Global measurements of human transcription factor occupancy: Insights into development and genome evolution

Andrew Ben Stergachis

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

John A. Stamatoyannopoulos

Michael J. MacCoss

James H. Thomas

Program Authorized to Offer Degree:

Genome Sciences

University of Washington

Abstract

Global measurements of human transcription factor occupancy:

Insights into development and genome evolution

Andrew Ben Stergachis

Chair of the Supervisory Committee:

Associate Professor John A. Stamatoyannopoulos

Department of Genome Sciences

Transcription factors (TFs) are a class of proteins that interact with the genome, dictating which parts of the genome are utilized by a given cell. Despite the central role TFs play in regulating the genome, technologies available to date have not permitted global measurement of TF occupancy within the cell. Consequently, our understanding of how TFs model cellular development and genome evolution has been limited. To address this, I focused my graduate studies on the development of genomic and proteomic tools that facilitates global measurements of TF occupancy within a cell, including the development of: (i) targeted proteomic assays to quantify TF protein abundances within the nucleus; (ii) a protein-centric method to map TF

protein occupancy in functionally distinct chromatin microenvironments; (iii) a genome-wide DNaseI footprinting method that enables construction of global maps of TF occupancy along the genome and core human regulatory networks; and (iv) a method for mapping the *in vivo* affinity of a TF genome-wide. Using a combination of these techniques, I have identified the sequence and chromatin features that direct TF occupancy within a cell and have characterized the mechanisms underlying TF occupancy dynamics during development and oncogenesis. These global TF occupancy maps have revealed novel insights into how TF occupancy shapes the evolution of both non-coding and coding genomic sequence. Specifically, it appears that TF affinity, not occupancy, plays the dominant role in shaping the evolution of TF binding elements. In addition, coding TF binding elements significantly contribute to protein evolution and codon usage biases in mammals. Together, these results depict the sequence, chromatin, developmental and evolutionary forces that shape TF occupancy within a cell—revealing unforeseen evolutionary constraints on the human genome.

Table of Contents

PAGE

Chapter 1 – Introduction

Every cell in a human body has essentially the same genetic blueprint, yet somehow different cell types in the body "read" this genetic blueprint in a unique manner to generate the diverse cell and tissue types that define the human body. This puzzle motivates my graduate work, and is the topic of this thesis. However, before I sink into my graduate research, I first want to give some background into what this genetic blueprint looks like and what within the cell is actually charged with 'reading' it.

1.1 – Transcription factors (TFs) and their role in gene regulation

The human genome contains ~20,000 genes that comprise just ~1.1% (34 Mb) of the total genome (The International Human Genome Sequencing Consortium, 2004). These 20,000 genes are coordinated by over 8 million short (4-40bp) regulatory DNA elements that together comprise over 5.1% (157 Mb) of the genome (Neph et al., 2012c), but each individual cell utilizes only 1.1 million of these regulatory DNA elements (Neph et al., 2012c).

The task of reading and regulating the genome is orchestrated by a class of proteins called transcription factors (TFs). TFs are sequence-specific DNA-binding proteins that interact with regulatory DNA elements throughout the genome, the result of which determines which genes are expressed in a given cell type. The human genome encodes ~1,400 TFs, two to three hundred of which are expressed in a given cell type (Vaquerizas et al., 2009). Although some TFs are ubiquitously expressed across all cell types and are thought to be involved in basal transcriptional processes (McKnight and Tjian, 1986), many TFs are expressed in a highly cell-selective manner and play a dominant role in determining the identity of a cell (Davis et al., 1987; Graf and Enver, 2009).

1.2 – Methodology for quantifying TF occupancy

Soon after Jacob and Monod first implicated a class of cellular molecules in transcriptional regulation (Jacob and Monod, 1961), the first transcription factor was isolated from the relatively simple bacteriophage (Ptashne, 1967; Pirrotta and Ptashne, 1969). Due to the considerably more complex genomic and proteomic architecture of humans, it took over two decades from Jacob and Monod's groundbreaking work to isolate the first human transcription factor (Dynan and Tjian, 1983). Although a considerable amount of progress has been made over the past several decades, the study of human transcription factors has been hamstrung by a lack of sensitive and accurate quantification methods for human TFs. Most TFs lack high-quality antibodies, and shotgun mass spectrometry is not sensitive enough to quantify TFs due to their low concentrations within the human nucleus (Washburn et al., 2001; Ghaemmaghami et al., 2003). Consequently, although the human genome encodes ~1,400 distinct transcription factors (Vaquerizas et al., 2009), we currently understand the occupancy of only ~100 human TFs in only a handful of cell types (ENCODE Project Consortium et al., 2012).

1.3 – Dissertation aims

Given the poor availability of high-quality TF antibodies and the sizable number of TFs and cellular states that need to be studied, I sought to develop alternative, antibody independent methods for generating global maps of human transcription factor occupancy. In collaboration with other members of the Stamatoyannopoulos and MacCoss labs, I developed a sensitive and scalable targeted proteomic method for quantifying transcription factor occupancy within the human nucleus, and a sensitive and scalable genomic method for quantifying transcription factor occupancy and affinity along the genome. Application of these methods to diverse human cell

and tissue types has exposed previously-hidden layers of nuclear TF organization, offering a new perspective on the evolution of genome-wide patterns of TF occupancy.

1.4 – Organization of thesis

In this thesis, I describe the development and application of different methodologies to obtain global measurements of human transcription factor occupancy. The first several chapters describe methodologies to generate global maps of transcription factor occupancy within the nucleus. The last several chapters describe the application of these approaches to understand the determinants of *in vivo* TF occupancy, the mechanisms underlying TF occupancy dynamics during development, and the evolutionary consequences of TF occupancy. More specifically:

Chapter 2 describes a targeted proteomic-based method for quantifying transcription factors. This approach is significantly more sensitive and quantitative than traditional 'shotgun' proteomic-based approaches, and is superior to antibody-based approaches, as hundreds of TFs can be readily quantified in the same experiment.

Chapter 3 describes an application of these targeted proteomic TF measurements for identifying the chromatin microenvironments in which a specific TF resides. This method enables the high-throughput characterization of TF occupancy patterns within and between cell types. Further, application of this method revealed that TF post-translational modifications (PTMs) and protein isoforms play a dominant role in directing TFs to different chromatin microenvironments.

Chapter 4 describes the application of *in vivo* DNaseI footprinting to map the genome-wide occupancy landscape of TFs within a cell. Application of this approach to 41 cell types exposed 45 million transcription factor occupancy events within regulatory regions, representing differential binding to 8.4 million distinct short sequence elements. Notably, high-resolution

DNase I cleavage patterns mirror nucleotide-level evolutionary conservation and track the crystallographic topography of protein–DNA interfaces.

Chapter 5 describes the use of *in vivo* DNaseI footprint maps to assemble extensive core human regulatory networks. This approach reveals that the circuitry of human transcription factor networks is highly cell-selective. Strikingly, in spite of their inherent diversity, all cell type regulatory networks independently converge on a common architecture that closely resembles the topology of living neuronal networks.

Chapter 6 describes a novel method for mapping the *in vivo* affinity of a TF genome-wide termed 'salted ChIP-seq.' Integration of this *in vivo* affinity map of CTCF with *in vivo* occupancy and chromatin accessibility measurements reveals that chromatin environment plays a central role in modeling the occupancy of TFs genome-wide. In contrast to previous assumptions, these data demonstrate that affinity, not occupancy, is largely dictating the extent of evolutionary sequence constraint at TF binding elements genome-wide.

Chapter 7 utilizes *in vivo* maps of chromatin accessibility across diverse human cell and tissue types to investigate the mechanisms underlying TF occupancy dynamics during development. This analysis revealed that the TF occupancy landscape of a cell encodes a memory of prior developmental fate decisions and that regulatory DNA elements that that are stably occupied throughout development are chiefly occupied by transcription factors that participate in autoregulatory feedback circuits. In contrast to normal cells, cancer regulatory landscapes feature both extensive reactivation of silenced embryonic stem cell regulatory elements and ectopic activation of regulatory DNA from alternative developmental programs external to the cell lineage from which the malignancy derives.

Chapter 8 studies the subset of TF occupancy events that occur within coding sequence. Utilizing DNaseI footprint maps for 81 diverse human cell and tissue types, I identified that 14% of coding bases and 87% of all genes are occupied by a TF in at least one cell type. Overall, coding TF occupancy has constrained amino acid divergence throughout human evolution and has systematically shaped codon choice, resulting in its appearance as the major determinant of codon usage bias in mammalian genomes. These results indicate that coding variation can significantly alter the wiring of superimposed TF binding elements.

The final chapter of this thesis (**Chapter 9**) discusses some future directions of this work and the field of gene regulation at large.

Chapter 2 – Development of targeted proteomic assays for human transcription factors

Note: This work was published in the November 2011 edition of *Nature Methods* as:

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

2.1 – ABSTRACT

We report a method for high-throughput, cost-efficient empirical discovery of optimal proteotypic peptides and fragment ions for targeted proteomics applications using in vitro–synthesized proteins. We demonstrate the approach using human transcription factors, which are typically difficult, low-abundance targets and empirically derived proteotypic peptides for 98% of the target proteins. We show that targeted proteomic assays developed using our approach facilitate robust in vivo quantification of human transcription factors.

2.2 – INTRODUCTION

Targeted proteomics is a powerful approach that enables quantitative analysis of peptides from complex biological samples with high sensitivity and specificity (Lange et al., 2008; Carr and Anderson, 2008). However, a major bottleneck limiting wider application of targeted proteomics has been the identification of optimal proteotypic peptides that are readily detectable by the mass spectrometer as well as the characteristic fragmentation patterns of these peptides.

Because of differences in physiochemical properties, different peptides from the same protein can produce drastically different signal intensities when measured with a mass spectrometer (Lange et al., 2008). Peptides are referred to as 'proteotypic' if they (i) are unique to a given protein, (ii) have good response characteristics in the mass spectrometer and (iii) have a fragmentation pattern with salient features to accurately detect and quantify. Traditional strategies for identifying proteotypic peptides and their fragmentation patterns have relied on the combination of experimental data with bioinformatic analyses. A common approach has been to use peptides catalogued in the course of 'shotgun' proteomic experiments conducted by data-dependent acquisition (Picotti et al., 2008; Prakash et al., 2009). This approach assumes that the peptides most frequently identified in shotgun experiments will produce the best response in a targeted proteomics setting. This assumption also underlies the application of machine-learning methods, which aim to predict proteotypic peptides (but not their fragmentation spectra) *de novo* (Mallick et al., 2006; Fusaro et al., 2009). Complicating these efforts, a large subset of the human proteome is absent from fragmentation spectra databases, and this deficit is particularly acute for low-abundance proteins such as transcription factors and kinases. To generate such peptide fragmentation data, large-scale efforts aim to synthesize predicted proteotypic peptides and empirically determine their fragmentation patterns (Picotti et al., 2010). However, which, if any, of these approaches is best suited for sensitive targeted proteomic analyses is unknown.

2.3 – RESULTS

Here we report an empirically driven approach for generating both optimal proteotypic peptides and their fragmentation patterns in a scalable, economical and generalizable fashion. Rather than relying on sparsely populated spectral databases (Picotti et al., 2008; Prakash et al.,

2009), prediction algorithms (Mallick et al., 2006; Fusaro et al., 2009), costly peptide synthesis (Picotti et al., 2010) or the costly purchase of full-length proteins (Keshishian et al., 2007), we leveraged the rich collection of tagged cDNA clones that are currently available for most human and model-organism proteins (Goshima et al., 2008; Ramachandran et al., 2008b) to generate *in vitro*–synthesized full-length protein samples, followed by tryptic digestion and mass spectrometry analysis using selected reaction monitoring (SRM). Because all monitored tryptic peptides for each protein originate from the same full-length protein molecules, we can compare the relative intensities of different peptides to identify those that provide the most sensitive proxy for the target protein. In addition to determining the relative peptide response, we can identify in parallel the fragmentation patterns of these peptides in a triple-quadruple mass spectrometer using SRM (**Fig. 2.M1**).

To demonstrate our approach, we studied transcription factors, a diverse class of low-abundance proteins with a paucity of spectral data in public databases (**Fig. 2.S1**). We selected 96 human transcription factor proteins spanning all major structural families (Vaquerizas et al., 2009) (**Fig. 2.M1a**). For each of these proteins, we obtained full-length cDNA clones contained in an *in vitro* transcription and translation–compatible vector with an in-frame C-terminal *Schistosoma japonicum* glutathione *S*-transferase (GST) tag (Rolfs et al., 2008). We then optimized *in vitro* protein production and purification in a 96-well plate format. We tested different protein production, capture, wash and digestion conditions to develop a protocol that gave maximal protein yield at the highest possible purity. To verify that enriched full-length proteins were produced, we performed silver-staining and western-bloting analyses for 46 of the 96 proteins (**Fig. 2.M1c** and **Fig. 2.S2**). For nearly all of the tested proteins, the target protein and the two endogenous glutathione-binding proteins GSTM3 and EEF1G were the top three

most intense bands as determined by silver staining, indicating that SRM signal contamination should be minimal. Of the tested clones, 96% (44 of 46 clones) produced highly enriched proteins with the correct molecular weight. The remaining two samples produced multiple species of different molecular weights, likely originating from alternative methionine start codons.

For each protein, we selected peptides and fragment ions to measure using Skyline (MacLean et al., 2010), an open-source application for building SRM methods and analyzing the resulting mass spectrometry data. We focused our analysis on predicted fully tryptic peptides with lengths of 7–23 amino acids. For each doubly charged monoisotopic precursor, we monitored singly charged monoisotopic y_3 to y_{n-1} product ions using a TSQ-Vantage triple-quadrupole mass spectrometer. We imported these measurements into Skyline to identify the relative peptide responses and their fragmentation patterns (**Fig. 2.M1d,e**).

To quantify the amount of each protein synthesized, we spiked heavy isotope–labeled forms of the schistosomal GST peptides LLLEYLEEK and IEAIPQIDK into each *in vitro* synthesis reaction. We measured the light-to-heavy isotope ratio of these two peptides and calibrated this ratio using an absolute quantification curve containing the same amount of the heavy isotope–labeled peptides but different known quantities of the light isotope–labeled peptide (**Fig. 2.S3**). Using this approach we determined that all of the 96 tested proteins produced at least 0.5 nM of product (**Fig. 2.M2a**).

To determine peptide-signal quality we manually analyzed chromatographic data for each peptide. Each peptide was given a quality score between 1 and 4, with 1 being the highest quality (Methods). For subsequent analysis we considered only peptides with a quality score of either 1 or 2. On average we identified eight peptides per protein with a quality score of 1 or 2. All but

two of the proteins assayed (CEBPG and HMGA1) had at least one peptide with a quality score of 1 or 2 (**Fig. 2.M2b**). Although sufficient quantities of both CEBPG and HMGA1 protein were produced using our *in vitro* approach (**Fig. 2.S2**) and the proteins were sufficiently digested as indicated by the mass spectrometry responses of the GST peptides, none of the monitored tryptic peptides from these two proteins gave a good response in the mass spectrometer. This suggests that a small minority of transcription factor proteins may not be amenable to proteomic analysis using trypsin-based digestion.

To determine our fragmentation-pattern quality, we compared our observed peptide fragmentation patterns with those contained in the US National Institute of Standards and Technology (NIST) spectral database. Of the 760 peptides in our dataset with a quality score of 1 or 2, only 18% (136 peptides) were represented in the NIST database (Methods). Of these, all had high spectral similarity scores, with 93% having dot-products >0.85 (**Fig. 2.S4**). This finding corroborates both our data and the NIST database and highlights the scarcity of proteotypic peptides in large spectral databases.

