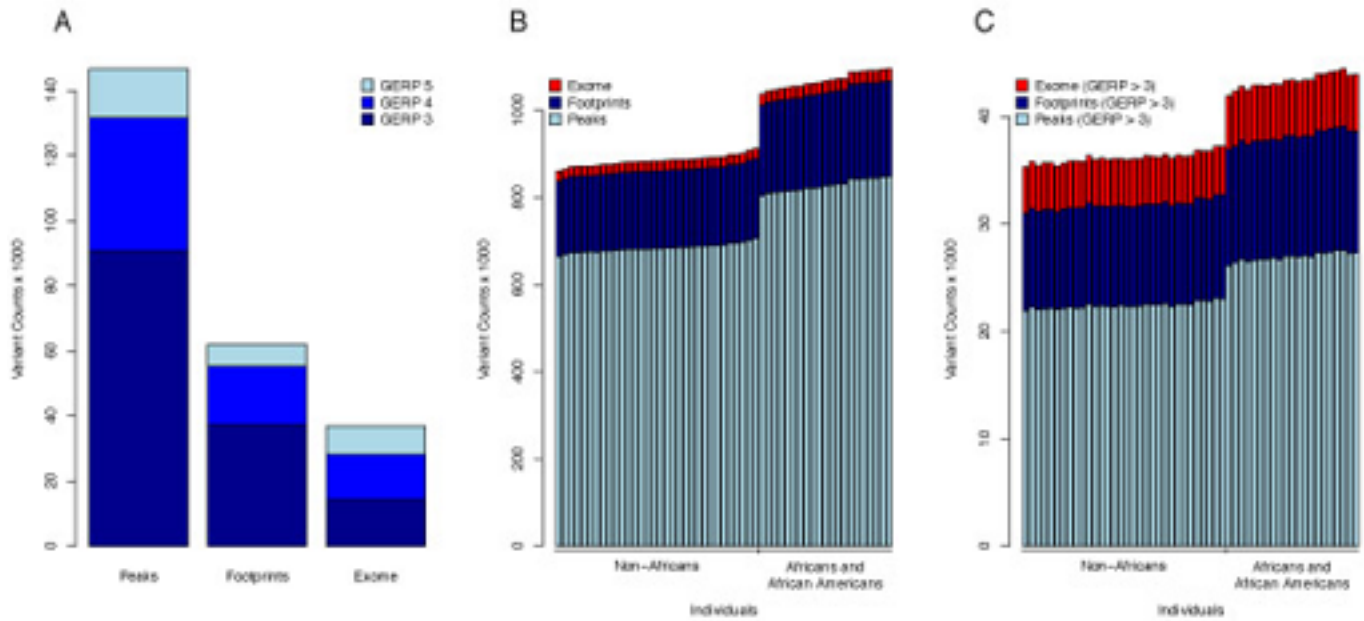


**Figure 2** | **Characteristics of regulatory variation among individuals.** (**a**) Total number of variants in DNase I peaks, footprints, and the exome stratified by GERP score. (**b**) Distribution of the number of variants per individual in DNase I peaks, footprints, and the exome. (**c**) Distribution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

an interaction linear model across the whole set of mapped CTCF binding sites revealed that conservation consistently weakens the relationship between PWM score and the binding intensity (p=1.9e-7; Figure 5E). Thus, CTCF binding to evolutionary conserved sites may generally have a reduced dependence on sequence.

**Selection**

To address this, we first analyzed the selective pressure on both TFs and their targets. We predominantly used non-synonymous SNP density from the 1000 Genomes Pilot[21] to determine selection amongst modern-day humans (SOM/J). We also verified our results using other measures of selection (i.e. derived allele frequency (DAF) and the pN/pS statistic (SOM/J). For selection over longer time scales, we calculated the ratio of non-synonymous to synonymous substitution in human-chimp ortholog alignments (dN/dS). We find significant negative correlation between the regulatory in-degree of target genes and both their non-synonymous SNP density and dN/dS values (Tables 1 and S6e). Thus, target genes regulated by more TFs are under stronger negative selection. Similarly, we find that there is a significant negative correlation between TF regulatory out-degree and non-synonymous SNP density (Tables 1 and S6d). We observe a consistent result with TF dN/dS values and other measures of selection, although these are not all as statistically significant (Table S6d and SOM/J). This shows that TFs regulating more targets tend to be under stronger negative selection. Moreover, within the TF hierarchy, we find that TFs at the top are under significantly stronger negative selection (Fig. 2c, Tables 1 and S6b).

Consistent with all these results relating connectivity with constraint, we find that genes tolerant of loss-of-function mutations[46] are under weaker negative selection and have a significantly lower total degree (I+O) than other genes (SOM/J).

