An expansive human regulatory lexicon encoded in transcription factor footprints

Shane Neph[1]*, Jeff Vierstra[1]*, Andrew B. Stergachis[1]*, Alex P. Reynolds[1]*, Eric Haugen[1], Benjamin Vernot[1], Robert E. Thurman[1], Richard Sandstrom[1], Audra K. Johnson[1], Matthew T. Maurano[1], Richard Humbert[1], Eric Rynes[1], Hao Wang[1], Shinny Vong[1], Kristen Lee[1], Daniel Bates[1], Morgan Diegel[1], Vaughn Roach[1], Douglas Dunn[1], Jun Neri[1], Anthony Schafer[1], R. Scott Hansen[1,2], Tanya Kutyavin[1], Erika Giste[1], Molly Weaver[1], Theresa Canfield[1], Peter Sabo[1], Miaohua Zhang[3], Gayathri Balasundaram[3], Rachel Byron[3], Michael J. MacCoss[1], Joshua M. Akey[1], Michael Bender[3], Mark Groudine[3], Rajinder Kaul[1,2], John A. Stamatoyannopoulos[1,4,#]

[1]*Department of Genome Sciences, University of Washington, Seattle, WA 98195*

[2]*Division of Medical Genetics, Department of Medicine, University of Washington, Seattle, WA 98195*

[3]*Basic Sciences Division, Fred Hutchison Cancer Research Center, Seattle, WA 98109*

[4]*Division of Oncology, Deparment. of Medicine, University of Washington, Seattle, WA 98195*

**Equal contributors*

[#]*Correspondence: jstam@uw.edu*

Keywords: chromatin, protein occupancy, DNaseI footprinting, ENCODE, regulation

Abstract

Regulatory factor binding to genomic DNA protects the underlying sequence from cleavage by DNaseI, leaving nucleotide-resolution footprints. Using genomic DNaseI footprinting across 41 diverse cell and tissue types, we detected 45 million factor occupancy events within regulatory regions, representing differential binding to 8.4 million distinct short sequence elements. Here we show that this small genomic sequence compartment, roughly twice the size of the exome, encodes an expansive repertoire of conserved recognition sequences for DNA-binding proteins that nearly doubles the size of the human *cis*-regulatory lexicon. We find that genetic variants affecting allelic chromatin states are concentrated in footprints, and that these elements are preferentially sheltered from DNA methylation. High-resolution DNaseI cleavage patterns mirror nucleotide-level evolutionary conservation and track the crystallographic topography of protein-DNA interfaces, indicating that transcription factor structure has been evolutionarily imprinted on the human genome sequence. We identify a stereotyped 50 base-pair footprint that precisely defines the site of transcript origination within thousands of human promoters. Finally, we describe a large collection of novel regulatory factor recognition motifs that are highly conserved in both sequence and function, and exhibit cell-selective occupancy patterns that closely parallel major regulators of development, differentiation, and pluripotency.

Introduction

Sequence-specific transcription factors (TFs) interpret the signals encoded within regulatory DNA. The discovery of DNaseI footprinting over 30 years ago[1] revolutionized the analysis of *cis*-regulatory sequences in diverse organisms, and directly enabled the discovery of the first human sequence-specific transcription factors[2]. Binding of TFs to regulatory DNA regions in place of canonical nucleosomes triggers chromatin remodeling, resulting in nuclease hypersensitivity[3]. Within DNaseI hypersensitive sites (DHSs), DNaseI cleavage is not uniform; rather, punctuated binding by sequence-specific regulatory factors occludes bound DNA from cleavage, leaving 'footprints' that demarcate TF occupancy at nucleotide resolution[1,4] (**Figure 1a**). DNaseI footprinting has been applied widely to study the dynamics of transcription factor occupancy and cooperativity within regulatory DNA regions of individual genes[5], and to identify cell- and lineage-selective transcriptional regulators[6].

Regulatory DNA is densely populated with DNaseI footprints

To map DNaseI footprints comprehensively within regulatory DNA, we adapted digital genomic footprinting[4] to human cells. The ability to resolve DNaseI footprints sensitively and precisely is critically dependent on the local density of mapped DNaseI cleavages (**Supplementary Figs. 1a-d**), and efficient footprinting of a large genome such as human requires substantial concentration of DNaseI cleavages within the small fraction (~1-3%) of the genome contained in DNaseI-hypersensitive regions. We selected highly enriched DNaseI cleavage libraries from 41 diverse cell types in which 53-81% of DNaseI cleavage sites localized to DNaseI-hypersensitive regions[7] (**Supplementary Table 1**), representing nearly 10-fold higher signal-to-noise ratio vs. prior results from yeast[4], and 2- to 5-fold greater enrichment than achieved using end-capture of single DNaseI cleavages[8,9]. We then performed deep sequencing of these libraries, and obtained 14.9 billion Illumina sequence reads, 11.2 billion of which mapped to unique locations in the human genome (**Supplementary Table 1**). We achieved an average sequencing depth of ~273 million DNaseI cleavages per cell type that enabled extensive and accurate discrimination of DNaseI footprints.

To detect DNaseI footprints systematically, we implemented a detection algorithm based on the original description of quantitative DNaseI footprinting[1] (**Supplementary Methods**). We identified an average of ~1.1 million high-confidence (FDR 1%) footprints per cell type (range 434,000 to 2.3 million; **Supplementary Table 1**), and collectively 45,096,726 6-40 bp footprint events across all cell types. We resolved cell-selective footprint patterns to reveal 8.4 million distinct footprinted elements, each occupied in one or more cell types. At least one footprint was found in >75% of DHSs (**Supplementary Figs. 1c,d** and **Supplementary Table 2**), with detection strongly dependent on the number of mapped DNaseI cleavages within each DHS. 99.8% of DHSs with >250 mapped DNaseI cleavages contained at least one footprint, indicating that DHSs are not simply open or nucleosome-free chromatin features, but are constitutively populated with DNaseI footprints. Modeling DNaseI cleavage patterns using empirically derived intrinsic DNA cleavage propensities for DNaseI showed that only a miniscule fraction (0.24%) of discovered FDR 1% footprints from cell and tissue samples could be caused by inherent DNaseI sequence specificity (**Supplementary Methods**).

DNaseI footprints were distributed throughout the genome, including intergenic regions (45.7%), introns (37.7%), upstream of transcriptional start sites (8.9%), and in 5' and 3' UTRs (1.4% and 1.3%, respectively; **Supplementary Figs. 2a,b**). DNaseI footprints were enriched in

promoters (3.6 fold; $P < 2.2x10^{-16}$; Binomial test) and 5' UTRs (2.4 fold; $P < 2.2x10^{-16}$; Binomial test), commensurate with high DNaseI cleavage densities observed in these regions. 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.

Quantitative markers of *in vivo* regulatory factor occupancy

We next examined the correspondence between DNaseI footprints and known regulatory factor recognition sequences within DNaseI hypersensitive chromatin. Comprehensive scans of DNaseI hypersensitive regions for high confidence matches to all recognized TF motifs in the TRANSFAC[10] and JASPAR[11] databases revealed striking enrichment of motifs within footprints ($P \approx 0$, Z-score = 204.22 for TRANSFAC; Z-score = 169.88 for JASPAR; **Fig. 1b** and **Supplementary Fig. 3**).

To quantify the occupancy at TF recognition sequences within DHSs genome-wide, we computed for each instance a footprint occupancy score (FOS) relating the density of DNaseI cleavages within the core recognition motif to cleavages in the immediately flanking regions (**Supplementary Methods**). The FOS can be used to rank motif instances by the 'depth' of the footprint at that position, and is expected to provide a quantitative measure of factor occupancy[1]. To examine this relationship for a well-studied sequence-specific regulator (NRF1[12]), we plotted DNaseI cleavage patterns surrounding all 4,262 NRF1 motifs contained within DNaseI hypersensitive sites and ranked these by FOS. While only a subset of these motif instances (2,351) coincided with high-confidence footprints, the vast majority of NRF1 motif instances in DNaseI footprints (89%) overlapped reproducible NRF1 ChIP-seq peaks (**Fig. 1c**). In parallel, we analyzed nucleotide-level evolutionary conservation patterns around NRF1 binding sites, revealing that FOS closely parallels phylogenetic conservation within the core motif region, suggesting strong selection on factor occupancy (**Fig. 1c**). We observed a nearly monotonic relationship between FOS and ChIP-seq signal intensities at NRF1 binding sites within K562 DNaseI footprints (**Fig. 1d**). 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 footprint occupancy and nucleotide-level conservation correlated for 80% of all TF 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. Taken together, these results indicate that footprint occupancy provides a quantitative measure of sequence-specific regulatory factor occupancy that closely parallels evolutionary constraint and ChIP-seq signal intensity.

To validate the potential for selective binding of footprints by factors predicted on the basis of motif-to-footprint matching, we developed an approach to quantify specific occupancy in the context of a complex TF milieu using targeted mass spectrometry (DNA interacting protein precipitation or DIPP; Methods). Using DIPP, we affirmed specific binding by several different classes of TFs (**Supplementary Figs. 5a-e**). Together with the analysis of ChIP-seq data described above, these results indicate that the localization of TF recognition motifs within DNaseI footprints can accurately illuminate the genomic protein occupancy landscape.

Footprints harbor functional variants and are sheltered from DNA methylation

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 harboring 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 analyzed their distribution relative to DNaseI footprints. This analysis revealed significant enrichment ($P < 2.2\text{x}10^{-16}$; Fisher's exact test) of such variants within DNaseI footprints (**Supplementary Fig. 6**). For example, rs4144593 is a common 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**).

Protein-DNA interactions are also sensitive to cytosine methylation[15,16]. Comparing DNaseI footprints and whole genome bisulfite sequencing methylation data from pulmonary fibroblasts (IMR90), we found that CpG dinucleotides contained within DNaseI footprints were significantly less methylated than CpGs in non-footprinted regions of the same DHS (Mann-Whitney test; $P < 2.2\text{x}10^{-16}$; **Fig. 2b**). Footprints therefore appear to be selectively sheltered from DNA methylation, suggesting a widespread connection between regulatory factor occupancy and nucleotide-level patterning of epigenetic modifications.

Transcription factor structure is imprinted on the human genome

We observed surprisingly heterogeneous base-to-base variation in DNaseI cleavage rates within the footprinted recognition sequences of different regulatory factors. And yet, the per site cleavage profiles for individual factors were highly stereotyped, with nearly identical local cleavage patterns at thousands of genomic locations (**Supplementary Fig. 7**). This raised the possibility that DNaseI cleavage patterns may provide information concerning the morphology of the DNA-protein interface. We obtained the available DNA-protein co-crystal structures for human transcription factors, and mapped aggregate DNaseI cleavage patterns at individual nucleotide positions onto the DNA backbone of the co-crystal model. **Fig. 3a** and **Supplementary Fig. 8a** show two examples, USF[17] and SRF[18]. For both factors, DNaseI cleavage patterns clearly parallel the topology of the protein-DNA interface, including a marked depression in DNaseI cleavage at nucleotides involved in protein-DNA contact, and increased cleavage at exposed nucleotides such as those within the central pocket of the leucine zipper. These data show that nucleotide-level aggregate DNaseI cleavage patterns reflect fundamental features of the protein-DNA interaction interface at unprecedented resolution.

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 vs. conservation across nearly all motifs examined (six representative examples are shown in **Fig. 3b** and **Supplementary Fig. 8b**). Surprisingly, 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.