We next determined the utility of predictor algorithms and shotgun analyses to identify optimal proteotypic peptides. A comparison of our empirical ranking of proteotypic peptides with peptide rank predictions from the ESPPredictor algorithm (Fusaro et al., 2009) revealed Spearman correlations from −0.45 to 0.85 with an average correlation of 0.47 (**Fig. 2.S5**). Similarly, roughly half of the optimal proteotypic peptides from our experiments were undetected in shotgun analyses of the identical samples (**Fig. 2.S6**). Although these approaches are better than selecting proteotypic peptides at random, our results suggest that current predictor algorithms and spectral counting approaches provide imperfect ranking and identification of

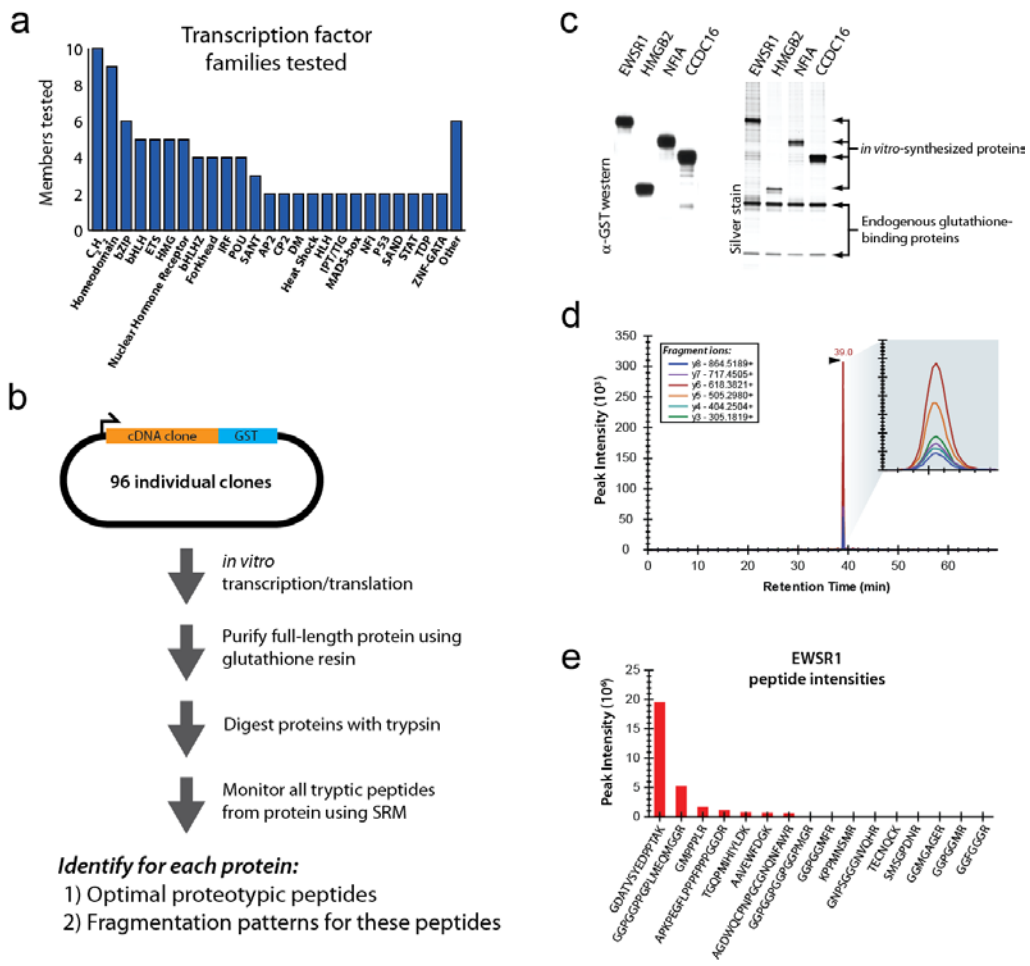
optimal proteotypic peptides, potentially limiting the utility of large-scale peptide-synthesis efforts that rely on such approaches as a first-round filter (Picotti et al., 2010).

Finally, we sought to confirm the utility of proteotypic peptides identified using our approach for *in vivo* analyses and how the *in vitro*–derived intensity rankings compared with those from complex biological samples. To test this, we first monitored all 12 of the quality score 1 and 2 peptides from the genomic master regulatory transcription factor CTCF in trypsin-digested nuclear lysate from erythroleukemia cells (K562). Using the fragmentation patterns identified *in vitro*, we identified corresponding chromatographic peaks for six of these CTCF peptides in K562 nuclear extract (**Fig. 2.M2c**). The relative intensity of these peptides *in vitro* and *in vivo* closely matched, confirming the relevance of the rank order of peptides identified empirically using *in vitro*–synthesized protein (**Fig. 2.M2d**). Next, we selected top-ranking peptides from four transcription factors and used these to generate nuclear abundance measurements of these factors across four distinct cell types (**Fig. 2.M2e**). The relative abundance measurements were consistent with previous reports on the tissue distribution of these transcription factors using RNA abundance (Lee et al., 1991; Klenova et al., 1993).

In summary, we demonstrated and validated a rapid and cost-efficient method for empirical identification of optimal proteotypic peptides and their fragmentation patterns using *in vitro*–synthesized proteins. Our method can be applied to generate assays to identify and quantify structurally diverse low-abundance proteins, such as human transcription factors, in unfractionated cellular extracts.

2.4 – FIGURES

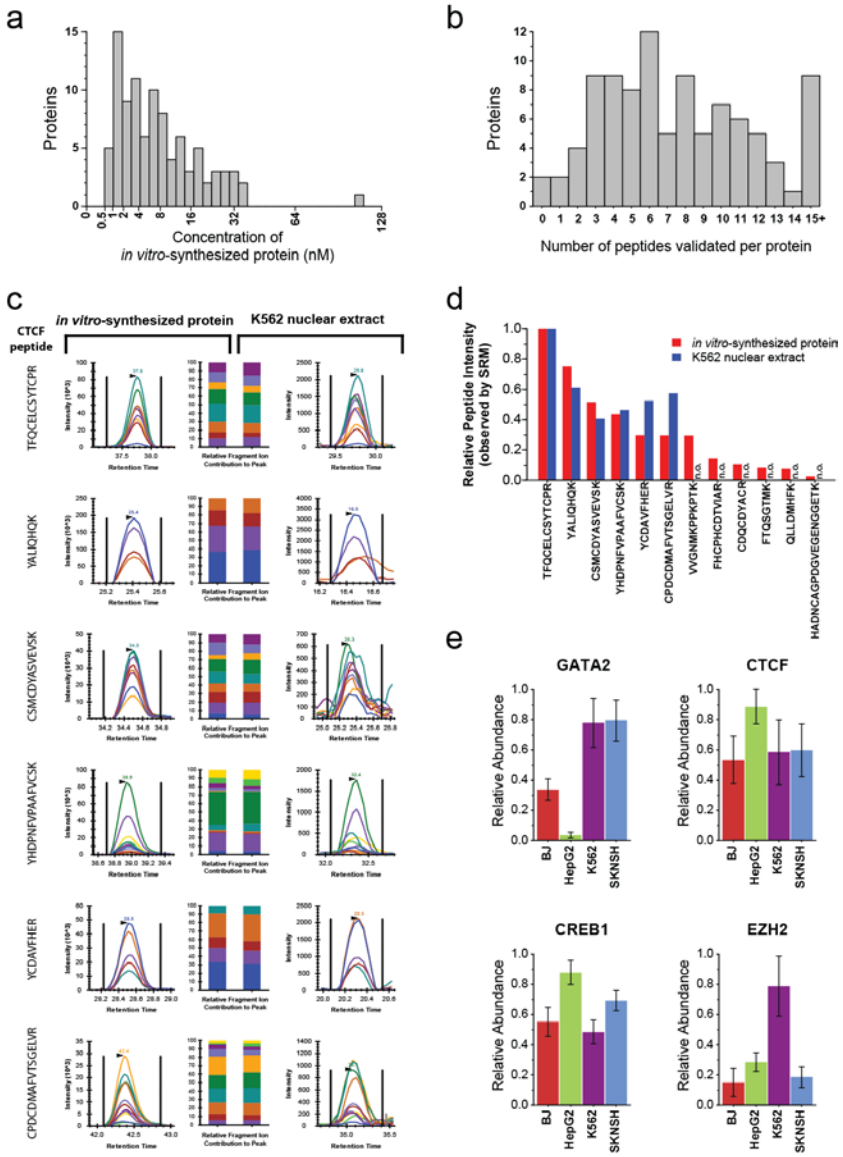
Figure 2.M1. Development of targeted proteomics assays using enriched in vitro–synthesized full-length proteins.



(**a**) Transcription factor family membership for the proteins for which targeted assays were built. (**b**) Schematic of the synthesis, enrichment, digestion and analysis of proteins to identify proteotypic peptides and their fragmentation patterns. (**c**) Target protein enrichment and purity were analyzed for 46 samples by immunodetection with an antibody to schistosomal GST (left) and silver staining (right). (**d**) SRM chromatographic traces from the NFIA peptide

EDFVLTVTGK. Insert, magnification of the chromatographic peak marked by the arrowhead. (**e**) EWSR1 peptide intensities (arbitrary units).

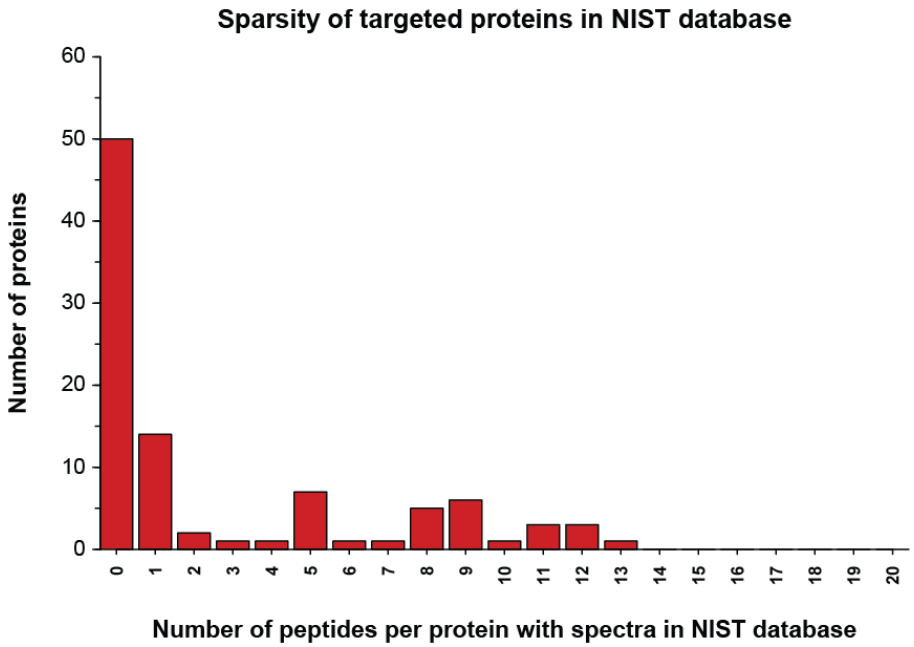
Figure 2.M2. Targeted assays can be efficiently developed using in vitro–synthesized proteins and applied to measure proteins in vivo.



(**a**) Absolute quantity of each *in vitro*–synthesized protein sample, as measured using a tryptic peptide contained in the C-terminal schistosomal GST tag. (**b**) Number of peptides per protein empirically assessed with salient features to accurately detect and quantify the target proteins (peptides with a quality score of either 1 or 2). (**c**) Proteotypic peptides identified using *in vitro*–synthesized CTCF were monitored in K562 nuclear extracts. The relative contribution of each

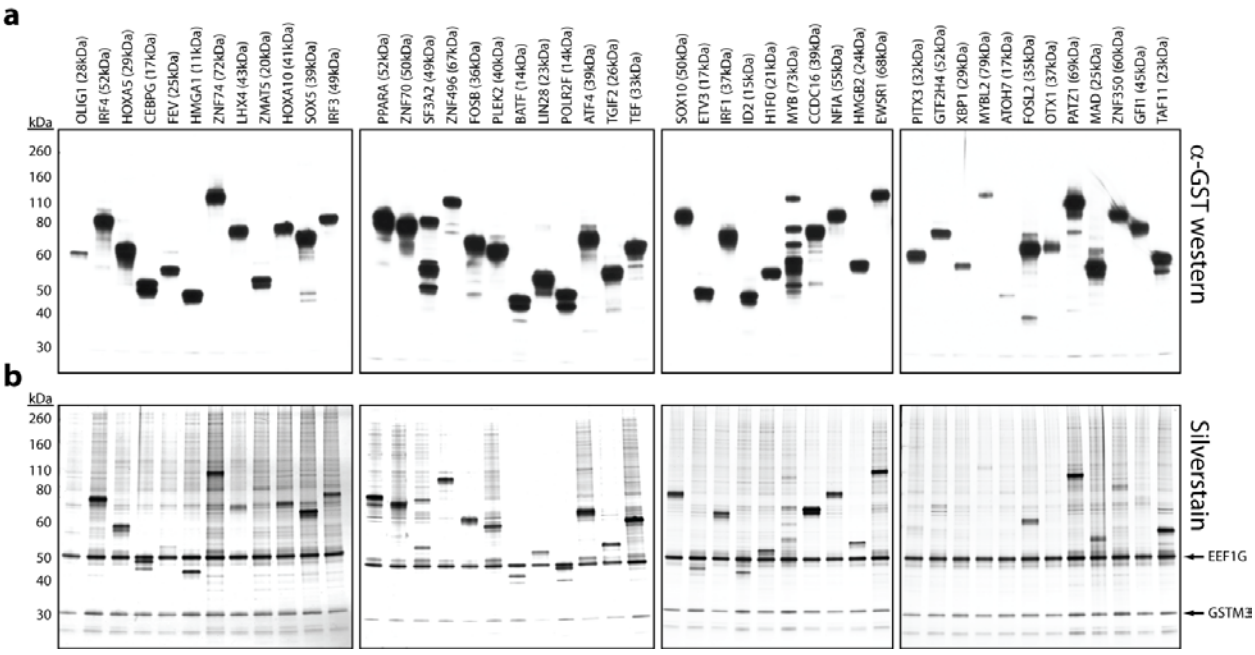
fragment ion to each peptide peak is displayed as different colors. (**d**) For each CTCF proteotypic peptide, the relative signal intensity observed using *in vitro*–synthesized protein is displayed alongside the relative signal intensity observed using K562 nuclear extract peptides not observed (n.o.) in K562 nuclear extracts are indicated. (**e**) Measured relative abundance of four transcription factors between the fibroblast (BJ), hepatic carcinoma (HepG2), erythroleukemia (K562) and neuroblastoma (SKNSH) human cell lines. Error bars, s. d. (n = 6).

Figure 2.S1. Poor coverage of target transcription factor proteins in NIST database



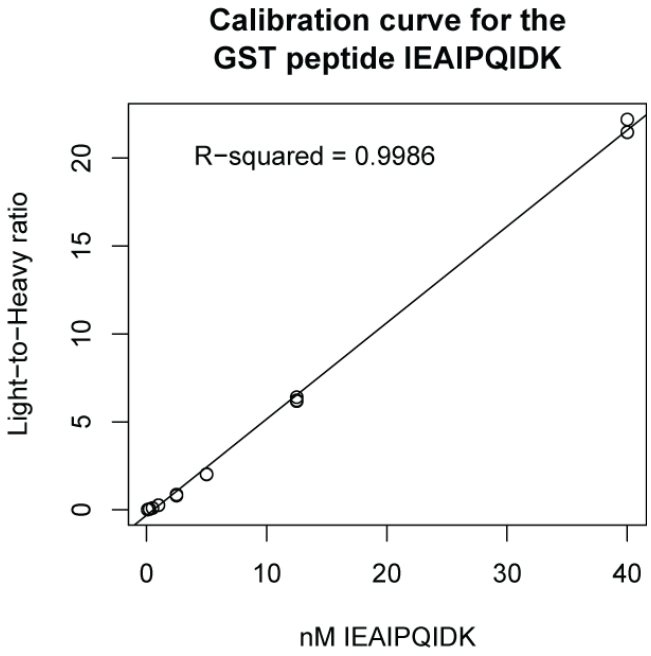
A histogram of the number peptide MS/MS spectra in the NIST database per target protein demonstrates how underrepresented transcription factors are from current peptide spectral libraries. A majority of our target proteins have no MS/MS spectra in the NIST spectrum library.