**Selection and Allelic Effects.**

Finally, we attempted to relate selection and allelic effects. We extracted TF binding peaks in promoters and gene bodies showing ASB, and compared the selective pressure in these against a control (binding peaks

within the same regions without ASB). We find that TF-binding peaks exhibiting allelic effects have higher SNP densities relative to the control (Fig. 5b). Moreover, binding peaks with no allelic effects show a skew in the DAF spectrum toward rarer SNPs, relative to ASB ones (Fig. 5b and S10c). The same trend holds true for indels and structural variations (Figs. 5b and S10b,c). Interestingly, these results indicate that allelic regulation appears to be under less selective constraint.

We studied the evolutionary selection on human pseudogenes by integrating the annotation with the variation data from the 1000 Genomes pilot project[47]. We computed the densities of SNPs, indels and structural variations (SVs) in pseudogene sequences and their respective derived allele frequencies (DAFs). The densities suggested a weak signal for differential selection on transcribed versus non-transcribed pseudogenes (Additional file 1, Figure s6). However, no significant differences were found in the DAF spectra (Fig. 7), and it is possible that the difference in the densities may be due to confounding factors such as variation in mutation rates in the genome. Thus, we cannot make a strong statement about selection in the human population on transcribed pseudogenes.

Next we analysed the pseudogenes' divergence using sequence identity to orthologs in the chimpanzee genome, where higher sequence identity implies lower divergence and negative selection. The distribution of pseudogenes' divergence was calculated and the results indicate that a fraction of the pseudogenes exhibiting lower divergence are under evolutionary constraint (Additional file 1, Figure s5).

Divergence and diversity results indicate that although pseudogenes, as a group, are under low selection pressure, a small subset may exhibit higher evolutionary constraint. To identify these pseudogenes, we analyzed the divergence to orthologs in the chimp and the mouse genome under the assumption that the conserved pseudogenes will show significantly lower divergence than neutral background (see Method). There are 1,019 conserved pseudogenes identified in the human genome. The conserved group is enriched with transcribed pseudogenes (195 conserved pseudogenes are transcribed, p-value $= 1.19 \times 10^{-35}$), strongly implying biological function.

We found that 2.0% of footprints localized within exons, raising the possibility that occupancy by DNA binding proteins could further restrict sequence diversity within coding DNA, thus superimposing an unexpected layer of constraint on codon usage.

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 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