A stereotyped 50 bp footprint localizes transcription initiation within promoters

Transcription initiation requires the binding of multi-protein complexes that position RNA polymerase II (PolII)[20-23]. Using a modified footprint detection algorithm designed to detect larger features (**Supplementary Methods**), we scanned the regions upstream from Gencode transcriptional start sites (TSSs) and identified highly stereotyped ~80bp chromatin structure comprising a prominent ~50 bp central DNaseI footprint, flanked symmetrically by ~15 bp regions

of uniformly elevated DNaseI cleavage (**Fig. 4a**). Alignment of per-nucleotide DNaseI cleavage profiles from 5,041 prominent footprints mapped in different K562 promoters highlights the homogeneous, nearly invariant nature of the structure (**Fig. 4b**).

Plotting evolutionary conservation in parallel with DNaseI cleavage revealed two distinct peaks in evolutionary conservation within the central footprint (**Fig. 4c)** compatible with binding sites for paired canonical sequence-specific TFs. The density of CAGE tags (**Fig. 4d**; green line) and 5' ends of expressed sequenced tags (ESTs) (**Fig. 4d**; orange line) relative to the central ~50 bp footprint revealed that, at the vast majority of promoters, RNA transcript initiation localized precisely within the stereotyped footprint. It is notable that the location of this footprint is often offset, typically 5', from many Gencode-annotated TSSs. This likely derives from the incomplete nature of many of the 5' transcript ends used to define TSSs[24].

These data together define a new high-resolution chromatin structural signature of transcription initiation and the interaction of the pre-initiation complex (PIC) with the core promoter. Indeed, chromatin occupancy of TATA-binding protein (TBP), a critical component of the PIC, is maximal precisely over the center of the 50bp footprint region (**Supplementary Fig. 9a**). Sequence analysis of the two conservation peaks within the 50bp footprint identified motifs for GC-box-binding proteins such as SP1 and, less frequently, other general transcription factors (though with the notable absence of TATA motifs) (**Supplementary Fig. 9b**), suggesting that TBP (and potentially other PIC components) interact preferentially with general transcriptional factors bound to GC-box-like features in the central footprinted region. The results are therefore consistent with a model in which a limited number of sequence-specific factors function both to prime the chromatin template for recruitment of RNA polymerase II and to guide transcriptional positioning.

Differentiating DNA binding vs. indirect occupancy by TFs

Many transcriptional regulators are posited to interact indirectly with the DNA sequence of some target sites though mechanisms such as tethering[25]. Approaches such as ChIP-seq detect chromatin occupancy, but cannot by themselves distinguish sites of direct DNA binding from non-canonical indirect binding. We therefore asked whether DNaseI footprint data could illuminate ChIP-seq-derived occupancy profiles by differentiating directly bound factors from indirect binding events. We first partitioned ChIP-seq peaks from each of 38 ENCODE transcription factors[26] mapped in K562 cells into three categories of predicted sites: ChIP-seq peaks containing a compatible

footprinted motif (directly bound sites); ChIP-seq peaks lacking a compatible motif or footprint (indirectly bound sites); and ChIP-seq peaks overlying a compatible motif lacking a footprint (indeterminate sites). Predicted indirect sites showed significantly reduced ChIP-seq signal compared with predicted directly bound sites (**Supplementary Fig. 10**), consistent with lack of direct cross-linking to DNA (and therefore reduced ChIP efficiency). Indeterminate sites exhibited low ChIP-seq signal and were therefore excluded from further analysis (**Supplementary Fig. 10**).

The fraction of ChIP-seq peaks predicted to represent direct vs. indirect binding varied widely between different factors, ranging from nearly complete direct sequence-specific binding (e.g., CTCF), to nearly complete indirect binding (e.g., TBP; **Supplementary Fig. 11**). In many cases factors that preferentially engage in direct DNA binding at distal sites show predominantly indirect occupancy in promoter regions and *vice versa* (**Supplementary Figs. 12a,b**),

Next, we analyzed the frequency with which indirectly bound sites of one transcription factor coincided with directly bound sites of a second factor, suggestive of protein-protein interactions (e.g., tethering). This analysis recovered many known protein-protein interactions, such as CTCF/YY1 and TAL1/GATA1[27], as well as many novel associations (**Fig. 5**). We observed enrichment for NFE2 indirect interactions at promoter bound USF2 sites, compatible with their known interaction[28]. At distal sites, we observed the opposite, with NFE2 predominantly directly bound accompanied by USF2 indirect peaks (**Supplementary Figs. 12a,b**), suggesting the possibility of a reciprocal or looping mechanism. Notably, directly bound promoter-predominant transcription factors were enriched for co-localization with indirect peaks compared to distal regions (**Supplementary Figs. 13a,b**). These results suggest that combining DNaseI footprinting with ChIP-seq has the potential to expose a previously unappreciated landscape of complex transcription factor occupancy modes.

Footprints encode an expansive *cis*-regulatory lexicon

Since the discovery of the first sequence-specific transcription factor[29], considerable effort has been devoted to identifying the cognate recognition sequences of DNA-binding proteins[30,31]. Despite these efforts, high-quality motifs are available for only a minority of the >1,400 human transcription factors with predicted sequence-specific DNA binding domains[32].

We reasoned that the genomic sequence compartment defined by DNaseI footprints in a given cell type ideally should contain much, if not all, of the factor recognition sequence information relevant for that cell type. Consequently, applying *de novo* motif discovery to the footprint

compartments gleaned from multiple cell types should greatly expand our current knowledge of biologically active TF-binding motifs.

We performed unbiased *de novo* motif discovery within the footprints identified in each of the 41 cell types that yielded 683 unique motif models (**Fig. 6a** and **Supplementary Methods**). We compared these models with the universe of experimentally-grounded motif models in the TRANSFAC, JASPAR, and UniPROBE[33] databases. Due to the redundancy of motif models contained within these databases, we first collapsed all duplicate models (**Supplementary Methods**). 394 of the 683 (58%) *de novo* motifs matched distinct experimentally-grounded motif models, accounting collectively for 90% of all unique entries across the three databases (**Fig. 6b** and **Supplementary Figs. 14a-c**). The wholesale *de novo* derivation of the vast majority of known regulatory factor recognition sequences from the small genomic compartment defined by DNaseI footprints highlights the dramatic concentration of regulatory information encoded within this sequence space.

Strikingly, 289 of the footprint-derived motifs were absent from major databases (**Fig. 6b** and **Supplementary Fig. 14d**). These novel motifs populate millions of DNaseI footprints (**Fig. 6c**), and show features of *in vivo* occupancy and evolutionary constraint similar to motifs for known regulators, including marked anti-correlation with nucleotide-level vertebrate conservation (**Figs. 6d,3**).

To test whether novel motifs were functionally conserved in a distant mammal, we analyzed DNaseI cleavage patterns around human novel motifs mapped within DHSs assayed in primary mouse liver tissue (**Figs. 6e,f** and **Supplementary Figs. 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 men.

Given the conservation of protein occupancy in a distant mammal, we assessed whether the novel motifs are under selection in human populations by analyzing 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 following 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 TF recognition sequences and novel motifs that are functionally conserved in mouse and bear strong signatures of ongoing selection in humans.

Novel motif occupancy parallels known regulators of pluripotency and cell fate

Cell-selective gene regulation is mediated by the differential occupancy of transcriptional regulatory factors at their cognate *cis*-acting elements. For example, the nerve growth factor gene *VGF* is selectively expressed only within neuronal cells (**Fig. 7a**), presumably due to the repressive action of the transcriptional regulator NRSF/REST at the VGF promoter in non-neuronal cell types[36]. Although *VGF* is expressed only in neuronal cells, its promoter is DNaseI-hypersensitive in most cell types (not shown). Examination of nucleotide-level cleavage patterns within the *VGF* promoter exposes its fundamental *cis*-regulatory logic, coordinated by the transcriptional regulators NRSF, SP1, USF1, and NRF1. Whereas the NRSF motif is tightly occupied in non-neuronal cells, in neuronal cells, NRSF repression is relieved, and recognition sites for the positive regulators USF1 and SP1 become highly occupied, resulting in *VGF* expression. These data collectively illustrate the power of genomic footprinting to resolve differential occupancy of multiple regulatory factors in parallel at nucleotide resolution.

We next extended this paradigm using genome-wide DNaseI footprints across 12 functionally distinct cell types to identify both known and novel factors showing highly cell-specific occupancy patterns. To calculate the footprint occupancy of a motif, we enumerated for each motif and cell type the number of motif instances encompassed within DNaseI footprints and normalized this by the total number of DNaseI footprints in that cell type. **Fig. 7b** shows a heatmap representation of cell-selective occupancy at motifs for 60 known transcriptional regulators and for 29 novel motifs. This approach appropriately identified a number of known cell-selective transcriptional regulators including; (1) the pluripotency factors OCT4, SOX2, KLF4, and NANOG in human embryonic stem cells[37]; (2) the myogenic factors MEF2A and MYF6 in skeletal myocytes[38]; and (3) the erythrogenic regulators GATA1, STAT1, and STAT5A in erythroid cells[39-41] (**Fig. 7b**).

Many of the footprint-derived novel motifs displayed markedly cell-selective occupancy patterns highly similar with the aforementioned well-established regulators. This suggests that many novel motifs correspond to recognition sequences for important but uncharacterized regulators of fundamental biological processes. Notably, both known and novel motifs with high cell-selective occupancy predominantly localized to distal regulatory regions (**Fig. 7c**), further highlighting the role of distal regulation in developmental and cell-selective processes[42,43].

Perspective

We describe an expansive map of regulatory factor occupancy at millions of precisely demarcated sequence elements across the human genome revealed by genomic DNaseI footprinting applied to a wide spectrum of cell types. These elements collectively define a highly information rich genomic sequence compartment which encodes the recognition landscape of hundreds of DNA binding proteins. This compartment has been extensively shaped by evolutionary forces to match closely the physical properties of its cognate interacting proteins. Mining footprint sequences for recognition motifs has nearly doubled the human *cis*-regulatory lexicon, exposing a previously hidden trove of elements with evolutionary, structural, and functional profiles that parallel the collections of experimentally-derived genomic regulators brought to light during the past 30 years. Because the ability to resolve footprints is dependent on sequencing depth, and the sequencing level of DNaseI cleavage events in most DHSs is not saturating (even in cell types with >500 million mapped unique DNaseI cleavages), the present study, while extensive in many respects, represents only an initial foray into this biologically rich space. Identification of the cognate DNA binding proteins for novel recognition sequences presents a significant challenge, though one which can be addressed with confidence using emerging technologies and our extensive experimental data demonstrating both occupancy *in vivo* and strong evolutionary signatures of function. On a broader level, the approach we describe here can, in principle, be applied to derive the *cis*-regulatory lexicon of any organism. We anticipate that the extensive new resources we describe, particularly in combination with other ENCODE data, will help to advance many aspects of human gene regulation research.

Methods Summary

DNaseI digestion and high-throughput sequencing were performed on intact human nuclei from various cell types, following published methods[4,44]. Briefly, roughly 10 million cells were grown in appropriate culture media and nuclei were extracted using NP-40 in an isotonic buffer. The NP-40 detergent was removed and the nuclei were incubated for 3 minutes at 37°C with limiting concentrations of the DNA endonuclease, deoxyribonuclease I (DNaseI) (Sigma) supplemented with Ca^{2+} and Mg^{2+}. The digestion was stopped with EDTA and the samples were treated with proteinase K. The small 'double-hit' fragments (<500 bp) were recovered by sucrose ultra-centrifugation, end-repaired and ligated with adapters compatible with the Illumina sequencing platform. High quality libraries from each cell type were sequenced on the Illumina platform to an average depth of 273 million uniquely mapping single-end tags. The sequencing tags were aligned to the human reference genome and per-nucleotide cleavage counts were generated by summing the 5' ends of the aligned sequencing tags at each position in the genome. FDR 1% DNaseI footprints were identified using an iterative search method based upon optimization of the footprint occupancy score. *De novo* motif discovery was performed using a full enumeration algorithm.