Figure 2.S2. In vitro-synthesized proteins are enriched full-length proteins



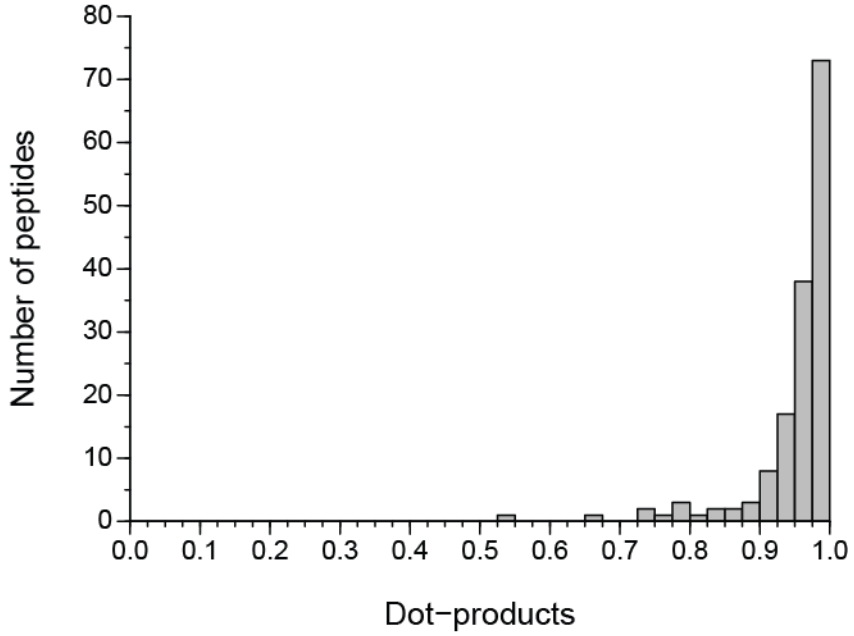
(a-b) Glutathione-enriched protein samples from 46 of the 96 reactions were run on a denaturing SDS-PAGE gel and subjected to either western blotting with an anti-schistosomal GST antibody **(a)** or silverstaining **(b)**. The endogenous glutathione-binding proteins EEF1G and GSTM3 were identified using 'shotgun' mass spectrometry. The molecular weight of the full length cDNA is indicated in parentheses. These weights do not include the 26kDa GST tag.

Figure 2.S3. Calibration curve for the schistosomal GST peptide IEAIPQIDK



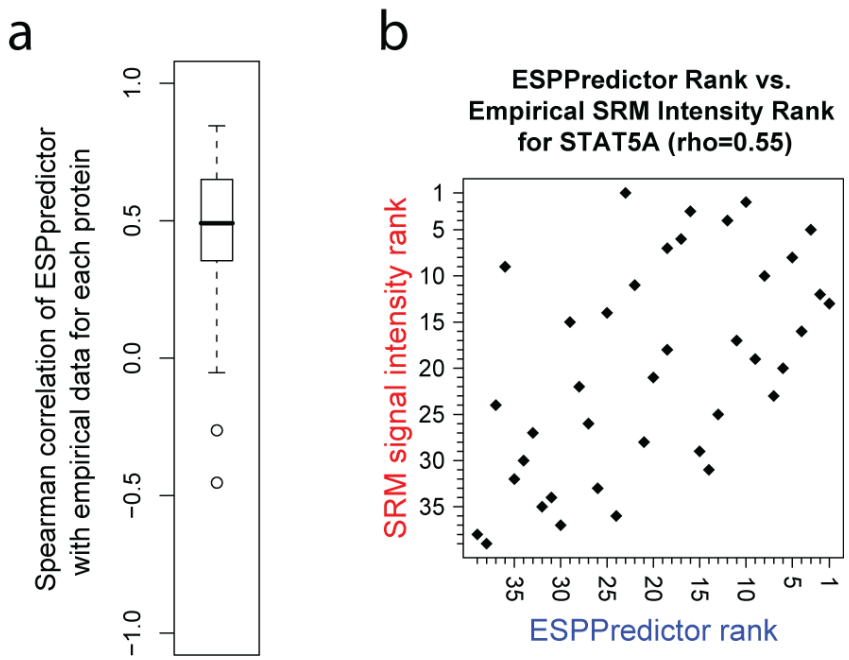
A dilution curve of the unlabeled schistosomal GST peptide IEAIPQIDK peptide standard was spiked with a constant amount of the heavy labeled IEAIPQIDK peptide. The unlabeled to labeled peak area was measured by LC-SRM-MS for each standard in triplicate.

Figure 2.S4. Histogram of dot-products for quality score 1 and 2 peptides



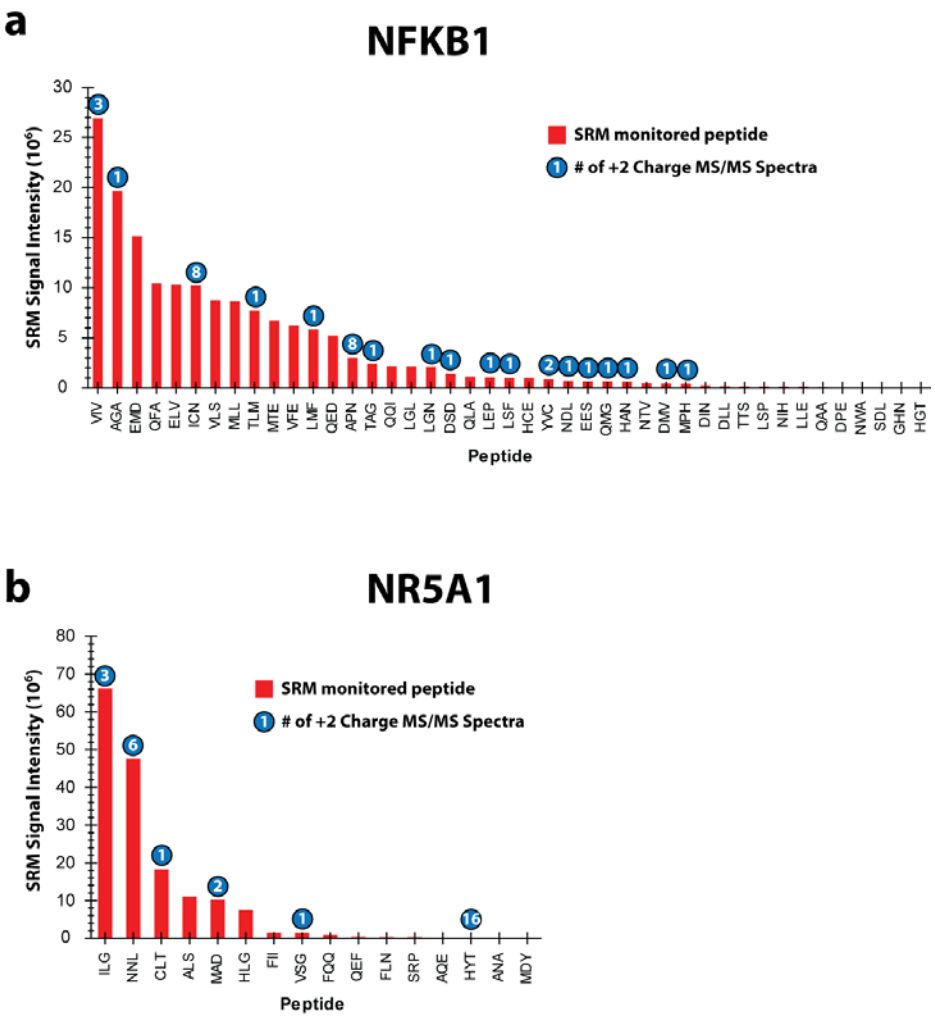
Dot-products were calculated using Skyline and the 2011_05_26 NIST release of the H. sapiens Ion Trap peptide spectral library.

Figure 2.S5. Spearman correlation of our empirical peptide ranking with ESPPredictor rankings



(a) Box-and-whisker plot showing the distribution of spearman correlations for the 75 proteins with both ESPPredictor scores and empirical SRM peptide signal intensities. The spearman correlation of these 2 rankings for each protein ranged from -0.45 to 0.85 with an average correlation of 0.47. **(b)** A representative comparison of our empirical signal intensity rank for the STAT5A peptides and the ESPPredictor score rank for the same peptides. This comparison shows a better than average correlation (rho = 0.55 versus an average of rho = 0.47).

Figure 2.S6. Peptide MS/MS spectrum counts are a poor predictor of targeted peptide signal intensity using selected reaction monitoring-mass spectrometry



(a-b) The NFKB1 **(a)** and NR5A1 **(b)** samples were subjected to 'shotgun' analysis using data-dependent acquisition. For each peptide, the number of +2 charge state spectra identified in this 'shotgun' run is indicated above the SRM signal intensity of that peptide.

Figure 2.S7. Reagent cost for generating in vitro-synthesized proteins from plasmids

Item	Manufacturer	Product #	Cost per unit	Units used per 96 reactions	Cost per 96 reactions
Human In Vitro Protein Expression Kit - DNA (50 reactions)	Pierce	88855	$515	2	$1,030.00
Glutathione Sepharose 4B (10mL)	GE	17-0756-01	$242	0.2	$48.40
PPS Silent Surfactant (5x1 mg vials)	Protein Discovery	21011	$199	1	$199.00
Oasis® MCX plate 30mg/60 µm 1/pkg	Waters	186000250	$326	1	$326.00
Sequencing Grade Modified Trypsin (100µg)	Promega	V5111	$80	0.4	$32.00
Lab reagents	Various	Various	$150	1	$150.00
				Total:	**$1,785.40**
				Cost per reaction:	**$18.60**

The reagent usage is based on our current working protocol (Methods). The prices used in this calculation are based on manufacture posted prices on May 24th 2011.

2.5 – METHODS

Clones and plasmids.

All of the clones used are from the pANT7_cGST clone collection distributed by the Arizona State University Biodesign Institute plasmid repository. These full-length cDNA clones contain a T7 transcriptional start sequence as well as an internal ribosome entry site (IRES), which is compatible with *in vitro* transcription-translation reagents (Rolfs et al., 2008). Additionally, each clone contains an in-frame fused C-terminal *Schistosoma japonicum* GST tag. Each bacterial stock clone was grown overnight in 5 ml of Luria broth with 100 μg ml^{-1} ampicillin (LB-amp). Plasmid DNA was extracted using the manufacture mini-prep protocol with the exception of an additional wash with PE buffer (Qiagen). All plasmid stocks were Sanger sequenced (University of Washington High Throughput Genomics Unit) using an M13 priming site upstream of the T7 promoter to confirm the identity of the insert and to ensure that there was no contamination of the plasmid stocks.

Peptides.

We obtained ~0.1 mg of FasTrack crude 'heavy' [$^{13}C_6{}^{15}N_2$]L-lysine–labeled LLLEYLEEK and IEAIPQIDK peptides to use as internal standards (Thermo). The LLLEYLEEK peptide was resuspended in 75% acetonitrile and 0.1% formic acid in H_2O. The IEAIPQIDK peptide was resuspended in 5% acetonitrile in H_2O. Unlabeled peptides provided at a concentration of 5 pmol μl^{-1} (assessed by the manufacturer by amino acid analysis) of AQUA Ultimate light LLLEYLEEK and IEAIPQIDK peptides were obtained to use as calibration standards (Thermo).

Protein production and purification.

Protein production and purification was optimized to be performed in 96-well plate format. Different protein production conditions, capture conditions, wash conditions and digestion conditions were tested to identify a protocol that gave maximal protein yield at the highest possible purity. The final protocol takes one person 2 d to transform a 96-well plate of plasmids into desalted peptide samples ready for mass-spectrometry analysis with a cost of less than $20 per protein (**Fig. 2.S7**).

Protein production conditions.

Proteins were synthesized from plasmid DNA using the Pierce Human In vitro Protein Expression kit (Thermo) according to the manufacturer's protocol with some slight modifications. Briefly, 1 μg of plasmid DNA was transcribed at 32 °C for 70 min in a 20-μl transcription reaction supplemented with 0.3 μl RNase inhibitors (Thermo). Two microliters of the transcription reaction was then added to a 23 μl translation reaction mix and incubated at 30 °C for 2 h. The translation reaction was then spiked with an additional 2 μl of the transcription reaction and incubated at 30 °C for an additional 2 h.

Protein capture conditions.

To enrich the GST-fusion protein, we used 2 ml of glutathione sepharose 4B beads (GE), washed 3 times with 15 ml 1× Dulbecco's phosphate-buffered saline (DPBS; Gibco) and resuspended in 12.5 ml of 1× DPBS. A 125-μl aliquot of the washed bead slurry was added to each well of eight 12-well strip-tubes such that each well received 20 μl of packed beads. Completed translation reactions were added to the beads and the bead-protein mixture was rocked end-over-end for 16 h at 4 °C.

Bead wash conditions.

Bead washing was staggered to ensure that only two 12-well strip-tubes were washed at a given time. By limiting the number of tubes washed at a time, it enabled the total wash time for each reaction to be reduced to less than 25 min. The bead-protein mixture was sedimented at 500*g* for 2 min using a swinging plate rotor. The supernatant was removed and 150 µl of wash buffer (1× DPBS supplemented with 863 mM NaCl) was added to the beads. The beads were mixed by inverting several times and sedimented at 500*g* for 2 min. The beads were washed twice with 150 µl wash buffer (1× DPBS supplemented with 863 mM NaCl) each and twice with 150 µl 50 mM ammonium bicarbonate (pH 7.8) each. After the last wash, the beads were resuspended in 100 µl Elution Buffer (0.05% PPS silent surfactant (Protein Discovery), ~5 nM heavy isotope–labeled GST peptide LLLEYLEEK (+8 Da) (Thermo), ~5 nM heavy isotope–labeled GST peptide IEAIPQIDK (+8 Da) (Thermo) and 50 mM ammonium bicarbonate (pH 7.8)) and stored at 4 °C until all eight 12-well strip-tubes had been washed. Ten microliters of each enriched protein sample was added to 4 µl 4× LDS buffer (Invitrogen) and saved for silver staining and western blotting.

Protein digestion.

Bead bound protein samples were boiled at 95 °C for 5 min, reduced with 5 mM dithiothreitol (DTT) at 60 °C for 30 min and alkylated with 15 mM iodoacetic acid (IAA) at 25 °C for 30 min in the dark. Proteins were then digested with 400 ng trypsin (Promega) at 37 °C for 2 h while shaking. Beads were then sedimented at 500*g* for 2 min and the supernatant, which contained the digested peptides, was transferred to a new 96-well collection plate. The beads were washed once with 150 µl 50 mM ammonium bicarbonate pH 7.8, and the supernatant from this wash was combined with the previous supernatant. The pH of the supernatant sample was

adjusted to <3.0 by 5 µl of 5 M HCl and incubated at 25 °C for 20 min. The digested samples were desalted using a 96-well Oasis MCX plate 30 mg per 60 µm (Waters) following the manufacturer's protocol with minor modifications. Briefly, the cartridge was conditioned using 1 ml methanol, 1 ml 10% ammonium hydroxide in H_2O, 2 ml methanol and finally 3 ml 0.1% formic acid in H_2O. The samples were then loaded onto the cartridge and washed with 1 ml 0.1% formic acid in H_2O and 1 ml of 0.1% formic acid in methanol. The peptides were eluted from the cartridge with 600 µl 10% ammonium hydroxide in methanol, collected in a 1 ml round-bottom 96-well collection plate and evaporated using a SpeedVac (Labconco) set to 50 °C. Peptide samples were evaporated down to 10–30 µl of volume then resuspended in 50 µl 0.1% formic acid in H_2O. These peptide samples were stored at −20 °C until injected into the mass spectrometer.

Silver staining and immunoblotting.

Undigested protein extract from each of the fractions was boiled in 1× LDS buffer (Invitrogen) and separated on a 4–12% bis-Tris denaturing and reducing SDS-PAGE (Invitrogen). Gels were then subjected to either silver staining (Invitrogen) or transferred onto a nitrocellulose membrane (Bio-Rad) for immunoblotting. Membranes were blocked with 5% non-fat dry milk (Safeway) in TBS-tween buffer and probed for schistosomal GST (GE 27-4577-01). All primary incubations were done at 4 °C overnight using a 1:1,000 dilution. Secondary incubations were performed in 5% non-fat dry milk in TBS-tween using 1:10,000 diluted peroxidase-conjugated rabbit anti-goat IgG (H+L) (Pierce). Membranes were visualized using an ECL plus western blotting kit (Amersham) and detected with radiographic film (Thermo).

Nuclear protein extraction.