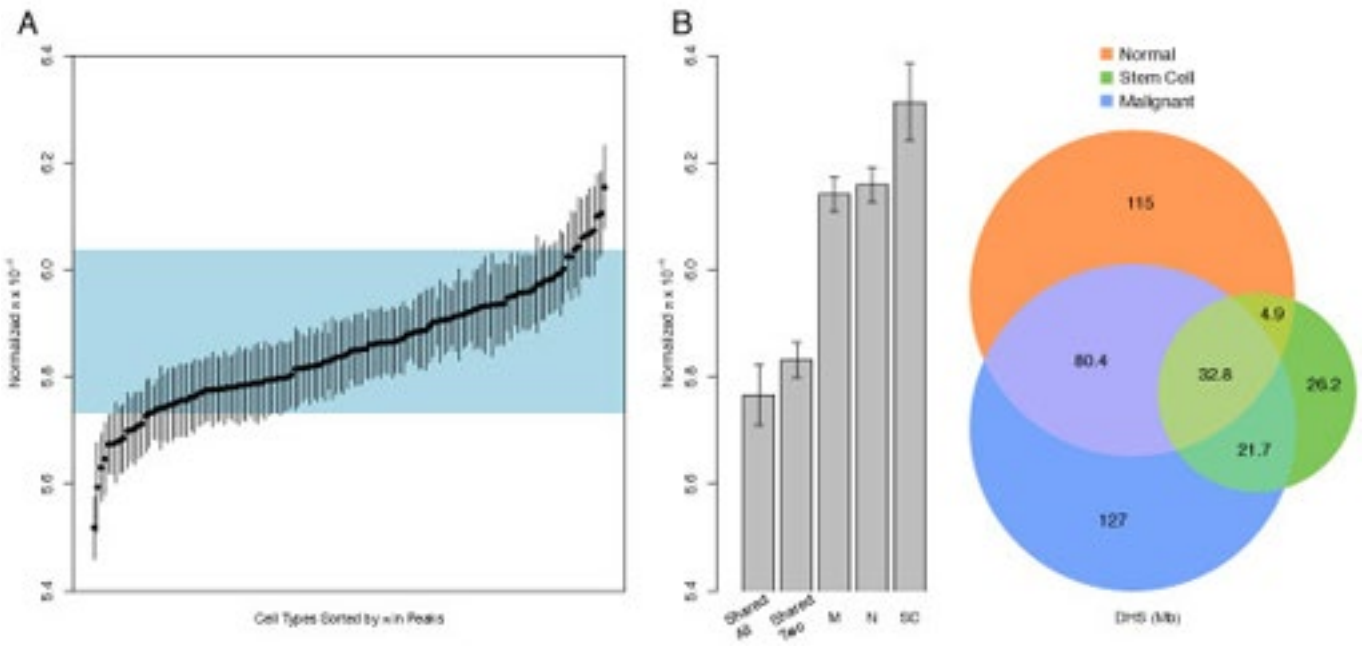


**Figure 4 | Heterogeneity of polymorphism across cell types. (b) Distribution of normalized nucleotide diversity (black points) across DNase I peaks in 138 cell types. Vertical bars around peaks indicate 95% confidence intervals obtained by bootstrapping. (Blue rectangle) 95% confidence interval for normalized nucleotide diversity in fourfold degenerate sites. (b) Venn diagram showing the amount of shared and unique sequence for DNase I peaks among normal/primary, malignant, and iPS/ES cell types. The barplot on the *left* shows average normalized diversity for several categories of peaks in the Venn diagram. Shared all and shared two denote peaks shared among all three categories and between any two categories, respectively. N, M, and SC denotes peaks specific to normal/primary, malignant, and stem cell (iPS/ES) cell types, respectively.**

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.

On average, individuals contain 24.2k $\pm$ 2.3k, 10.1k $\pm$ 0.92k, and 4.7 k$\pm$ 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.

We next tested the hypothesis that levels of functional constraint acting on regulatory DNA varied across cell types. To this end, we calculated normalized $\pi$ averaged across all DNase I peaks for each of the 138 cell lines (normalized to mutation rate - see Methods). We found marked differences in normalized diversity between cell lines ($P < 10$^-4) (Fig. 4A), which ranged from a low of 5.52 $\times$ 10^-4 in primary hepatocytes to a high of 6.15 $\times$ 10^-4 in the immortalized B-lymphoblastoid cell line GM12864. The majority of cell types exhibited average levels of normalized diversity that are within the range of fourfold degenerate sites (Fig. 4A). Note, as we are

averaging over many megabases of sequence in each cell type, this does not mean that specific sites, such as motifs embedded within peaks, are evolving neutrally. Six cell types (retinal pigment epithelial, neuroblastoma, primary liver, skeletal muscle myoblast, umbilical vein endothelial, and prostate adenocarcinoma cells, corresponding to cell lines HRPEpiC, SK-N-SH, Hepatocyte, Hsmm, Huvec, LNCaP, respectively) exhibited average levels of normalized diversity that are significantly lower (ranging from $P = 0.024$ to $P < 10^{-4}$) than fourfold degenerate sites, indicative of stronger functional constraint.

We next investigated differences in normalized diversity between "core" DHS and DHS found in only one category of cell types. To this end, all of the cell types can be grouped into one of three categories: normal/primary, iPS/ES, and malignant. To minimize potential contributions from experimental noise, we focused on a subset of 92 cell types with high-quality DNase I data in which >40% all sequence tags map within DHSs (equivalent to average signal-to-noise of 100:1) (Thurman *et al.* 2011) and calculated normalized $\pi$ in DNase I peaks that are shared and unique to each cell type category (Fig. 4B). Eight percent of peaks are found in all three categories, whereas 6.4%, 31.1%, and 28.2% of peaks are unique to iPS/ES, malignant, and normal/primary cell types, respectively (Fig. 4B). Overall, there is significant variation ($P < 10^{-4}$) in normalized diversity among peaks shared between cell type categories versus those found in a single category (Fig. 4B). In particular, DNase I peaks shared by two or three categories of cell types exhibit the lowest levels of normalized diversity (Fig. 4B), consistent with stronger selective constraint. Conversely, peaks found in only one cell category have significantly higher normalized diversity than shared peaks, (Fig. 4B). These results suggest that the "core" set of DHSs, present in more than one cell type category, is subject to stronger purifying selection selective pressure because they are necessary for proper transcriptional programs in multiple cell types.