Acknowledgements

This work was supported by NIH grants HG004592 (J.A.S.) and RC2HG005654 (J.A.S. and M.G.). J.V. is supported by a National Science Foundation Graduate Research Fellowship. This work was supported in part by the University of Washington Proteomics Resource (UWPR95794). We thank Sam John and Fyodor Urnov for critical readings of the manuscript and many helpful discussions, and Sean Thomas for many helpful insights.

Author Contributions

J.A.S., A.B.S., S.N., M.T.M., B.V., and J.V. designed the experiments, S.N., J.V., A.B.S., A.P.R., B.V., M.T.M., R.E.T., E.H. and R.S. carried out the analysis, J.A.S., J.V., A.B.S., S.N., and A.P.R. wrote the paper, and all others carried out various aspects of data collection. The authors declare no competing interests.

Data Availability

All genomic DNaseI footprinting data are are available through the NCBI Gene Expression Omnibus (GEO) data repository (accessions GSE26328 and GSE18927), and also through the UCSC browser

under the Digital Genomic Footprinting (DGF) table designation. *De novo* motif models are available through the ENCODE Consortium data release website.

References

1. Galas, D. J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).

2. Dynan, W. S. & Tjian, R. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell* **35**, 79–87 (1983).

3. Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**, 159–197 (1988).

4. Hesselberth, J. R. *et al.* Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6**, 283–289 (2009).

5. Thanos, D. & Maniatis, T. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell* **83**, 1091–1100 (1995).

6. Tsai, S. F. *et al.* Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. *Nature* **339**, 446–451 (1989).

7. Thurman, R. E. *et al.* The accessibe chromatin landscape of the human genome. *Nature*

8. Sabo, P. J. *et al.* Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16837–16842 (2004).

9. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).

10. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–10 (2006).

11. Bryne, J. C. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* **36**, D102–6 (2008).

12. Chan, J. Y., Han, X. L. & Kan, Y. W. Cloning of Nrf1, an NF-E2-related transcription factor, by genetic selection in yeast. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11371–11375 (1993).

13. Rockman, M. V. & Wray, G. A. Abundant raw material for cis-regulatory evolution in humans. *Mol. Biol. Evol.* **19**, 1991–2004 (2002).

14. 1000 Genomes Project Consortium A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

15. Tate, P. H. & Bird, A. P. Effects of DNA methylation on DNA-binding proteins and gene expression. *Curr. Opin. Genet. Dev.* **3**, 226–231 (1993).

16. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462**, 315–322 (2009).

17. Ferré-D'Amaré, A. R., Pognonec, P., Roeder, R. G. & Burley, S. K. Structure and function of the b/HLH/Z domain of USF. *EMBO J.* **13**, 180–189 (1994).

18. Pellegrini, L., Tan, S. & Richmond, T. J. Structure of serum response factor core bound to DNA. *Nature* **376**, 490–498 (1995).

19. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).

20. Pugh, B. F. & Tjian, R. Transcription from a TATA-less promoter requires a multisubunit TFIID complex. *Genes Dev.* **5**, 1935–1945 (1991).

21. Kim, T. H. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).

22. Buratowski, S., Hahn, S., Guarente, L. & Sharp, P. A. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell* **56**, 549–561 (1989).

23. Kim, T. K. *et al.* Trajectory of DNA in the RNA polymerase II transcription preinitiation complex. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 12268–12273 (1997).

24. Project, A. C. S. H. E. T. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs. *Nature* **457**, 1028–1032 (2009).

25. Biddie, S. C. *et al.* Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. *Mol. Cell* **43**, 145–155 (2011).

26. Consortium, T. E. P. An integrated encyclopedia of DNA elements in the human genome.

27. Wadman, I. A. *et al.* The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *EMBO J.* **16**, 3145–3157 (1997).

28. Zhou, Z. *et al.* USF and NF-E2 cooperate to regulate the recruitment and activity of RNA polymerase II in the beta-globin gene locus. **285**, 15894–15905 (2010).

29. Gilbert, W. & Müller-Hill, B. Isolation of the lac repressor. *Proc. Natl. Acad. Sci. U.S.A.* **56**, 1891–1898 (1966).

30. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).

31. Mukherjee, S. *et al.* Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.* **36**, 1331–1339 (2004).

32. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263 (2009).

33. Newburger, D. E. & Bulyk, M. L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **37**, D77–82 (2009).

34. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).

35. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U.S.A.* **76**, 5269–5273 (1979).

36. Schoenherr, C. J. & Anderson, D. J. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science* **267**, 1360–1363 (1995).

37. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–872 (2007).

38. Yun, K. & Wold, B. Skeletal muscle determination and differentiation: story of a core regulatory network and its context. *Curr. Opin. Cell Biol.* **8**, 877–889 (1996).

39. Pevny, L. *et al.* Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. *Nature* **349**, 257–260 (1991).

40. Socolovsky, M. *et al.* Ineffective erythropoiesis in Stat5a(-/-)5b(-/-) mice due to decreased survival of early erythroblasts. *Blood* **98**, 3261–3273 (2001).

41. Halupa, A. *et al.* A novel role for STAT1 in regulating murine erythropoiesis: deletion of STAT1 results in overall reduction of erythroid progenitors and alters their distribution. *Blood* **105**, 552–561 (2005).

42. Treisman, R. & Maniatis, T. Simian virus 40 enhancer increases number of RNA polymerase II molecules on linked DNA. **315**, 73–75 (1985).

43. Grosveld, F., van Assendelft, G. B., Greaves, D. R. & Kollias, G. Position-independent, high-level expression of the human beta-globin gene in transgenic mice. **51**, 975–985 (1987).

44. Sabo, P. J. *et al.* Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat. Methods* **3**, 511–518 (2006).

Figure Legends

Figure 1. Parallel profiling of genomic regulatory factor occupancy across 41 cell types. a, DNaseI footprinting of K562 cells identifies the individual nucleotides within the MTPN promoter that are bound by NRF1. **b,** Example locus harboring eight clearly defined DNaseI footprints in Th1 and SK-N-SH_RA cells, with TRANSFAC database motif instances indicated below. **c,** Heatmaps showing per-nucleotide DNaseI cleavage (left) and vertebrate conservation by phyloP (right) for 4,262 NRF1 motifs within K562 DHSs ranked by the local density of DNaseI cleavages. Green ticks indicate the presence of DNaseI footprints over motif instances. Blue ticks indicate the presence of ChIP-seq peaks over the motif instances. **d,** Lowess regression of NRF1, USF, NFE2, and NFYA K562 ChIP-seq signal intensities versus DNaseI footprinting occupancy (footprint occupancy score) at K562 DNaseI footprints containing NRF1, USF, NFE2, and NFYA motifs.

Figure 2. DNaseI footprints mark sites of *in vivo* protein occupancy. a, Schematic and plots showing the effect of T/C SNV rs4144593 on protein occupancy and chromatin accessibility. Bar graph y-axis is the number of DNaseI cleavage events containing either the T or C allele. Middle plots show T or C allele-specific DNaseI cleavage profiles from 10 cell lines heterozygous for the T/C alleles at rs4144593. Right plots show DNaseI cleavage profiles from 18 cell lines homozygous for the C allele at rs4144593 and 1 cell line homozygous for the T allele at rs4144593. Cleavage plots are cut off at 60% cleavage height. **b,** The average CpG methylation within IMR90 DNaseI footprints, IMR90 DHSs (but not in footprints) and non-hypersensitive genomic regions in IMR90 cells. CpG methylation is significantly depleted in DNaseI footprints *($P < 2.2 \times 10^{-16}$*, Mann-Whitney test).

Figure 3. Footprint structure parallels TF structure and is imprinted on the human genome. **a,** The co-crystal structure of Upstream Stimulatory Factor (USF) bound to its DNA ligand is juxtaposed above the average nucleotide-level DNaseI cleavage pattern (blue) at motif instances of USF in DNaseI footprints. Nucleotides that are sensitive to cleavage by DNaseI are colored as blue on the co-crystal structure. The motif logo generated from USF DNaseI footprints is displayed below

the DNaseI cleavage pattern. Below is a randomly ordered heatmap showing the per-nucleotide DNaseI cleavage for each motif instance of USF in DNaseI footprints. **b,** The per-base DNaseI hypersensitivity (blue) and vertebrate phylogenetic conservation (red) for all DNaseI footprints in dermal fibroblasts matching three well annotated transcription factor motifs. The white box indicates width of consensus motif. The number of motif occurrences within DNaseI footprints in indicated below each graph.

Figure 4. A highly stereotyped chromatin structural motif marks sites of transcription initiation in human promoters. a, A 35-55 base-pair footprint is the predominant feature of many promoter DHSs and is in tight spatial coordination with the transcription start site. **b,** Heatmap of the per-nucleotide DNaseI cleavage pattern at 5,041 instances of this stereotypical footprint in K562 cells. **c,** Aggregate per-base DNaseI cleavage profile (blue line) and mean per-nucleotide conservation score (phyloP) surrounding instances of this stereotypical footprint in K562 cells (red dashed line). **d,** Aggregate strand corrected CAGE sequencing data (green line) and the average nearest 5' end of a spliced EST (orange line) surrounding instances of this stereotypical footprint in K562 cells.

Figure 5. Distinguishing direct and indirect binding of transcription factors. Heatmap of the enrichment of pairs of transcription factors in a direct-indirect association. Direct peaks are defined by ChIP occupancy accompanied by a footprint overlapping a compatible motif. Indirect peaks do not have a compatible motif. The color of each cell is determined by the fraction of indirect peaks that co-localize with the direct peaks of another factor.

Figure 6. *De novo* motif discovery expands the human regulatory lexicon. a, Overview of *de novo* motif discovery using DNaseI footprints. **b,** Annotation of the 683 *de novo*-derived motif models using previously identified transcription factor motifs. 394 of these *de novo*-derived motifs match a motif annotated within the TRANSFAC, JASPAR or UniPROBE databases, whereas 289 are novel motifs (pie chart). The de novo consensus matching TRANSFAC, JASPAR or UniPROBE sequences cover the majority of each database (bar chart) **c,** Example of a DNaseI footprint found in

multiple cell types that is annotated solely by one of the novel *de novo*-derived motifs. **d,** Box-and-whisker plot comparing the average nucleotide diversity at instances of the 289 novel *de novo*-derived motif models to instances of motifs present in databases of known specificities (x-axis). The blue bar indicates the average nucleotide diversity (π) at 4-fold degenerate coding sites (width is equal to 95% confidence interval); gold bar indicates π at all coding sites (width is equal to 95% confidence interval). **e,** Phylogenetic conservation (red dashed) and per-base DNaseI hypersensitivity (blue) for all DNaseI footprints in dermal fibroblast cells matching two novel *de novo*-derived motifs. The white box indicates width of consensus motif. **f,** Per-nucleotide mouse liver DNaseI cleavage patterns at occurrences of the motifs in (e) at DNaseI footprints identified in mouse liver.

Figure 7. Multi-lineage DNaseI footprinting reveals cell-selective gene regulators. a, Comparative footprinting of the nerve growth factor gene (VGF) promoter in multiple cell types reveals both conserved (NRF1, USF and SP1) and cell-selective (NRSF) DNaseI footprints. **b,** Shown is a heatmap of footprint occupancy computed across 12 cell types (columns) for 89 motifs (rows), including well-characterized cell/tissue-selective regulators, and novel *de novo*-derived motifs (red text). The motif models for some of these novel *de novo*-derived motifs are indicated next to the heatmap. **c,** The proportion of motif instances in DNaseI footprints within distal regulatory regions for known (black) and novel (red) cell-type specific regulators in (b) is indicated. Also noted are these values for a small set of known promoter-proximal regulators (green).