Nuclear proteins from K562, HepG2 and SKNSH cancer cell lines and the BJ fibroblast cell line were isolated in three biological replicates as previously described (Dorschner et al., 2004). BJ cells were grown in MEM (Gibco) supplemented with 10% fetal bovine serum (PAA), non-essential amino acids (Gibco), sodium pyruvate (Gibco), 1.5 mg ml^{-1} $NaHCO_3$, penicillin and streptomycin (Gibco). HepG2 cells were grown in MEM supplemented with 10% FBS, non-essential amino acids, sodium pyruvate, penicillin and streptomycin. K562 and SKNSH cells were grown in RPMI (Gibco) supplemented with 10% FBS, sodium pyruvate, L-glutamine (Gibco), penicillin and streptomycin. SKNSH cells were treated with 6 μM retinoic acid for 48 h before they were collected. K562 nuclear extraction was performed by resuspending cells at 2.5 × 10^6 cells per ml in buffer A (15 mM Tris pH 9.0, 15 mM NaCl, 60 mM KCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 0.5 mM spermidine) containing 0.05% NP-40 (Roche). After an 8-min incubation on ice, nuclei were pelleted at 400*g* for 7 min and washed once with buffer A. SKNSH, HepG2 and BJ nuclei were isolated in a similar fashion, but with the use of cell line–specific NP-40 concentrations and cytoplasmic lysis times (SKNSH was 0.05% NP-40 for 5 min; HepG2 was 0.1% NP-40 for 8 min; and BJ was 0.5% NP-40 for 40 min). Nuclei were then resuspended in buffer A containing 0.2% NP-40, sonicated at setting output 3 for 30 s, digested with benzonase for 15 min at 4 °C, digested with DNaseI for 15 min at 37 °C and finally digested with trypsin. Samples were brought to 6 mM $MgCl_2$ before digestion with 0.375 U μl^{-1} benzonase (Fisher Scientific). Samples were brought to 6 mM $CaCl_2$ and 90 mM NaCl before digestion with DNaseI (Sigma). DNaseI digestion reactions were stopped using 50 mM EDTA. Nuclear protein samples were digested with trypsin as described above using a 50:1

protein:trypsin ratio. After digestion and MCX cleanup, each sample was resuspended in 0.1% formic acid in H_2O to a final concentration of ~10,000 nuclei per µl.

Targeted proteomic mass spectrometry.

Peptide samples were analyzed with a TSQ-Vantage triple-quadrupole instrument (Thermo) using either a nanoLC separation system (Eksigent) or a nanoACQUITY UPLC (Waters). A 5-µl aliquot of each sample was separated on a 16-cm-long 75 µm inner diameter packed column (Polymicro Technologies) using Jupiter 4u Proteo 90A reverse-phase beads (Phenomenex). Peptides were separated using a 27.5-min gradient from 2% acetonitrile in 0.1% formic acid to 23% acetonitrile in 0.1% formic acid. The gradient was followed by a wash for 10.5 min at 80% acetonitrile in 0.1% formic acid and a column re-equilibration at 2% acetonitrile in 0.1% formic acid for 12 min. Ions were isolated in both Q1 and Q3 using 0.7 FWHM resolution. Peptide fragmentation was performed at 1.5 mTorr in Q2 using calculated peptide specific collision energies (Maclean et al., 2010). Data was acquired using a scan width of 0.002 mass to charge ratio (*m/z*) and a dwell time of 10 ms.

Each protein sample was injected separately. For the target protein, all monoisotopic, +2 charge state, fully tryptic peptides from 7 to 23 amino acids in length were tested. In addition, the heavy and light forms of the schistosomal GST peptides LLLEYLEEK and IEAIPQIDK and the light form of the endogenous glutathione-binding protein GSTM3 peptide IAAYLQSDQFCK were tested. Peptides that flanked the target and fusion protein were not tested. For all peptides, the monoisotopic, +1 charge state y_3 to y_{n-1} fragment ions were monitored. All cysteines were monitored as carbamidomethyl cysteines. All methods were designed such that no more than 240 transitions were monitored in a given run. Quality control runs were acquired after every ~8 injections to monitor column stability.

SRM data analysis.

Targeted proteomic data were analyzed using the software package Skyline (MacLean et al., 2010). Chromatographic data from each peptide were manually analyzed to determine the quality of the peptide signal. Scoring of peptide quality was done by assessing the following requirements: (A) a prominent chromatographic peak with a signal intensity of at least 60,000 (total peak area under the curve (AUC) for all contributing fragment ions (arbitrary units)); (B) two or more data points were collected across the peak; (C) three or more fragment ions not including y_3 co-eluted to contribute to this peak signal; and (D) chromatographic peak had a Gaussian elution profile. Based on these requirements, peptides were given a quality score between 1 and 4 with 1 being the highest score. Peptides that had chromatographic traces that met all of these requirements were given a quality score of 1. Peptides were given a quality score of 2 if they met requirement A with a signal intensity of at least 20,000 and requirement B but either had only three fragment ions including y_3 contributing to the peak or had an abnormal peak shape. Peptides were given a quality score of 3 if (i) more than one chromatographic peak was detected that met requirements B, C and D; (ii) requirement B was not met; or (iii) if requirements A and D were not met. Peptides not classified as having a quality score of 1, 2 or 3 were given a quality score of 4.

Chromatographic peak intensities from all monitored transitions of a given peptide were integrated and summed to give a final peptide peak height. Fragment ion chromatographic traces that were clearly contaminated by some other ion were removed from this analysis.

Absolute quantification.

Absolute quantification of the GST peptides LLLEYLEEK and IEAIPQIDK was performed using a calibration curve of light isotope–labeled peptides with each sample

containing the same amount of heavy peptide. The calibration points used were 40 nM, 12.5 nM, 5 nM, 2.5 nM, 1 nM, 0.5 nM, 0.25 nM and 0.1 nM each of the light LLLEYLEEK and IEAIPQIDK peptides. All peptide standards were mixed with identical quantities of heavy isotope–labeled LLLEYLEEK and IEAIPQIDK peptides (~5 nM each) and a Bovine QC standard mix (25 nM) (Michrome) in 2% acetonitrile and 0.1% formic acid in water. Peptide standards were measured in triplicate and a linear regression of the data points was used to calibrate the GST peptide light-to-heavy ratio of all other samples.

Relative quantification between nuclei.

Six replicate measurements comprising two technical replicates each of three biological replicates were made for each protein in each of the four cell types. The peptides monitored were the HIST4 peptide DNIQGITKPAIR, the GATA2 peptide GAECFEELSK, the CTCF peptide CPDCDMAFVTSGELVR, the CREB1 peptide ILNDLSSDAPGVPR and the EZH2 peptide EFAAALTAER. For each replicate, the intensity of the target peptide was normalized to the intensity of the HIST4 peptide to control for any variance in the autosampler and/or chromatography. Additionally, as the amount of histone 4 protein should be constant between the four cell types, this normalization should correct for any errors in the measurement of nuclei used for digestion. The mean and s.d. of these normalized intensities were then calculated for each protein in each of the cell types.

Shotgun proteomic mass spectrometry.

Peptide samples were analyzed with an LTQ-VELOS instrument (Thermo) using an 1100 binary pump and autosampler (Agilent). Five microliters of each sample was separated on a 16-cm-long 75-μm inner diameter packed column (Polymicro Technologies) using Jupiter 4u Proteo 90A reverse-phase beads (Phenomenex). Peptides were separated using a 25-min gradient from

8.75% acetonitrile in 0.1% formic acid to 33% acetonitrile in 0.1% formic acid. The gradient was followed by a wash for 15 min at 65% acetonitrile in 0.1% formic acid and a column re-equilibration at 8.75% acetonitrile in 0.1% formic acid for 15 min. Spectra were acquired in data-dependent acquisition (DDA) mode. Raw spectral files were searched using the Sequest algorithm (Eng et al., 1994) and spectra identified using Percolator (Kall et al., 2007) with a false discovery rate cutoff of 1% were used for analysis.

Database spectra analysis.

The 2011_05_26 release of the *Homo sapiens* Ion Trap library of peptide tandem mass spectra was downloaded from NIST (downloaded from http://peptide.nist.gov/ on 12 September 2011). Of the 1,421 peptides monitored in our dataset, 189 had spectra available in the NIST database. Dot-products were calculated using Skyline for all peptides with four or more monitored fragment ions (186 peptides) (Frewen et al., 2006).

ESPPredictor scores.

ESPPredictor scores (Fusaro et al., 2009) were calculated using the Gene Pattern web-tools interface (http://www.broadinstitute.org/cancer/software/genepattern/modules/ESPPredictor.html). Proteins that had eight or more peptides, with one-third of these peptides having a quality score of 1 or 2, were analyzed using the ESPPredictor algorithm (75 total proteins).

Chapter 3 – Protein-centric mapping of nuclear transcription factor occupancy

3.1 – ABSTRACT

The nucleus harbors multiple functionally distinct chromatin environments that impact the functional behaviors of transcription factors (TFs) and other key regulatory proteins. However, current approaches for studying TF-chromatin interactions are predominantly DNA-centric, focusing on measurement of TF occupancy along the genome. Here we present a simple, rapid, protein-centric approach for profiling TF occupancy across diverse nuclear chromatin compartments that couples classical salt fractionation with targeted proteomic analysis. Application to human nuclei reveals that most TFs are surprisingly tightly partitioned across well-defined biochemically and functionally distinct chromatin 'niches'. We find that specific sub-populations of the same TF may concurrently occupy both heterochromatin and euchromatin, and that the nuclear occupancy profiles of transcriptional activators vs. repressors differ markedly, enabling novel functional insights into TFs about which currently little is known. Both splicing and post-translationally modified TF isoforms are extensively compartmentalized within the nucleus, providing a possible mechanism for how TFs can occupy functionally distinct chromatin types. Protein-centric profiling of the nuclear occupancy provides a powerful approach for exposing previously-hidden layers of nuclear TF organization, providing a new vista on transcription factor-chromatin interactions.

3.2 – INTRODUCTION

Within the nuclear milieu, transcription factors (TFs) coordinate the regulation of the genome through their interactions with a variety of distinct regulatory elements and chromatin structures. These chromatin structures, such as euchromatin and heterochromatin, self-organize within the nucleus (Jackson et al., 1993; Lamond and Earnshaw, 1998; Misteli, 2001) and are closely connected with genome function and gene expression (Weintraub and Groudine, 1976). Chromatin structures can be biochemically isolated from nuclei with the use of salt, as protein-nucleic acid interactions are nearly entirely electrostatic in nature (Privalov et al., 2011), and protein-protein interactions often have a large electrostatic component (Sheinerman et al., 2000). Therefore, varying salt concentrations can be used to selectively disrupt the interaction of a DNA-binding protein with its surrounding nuclear material (Weisbrod and Weintraub, 1979; Dignam et al., 1983; Burton et al., 1978; Hyde and Walker, 1975). Biochemical fractionation of nuclear chromatin using salt gradients represents a gold-standard for the physical separation of nuclear structures and chromatin species (Berezney and Coffey, 1974; Bloom and Anderson, 1978; Sanders, 1978; Rocha et al., 1984). Classical heterochromatin and euchromatin are readily separable (Bloom and Anderson, 1978; Sanders, 1978), as are nucleoli (Monty et al., 1956) and structural protein scaffolds (Berezney and Coffey, 1974). Major core histones, linker histones, and some components of the transcriptional machinery are known to partition into specific salt fractions (**Fig. 3.M1a**). This paradigm can also be applied to probe how different nuclear microenvironments align with major genomic features (Henikoff et al., 2009). However, how most TFs and core chromatin components are distributed relative to different biochemically defined chromatin types is unknown.

Current approaches to the study of TF-chromatin interactions rely on chromatin immunoprecipitation (ChIP), which can effectively localize sites of transcription factor occupancy along the genome. However, determining how genomic occupancy sites are distributed relative to major chromatin types is largely inferential. ChIP of histone modifications has been extensively employed to define linearly ordered chromatin 'states' along the genome (Ernst et al., 2011; Filion et al., 2010), yet the relationship between these and classical chromatin species is not clear. Additional deficits of DNA-centric approaches include a general inability to discriminate TF isoforms, binding partners, or specific functions associated with different genomic locations or chromatin environments. DNA-centric approaches also overlook sub-populations of nuclear TFs that are not involved in sequence-specific interactions with the genome, which may be substantial.

Despite the potential insights provided by chromatin fractionation, a major hurdle impeding movement beyond DNA-centric approaches has been the lack of sensitive and accurate quantification methods for TFs. Most TFs lack high-quality antibodies, and shotgun mass spectrometry is not sensitive enough to quantify TFs due to their low concentrations within the human nucleus (Washburn et al., 2001; Ghaemmaghami et al., 2003). By contrast, targeted proteomics, which relies on the use of ‘proteotypic’ peptides that sensitively and uniquely distinguish a specific protein, is significantly more sensitive than shotgun mass spectrometry (Lange et al., 2008; Stergachis et al., 2011). Targeted proteomics can be flexibly applied to virtually any protein, and has been shown to be capable of quantifying TFs and other low abundance proteins within the highly complex human nuclear proteome (Stergachis et al., 2011).

Here we present a simple and practical approach to profiling nuclear TF occupancy patterns within different biochemically-defined chromatin species by combining classical

chromatin fractionation with targeted proteomic analysis. The approach is capable of readily generating in parallel a complete chromatin occupancy pattern for large numbers of TFs (or other proteins of interest). We demonstrate the method on different model human cell types, exposing a diverse landscape of functionally distinct TF compartments within the nucleus, including unexpected cell-selective differences in chromatin occupancy patterns for major TFs, as well as isoform-specific behaviors. Among the findings is that the nuclear populations of a surprisingly large number of TFs are partitioned between diverse chromatin compartments, indicating that DNA-centric methods alone overlook a significant amount of the functional behaviors of regulatory proteins and core chromatin components.

3.3 – RESULTS

Segregation of TFs into biochemically defined chromatin compartments

To sensitively quantify the occupancy of TFs in nuclear chromatin, we developed targeted proteomic assays for diverse TFs using shotgun mass spectrometry derived spectral libraries (Prakash et al., 2009). Due to a systemic lack of peptide spectra from human TFs in publicly available databases, we initially performed traditional shotgun proteomic analysis of bulk and fractionated nuclear extracts from multiple cell types (**Methods**). Using this approach, we identified peptide fragmentation spectra from 387 TFs, and validated peptides from 67 of these using selected reaction monitoring (SRM) on nuclear extracts.

We initially studied the distribution of 14 diverse TFs (tissue-specific, ubiquitous, activating, and repressing) across six chromatin fractions obtained from two well-studied human erythroid (K562) and hepatic (HepG2) model cell lines, and compared this to the distribution of 5 histone proteins, a well-studied integral nuclear envelope protein (NUP210), and a non-

chromatin translation factor (EIF3A) that is abundantly detectable in nuclei (Severin et al., 1997; Chudinova et al., 2004). We observed clean segregation of the large nuclear pore complex member NUP210 into the 600mM salt-stable fraction, and the soluble translational regulator EIF3A into the isotonic-soluble niche (**Fig. 3.M1b**). In addition, the linker histone proteins H1.2;H1.3;H1.4 and the core histone proteins HIST4 and H2AFZ segregated as expected based on prior reports (**Fig. 3.M1a,b**), confirming the performance of the gradients.

In contrast to the structural proteins, the nuclear populations of all 14 transcriptional regulators were partitioned into two or more distinct chromatin compartments (**Fig. 3.M1b**). For example, nuclear c-Jun and RREB1 populations were partitioned between three different compartments: isotonic soluble, 350mM solubilized, and the 600mM salt-stable. These patterns were nearly identical between chromatin fractions obtained from K562 and HepG2 cells, in spite of the fact that these cells differ markedly in their morphology, growth properties, and cadre of expressed genes (The ENCODE Project Consortium, 2011). Two notable exceptions were the erythroid-selective TF NFE2 and the Histone H1 variant H1.5. Measurements performed on whole nuclei revealed that the absolute abundance of NFE2 is significantly greater in K562 vs. HepG2, while H1.5 is more abundant in HepG2 (**Fig. 3.S1**).

Patterning of TF occupancy across specific chromatin niches

We next explored the patterning of a much larger set of TFs (n=96) across a set of 16 chromatin 'niches' defined by fine incremental gradations in salt stability. We developed targeted proteomic assays for these TFs using purified full-length proteins (Stergachis et al., 2011) and validated these using SRM on K562 nuclear extracts (**Methods**). Together these represent roughly half of the TFs predicted to be expressed in K562 cells (Vaquerizas et al.,

2009). We then quantified the relative occupancy of each of the 96 TFs within each of the 16 chromatin niches (total 1,536 TF x niche combinations; **Fig. 3.M2b,c** and **Fig. 3.S2**). Most TFs showed a graded elution profile between 80-360mM NaCl (**Fig. 3.M2b,c**), likely reflecting the disruption of protein-DNA interactions of differing strengths at different salt concentrations. However, occupancy within the readily solubilized isotonic/nucleoplasmic niche varied widely between different TFs, with only half of the tested TFs showing significant nucleoplasmic occupancy. This finding indicates that *in vivo*, most human TFs greatly diverge from the so-called "50/50 rule" which holds that maximal binding rate is achieved when a TF spends half of its time in solution and half scanning along the DNA (Slutsky and Mirny, 2004).