**Signatures of Positive Selection**

The large compendium of experimentally characterized regulatory regions provides a unique data set to interrogate for signatures of recent positive selection. To this end, we performed a population genomics analysis to identify DNase I peaks that contain variants with large allele frequency differences between populations relative to the genome at-large, which is a signature of geographically restricted selection (Akey *et al.* 2002; Akey *et al.* 2004). Specifically, we calculated locus-specific branch lengths (LSBLs) (Shriver *et al.* 2004) for variants in DNase I peaks in Africans, Asians, and Europeans. LSBL is a function of pairwise $F_{ST}$ between populations (see Methods) and helps isolate the direction of allele frequency change (Shriver *et al.* 2004). To reduce the stochasticity inherent in summary statistics of population differentiation, we averaged LSBL across all variants in a peak.

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 ($n = 3372$, 3224, and 3099 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).

Second, to develop a more refined list of putative targets of recent adaptive evolution, we focused on the most differentiated 1% of DHSs that also contain one or more highly differentiated variants with a GERP > 3. In total, 323, 349, and 313 DHSs meet these criteria in Africans, Asians, and Europeans, respectively. We identified

genes located within 50 kb of each of these peaks and identified 187, 174, and 179 genes in Africans, Asians, and Europeans, respectively (Supplemental Text 1). Notably, included in this set of peaks is the well-documented promoter variant in *DARC* that results in malaria resistance in African populations (Hamblin *et al.* 2002), which demonstrates the potential power of this data set to fine-scale map signatures of selection and identify selected alleles. Moreover, 61, 40, and 51 of these candidate selection genes in Africans, Asians, and Europeans, respectively, overlap previously reported signatures of selection collected by Akey (2009), which is significantly more than expected by chance ($P < 10^{-6}$ for all populations). Thus, these observations suggest that the loci identified here are enriched for targets of recent positive selection. Particularly interesting examples of novel targets of selection include the vitamin D receptor (*VDR*) in Africans and the fat mass and obesity associated gene (*FTO*) in Europeans (Figs. 6C, 5D). The list of all candidate selection genes located within 50 kb of highly differentiated peaks is provided in Supplemental Text 1, which provides a powerful framework for more detailed analyses into recent adaptation of noncoding DNA in humans.

Similarly strong correlations between footprint occupancy and either ChIP-seq signal or phylogenetic conservation were evident for diverse factors (Fig. 1d and Supplementary Figs 4a-d). We found that footprint occupancy and nucleotide-level conservation correlated for 80% of all transcription factor motifs in the TRANSFAC database, of which 50% were statistically significant ($P < 0.05$; Supplementary Methods). This relationship between footprint occupancy and conservation is most readily explained by evolutionary selection on factor occupancy, with higher conservation of higher affinity binding sites.

We next asked how these patterns related to evolutionary conservation. Plotting nucleotide-level aggregate DNaseI cleavage in parallel with per-nucleotide vertebrate conservation calculated by phyloP[19] revealed striking antiparallel patterning of cleavage versus conservation across nearly all motifs examined (six representative examples are shown in Fig. 3b and Supplementary Fig. 8b). Notably, conservation is not limited to only DNA contacting protein residues, but exhibits graded changes that mirror DNaseI accessibility across the entirety of the protein-DNA interface (Supplementary Figs 8c, d). Taken together, these results imply that regulatory DNA sequences have evolved to fit the continuous morphology of the transcription factor-DNA binding interface.

To test whether novel motifs were functionally conserved in an evolutionarily distant mammal, we analysed DNaseI cleavage patterns around human novel motifs mapped within DHSs assayed in primary mouse liver tissue (Fig. 6e, f and Supplementary Fig. 15a, b). This analysis demonstrated that many novel motifs show nearly identical DNaseI footprint patterns in both human cells and mouse liver, indicating that these novel motifs correspond to evolutionarily conserved transcriptional regulators that are functional in both mice and human.

Given the conservation of protein occupancy in a distant mammal, we assessed whether the novel motifs are under selection in human populations by analysing nucleotide diversity across all motif instances found within accessible chromatin. Using high-quality genomic sequence data from 53 unrelated individuals[34] (Supplementary Table 4), we calculated the average nucleotide diversity[35] for each individual motif space (Supplementary Fig. 15c). Reduced diversity levels are indicative of functional constraint, through the elimination of deleterious alleles from the population by natural selection. We found that novel motifs are collectively under strong purifying selection in human populations. On average, the new motifs are more constrained than most motifs found in the major databases (Fig. 6d and Supplementary Fig. 15c), even after exclusion of motifs containing highly mutable CpG dinucleotides, which underlie the marked increase in nucleotide diversity seen with a subset of known motifs (Supplementary Fig. 15c, right). Collectively, these results demonstrate that DNaseI footprints encode an expansive *cis*-regulatory lexicon encompassing both known transcription factor recognition sequences and novel motifs that are functionally conserved in mouse and bear strong signatures of ongoing selection in humans.