Supplementary Methods

Section	Title	Main Text Figure(s)	Supp. Figure(s)
0.1	Data downloads	-	-
0.2	Cell types used for digital genomic footprinting	-	-
1.1	Identification of DNaseI footprints	1a,b	-
1.2	Footprinting vs. tag levels	-	1a,b
1.3	FDR 1% DNaseI hypersensitive sites	-	1c,d
1.4	Annotation of footprints	-	2a,b
1.5	Putative motif binding sites and footprints	-	3
1.6	DNaseI cleavages vs. ChIP-seq	1c	4a,b
1.7	Footprint strength vs. ChIP-seq signal intensity	1d	-
1.8	Footprint strength vs. evolutionary conservation	-	4c,d
2.1	DNA Interacting Protein Precipitation (DIPP) experiments	-	5a-e
2.2	Allelic imbalance in footprints	2a	6
2.3	CpG methylation calculation within footprints, DHSs and non-DHSs	2b	-
3.1	Rendering of crystallography data showing DNA-protein complexes	3a	-
3.2	Visualization of DNaseI cleavage profiles by motif	3b	7;8

4.1	Analysis of stereotyped TSS-linked footprint	4	-
4.2	Determining direct and indirect transcription factor binding	5	9;10;11;12;13
5.1	*De novo* motif discovery	6a	
5.2	Motif matching	6b-c	14
5.3	Mouse scans of novel human motifs	6f	15b
5.4	Nucleotide diversity in DNaseI footprints	6d	15c
6.1	Cell type predominance - motifs within footprints	7a,b	-
6.2	Proximal vs. distal regulators	7c	-

0.1 – Data downloads

DNaseI-seq production data for Digital Genomic Footprinting (DGF) are available through the NCBI's Gene Expression Omnibus (GEO) data repository (accessions GSE26328 and GSE18927), and also through the table browser from University of California at Santa Cruz[45] (see http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeUwDgf).

Supplementary data too large to include in this supplement are being made available via the ftp server at ebi.ac.uk which contains an organized file structure with the ENCODE data. Analysis datasets are located at:

ftp://ftp-private.ebi.ac.uk/ (Login:encode-box-01 Password: enc*deDOWN)

in the subdirectories of byDataType.

0.2 – Cell types used for digital genomic footprinting

The following human cell types were subjected to DNaseI digestion and high-throughput sequencing, following previous methods[4,46] at the 36mer or 27mer* level: AG10803, AoAF, CD20+, CD34+ Mobilized, fBrain, fHeart, fLung, GM06990*, GM12865, HAEpiC, HA-h, HCF, HCM, HCPEpiC, HEEpiC, HepG2*, H7-hESC, HFF, HIPEpiC, HMF, HMVEC-dBl-Ad, HMVEC-dBl-Neo, HMVEC-dLy-Neo, HMVEC-LLy, HPAF, HPdLF, HPF, HRCEpiC, HSMM, Th1*, HVMF, IMR90, K562*, NB4, NH-A, NHDF-Ad, NHDF-neo, NHLF, SAEC, SKMC, and SK-N-SH RA*.

Tags were aligned to the reference genome, build GRCh37/hg19 (specified by ENCODE http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/referenceSequences/) using Bowtie[47], version 0.12.7 with parameters: *--mm -n 3 -v 3 -k 2*, and *--phred33-quals* for Illumina HiSeq sequencer runs or *--phred64-quals* for Illumina GAII sequencer runs.

1.1 – Identification of DNaseI footprints

For each cell type, we computed the DNaseI cleavage per nucleotide by assigning to each base of the human genome an integer score equal to the number of uniquely mappable sequence tags with 5' ends mapping to that position. To identify DNaseI footprints comprehensively across the genome, we used an improved and conceptually simplified approach versus that applied previously to the yeast genome[4]. We focused on high cleavage density regions, hotspot regions as identified by the *hotspot* algorithm[46], within each cell type. We scanned the genome for 6-40 nt stretches of successive nucleotides with low DNaseI cleavage rates relative to the immediately flanking regions, the signature of localized protection from DNaseI cleavage[1,4]. We then filtered findings to those occurring within the hotspot regions.

A priori, footprints comprise three components: a central area of direct factor engagement, and an immediately flanking component to each side. Upon factor engagement, local DNA architecture is distorted, frequently resulting in enhanced cleavage rates for flanking nucleotides outside of the factor recognition sequence. Greater disparity between the central and flanking components is indicative of higher factor occupancy.

To quantify this, we applied a simple footprint occupancy score (FOS) such that

$$FOS = (C+1)/L + (C+1)/R$$

Where C represented the average number of tags in the central component, L represented the average number of tags in the left flanking component, R represented the average number of tags in the right flanking component, and a smaller FOS value indicated greater average contrast levels between the central component and its flanking regions.

We sought to optimize the statistic across a range of central component (6-40 nt) and flanking component (3-10 nt) sizes. The output of the algorithm was the set of footprints with optimal FOS scores, subject to the criteria that L and R were greater than C, and all central components were disjoint and non-adjoining. When two or more potential footprints (those with L and R greater than C) had overlapping or abutting central components, we selected the one with the lowest FOS (or, in rare cases of identical scores, the 5'-most footprint relative to the forward strand). We then rescanned the entire local region to identify additional footprints. A local region was defined as the smallest genomic segment to contain all potential footprints of shared bases (by transitivity). No newly identified footprint consisted of a central component that overlapped or abutted the central component of any previously selected footprint. The rescan process was iterated until no new footprint was identified within the local region.

Human genomic positions uniquely mappable using 36 nt (and 27 nt as appropriate) sequence reads were computed using the same algorithm previously applied to yeast[4]. Any computed footprint whose central component consisted of non-uniquely mappable bases (thus having no mapped cleavage events by definition) that covered at least 20% of its length was discarded. Typically, fewer than 1% of unthresholded footprints were discarded during this process.

False discovery rate threshold

Due to the large number of tests for footprints performed over the genome, it was necessary to control for the expected number of false positives that arose due to chance through multiple testing[48]. We applied a false discovery rate (FDR) measure, defined as the expected value of the fraction of truly null features called significant divided by the total number of features called significant. To estimate FDR, we first generated a null set of pseudo-cleavages. For each hotspot in one cell type, we randomly reassigned the same number of tags found within the region to uniquely mappable positions within the same genomic interval. Analogous with experimental data, each base received an *in silico* cleavage score equal to the number of tags with 5' ends mapped to that base. We then considered the identical footprint

positions under the randomized scenario that were derived as output for the non-thresholded experimental data, thus encompassing the same number of footprint calls for FDR calculation purposes. We computed the maximum FOS threshold at which the number of footprints in the null set divided by the number of footprints in the observed set was less than or equal to 1%. The 1% FDR estimates were computed separately for all 41 cell types, covering a wide range of total tag levels and number of hotspot regions, to produce an average FOS threshold of 0.95 with a standard deviation of 0.02. We applied a final FOS threshold of 0.95 to footprints across all cell types. The central components of these FDR thresholded footprints, henceforth footprints, made up the final output of the procedure.

We tested whether DNaseI sequence bias contributed significantly to our FDR thresholded footprint sets. We digested purified genomic DNA with DNaseI enzyme, and sequenced resulting DNaseI cleaved fragments of size 1 kbp or below. The data were used to build a model that describes relative cut rate biases among all 6-mer subsequences (H. Bussemaker, personal communication). We visited each FDR thresholded footprint in the SKMC cell type and counted the total number of mapped tags falling in its central, left, and right flanking regions. We then randomly assigned the same number of simulated tags to positions within these regions, using probabilities proportional to the model's DNaseI cut-rate bias for the sequence context surrounding each position. A new FOS was calculated over the same L, C, and R regions as before and compared to the FOS value of the original footprint to see which footprints could be explained by sequence bias alone.

Combining Footprints Across Cell Types

We computed the multiset union of all footprints across all cell types. For each element of the union, we collected all significantly overlapping footprints, which were defined as those footprints with 65% or more of their bases in common with the element. A footprint's genomic coordinates were redefined to the minimum and maximum coordinates from its overlap set, which always included the footprint itself[49]. All redefined footprints from the union then passed through a subsumption and uniqueness filter: when a footprint was genomically contained within another, the filter discarded the smaller of the two or selected just one footprint if identical. Footprints passing through the filter comprised the final set of 8.4 million combined footprints across all cell types. Unlike footprints from any single cell type, the combined set included overlapping footprints.

1.2 – Footprinting vs. tag levels

Random subsamples (sampling without replacement) of the 543 million uniquely mappable DNaseI-seq tags from SKMC were generated. Increasing sample sizes utilized tags generated from smaller samples in addition to new tags generated from the randomized process. Footprints were called at each subsampled tag level.

1.3 – FDR 1% DNaseI hypersensitive Sites

We counted the number of footprints falling within every DNaseI hypersensitive sites (DHS, defined as 150 nt in length)[46] and grouped peaks by their number of footprints. Any peak containing more than 10 footprints was grouped with peaks containing exactly 10 footprints. The analysis was performed in every cell type separately, and then results were combined. We also decile-partitioned the DHSs by the number of sequencing tags mapped to them. For each partition, we drew a box plot to indicate the distribution of the number of footprints falling within the DHSs. We also determined the average number of footprints falling in DHSs (Supplementary Table 2).

1.4 – Annotation of footprints

We counted and summarized the number of combined footprints (8.4 million) falling into common genomic element categories (defined by at least 1 nt of overlap), such as those overlapping introns, coding elements, and intergenic regions. We utilized annotations from Gencode, version 7. Promoter regions were defined as within +/- 2.5 kb from a transcriptional start site (TSS). Regions within +/- 2.5 kb of transcriptional end sites were categorized as 3'-proximal. Other feature categories, such as Coding, 5'-UTR, 3'-UTR, and Introns were derived directly from Gencode annotations using transcriptional and coding start and stop site information, as well as exon boundary coordinates. When a footprint satisfied more than one category's condition (for example, when a footprint was found near more than one annotated transcript), we assigned it to only a single category. The order of category assignment in such cases was: coding, 5'-UTR, 3'-UTR, promoter, 3'-proximal, intronic, and intergenic.

1.5 – Putative motif binding sites and footprints

Genome Structure Correction

We determined the significance of overlap between footprints and predicted motifs within hotspot regions utilizing the Genome Structure Correction (GSC) test[50]. Merged genomic hotspot regions across all 41 cell types made up the domain. The multiset union of all footprints, part of the domain by

definition, as well as motif predictions within the domain (FIMO[51]; $P < 1x10^{-5}$ using TRANSFAC[10] and JASPAR Core[11], separately) were used as inputs to GSC. Program parameters were: *-n 10000*, *-s 0.1*, *-r 0.1*, and *-t m*. Significance was reported as a Z-score (empirical *p*-value was 0).

Average Motif Density Per-nucleotide

We determined the average per-nucleotide number of overlapping motif instances over segments of a genome-wide partition. We separately merged the hotspot regions and footprint regions across the 41 cell types. Using genome-wide FIMO scan predictions over TRANSFAC (*P* < 1e-5), we counted the number of motif scan bases contained within the merged footprint partition and divided by the total number of bases within the partition. Similarly, we found the average over the genomic complement between merged hotspots and merged footprints. Finally, we found a genome-wide average outside of hotspots and divided by the number of nucleotides with known base labels (A,C,G,T), thereby ignoring large centromeric and telemeric regions.