Many TFs are poorly solubilized by widely-used extraction methods

With over 10,000 citations, the Dignam nuclear extract (Dignam et al., 1983) is perhaps the most methodologically important nuclear extraction protocol. The Dignam procedure utilizes ~420mM NaCl extracted chromatin. Although most TFs we analyzed contained sub-populations that were solubilized under these conditions, nearly half of all TFs showed significant occupancy within the 600mM salt-stable chromatin niche (**Fig. 3.M2b**). This indicates that many transcription factors, and their corresponding protein complexes, are either partly or wholly invisible to studies that rely on the Dignam et al. conditions to prepare nuclear protein extract.

TFs can occupy both euchromatic and heterochromatic niches

To characterize the occupancy of TFs within niches corresponding to classically defined euchromatin and heterochromatin, we applied the salt fractionation approach described above in combination with enzymatic disruption of nuclear chromatin using micrococcal nuclease

(MNase)(Bloom and Anderson, 1978; Sanders, 1978; Rocha et al., 1984; Henikoff et al., 2009) (**Fig. 3.M3a**). MNase action rapidly solublizes accessible euchromatin from the nucleus, whereas heterochromatin remains condensed and physically intact even when its underlying DNA framework is digested (Mazia and Jaeger, 1939; Weintraub, 1984) (**Fig. 3.M3b** and **Fig. 3.S3**). Consequently, heterochromatin remains insoluble after MNase treatment until linker histones are displaced and other interactions that support heterochromatin compaction are disrupted under high salt conditions (Weintraub, 1984; Thoma et al., 1979), following which only very large nuclear protein complexes remain. An intermediate state comprising decondensed inactive chromatin is also solubilized by low salt from MNase-treated nuclei. This population of chromatin is thought to arise from H1-depleted nucleosome core particles juxtaposed between insoluble filamentous heterochromatic structures (Weintraub, 1984), and is depleted of active histone acetylation marks (Hebbes et al., 1988).

Unexpectedly, many human TFs do not exclusively occupy classically defined euchromatin (**Fig. 3.M3c** and **Fig. 3.S4**). Widely-studied TFs involved in gene activation and/or binding within active regulatory DNA (e.g. c-Jun, SP-1, GATA-2, CREB1, and CTCF) do exhibit a predominantly euchromatic occupancy profile (**Fig. 3.M3c,** top). However, TFs implication in transcriptional repression, such as NKRF, RREB1, and EZH2, appear to occupy both euchromatic and heterochromatic chromatin niches (**Fig. 3.M2c**, middle). The fact that the chromatin occupancy patterns of well-studied activators vs. repressors differ considerably suggests that chromatin niche occupancy signatures are closely connected with functional regulatory behaviors, which are currently undefined for the majority of annotated TFs.

We also observed a third class of TFs comprising the chromatin repressor HIC2, the uncharacterized putative TF ZNF512 and the erythroid regulator NFE2, which predominantly

occupied the high salt-stable chromatin niche in HepG2 cells (**Fig. 3.M3c**, bottom). Notably, NFE2 exhibits a predominantly euchromatic profile in K562 nuclei (**Fig. 3.S4**), where it is highly expressed and plays a major role in erythroid gene regulation. By contrast, in HepG2 nuclei, where it is expressed at a much lower level with unclear function (Andrews et al., 1993), it largely occupies the high-salt-stable chromatin niche (**Fig. 3.S1**). These results demonstrate that human TFs are capable of occupying functionally distinct chromatin niches within the same nuclear environment, and that chromatin occupancy patterns may differ radically between cell types.

Euchromatic occupancy by a linker histone (H1.x)

The linker histone H1 variants comprise an 11 gene family (Izzo et al., 2008), which differ in their ability to bind to and condense chromatin *in vitro* (Orrego et al., 2007; Clausell et al., 2009), and show distinct *in vivo* kinetics in over-expression studies (Misteli et al., 2000; Lever et al., 2000; Th'ng et al., 2005; Takata et al., 2007). However, the endogenous behaviors of specific H1 variants are largely unknown due to the prior lack of reagents for distinguishing these highly similar proteins *in vivo*. Use of linker histone gene-specific targeted proteomic reagents revealed that, while H1.5 and H1.2/3/4 were found only in the heterochromatic niche (**Fig. 3.M3d**), H1.x was found in both the heterochromatic and euchromatic niches in K562 and HepG2 nuclei (**Fig. 3.M3d** and **Fig. 3.S4**). This novel association of H1.x with biochemically defined euchromatin is consistent with the recent finding that H1.x incorporates into the NANOG promoter upon gene activation (Shahhoseini et al., 2010) and can potentially explain why H1.x has a shorter residence time on nuclear chromatin than the H1.2 variant (Takata et al., 2007). Although H1.x appears to interact *in vivo* with this less condensed euchromatic fraction,

H1.x also elutes in the heterochromatic fraction. As such, under native conditions, linker histones are capable of occupying multiple chromatin niches *in vivo*.

Restriction of individual TF isoforms to distinct chromatin niches

Many TFs are alternatively spliced, generating functional isoforms that contain altered DNA and protein binding capacities(Duma et al., 2006). We therefore asked whether different protein isoforms of the same factor display similar chromatin niche occupancy patterns. We first examined KRAB-associated protein 1 (KAP1), a key regulator of KRAB zinc finger function which is known to encode multiple isoforms (**Fig. 3.M4a**). The full-length isoform 1 of KAP1 occupies multiple chromatin niches within K562 nuclei (**Fig. 3.M4b**). However, immunoblotting and targeted proteomic experiments revealed that isoform 3 of KAP1 was restricted to the 600mM salt-stable chromatin niche within K562 nuclei (**Fig 3.M4b,d**), and was not present in HepG2 nuclei (**Fig. 3.M4c,e**). This suggests that KAP1 isoform 3 is only involved in only a subset of the functions performed by full-length KAP1 isoform 1.

In addition to KAP1, we observed distinct compartmentalization of protein isoforms for the transcriptional regulators NKRF (**Fig. 3.S5a,b**), EZH2 (**Fig. 3.M2c**), PML (**Fig. 3.S6a,b**), ZNF512 (**Fig. 3.S6c,d**) and HIC2 (**Fig. 3.S6e,f**). We also observed two different post-translationally modified isoforms of the pleiotropic genomic regulator CTCF – a major isoform occupying all soluble fractions in both K562 and HepG2 nuclei (**Fig. 3.M2c** and **Fig. 3.S5c,d**), and a distinct post-translationally modified form residing in the 600mM salt-stable fraction from intact nuclei (**Fig. 3.M2c**) and the 80mM salt-stable fraction from MNase-treated nuclei (**Fig. 3.S5e**). Although the precise nature of the modifications could not be readily determined, CTCF is known to be poly(ADP-ribosyl)ated, which results in a protein isoform of the same molecular

weight as the observed band (Yu et al., 2004). Of note, the insulator activity of CTCF is thought to be dependent upon poly(ADP-ribosyl)ation status (Yu et al., 2004), indicating that the compartmentalization of CTCF within the nucleus likely reflects the actions of different functional microenvironments. Overall, these findings suggest that restriction of both splice and posttranslationally modified TF isoforms to distinct chromatin niches is a frequent phenomenon with important functional implications.

3.4 – DISCUSSION

The protein-centric approach for quantifying nuclear transcription factor occupancy we describe clearly shows that nuclear chromatin is partitioned into a series of well-defined biochemical niches occupied by distinct sub-populations of TFs and chromatin structural proteins (**Fig. 3.M5a**). Protein-centric profiling of chromatin niches can be performed very rapidly, with several dozen TFs analyzed in parallel.

A clear and striking implication of our results is that the functional nuclear environments experienced by major transcriptional regulators are far more complex than classical euchromatin-vs-heterochromatin or 'open'-vs-'closed' chromatin paradigms have suggested. Patterned partitioning across different chromatin niches appears to be a general property of human TFs, as it is observed for diverse structural and functional classes and is conserved across cell types. This partitioning is likely to be a direct consequence of diverse protein-protein and protein-DNA interactions maintained *in vivo* (Ravasi et al., 2010; Malovannaya et al., 2011). Additional complexity is imparted by post-translational modifications and by TF isoforms, some of which are restricted to specific niches. Although precise and targeted to individual proteins, our measurements likely reflect steady-state concentrations within a dynamic system. TFs are highly

mobile proteins (McNally et al., 2000; Becker et al., 2002; Phair et al., 2004), and it is possible that the same TF protein molecule cycles between multiple chromatin niches.

The existence of discrete, biochemically separable chromatin niches provides an explanation for many (in some cases contradictory) observed behaviors of regulatory factors. For example, KAP1 is a transcriptional co-repressor that contains a number of domains involved in protein complex formation and repression (Friedman et al., 1996). We observe that the full-length isoform 1 of KAP1 occupies multiple chromatin niches within K562 nuclei, whereas, the TRIM domain lacking KAP1 isoform 3 is restricted to a single niche (**Fig. 3.M4**). This suggests that KAP1 isoform 3 is only involved in a subset of the functions performed by full-length KAP1 isoform 1. This prediction was recently validated using genomic profiling of full length KAP1 and KAP1 deletion constructs (Iyengar et al., 2011). Whereas full-length KAP1 was found bound at both 3' UTRs and promoters, the KAP1 TRIM domain deletion construct, which is nearly identical to isoform 3, is bound only at promoters. Together, these findings argue that KAP1 isoform 1 occupies at least two functionally and physically distinct chromatin niches, whereas KAP1 isoform 3 is restricted to only one of these niches (**Fig. 3.M5b**). Our results thus provide an explanation for what may appear to be discordant or overly complex protein interaction landscapes emanating from measurements on bulk nuclear extracts.

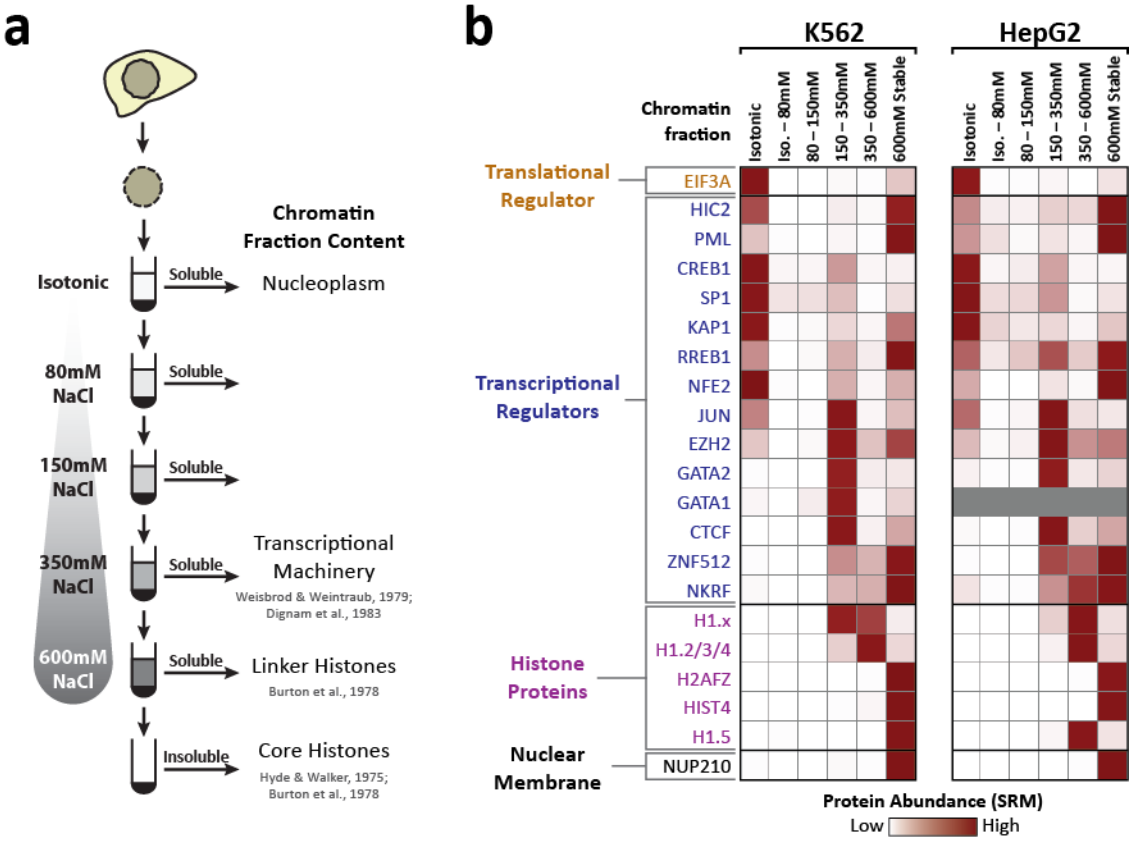
The chromatin niche profiling approach we describe can also illuminate novel regulatory functions for TFs. For example, little is known about ZNF512. However, our results suggest that ZNF512 is predominantly found within heterochromatin and chromatin-independent complexes within the nucleus, suggesting a repressive and/or chromatin-independent function. Furthermore, the erythroid regulator NFE2 appears to play functionally distinct roles in erythroid (K562) and hepatic (HepG2) cell line models, as it predominantly occupies euchromatin in K562 cells and

chromatin-independent complexes in HepG2 cells. Chromatin niche profiling therefore provides a powerful and generic tool for assigning TFs with basic repressive vs. activating functions, and for the discernment of variable TF function under conditions of variable abundance or in different cellular environments. TFs can be analyzed *en masse* using this approach, and additional factors can easily be added using the emerging databases of TF proteotypic peptides (Stergachis et al., 2011).

By providing a generic tool to quantify the nuclear compartmentalization of TFs, the approach we describe can be systematically applied in a wide variety of experimental contexts to expand greatly our understanding of TF function and organization within the nucleus.

3.5 – FIGURES

Figure 3.M1. Segregation of transcription factors into discrete chromatin compartments

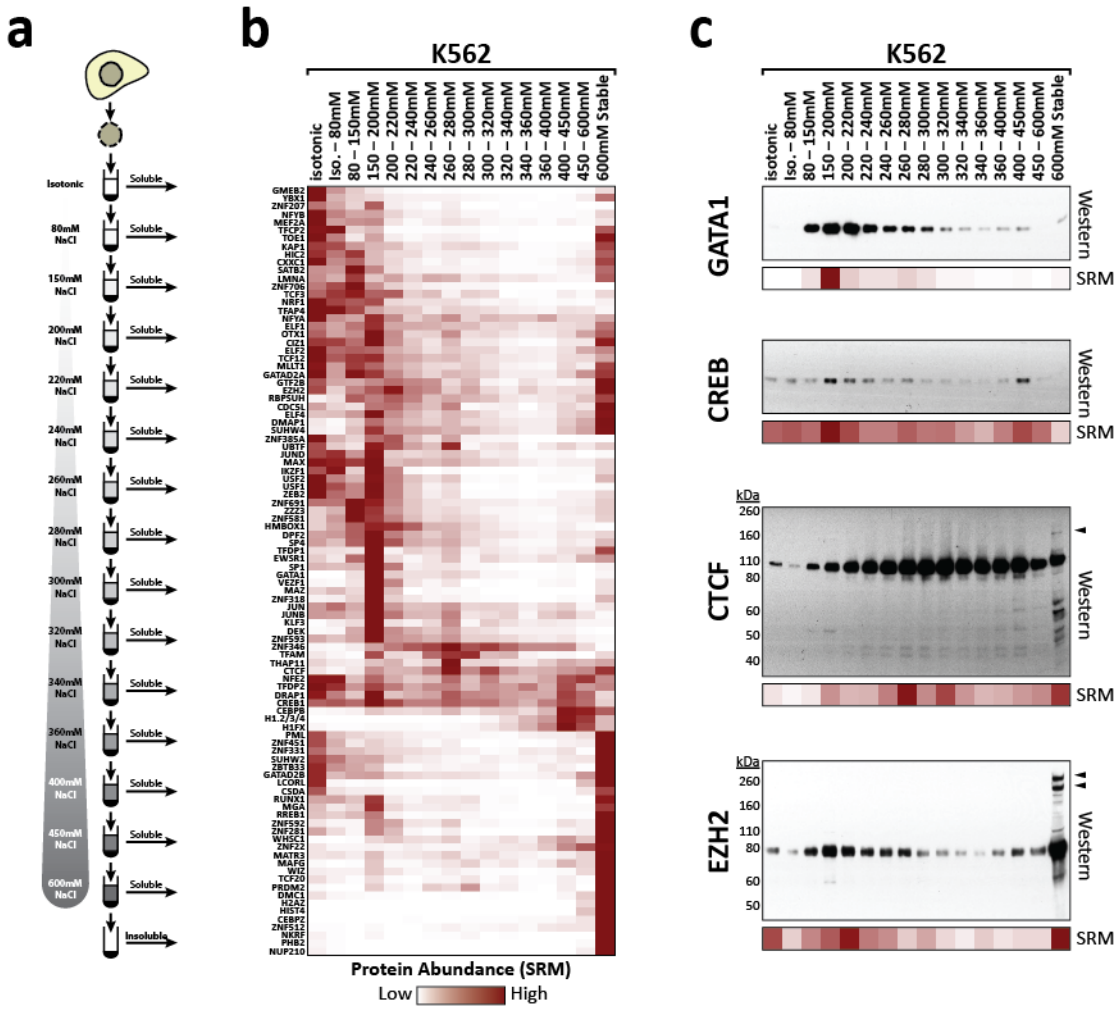


(a) Isolated nuclei were extracted with buffers containing increasing concentrations of NaCl and the proteins solubilized by each wash were analyzed. The displayed fraction contents are based on previous reports.