1.6 – DNaseI cleavages vs. ChIP-seq

Motif models (from TRANSFAC, version 2011.1, JASPAR Core, and UniPROBE[33]) were used in conjunction with the FIMO motif scanning software, version 4.6.1 using a *P* < 1e-5 threshold, to find all motif instances within DNaseI hotspots of the K562 cell line. We buffered (+/- 35 nt) a discovered motif instance and counted at each base position the number of uniquely mapping DNaseI sequencing tags with 5' ends mapping to the position. We sorted buffered motif instances by their total counts, and then normalized each instance's counts to a mean value of 0 and variance 1. A heatmap, with 1 row per motif instance, was generated using matrix2png[59], version 1.2.1. A phyloP evolutionary conservation[19] score heatmap over the same ordered motif instances and bases was generated using the same processing techniques. Motif instances that overlapped footprints by at least 3 nt were annotated. Uniformly processed hg19 K562 ChIP-seq peaks generated from experiments as part of the ENCODE Consortium were downloaded from the UCSC Table Browser[53]. Motif instances overlapping ChIP-seq peaks by at least 1 nt were also annotated.

1.7 – Footprint strength vs. ChIP-seq signal intensity

For a given ChIP-seq factor, we collected footprints that overlapped putative binding sites within hotspot regions by at least 3 nt. We calculated the summed ChIP-seq signal density over each region,

after buffering by +/- 50 nt from footprint centroid. Footprints were ordered by their FOS values, and signal data were plotted using lowess curve fitting with a span of 25%. ChIP-seq data (raw tag counts) included those from first replicates only. Average tag count numbers replaced cases where multiple measurements over the same genomic coordinates existed in the ChIP-seq data.

1.8 – Footprint strength vs. evolutionary conservation

We additionally calculated the maximum phyloP evolutionary conservation score over the same set of footprints. The maximum score was derived over the core footprint region (no buffering), with ten percent of outlying scores removed. As before, footprints were ordered by their FOS values, and signal data were plotted using loess curve fitting with a span of 25%. We applied a linear regression model with R statistical software (http://www.r-project.org) collecting the associated F-test's p-value.

2.1 – DNA Interacting Protein Precipitation (DIPP) experiments

Protein extraction for DNA Interacting Protein Precipitation (DIPP)

Nuclei were isolated using a standard protocol previously described[8]. Briefly, K562 cells were grown in RPMI (GIBCO) supplemented with 10% Fetal Bovine Serum (PAA), sodium pyruvate (GIBCO), L-glutamine (GIBCO), penicillin and streptomycin (GIBCO), and washed once with 1xDPBS (GIBCO). Nuclear extraction was performed by resuspending cells at $2.5x10^6$ cells/mL in 0.05% NP-40 (Roche) in Buffer A (15mM Tris pH 8.0, 15mM NaCl, 60mM KCl, 1mM EDTA pH 8.0, 0.5mM EGTA pH 8.0, 0.5mM Spermidine). After an 8 minute incubation on ice, nuclei were pelleted at 400 rcf for 7 minutes and washed once with Buffer A. Nuclei were then transferred to a 37°C water bath and resuspended at $1.25x10^7$ nuclei/mL in Extraction Buffer (10mM Tris pH 8.0, 600mM NaCl, 1.5mM EDTA pH 8.0, 0.5mM Spermidine). After 3 minutes at 37°C the sample was transferred to ice and rocked at 4°C for 2 hours. The soluble and insoluble fractions were separated by centrifugation at 3,220g for 15 minutes. The soluble fraction was then dialyzed for 2 hours at 4°C using a 3,500 Da molecular weight cut off (MWCO) cartridge (Pierce) against 500mL Dialysis Buffer (15mM Tris pH 7.5, 15mM NaCl, 60mM KCl, 5µM $ZnCl_2$, 6mM $MgCl_2$, 1 mM DTT, 0.5mM Spermidine, 40% Glycerol). The dialysis buffer was refreshed after 1 hour of dialysis. Dialyzed protein samples were quantified using a BCA assay (Pierce), flash frozen using liquid nitrogen and stored at -80°C until use.

DNA probe construction for DNA Interacting Protein Precipitation (DIPP)

Three genomic loci were targeted that demonstrated varying footprinting strengths. These footprints included (in hg19 coordinates) a MAX footprint (chr22:39707228-39707245) and two AP1 footprints – AP1 site 1 footprint (chr11:5301978-5302005) and AP1 site 2 footprint (chr5:75668604-75668626). For each of these sites, a 70-85 base pair region of DNA centered on the DNaseI footprint was selected. The selected DNA regions, in hg19 coordinates, were; chr22:39707201-39707270 for the MAX site; chr11:5301945-5302029 for the AP1 site 1; and chr5:75668577-75668646 for the AP1 site 2. DNA oligos were ordered for the forward and reverse strand for each of these sites, with the forward strand oligo containing a 5' biotin modification (Integrated DNA Technologies). For each of these sites, we also shuffled the footprinting sequence and ordered DNA oligos that contained this shuffled footprinting sequence along with the same flanking sequence as for the oligos above (Integrated DNA Technologies). The sequences of each of the probes can be found in **Supplementary Table 3**.

Generation of dsDNA bound beads for DNA Interacting Protein Precipitation (DIPP)

For each probe set, 500 picomoles of the forward strand biotinylated DNA oligo was mixed with 1 nanomoles of the reverse strand DNA oligo in Annealing Buffer (20mM Tris pH 8.0, 100mM KCl, 10mM $MgCl_2$). The reaction was denatured at 90°C for 5 minutes, slowly cooled to 65°C over 10 minutes, held at 65°C for 5 minutes and then cooled to 25°C. For each reaction, 100µl of Dynabeads MyOne Streptavidin T1 beads (Invitrogen) were washed twice with 0.75 mL of Bead Buffer (20mM Tris pH 8.0, 2M NaCl, 0.5mM EDTA, 0.03% NP-40) and resuspended in 0.8mL Bead Buffer similar to how previously described[54]. Annealed dsDNA probes were then added to the beads and rocked at room temperature for 1 hour. Beads were then washed twice with 0.8mL Bead Buffer to remove unbound oligos. 1 mL of Blocking Buffer (20mM Hepes pH 7.9, 300mM KCl, 50µg/mL bovine serum albumin (BSA), 50µg/mL glycogen, 5mg/mL polyvinylpyrrolidone (PVP), 2.5mM DTT, 0.02% NP-40) was added to each bead reaction and incubated at room temperature for 2 hours. Beads were then washed twice with 0.75mL of Binding Buffer (20mM Tris-HCl pH 7.3, 5µM $ZnCl_2$, 100mM KCl, 0.2 mM EDTA pH 8.0, 10mM potassium glutamate, 2mM DTT, 0.04% NP-40, 10% glycerol).

Pre-clearing protein extract for DNA Interacting Protein Precipitation (DIPP)

60µl of fresh Dynabeads MyOne Streptavidin T1 beads (Invitrogen) were washed twice with 0.3 mL of Bead Buffer and once with 0.3 mL of Binding Buffer and then added to 80µg of 600mM soluble K562 nuclear protein extract and 80µg of poly [d(I-C)] (Roche) in a 400µl total reaction volume with Binding

Buffer. This reaction was incubated at 4°C for 1.5 hours, the beads were removed and the buffered protein extract was cleared by centrifugation at 10,000 g for 8 minutes at 4°C.

DNA Interacting Protein Precipitation (DIPP) reaction and digestion

To each of the washed dsDNA bound bead reactions, 200µl of the pre-cleared buffered protein extract was added. This was incubated at 4°C for 2 hours then washed 3 times with 1 mL Binding Buffer, twice with 0.5 mL 50mM Ammonium Bicarbonate pH 7.8 and resuspended in 100µl 0.1% PPS Silent Surfactant (Protein Discovery) in 50mM Ammonium Bicarbonate pH 7.8. Bead bound proteins were boiled at 95°C for 5 minutes, reduced with 5 mM DTT at 60°C for 30 minutes and alkylated with 15 mM iodoacetic acid (IAA) at 25°C for 30 minutes in the dark. Proteins were then digested with 2µg Trypsin (Promega) at 37°C for 1.5 hours while shaking. The supernatant, which now contains digested peptides, was then transferred to a new tube, the pH was adjusted to <3.0 by 5 µl of 5 M HCl and incubated at 25°C for 20 minutes and then cleared by centrifugation at 20,817g for 10 minutes. The digested samples were desalted using an Oasis MCX cartridge 30 mg/60 µm (Waters) as previously described[55]. Peptide samples were then resuspended in 30µl 0.1% formic acid in H2O. These peptide samples were stored at -20°C until injected on the mass spectrometer.

Targeted Proteomic Mass Spectrometry on DIPP samples

Proteotypic peptides for c-Jun, MAX and CTCF were identified as previously described[55]. These peptides were; CPDCDMAFVTSGELVR and TFQCELCSYTCPR for CTCF; NSDLLTSPDVGLLK and NVTDEQEGFAEGFVR for c-Jun; and QNALLEQQVR and ATEYIQYMR for MAX. For each doubly charged monoisotopic precursor, we monitored singly charged monoisotopic y3 to yn-1 product ions. All cysteines were monitored as carbamidomethyl cysteines. Ions were isolated in both Q1 and Q3 using 0.7 FWHM resolution. Peptide fragmentation was performed at 1.5mTorr in Q2 using calculated peptide specific collision energies[56]. Data was acquired using a scan width of 0.002 m/z and a dwell time of 40ms.

Peptide samples were analyzed with a TSQ-Vantage triple-quadrupole instrument (Thermo) using a nanoACQUITY UPLC (Waters). A 5µl aliquot of each sample was separated on a 20cm long 75µm I.D. packed column (Polymicro Technologies) using Jupiter 4u Proteo 90A reverse-phase beads (Phenomenex) and chromatography conditions as previously described[54]. The injection order for each sample was randomized, and each sample was measured in three separate replicate injections.

Targeted measurements were imported into Skyline for analysis[57]. Chromatographic peak intensities from all monitored product ions of a given peptide were integrated and summed to give a final peptide peak height. For each peptide, peak heights from different samples and replicate runs were normalized such that the injection with the highest intensity was given a value of 1. Final peptide data were generated by taking the average normalized value of a peptide across replicates of a sample.

2.2 – Allelic imbalance in footprints

Read counts and genotype calls

A set of known autosomal single nucleotide variants (SNVs) was downloaded from the 1000 Genomes Project[14]. To avoid positions subject to mapping bias, SNVs were filtered to exclude any two within a read length (up to 36 nt) of one another. Allele counts used the same DNaseI-seq alignments from which the cut-counts were derived. For each cell type, reads overlapping each SNV were queried from the alignment in BAM format using the SAMtools[58]. Reads supporting a base call were counted only if they were mapped with no more than one mismatch excluding the SNV position being counted. If more than one read from a library was mapped at the same chromosome offset and strand, a single read was sampled at random to avoid overcounting from possible PCR duplicates. In order to call an individual heterozygous at a SNV conservatively, both alleles observed by 1000 Genomes had to be supported by at least four distinct reads. To call homozygotes conservatively, one of the known alleles had to be supported by at least 10 reads, and there had to be no reads supporting the other known allele, but a single read supporting another base was tolerated as a sequencing error where total read depth exceeded 50.

Allele-specific cut-count profiles

In the vicinity of each SNV (36 nt), DNaseI cut-counts from individuals homozygous for the same allele were added together, using the same genomic cut-count tracks used for calling footprints. In heterozygous individuals, reads overlapping the SNV were queried from the alignment BAM files but not subjected to the mismatch and duplicate filters used to obtain unbiased counts. The cut position represented by each read was reported as the aligned genomic position of the first base of the read, so cut-counts from reads aligning to the negative genomic strand may be offset by 1 nt, relative to the convention normally used for genomic cut counts. For each allele, the phased cut-counts for that allele from all heterozygous individuals were then added together.