(b) Heatmap showing the relative protein abundance of 21 nuclear proteins between the 6 different biochemical fractions taken from K562 nuclei (left) or HepG2 nuclei (right). Protein abundance was measured using selected reaction monitoring (SRM) for 2-3 peptides per protein and is row normalized to the maximum value. GATA1 protein was not sufficiently abundant in

HepG2 fractions to quantify using targeted mass spectrometry. The peptides used cannot uniquely distinguish the Histone H1 variants H1.2, H1.3 and H1.4.

Figure 3.M2. Patterned partitioning of TFs across nuclear microenvironments

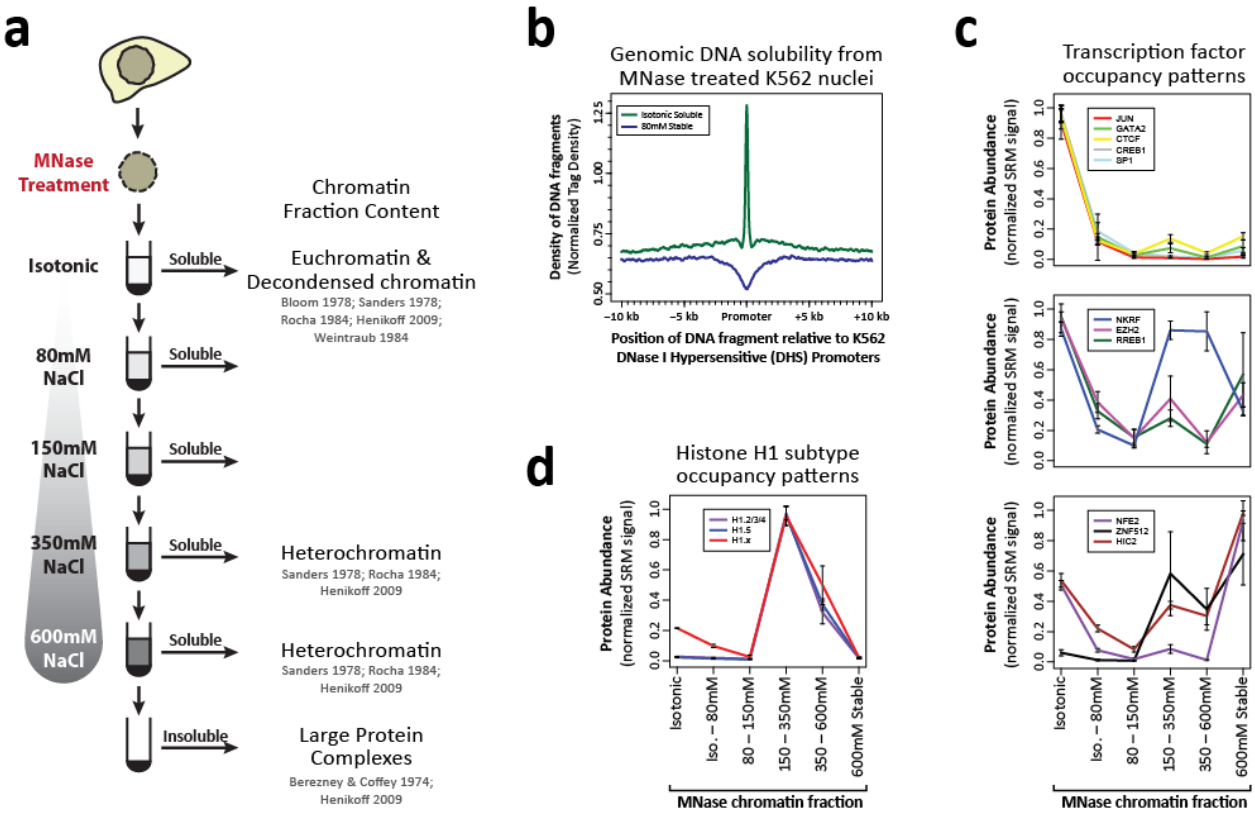


(a) Fine-scale fractionation of K562 nuclei using 20mM NaCl steps.

(b) Heatmap showing the relative abundance of 96 transcriptional regulators between these 16 biochemically defined chromatin niches.

(c) Western blotting and SRM data showing the nuclear occupancy pattern for four transcription factors. Arrowheads mark major protein isoforms that appear selectively within certain chromatin niches.

Figure 3.M3. TFs and linker histones occupy both euchromatic and heterochromatic niches



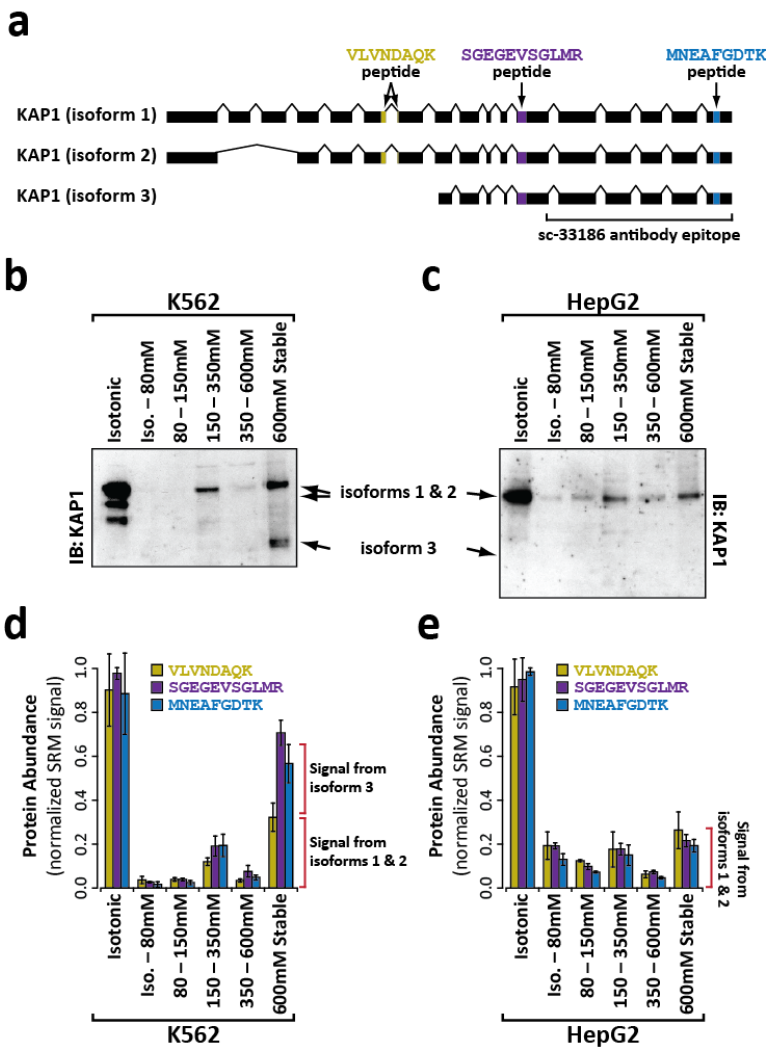
(a) Isolated nuclei were treated with micrococcal nuclease (MNase) and extracted with buffers containing increasing concentrations of NaCl and the proteins solubilized by each wash were analyzed. The displayed fraction contents are based on previous reports.

(b) *DNaseI hypersensitive (DHS) promoters and the surrounding chromatin are readily solubilized from MNase-treated nuclei.* DNA from the 10mM soluble and 80mM salt-stable K562 fractions were isolated and subjected to high-throughput sequencing. For both samples the normalized tag density is centered on DNaseI hypersensitive sites in the promoter regions of K562 cells.

(c) Solubility profile of transcription factors from MNase-treated HepG2 nuclei as measured by SRM. Active chromatin binding factors tend to occupy the low-salt compartment (top), whereas transcriptional repressors tend to occupy both the low-salt and high-salt compartments (middle and bottom). All data points are mean ±95% confidence intervals.

(d) The Histone H1 variant H1.x shows a unique chromatin distribution when compared to the other Histone H1 variants in HepG2 nuclei.

Figure 3.M4. Transcription factor isoforms guide chromatin niche occupancy

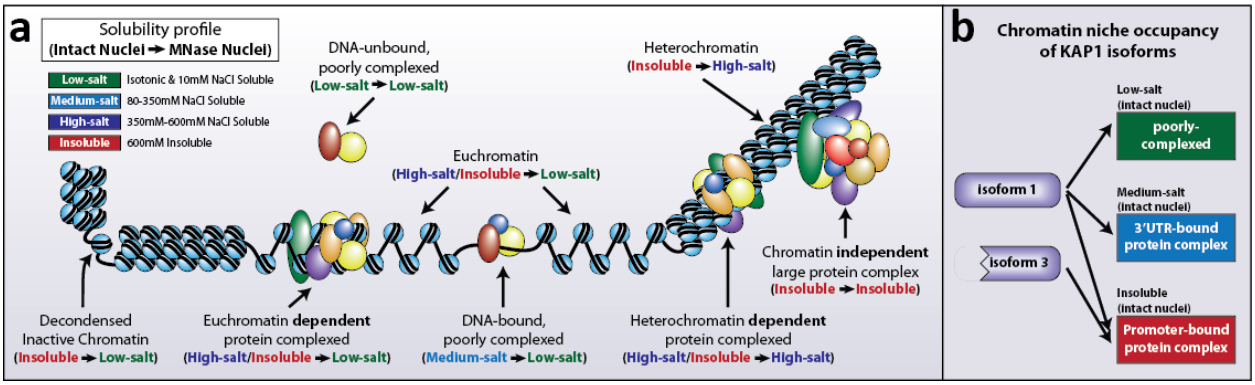


(a) Targeted proteomics and immunoblotting are able to distinguish the 3 unique splicing isoforms of KAP1. The peptides SGEGEVSGLMR and MNEAFGDTK and the antibody sc-33186 can detect all 3 isoforms. However, the peptide VLVNDAQK cannot detect isoform 3.

(b and **c)** KAP1 isoform 3 shows specific occupancy in the K562 600mM salt-stable fraction by western blot of biochemical fractions from K562 (**b**) and HepG2 (**c**) nuclei with an anti-KAP1 antibody (sc-33186).

(d and **e)** KAP1 isoform 3 shows specific occupancy in the K562 600mM salt-stable fraction by targeted proteomics of biochemical fractions from K562 (**d**) and HepG2 (**e**) nuclei. In the K562 600mM salt-stable fraction the KAP1 isoform 3 specific peptides SGEGEVSGLMR and MNEAFGDTK are enriched over KAP1 isoforms 1 and 2 specific peptide VLVNDAQK (**d**). Peptide data is an average of 4 replicates. All data points are mean ±95% confidence intervals.

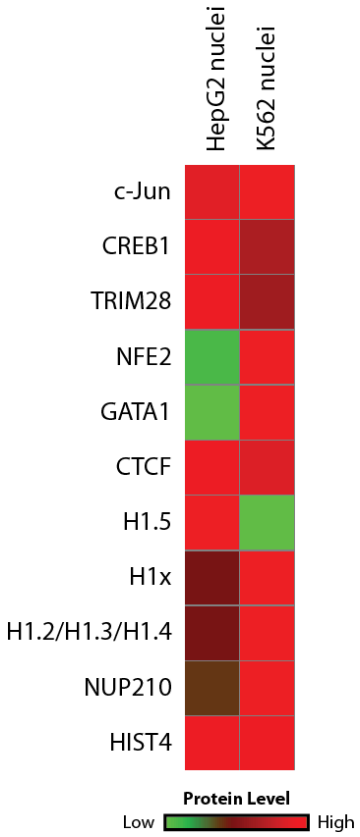
Figure 3.M5. Assortment of human transcription factor into chromatin niches



(a) Schematic representation of the protein components solubilized by different salt concentrations from intact and MNase-treated nuclei. The components of each class are based on findings from this paper as well as previous reports.

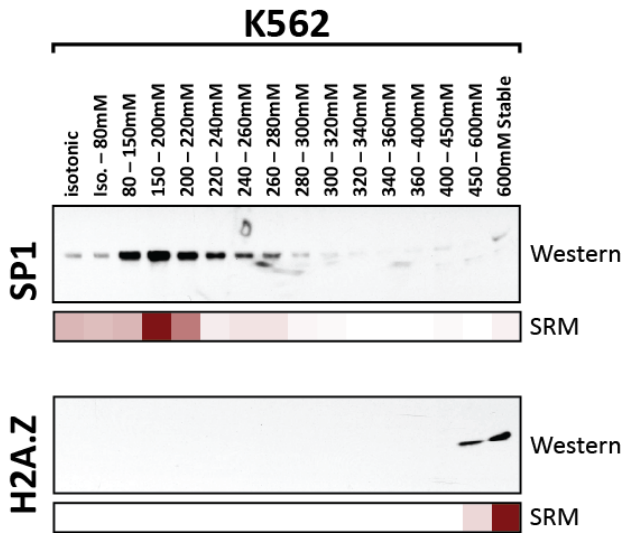
(b) The segregation of KAP1 isoforms into biochemically and functionally defined chromatin niches.

Figure 3.S1. Relative protein abundance between HepG2 and K562 nuclei



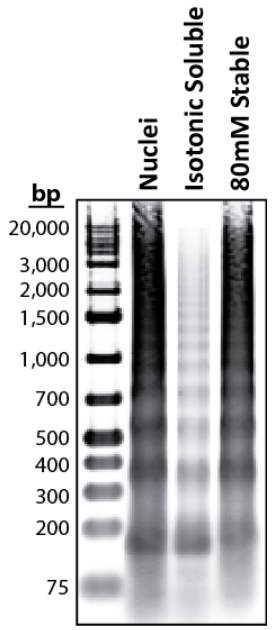
Whole K562 and HepG2 nuclei were digested with trypsin and peptides from each protein were monitored using SRM. The signal from each protein was normalized to the HIST4 signal.