Test for difference in the prevalence of allelic imbalance

At each SNV, the reads supporting each allele from all individuals heterozygous at the SNV were added together. Heterozygous sites were divided into two sets, those within the merged FDR 1% footprints across all cell types and those outside. A read-depth distribution was derived from each set, and the intersection was determined to generate a read-depth-matched random sample as large as possible. At each particular read depth, all sites from the set with fewer instances of that depth were included, and a random sample without replacement was taken from the set with more instances. Finally, we counted sites in each set showing allelic imbalance with two-sided binomial test $P < 0.01$. The difference between these counts was tested for significance with a one-sided Fisher's exact test.

2.3 – CpG methylation calculation within footprints, DHSs, and non-DHSs

IMR90 methylation calls[16] were filtered to CpGs covered by at least 40 reads. Methylation at each CpG is defined as the count of reads showing methylation (protection from bisulfite conversion) divided by the total read depth. We generated three sets of genomic coordinates with this signal: IMR90 FDR 1% footprints, IMR90 DNaseI peaks (subtracting overlapping footprint bases), and locations of CpGs in the GRCh37/hg19 genome reference sequence, removing elements that overlap IMR90 DNaseI hotspots. For each contiguous region in these datasets, we took the mean methylation of all overlapping CpGs that passed the 40-read coverage threshold. Regions with no such overlap were ignored. To compute *p*-values, vectors of mean methylation values were compared using a two-sided Mann-Whitney test.

3.1 – Rendering of DNA-protein complexes

Crystallography data showing DNA-protein complexes for selected factors[17,18] were obtained from the Protein Data Bank and rendered with MacPyMOL (http://www.pymol.org), version 1.3. Nucleotide residues were colored from white to blue, indicating increasing relative DNaseI cleavage propensity as aggregated across all motif instances.

Heatmap of DNaseI cleavages per-nucleotide

We buffered (+/- 35 nt) every motif instance of a motif model found within hotspot regions, and counted the number of uniquely mappable sequencing tags with 5' ends mapping at each base position. We sorted motif instances by their total counts, and then normalized each instance's counts to a mean value of 0 and variance 1. A heatmap, with 1 row per motif instance, was generated using matrix2png[59].

3.2 – Visualization of DNaseI cleavage profiles by motif occurrence

Motif models (from TRANSFAC, JASPAR Core, and UniPROBE) were used in conjunction with the FIMO motif scanning software, version 4.6.1 using a $P < 1e\text{-}5$ threshold, to find all motif instances within DNaseI hotspots of each cell type. The left and right coordinates of each motif instance were padded by 35 nt. Using the bedmap tool from the BEDOPS suite[49], version 1.2, the per-nucleotide DNaseI cleavage values from deeply sequenced DNaseI-seq libraries were recovered for each motif occurrence. A similar approach was used for phyloP vertebrate conservation. Aggregate plots were made by averaging over all strand-oriented motif occurrences the number of DNaseI cleavages and per-base conservation scores. All palindromic and near-palindromic motif occurrences were left in the dataset, reasoning that a transcription factor may bind to either orientation of the genomic region and binding events on either strand result in conformal changes to DNA that result in strand-specific cleavage patterns. Sequence logos were generated by assessing the information content of the oriented genomic sequences from all motif occurrences[60].

4.1 – Analysis of stereotyped TSS-linked footprint

The cleavage profiles +/-500 nt of all GENCODE V7 (level 1 and 2; manual curation) transcription start sites were used as regions to search for a 35-55 base-pair footprint following the method outline above with modifications. To amplify the signal in regions of low tag density and to remove noise in the data, the DNase I cutcounts were squared (x^2). The FOS score was then calculated for every segment 35-55 base-pairs in width using a fixed flank width of 10 base-pairs (left and right). The scored segments were ranked in ascending order (low FOS to high FOS) and the top non-overlapping segments were collected until no segments remained. Finally, a FOS threshold was selected (0.75, uniformly across 41 cell types) and these putative footprints were used in the subsequent analysis.

Graphical profiles were generated by enumerating the per-nucleotide DNaseI cleavages and phyloP conservation in a 250 base-pair window centered on the footprint. The heatmap representation was created using matrix2png.

Analysis of transcription initation. CAGE tags from the nuclear poly-A fraction (replicate 1) generated by RIKEN was downloaded from the UCSC Browser and the 5' stranded oriented ends were summed per-base. The footprint was stranded oriented to the nearest GENCODE V7 TSS. We enumerated the per-base CAGE tags in an 800 base-pair window centered on the footprint. To evaluate the spatial

relationship of transcription we calculated the distance to the nearest spliced EST curated from GenBank[61].

4.2 – Determining direct and indirect transcription factor binding

Uniformly processed hg19 K562 ChIP-seq peaks generated from experiments as part of the ENCODE Consortium were downloaded from the UCSC Genome Browser. Peaks overlapping DNaseI hypersensitive hotspot regions[46] by at least 20% were stratified into three categories: direct peaks, indirect peaks and indeterminate peaks. Direct peaks contained an appropriate motif instance (FIMO scan software, version 4.6.1, using $P < 1e\text{-}5$ threshold and motifs from TRANSFAC, version 2011.1) that overlapped a DNaseI footprint by at least 1 nt. Indirect peaks did not contain a cognate motif and indeterminate peaks were ambiguous (contained a motif which did not overlap a footprint). To identify enriched direct/indirect binding pairs, we counted the number of overlapping occurrences of all possible direct/indirect combinations. We normalized each ChIP-seq peak-pair count by the total number of indirect peaks for the indirectly bound factor, in order to reduce the effect of noise (due to incomplete motif models, insufficient DNaseI coverage, and/or non-specific antibodies).

5.1 – *De novo* motif discovery

We created different footprint subsets for each cell type for the purpose of *de novo* motif discovery. A proximal subset was defined as all footprints within 2000 nt of the canonical transcriptional start site of genes[61], a non-proximal set was defined as all footprints not in the proximal subset, a distal set was defined as all footprints more than 10,000 nt from any transcriptional start site, and cell-type-specific footprints were those footprints found within cell-type-specific DHSs. Cell-type-specific DHSs and constituent footprints were those found in only a single cell type.

We developed an exhaustive motif discovery procedure for inputs consisting of millions of genomic regions. To accomplish the exhaustive search, several simple heuristic filtering and clustering techniques were employed, along with a compute cluster. *De novo* motif discovery was performed separately for every cell type and on every footprint subset. For each subset, we symmetrically padded the central components of footprints by 4 nt and extracted genomic sequence information to create target regions for *de novo* discovery. We counted the number of target regions within which each subsequence pattern occurred, separately considering every 8 nt permutation over the 4-letter DNA nucleotide alphabet, with up to 8 intervening IUPAC 'N' degenerate symbols. For background estimates, nucleotide labels within

every target region were randomly shuffled, thereby maintaining local nucleotide label compositions. The number of regions within which each pattern existed was determined after each of 1000 shuffling operations in order to establish sample mean and variance values for expectation. These estimates for patterns further served as conservative estimates for longer patterns in the background case. For example, the estimates for 'acgttacc' also served as estimates for the 'acgNttacc' pattern. A Z-score was computed for each observed subsequence pattern by subtracting the mean background frequency estimate from the observed frequency and then dividing by the estimated standard deviation. Patterns with Z-score of at least 14 were listed in descending Z-score order and then further filtered and clustered to remove redundant motifs. Initially, the highest Z-score pattern was added to an output list, and each subsequent pattern was compared to every entry in the list. If a similar entry was found, the pattern was discarded; otherwise, the pattern was added to the bottom of the output list. Pattern similarities were determined by sequentially comparing characters. When two patterns were the same length and their 'N' placeholders aligned, they were considered similar if they had one character difference; otherwise, they were declared similar if they had up to two character differences. The reverse character sequence of every pattern then underwent the same filtering. The re-tuned motif list underwent a similar second stage filter that included all alignment possibilities and reverse complement combinations. Sequence patterns were converted to positional weight matrices (PWMs) by scanning all target sequences and normalizing over the nucleotide alphabet. Only exact matches to a subsequence pattern, ignoring all 'N' placeholders, were considered during PWM construction, which underwent further filtering. The PWM corresponding to the highest Z-score pattern was added to an output list and a comparison list. PWMs for subsequent patterns, still in descending Z-score order, were compared to every entry in the comparison list and then added to the bottom of that list. If no similar entry was found, the PWM was also added to the output list. During comparisons, Pearson correlation coefficients were calculated over all alignment possibilities and reverse complement combinations. We converted PWMs into 1-dimensional vector representations. Vectors were temporarily padded using samples from the genome-wide background nucleotide frequency distribution and renormalized for various alignments as needed. If a correlation value of at least 0.75 was found, two PWMs were considered similar. We reverted PWMs to their subsequence pattern forms and rescanned target regions, allowing up to one nucleotide mismatch from the pattern's subsequence representation. PWM filtering comparisons were performed as before, and PWM outputs from this stage formed the output.

The *de novo* discovery results for all footprint subsets and cell types were combined, clustered, and filtered further into a final set of 683 motifs. The PWM representations were converted to their subsequence pattern forms and combined in descending Z-score order. The first pattern was added to the output list. Each subsequent pattern was compared to every entry of the output list. If no similar entry was found, the pattern was added to the bottom of the list. Pattern comparisons included all alignment possibilities and reverse complement combinations. For a given alignment, the patterns were compared sequentially, character by character. In the event that all 'N' placeholders aligned, two patterns were declared similar if they had up to one character difference; otherwise, they were declared similar with up to two character difference.

For the final stage of clustering, we determined the proportion of instances of one pattern that genomically overlapped instances from another pattern. All pairwise combinations between patterns were considered. We scanned twice for every pattern's instances. The first scan included only those instances that do not deviate from their motif pattern. The second included all instances that have up to one mismatch. Scanning occurred over all padded footprints, merged across all cell types. If the proportion of overlapping instances between two patterns was 0.1 or more in the first case and 0.33 in the second case, in either motif comparison direction, we discarded the pattern of lower Z-score. We considered all cases with any amount of overlap (at least 1 nt). For example, if two patterns' instances overlapped at one part of the genome by 5 nt, and two more instances overlapped in another part of the genome by 2 nt, we conservatively counted both cases toward the proportion of overlaps (in contrast to the potential requirement of counting overlapping proportions at fixed offsets between instances). All patterns passing through this step made up the set of final motif models.

5.2 – Motif matching

We compared *de novo* motifs to motifs available as part of various databases, including TRANSFAC, version 2011.1, JASPAR Core, and UniPROBE using the TOMTOM software[52], version 4.6.1. We filtered TRANSFAC and JASPAR Core for motifs annotated to the human genome, and mouse motifs in UniPROBE. Redundant motifs were filtered per database to a single motif using redundant motif-name heuristics (for example, CTCF_01 and CTCF_02 are highly similar in TRANSFAC). TOMTOM parameters were set to their default values during motif comparisons. When partitioning the *de novo* motifs, assigning each to a single category, the order of match assignment preference was to TRANSFAC, JASPAR Core, UniPROBE, and then to the novel motif category.

5.3 – Mouse scans of novel human motifs

Novel *de novo* motifs (those with no motif match to entries of the TRANSFAC, JASPAR Core, and UniPROBE databases) were scanned across DNaseI hotspot regions of the mouse genome (build NCBI37/mm9) using FIMO at $P < 1e\text{-}5$. Average cleavage profiles were generated and compared to analogous profiles of the human genome.

5.4 – Nucleotide diversity in DNaseI footprints

To quantify the nature of selection operating on regulatory DNA, we surveyed nucleotide diversity (π) in footprint calls. Population genetics analyses were performed on 53 unrelated, publicly available human genomes (**Supplementary Table 4**) released by Complete Genomics, version 1.10[34]. Relatedness was determined both by pedigree and with KING[62]. Two Maasai individuals in the public dataset (NA21732 and NA21737) were not reported as related, but were found with KING to be either siblings or parent-child. NA21737 was removed from the analysis.

We defined four-fold degenerate sites using NCBI-called reading frames and the NimblegenSeqCapEZ Exome version 2.0 definition, downloaded from the NimbleGen website (http://www.nimblegen.com/products/seqcap/ez/v2/). Repeats were defined by RepeatMasker, downloaded from the UCSC Genome Browser, version 29Jan2009/open-3-2-7 (http://www.repeatmasker.org). Exome and repeats were removed from all footprints prior to analysis.

π calculation

π for a single variant is $2pq$, where p = major allele frequency and q = minor allele frequency. π was calculated for each cell type by summing π for all variants and dividing by total number of bases considered. Variant sites were filtered by coverage (>20% of individuals must have calls). Additionally, Complete Genomics makes partial calls at some sites (*i.e.*, one allele is A and the other is N). These were counted as fully missing.

6.1 – Cell type predominance - motifs within footprints

We scanned hotspot regions for motifs in each cell type using the FIMO software tool with a maximum *p*-value threshold of 1e-5 and defaults for other parameters. Scans included motif templates from TRANSFAC, JASPAR Core, UniPROBE, and novel *de novo* (those with no match to motifs in the

aforementioned databases). We filtered predicted motifs to those that overlapped footprints by at least 1 nt. For each cell type, we counted the number of discovered motif instances for a motif template and normalized to the total number of bases within footprints. We created a row-normalized heatmap over results in selected cell types using the matrix2png program.

6.2 – Proximal vs. distal regulators

For every motif template, we quantified the number of gene-distal and gene-proximal instances overlapping footprints by at least 1 nt, with proximal defined as within 2500 nt of the TSSs of genes in the reference sequence (NCBI RefSeq). The number of motifs found within a partition was scaled by the number of bases covered by footprints in that partition. Finally, we rescaled the partition values to proportions that summed to one.

7 – Supplementary References

45. Rosenbloom, K. R. *et al.* ENCODE whole-genome data in the UCSC Genome Browser: update 2012. *Nucleic Acids Res.* **40**, D912–D917 (2011).

46. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).

47. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

48. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300 (1995).

49. Neph, S. *et al.* BEDOPS: High performance genomic feature operations. *Bioinformatics*, In Press.

50. ENCODE Project Consortium *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).

51. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–8 (2009).

52. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).

53. Rosenbloom, K. R. *et al.* ENCODE whole-genome data in the UCSC Genome Browser. *Nucleic Acids Res.* **38**, D620–D625 (2009).

54. Mittler, G., Butter, F. & Mann, M. A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Res.* **19**, 284–293 (2009).

55. Stergachis, A. B., MacLean, B., Lee, K., Stamatoyannopoulos, J. A. & MacCoss, M. J. Rapid empirical discovery of optimal peptides for targeted proteomics. *Nat. Methods* **8**, 1041–1043 (2011).

56. MacLean, B. *et al.* Effect of Collision Energy Optimization on the Measurement of Peptides by Selected Reaction Monitoring (SRM) Mass Spectrometry. *Anal. Chem.* **82**, 10116–10124 (2010).

57. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).

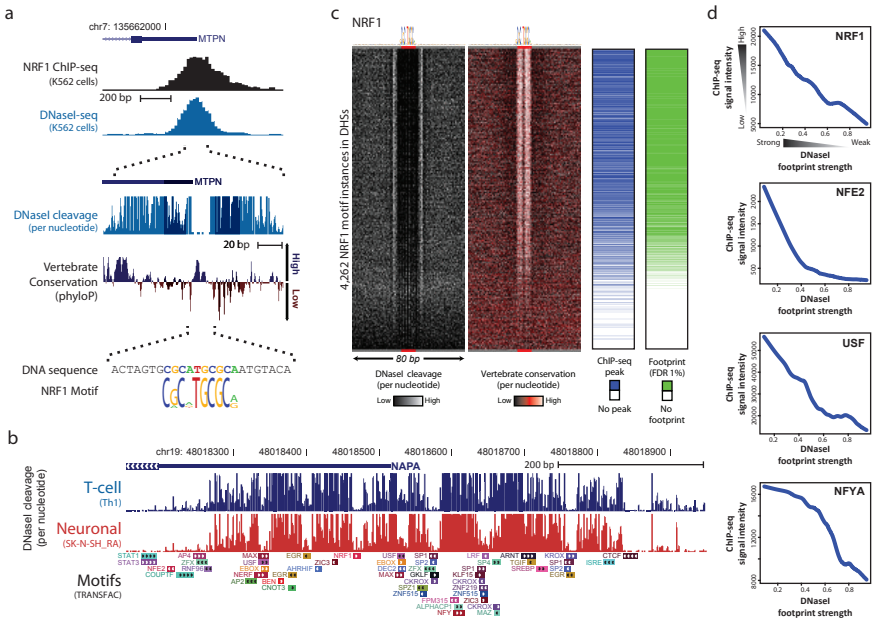
58. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

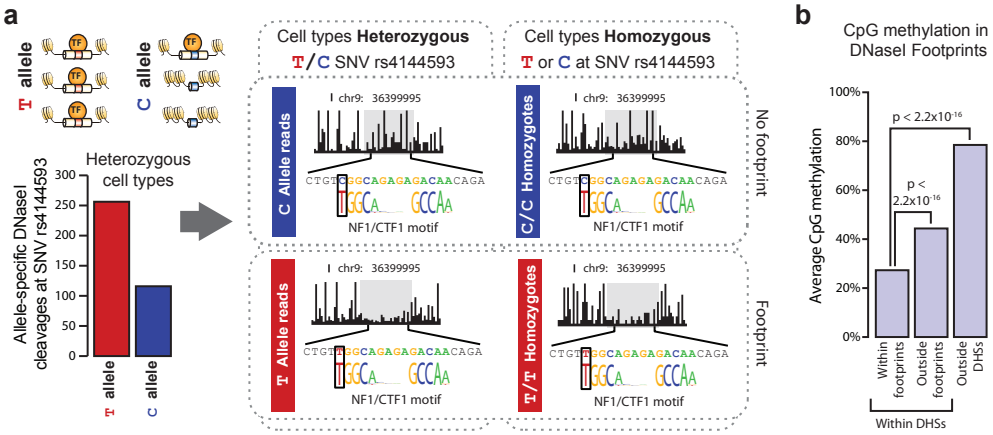
59. Pavlidis, P. & Noble, W. S. Matrix2png: a utility for visualizing matrix data. *Bioinformatics* **19**, 295–296 (2003).

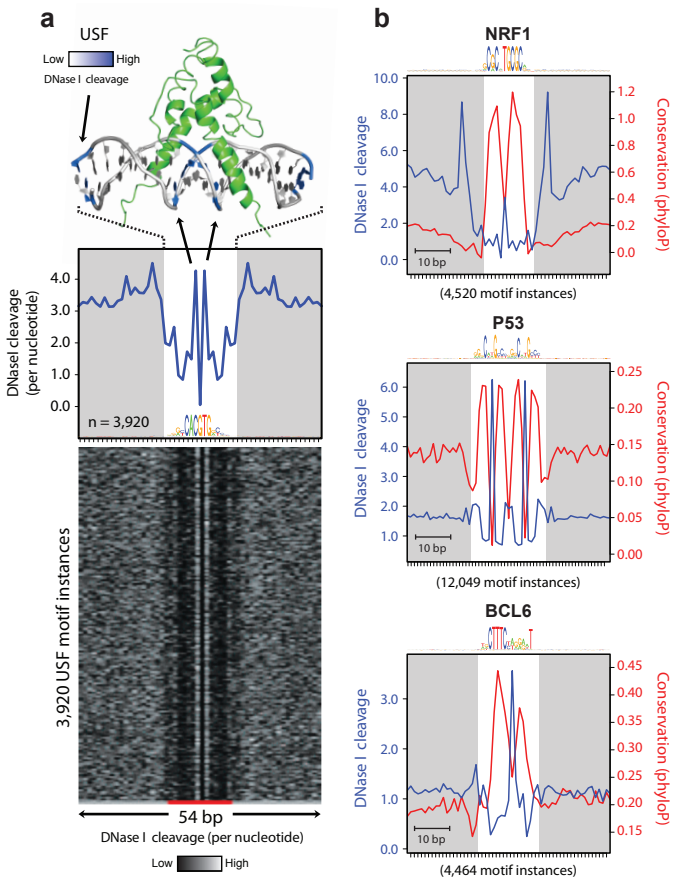
60. Crooks, G. E. WebLogo: A Sequence Logo Generator. *Genome Res.* **14**, 1188–1190 (2004).

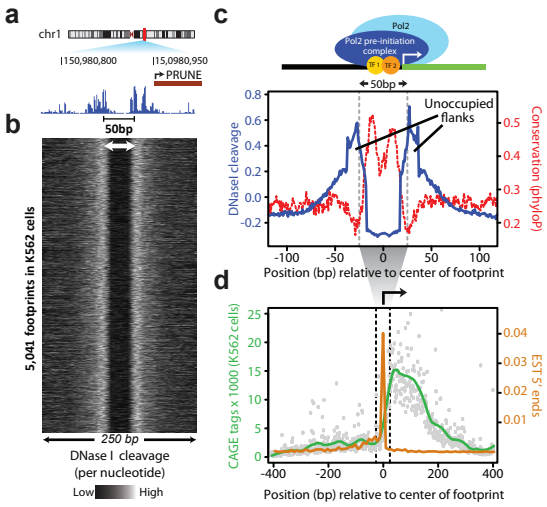
61. Pruitt, K. D., Tatusova, T., Klimke, W. & Maglott, D. R. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.* **37**, D32–D36 (2009).