Figure 3.S2. Fine-scale mapping of transcription factor chromatin compartments



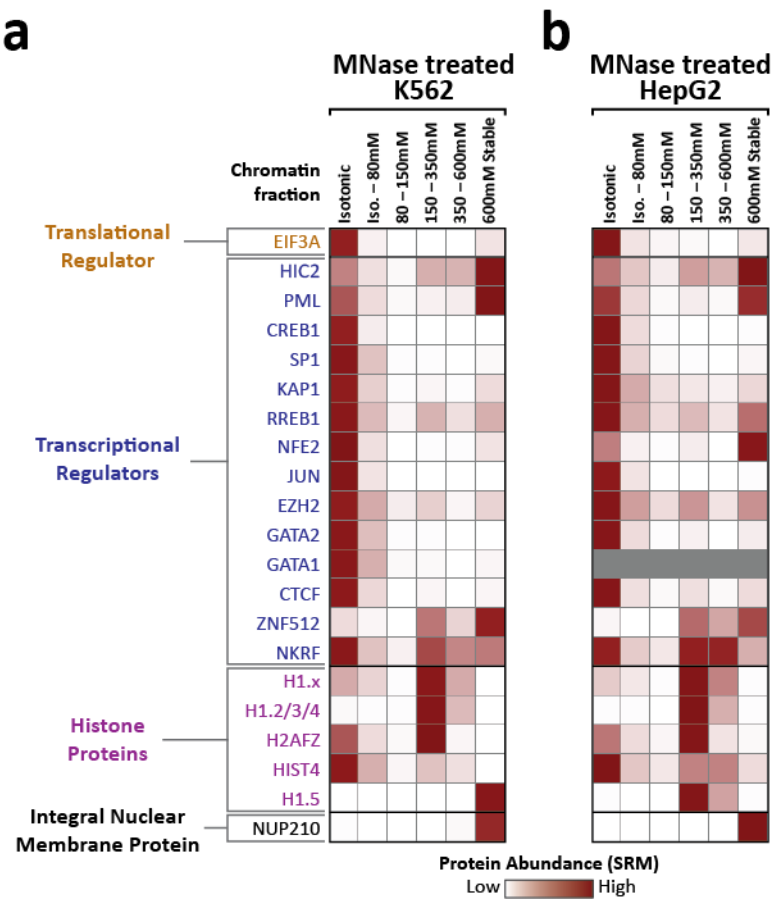
Western blotting and SRM data showing the nuclear occupancy pattern of the transcription factor SP1 and the active histone variant H2A.Z.

Figure 3.S3. Digestion patterns of MNase soluble and stable chromatin



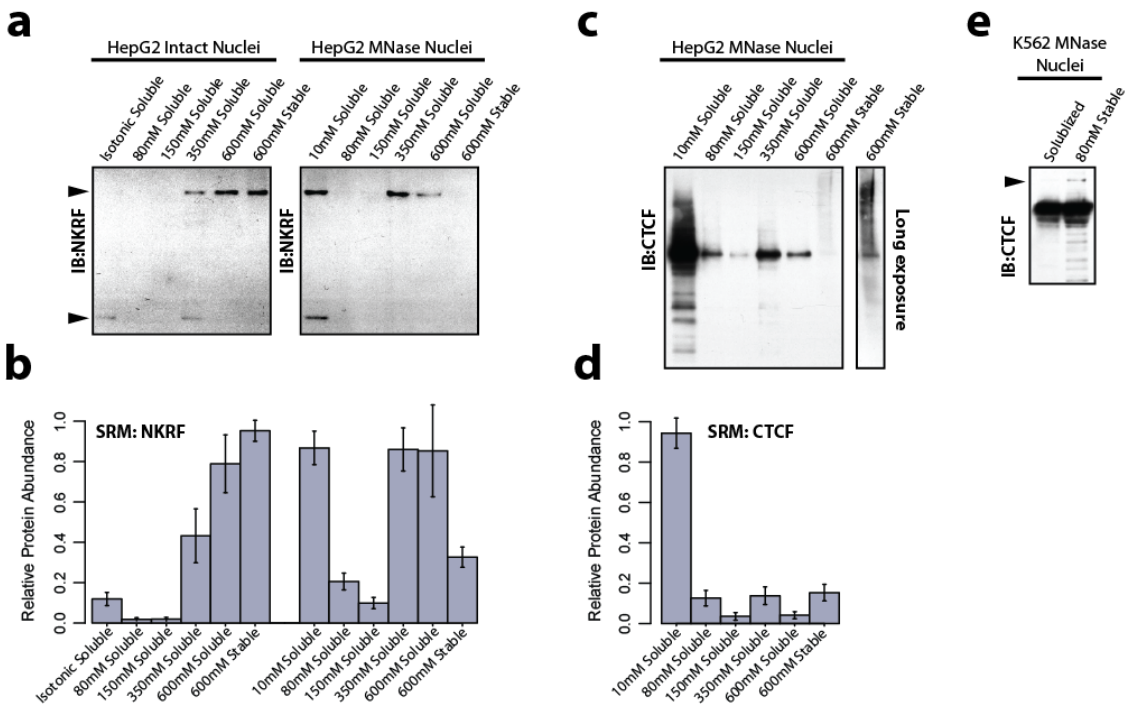
DNA gel showing the size distribution of DNA fragments solubilized from nuclei by low salt and MNase (10mM Soluble). DNA fragments that remain within the nucleus after such treatment are also displayed (80mM stable).

Figure 3.S4. Segregation of transcription factors into discrete chromatin compartments from MNase treated nuclei



(a and **b)** Heatmap showing the relative protein abundance of 21 nuclear proteins between 6 the different biochemical fractions taken from MNase treated K562 (a) or HepG2 nuclei (b). Protein abundance was measured using SRM for 2-3 peptides per protein and is row normalized to the maximum value. GATA1 protein was not sufficiently abundant in HepG2 fractions to quantify using SRM. The peptides used cannot uniquely distinguish the Histone H1 variants H1.2, H1.3 and H1.4.

Figure 3.S5. NKRF and CTCF isoforms occupy distinct chromatin niches



(a) Two major isoforms of the transcriptional repressor NKRF show distinct compartmentalization in HepG2 nuclei (arrowheads).

(b) NKRF relative abundance in the twelve biochemical fractions as measured by SRM using three peptides unique to the protein NKRF.

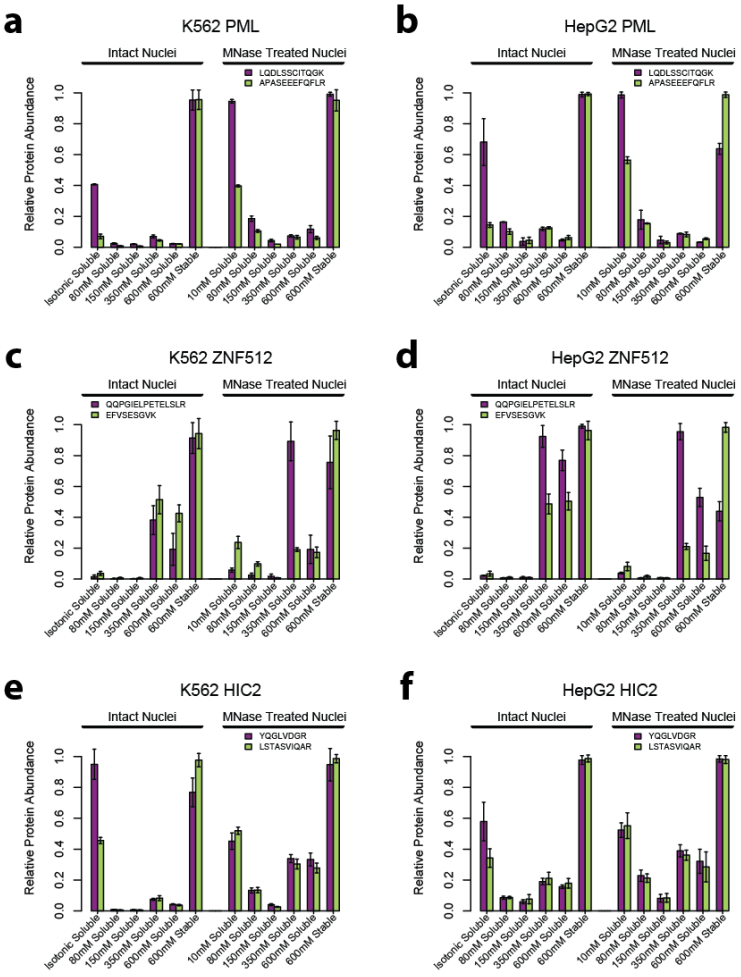
(c) A single major isoform of CTCF occupies all of the soluble fractions from HepG2 MNase treated nuclei. However, CTCF protein appears heavily post-translationally modified in the 600mM salt-stable fraction. “Long exposure” data originated from the same immunoblot.

(d) CTCF relative abundance in the chromatin compartments displayed in (c) was measured using SRM, verifying its occupancy within the 600mM salt-stable compartment. Protein data is

an average of 3 peptides and 4 replicate measurements of each peptide. All data points are mean ±95% confidence intervals.

(e) Immunoblot showing the relative abundance of CTCF in the K562 MNase soluble and 80mM stable fractions. Arrowhead marks the putative poly(ADP-ribosyl)ated CTCF isoform (180kDa) (Yu et al., 2004).

Figure 3.S6. PML, ZNF512 and HIC2 isoforms occupy distinct chromatin niches, as measured by SRM



(a and **b)** Peptides originating from PML show a large discordance in their signal intensity between the different fractions from K562 (a) and HepG2 (b) nuclei. This indicates that at least one of the monitored peptides is modified or enriched/depleted in one of the fraction as a result of fraction specific isoforms or PTMs. SRM peptide abundance plots are an average of 2 replicates.

(c and **d)** Same as (a) and (b), but for ZNF512. SRM peptide abundance plots are an average of 4 replicates.

(e and **f)** Same as (a) and (b), but for HIC2. SRM peptide abundance plots are an average of 4 replicates.

All data points are mean ±95% confidence intervals.

3.6 – METHODS

SRM method development

Peptides suitable for SRM analysis were identified in a similar fashion as described by Prakash et al.(Prakash et al., 2009) and Stergachis et al.(Stergachis et al., 2011).

Due to a systemic lack of peptide spectra from human transcription factors in publicly available databases, MS/MS spectral libraries were generated from 'shotgun' proteomics data collected internally on an LTQ-VELOS (Thermo) and an LTQ (Thermo). Raw spectral files were searched using the Sequest algorithm(Eng et al., 1994). Spectra were identified using Percolator(Kall et al., 2007) with an FDR cut off of 1% were assembled into spectral libraries using the software tool Skyline(MacLean et al., 2010). Spectra from +2 charge state peptides that uniquely mapped to the intended protein were further analyzed. Peptides identified as being potentially useful for SRM were further validated with a TSQ-Quantum triple-quadrupole mass spectrometer (Thermo). The monoisotopic, singly-charged y(3) to y(n-1) product ions from these +2 charge state monoisotopic precursors were monitored in nuclear extracts. Peptides that gave good chromatographic peaks and had a similar ratio of transition ions as seen in the MS/MS spectra were used in all subsequent analyses (**Supplementary Table 1** and **Supplementary Table 2**).

In addition to the approach described above, proteotypic peptides for transcription factors were also identified using the approach described by Stergachis et al.(Stergachis et al., 2011). Briefly, plasmid samples for full-length transcription factor cDNA clones contained in an *in vitro* transcription and translation compatible vector with an in-frame c-terminal GST tag were subjected to the Pierce Human In vitro Protein Expression kit (Thermo). GST-tagged full length proteins were purified using glutathione sepharose 4B beads (GE), digested with trypsin

(Promega) and analyzed using a TSQ-Vantage triple-quadrupole instrument (Thermo) using either a nanoLC separation system (Eksigent) or a nanoACQUITY UPLC (Waters). Targeted proteomic data were analyzed using the software package Skyline(MacLean et al., 2010) to identify proteotypic peptides and their fragmentation patterns (**Supplementary Data 1**). Proteotypic peptides from these proteins were tested using nuclear extracts from K562 cells to identify peptides sensitive enough to detect endogenous levels of the protein (**Supplementary Table 2**).

Nuclear extraction and salt-fractionation

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 similarly to Dorschner et al.(Dorschner et al., 2004) by resuspending cells at $2.5*10^6$ 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 400xg for 7 minutes and washed once with Buffer A. HepG2 cells were handled similarly except that they were grown in MEM (GIBCO) supplemented with 10% Fetal Bovine Serum (PAA), non-essential amino acids (GIBCO), sodium pyruvate (GIBCO), penicillin and streptomycin (GIBCO) and released from the culture flask with trypsin before nuclei were isolated. HepG2 nuclei were isolated in a 0.1% NP-40 Buffer A solution for 8 minutes before centrifugation.

For the nuclear fractionation protocol schematized in Figure 1a, nuclei were transferred to a 37°C water bath and resuspended at $1.25*10^7$ nuclei/mL in Isotonic Buffer (10mM Tris pH 8.0, 15mM NaCl, 60mM KCl, 6mM $CaCl_2$, 0.5mM Spermidine). After 3 minutes at 37°C,

EDTA was added to a final concentration of 15mM and samples were transferred to ice. The soluble and insoluble fractions were separated by centrifugation at 400xg for 7 minutes. The pellet was resuspended in 80mM Buffer B (10mM Tris pH 8.0, 80mM NaCl, 1.5mM EDTA pH 8.0, 0.5mM Spermidine), incubated at 4°C for 1 hour while rocking and then centrifuged at 2000xg for 8 minutes. The pellet was then washed sequentially for 1 hour each with 150mM Buffer B, 350mM Buffer B and 600mM Buffer B in a similar manner as the 80mM Buffer B wash except that the concentration of NaCl in Buffer B was adjusted. All supernatant fractions were cleared by centrifugation at 10,000xg for 10 minutes and any insoluble material was discarded. The 600mM insoluble pellet was sonicated with a Diagenode bioruptor at setting output 5 for 5 minutes to solubilize the proteins.

For the nuclear fractionation protocol schematized in Figure 2a, nuclei were resuspended at $1.25*10^7$ nuclei/mL in 4.2 mL isotonic Buffer (10mM Tris pH 8.0, 15mM NaCl, 60mM KCl, 1.5 mM EDTA pH 8.0, 0.5mM Spermidine) and rocked at 4°C for 35 minutes. The soluble and insoluble fractions were separated by centrifugation at 2,000xg for 7 minutes and the pellet was resuspended in 4.2mL 80mM Buffer B (10mM Tris pH 8.0, 80mM NaCl, 1.5mM EDTA pH 8.0, 0.5mM Spermidine), incubated at 4°C for 35 minutes while rocking and then centrifuged at 2000xg for 7 minutes. The pellet was then washed sequentially for 35 minutes each with 4.2 mLs of 150mM Buffer B, 200mM Buffer B, 220mM Buffer B, 240mM Buffer B, 260mM Buffer B, 280mM Buffer B, 300mM Buffer B, 320mM Buffer B, 340mM Buffer B, 360mM Buffer B, 400mM Buffer B, 450mM Buffer B and 600mM Buffer B in a similar manner as the 80mM Buffer B wash except that the concentration of NaCl in Buffer B was adjusted. All supernatant fractions were cleared by centrifugation at 10,000xg for 8 minutes and any insoluble material was discarded. The 600mM insoluble pellet was sonicated with a Diagenode bioruptor at setting

output 5 for 5 minutes to solubilize the proteins and the volume was brought to 4.2mL using isotonic Buffer.

For the nuclear fractionation protocol schematized in Figure 3a, nuclei were transferred to a 37°C water bath and resuspended at $1.25*10^7$ nuclei/mL in MNase Buffer (25 U/mL MNase [Worthington], 10mM Tris pH 7.5, 10mM NaCl, 1mM CaCl2, 3mM MgCl2, 0.5mM Spermidine). After 3 minutes at 37°C, EDTA was added to a final concentration of 15mM and samples were transferred to ice. The soluble and insoluble fractions were separated by centrifugation at 400xg for 7 minutes. The pellet was resuspended in 80mM Buffer B (10mM Tris pH 8.0, 80mM NaCl, 1.5mM EDTA pH 8.0, 0.5mM Spermidine), incubated at 4°C for 1 hour while rocking and then centrifuged at 2000xg for 8 minutes. The pellet was then washed sequentially for 1 hour each with 150mM Buffer B, 350mM Buffer B and 600mM Buffer B in a similar manner as the 80mM Buffer B wash except that the concentration of NaCl in Buffer B was adjusted. All supernatant fractions were cleared by centrifugation at 10,000xg for 10 minutes and any insoluble material was discarded. The 600mM insoluble pellet was sonicated with a Diagenode bioruptor at setting output 5 for 5 minutes to solubilize the proteins.

Protein Digestion

Protein samples were boiled at 95°C for 5 minutes, reduced with 5mM DTT at 60°C for 30 minutes and alkylated with 15mM iodoacetic acid (IAA) at 25°C for 30 minutes in the dark. Proteins were then digested with Trypsin (Promega) at 37°C for 2 hours while shaking (1µg trypsin per sample) and the pH of the sample was subsequently adjusted to <3.0 using HCl. The digested samples were desalted using a Oasis MCX 30 mg/60 µm cartridges (Waters) following the manufacturer's protocol with minor modifications. Briefly, the cartridge was conditioned using 1 mL methanol, 1 mL 10% ammonium hydroxide in H_2O, 2 mL methanol and finally 3 mL

0.1% formic acid in H2O. The samples were then loaded onto the cartridge and washed with 1 mL 0.1% formic acid in H2O and 1 mL of 0.1% formic acid in methanol. The peptides were eluted from the cartridge with 1 mL 10% ammonium hydroxide in methanol and evaporated down to 10-30µl of volume using a SpeedVac (Labconco) set to 50°C. Peptide samples were resuspended in 50µl 0.1% formic acid in H2O and stored at -20°C until injected on the mass spectrometer.