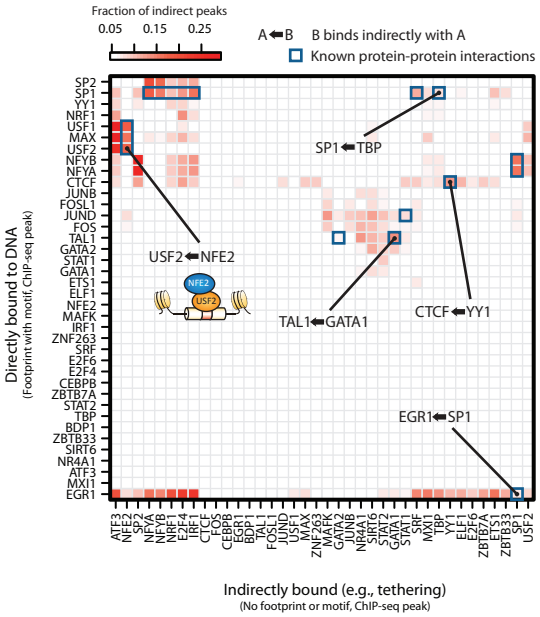
62. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

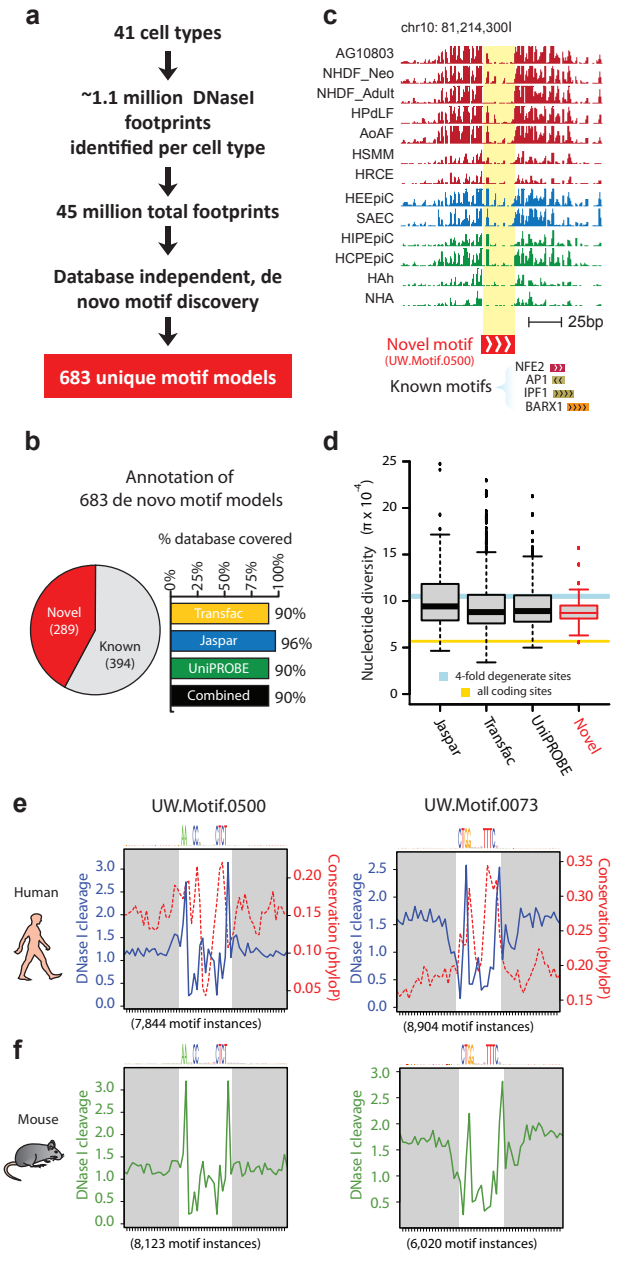

a
chr7: 135662000
MTPN
NRF1 ChIP-seq
(K562 cells)
200 bp
DNaseI-seq
(K562 cells)
MTPN
DNaseI cleavage
(per nucleotide)
20 bp
High
Low
Vertebrate
Conservation
(phyloP)
DNA sequence
ACTAGTGCGCATGCGCAATGTACA
NRF1 Motif
c
NRF1
4,262 NRF1 motif instances in DHSs
80 bp
DNaseI cleavage
(per nucleotide)
Low
High
Vertebrate conservation
(per nucleotide)
Low
High
ChIP-seq
peak
No peak
Footprint
(FDR 1%)
No
footprint
d
NRF1
NFE2
USF
NFYA
ChIP-seq
signal intensity
Low
High
Strong
Weak
DNaseI
footprint strength
b
chr19: 48018300
48018400
48018500
48018600
48018700
48018800
48018900
NAPA
200 bp
DNaseI cleavage
(per nucleotide)
T-cell
(Th1)
Neuronal
(SK-N-SH_RA)
Motifs
(TRANSFAC)

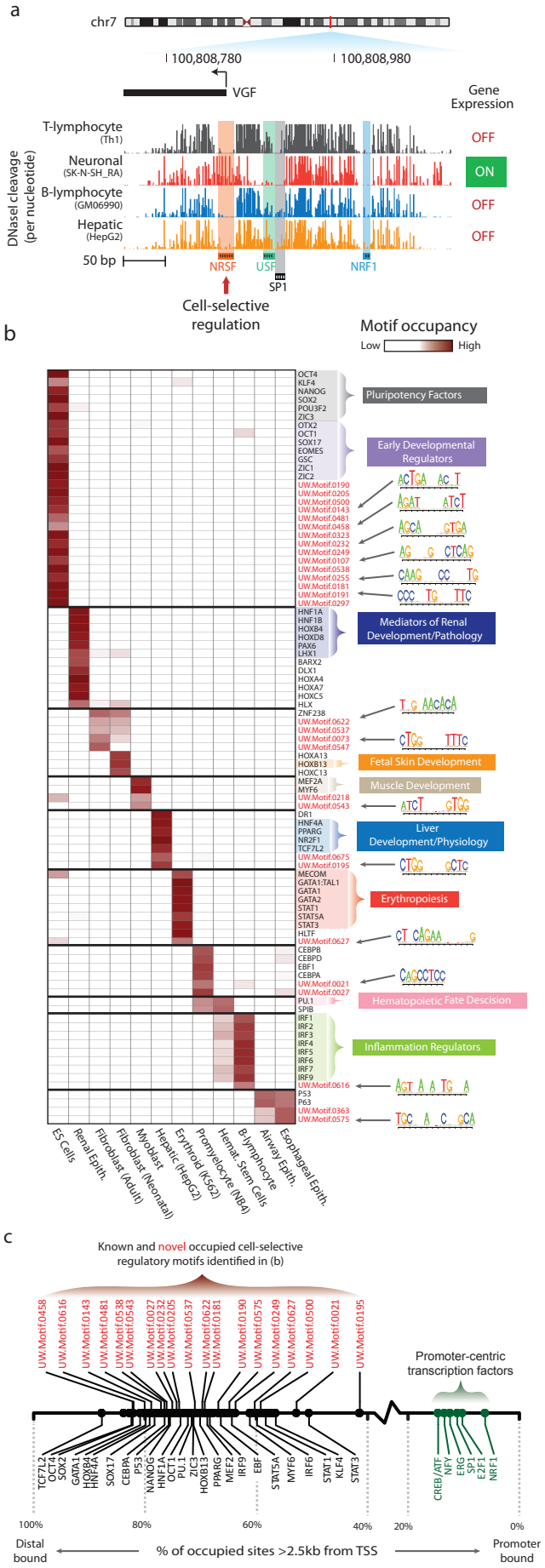

a
T allele
C allele
Heterozygous cell types
Allele-specific DNaseI cleavages at SNV rs4144593
300
250
200
150
100
50
0
T allele
C allele
Cell types Heterozygous
T/C SNV rs4144593
Cell types Homozygous
T or C at SNV rs4144593
C Allele reads
C/C Homozygotes
T Allele reads
T/T Homozygotes
chr9: 36399995
CTGTCGGCAGAGAGACAACAGA
NF1/CTF1 motif
No footprint
Footprint
b
CpG methylation in DNaseI Footprints
Average CpG methylation
100%
80%
60%
40%
20%
0%
p < 2.2x10-16
p < 2.2x10-16
Within footprints
Outside footprints
Outside DHSs
Within DHSs


a
USF
Low
High
DNase I cleavage
DNaseI cleavage (per nucleotide)
4.0
3.0
2.0
1.0
0.0
n = 3,920
3,920 USF motif instances
54 bp
DNase I cleavage (per nucleotide)
Low
High
b
NRF1
DNase I cleavage
Conservation (phyloP)
10.0
8.0
6.0
4.0
2.0
0.0
1.2
1.0
0.8
0.6
0.4
0.2
0.0
10 bp
(4,520 motif instances)
P53
DNase I cleavage
Conservation (phyloP)
6.0
5.0
4.0
3.0
2.0
1.0
0.25
0.20
0.15
0.10
0.05
0.00
10 bp
(12,049 motif instances)
BCL6
DNase I cleavage
Conservation (phyloP)
3.0
2.0
1.0
0.0
0.45
0.40
0.35
0.30
0.25
0.20
0.15
10 bp
(4,464 motif instances)


a
chr1
150,980,800
15,0980,950
PRUNE
50bp
b
5,041 footprints in K562 cells
250 bp
DNase I cleavage
(per nucleotide)
Low
High
c
Pol2
Pol2 pre-initiation
complex
TF1
TF2
50bp
Unoccupied
flanks
DNaseI cleavage
Conservation (phyloP)
Position (bp) relative to center of footprint
d
CAGE tags x 1000 (K562 cells)
EST 5' ends
Position (bp) relative to center of footprint


Fraction of indirect peaks
0.05
0.15
0.25
A ⬅ B
B binds indirectly with A
Known protein-protein interactions
SP1 ⬅ TBP
USF2 ⬅ NFE2
TAL1 ⬅ GATA1
CTCF ⬅ YY1
EGR1 ⬅ SP1
NFE2
USF2
Directly bound to DNA
(Footprint with motif, ChIP-seq peak)
Indirectly bound (e.g., tethering)
(No footprint or motif, ChIP-seq peak)
SP2
SP1
YY1
NRF1
USF1
MAX
USF2
NFYB
NFYA
CTCF
JUNB
FOSL1
JUND
FOS
TAL1
GATA2
STAT1
GATA1
ETS1
ELF1
NFE2
MAFK
IRF1
ZNF263
SRF
E2F6
E2F4
CEBPB
ZBTB7A
STAT2
TBP
BDP1
ZBTB33
SIRT6
NR4A1
ATF3
MXI1
EGR1
ATF3
NFE2
SP2
NFYA
NFYB
NRF1
E2F4
IRF1
CTCF
FOS
CEBPB
EGR1
BDP1
TAL1
FOSL1
JUND
USF1
MAX
ZNF263
MAFK
GATA2
JUNB
NR4A1
SIRT6
STAT2
GATA1
STAT1
SRF
MXI1
TBP
YY1
ELF1
E2F6
ZBTB7A
ETS1
ZBTB33
SP1
USF2


a
41 cell types
~1.1 million DNaseI footprints identified per cell type
45 million total footprints
Database independent, de novo motif discovery
683 unique motif models
b
Annotation of 683 de novo motif models
% database covered
0%
25%
50%
75%
100%
Novel (289)
Known (394)
Transfac 90%
Jaspar 96%
UniPROBE 90%
Combined 90%
c
chr10: 81,214,300l
AG10803
NHDF_Neo
NHDF_Adult
HPdLF
AoAF
HSMM
HRCE
HEEpiC
SAEC
HIPEpiC
HCPEpiC
HAh
NHA
25bp
Novel motif
(UW.Motif.0500)
Known motifs
NFE2
AP1
IPF1
BARX1
d
Nucleotide diversity (π x 10-4)
4-fold degenerate sites
all coding sites
Jaspar
Transfac
UniPROBE
Novel
e
UW.Motif.0500
UW.Motif.0073
Human
DNase I cleavage
Conservation (phyloP)
(7,844 motif instances)
(8,904 motif instances)
f
Mouse
DNase I cleavage
(8,123 motif instances)
(6,020 motif instances)


a
chr7
100,808,780
100,808,980
VGF
Gene Expression
T-lymphocyte (Th1)
Neuronal (SK-N-SH_RA)
B-lymphocyte (GM06990)
Hepatic (HepG2)
DNaseI cleavage (per nucleotide)
OFF
ON
OFF
OFF
50 bp
NRSF
USF
SP1
NRF1
Cell-selective regulation
b
Motif occupancy
Low
High
Pluripotency Factors
Early Developmental Regulators
Mediators of Renal Development/Pathology
Fetal Skin Development
Muscle Development
Liver Development/Physiology
Erythropoiesis
Hematopoietic Fate Decision
Inflammation Regulators
ES Cells
Renal Epith.
Fibroblast (Adult)
Fibroblast (Neonatal)
Myoblast
Hepatic (HepG2)
Erythroid (K562)
Promyelocyte (NB4)
Hemat. Stem Cells
B-lymphocyte
Airway Epith.
Esophageal Epith.
c
Known and novel occupied cell-selective regulatory motifs identified in (b)
Promoter-centric transcription factors
% of occupied sites >2.5kb from TSS
Distal bound
Promoter bound
100%
80%
60%
40%
20%
0%