Mass Spectrometry

Peptide samples were analyzed with either a TSQ-Quantum (Thermo) or a TSQ-Vantage (Thermo) triple-quadrupole instrument using either a nanoLC separation system (Eksigent) or a nanoACQUITY UPLC (Waters). The same volume of each sample was injected. Samples were separated on a 16-20cm 75µm I.D. packed column (Polymicro Technologies) using Jupiter 4u Proteo 90A reverse-phase beads (Phenomenex). A 60 minute gradient from 5% acetonitrile in 0.1% formic acid to 35% acetonitrile in 0.1% formic acid was used to separate the samples. The gradient was followed by a wash for 10 minutes at 80% acetonitrile in 0.1% formic acid and a column re-equilibration at 5% acetonitrile in 0.1% formic acid for 15 minutes. SRM methods were scheduled using Skyline such that all transitions were allotted at least a 40 milliseconds dwell time with a duty cycle of 2.5 seconds or less. This allowed for ~10 data points across the chromatographic peak. The order of sample injections was randomized and quality control runs were staggered every 6-8 injections.

Mass Spectrometry data analysis

Triple-quadrupole data was analyzed using the software package Skyline(MacLean et al., 2010). Chromatographic peak intensities from all monitored transitions of a given peptide were integrated and summed to give a final peptide peak height. For each experiment and peptide,

peak heights from different samples and replicates run on the same column 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 fraction. Final protein data were generated by averaging the fraction specific values from each of the peptides that contribute to that protein. The spectral concordance of each peptide measurement (dot-product) was also calculated using Skyline and any peptides that showed poor dot-product scores were removed from analysis.

Immunoblotting

Undigested protein extract from each of the fractions was boiled in 1x LDS buffer (Invitrogen) and separated on a 4-12% Bis-Tris denaturing and reducing SDS-PAGE gel (Invitrogen). Gels were then transferred using a GENIE blotter (Idea Scientific Co.) to a nitrocellulose membrane (Bio-Rad) for immunoblotting. Membranes were blocked with 5% non-fat dry milk (Safeway) in TBS-tween buffer and probed for NKRF (Santa Cruz Biotechnology, sc-130652), KAP1 (Santa Cruz Biotechnology, sc-33186), EZH2 (Active Motif, #39639), CREB1 (Cell Signaling Technology, #48H2), GATA1 (Cell Signaling Technology, #D52H6), H2A.Z (Cell Signaling Technology, #2718), SP1 (Cell Signaling Technology, #5931) or CTCF (Cell Signaling Technology, #2899). All primary incubations were done at 4°C overnight using manufacture recommended dilutions. Secondary incubations were performed in 5% non-fat dry milk in TBS-tween using 1:10,000 diluted HRP conjugated goat anti-rabbit antibody (Biorad) or HRP conjugated goat anti-mouse antibody (Amersham). Membranes were visualized using an ECL plus western blotting kit (Amersham) and detected with radiographic film (Thermo).

MNase solubilized and insoluble DNA analysis

K562 nuclei were treated with 25 U/mL MNase as described above, centrifuged at 400xg, and the supernatant collected and cleared by an additional centrifugation at 10,000xg for 10 minutes. This MNase solubilized fraction was then treated with 200µg of proteinase K (Invitrogen) at 55°C overnight in a PK buffer (10mM Tris pH 8.0, 1mM EDTA pH 8.0, 0.5% SDS). DNA was then extracted using phenol/chloroform, sonicated to obtain a more consistent size for Illumina sequencing, and finally ethanol precipitated. Sequencing libraries were generated and sequenced on an Illumina Genome Analyzer IIx (Illumina, Hayward, CA) by the High-Throughput Genomics Unit (University of Washington) following a standard protocol. The insoluble pellet after 80mM Buffer B extraction of 25 U/mL MNase treated K562 nuclei was handled in a similar fashion except that sonication was performed before proteinase K digestion.

Thirty-six base pair Illumina sequence reads were mapped to the human genome (UCSC hg19) allowing up to 2 mismatches using bowtie. Reads mapping to more than one location were then discarded. Smoothed density tracks were generated by counting the number of tags overlapping a sliding 150-bp window, with a step width of 20-bp. Density tracks were normalized for sequencing depth by dividing the tag count by the number of mapped tags, then multiplying by 10 million.

We identified 31,106 K562 DNaseI Hypersensitive sites within +/-2kb of a RefSeq TSS using publicly available ENCODE data(The ENCODE Project Consortium, 2011). To reduce the effect of outliers, we excluded 171 regions with a density higher than the 99.975th percentile (24 tags) in the MNase soluble or insoluble sample. The tag density +/-50kb of the remaining regions was averaged using a simple mean.

Chapter 4 – Genome-wide DNaseI footprints encode global maps of transcription factor occupancy

Note: This work was published in the September 2012 edition of *Nature* as*:*

*Neph, S., *Vierstra, J., *Stergachis, A. B., *Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K., et al. (2012). **An expansive human regulatory lexicon encoded in transcription factor footprints.** *Nature* 489, 83–90.

** indicates co-first authors*

4.1 – ABSTRACT

Regulatory factor binding to genomic DNA protects the underlying sequence from cleavage by DNase I, leaving nucleotide-resolution footprints. Using genomic DNase I footprinting across 41 diverse cell and tissue types, we detected 45 million transcription 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 DNase I 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. 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.

4.2 – INTRODUCTION

Sequence-specific transcription factors interpret the signals encoded within regulatory DNA. The discovery of DNase I footprinting over 30 years ago (Galas and Schmitz, 1978) revolutionized the analysis of cis-regulatory sequences in diverse organisms, and directly enabled the discovery of the first human sequence-specific transcription factors (Dynan and Tjian, 1983). Binding of transcription factors to regulatory DNA regions in place of canonical nucleosomes triggers chromatin remodelling, resulting in nuclease hypersensitivity (Gross and Garrard, 1988). Within DNase I hypersensitive sites (DHSs), DNase I cleavage is not uniform; rather, punctuated binding by sequence-specific regulatory factors occludes bound DNA from cleavage, leaving footprints that demarcate transcription factor occupancy at nucleotide resolution (Galas and Schmitz, 1978; Hesselberth et al., 2009) (**Fig. 4.M1a**). DNase I footprinting has been applied widely to study the dynamics of transcription factor occupancy and cooperativity within regulatory DNA regions of individual genes (Thanos and Maniatis, 1995), and to identify cell- and lineage-selective transcriptional regulators (Tsai et al., 1989).

4.3 – RESULTS

Regulatory DNA is populated with DNase I footprints

To map DNase I footprints comprehensively within regulatory DNA, we adapted digital genomic footprinting (Hesselberth et al., 2009) to human cells. The ability to resolve DNase I footprints sensitively and precisely is critically dependent on the local density of mapped DNase

I cleavages (**Fig. 4.S1a–d**), and efficient footprinting of a large genome such as human requires substantial concentration of DNase I cleavages within the small fraction (1-3%) of the genome contained in DNase I-hypersensitive regions. We selected highly enriched DNase I cleavage libraries from 41 diverse cell types in which 53-81% of DNase I cleavage sites localized to DNase I-hypersensitive regions (Thurman et al., 2012), representing nearly tenfold higher signal-to-noise ratio than previous results from yeast (Hesselberth et al., 2009), and two- to fivefold greater enrichment than achieved using end-capture of single DNase I cleavages (Sabo et al., 2004; Boyle et al., 2008). 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. We achieved an average sequencing depth of ~ 273 million DNase I cleavages per cell type that enabled extensive and accurate discrimination of DNase I footprints.

To detect DNase I footprints systematically, we implemented a detection algorithm based on the original description of quantitative DNase I footprinting1 (Methods). We identified an average of 1.1 million high-confidence (false discovery rate (FDR) of 1%) footprints per cell type (range 434,000 to 2.3 million), and collectively 45,096,726 640-base pair (bp) footprint events across all cell types. We resolved cell-selective footprint patterns to reveal 8.4 million distinct elements with a footprint, each occupied in one or more cell type. At least one footprint was found in 75% of DHSs (**Fig. 4.S1c,d**), with detection strongly dependent on the number of mapped DNase I cleavages within each DHS. 99.8% of DHSs with 250 mapped DNase I cleavages contained at least one footprint, indicating that DHSs are not simply open or nucleosome-free chromatin features, but are constitutively populated with DNase I footprints. Modelling DNase I cleavage patterns using empirically derived intrinsic DNA cleavage propensities for DNase I showed that only a miniscule fraction (0.24%) of discovered 1% FDR

footprints from cell and tissue samples could be caused by inherent DNase I sequence specificity (Methods).

DNase I footprints were distributed throughout the genome, including intergenic regions (45.7%), introns (37.7%), upstream of transcriptional start sites (TSSs, 8.9%), and in 5′ and 3′ untranslated regions (UTRs, 1.4% and 1.3%, respectively; **Fig. 4.S2a,b**). DNase I footprints were enriched in promoters (3.6-fold; $P < 2.2 \times 10{-}16$; Binomial test) and 5′ UTRs (2.4-fold; $P < 2.2 \times 10{-}16$; Binomial test), commensurate with high DNase I 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.

Footprints are quantitative markers of factor occupancy

We next examined the correspondence between DNase I footprints and known regulatory factor recognition sequences within DNase I hypersensitive chromatin. Comprehensive scans of DNase I hypersensitive regions for high-confidence matches to all recognized transcription factor motifs in the TRANSFAC (Matys et al., 2006) and JASPAR (Bryne et al., 2008) databases revealed a striking enrichment of motifs within footprints ($P \approx 0$, z-score = 204.22 for TRANSFAC; z-score = 169.88 for JASPAR; **Fig. 4.M1b** and **Fig. 4.S3**).

To quantify the occupancy at transcription factor recognition sequences within DHSs genome-wide, we computed for each instance a footprint occupancy score (FOS) relating the density of DNase I cleavages within the core recognition motif to cleavages in the immediately flanking regions (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

(Galas and Schmitz, 1978). To examine this relationship for a well-studied sequence-specific regulator (NRF1), we plotted DNase I cleavage patterns surrounding all 4,262 NRF1 motifs contained within DHSs and ranked these by FOS. Whereas only a subset of these motif instances (2,351) coincided with high-confidence footprints, the vast majority of NRF1 motif instances in DNase I footprints (89%) overlapped reproducible sites of NRF1 occupancy identified by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) (**Fig. 4.M1c**). In parallel, we analysed nucleotide-level evolutionary conservation patterns around NRF1-binding sites, revealing that FOS closely parallels phylogenetic conservation within the core motif region, indicating strong selection on factor occupancy (**Fig. 4.M1c**). We observed a nearly monotonic relationship between FOS and ChIP-seq signal intensities at NRF1-binding sites within DNase I footprints of K562 cells (**Fig. 4.M1d**). Similarly strong correlations between footprint occupancy and either ChIP-seq signal or phylogenetic conservation were evident for diverse factors (Fig. 1d and Supplementary Fig. 4a–d). We found that footprint occupancy and nucleotide-level conservation correlated for 80% of all transcription factor motifs in the TRANSFAC database, of which 50% were statistically significant ($P < 0.05$; 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 transcription factor milieu using targeted mass spectrometry (DNA

interacting protein precipitation or DIPP; Methods). Using DIPP, we affirmed specific binding by several different classes of transcription factor (**Fig. 4.S5a–e**). Together with the analysis of ChIP-seq data described above, these results indicate that the localization of transcription factor recognition motifs within DNase I footprints can accurately illuminate the genomic protein occupancy landscape.

Footprints harbour functional SNVs and lack methylation

The potential for single nucleotide variants (SNVs) within a transcription factor recognition sequence to abrogate binding of its cognate factor is well known (Rockman and Wray, 2002). The depth of sequencing performed in the context of our footprinting experiments provided hundreds- to thousands-fold coverage of most DHSs, enabling precise quantification of allelic imbalance within DHSs harbouring heterozygous variants. We scanned all DHSs for heterozygous SNVs identified by the 1000 Genomes Project (Durbin et al., 2010) and measured, for each DHS containing a single heterozygous variant, the proportion of reads from each allele. We identified likely functional variants conferring significant allelic imbalance in chromatin accessibility and analysed their distribution relative to DNase I footprints. This analysis revealed significant enrichment ($P < 2.2 \times 10{-}16$; Fishers exact test) of such variants within DNase I footprints (**Fig. 4.S6**). For example, rs4144593 is a common T-to-C (T/C) variant that lies within a DHS on chromosome 9. This variant falls on a high-information position within a footprint containing an NF1/CTF1 motif and substantially disrupts footprinting of this motif, resulting in allelic imbalance in chromatin accessibility (**Fig. 4.M2a**).

Protein–DNA interactions are also sensitive to cytosine methylation (Tate and Bird, 1993; Lister et al., 2009). Comparing DNase I footprints and whole-genome bisulphite

sequencing methylation data from pulmonary fibroblasts (IMR90), we found that CpG dinucleotides contained within DNase I footprints were significantly less methylated than CpGs in non-footprinted regions of the same DHS (MannWhitney U-test; $P < 2.2 \times 10{-}16$; **Fig. 4.M2b**). Footprints therefore seem to be selectively sheltered from DNA methylation, indicating a widespread connection between regulatory factor occupancy and nucleotide-level patterning of epigenetic modifications.

Transcription factor structure is imprinted on the genome

We observed surprisingly heterogeneous base-to-base variation in DNase I 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 (**Fig. 4.S7**). This raised the possibility that DNase I 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 DNase I cleavage patterns at individual nucleotide positions onto the DNA backbone of the co-crystal model. **Figure 4.M3a** and Fig. **4.S8a** show two examples: USF1 and SRF (Ferré-D'Amaré et al., 1994; Pellegrini et al., 1995). For both factors, DNase I cleavage patterns clearly parallel the topology of the protein–DNA interface, including a marked depression in DNase I 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 DNase I 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 DNase I cleavage in parallel with per-nucleotide vertebrate conservation calculated by phyloP (Pollard et al., 2010) revealed striking antiparallel patterning of cleavage versus conservation across nearly all motifs examined (six representative examples are shown in **Fig. 4.M3b** and **Fig. 4.S8b**). Notably, conservation is not limited to only DNA contacting protein residues, but exhibits graded changes that mirror DNase I accessibility across the entirety of the protein–DNA interface (**Figs 4.S8c,d**). Taken together, these results imply that regulatory DNA sequences have evolved to fit the continuous morphology of the transcription factor-DNA binding interface.

Footprints encode an expansive cis-regulatory lexicon

Since the discovery of the first sequence-specific transcription factor (Gilbert and Müller-Hill, 1967), considerable effort has been devoted to identifying the cognate recognition sequences of DNA-binding proteins (Xie et al., 2005; Mukherjee et al., 2004). 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 (Vaquerizas et al., 2009).

We reasoned that the genomic sequence compartment defined by DNase I 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 transcription factor 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. 4.M4a**). We compared these

models with the universe of experimentally grounded motif models in the TRANSFAC, JASPAR and UniPROBE (Newburger and Bulyk, 2009) databases. Owing to the redundancy of motif models contained within these databases, we first collapsed all duplicate models (Methods). A total of 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. 4.M4b** and **Fig. 4.S9a–c**). The wholesale *de novo* derivation of the vast majority of known regulatory factor recognition sequences from the small genomic compartment defined by DNase I footprints highlights the marked concentration of regulatory information encoded within this sequence space.

Notably, 289 of the footprint-derived motifs were absent from major databases (**Fig. 4.M4b** and **Fig. 4.S9d**). These novel motifs populate millions of DNase I footprints (**Fig. 4.M4c**), 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 4.M3b**, **4.M4e**, **4.S8** and **4.S10a**).

To test whether novel motifs were functionally conserved in an evolutionarily distant mammal, we analyzed DNase I cleavage patterns around human novel motifs mapped within DHSs assayed in primary mouse liver tissue (**Fig. 4.M4e,f** and **Fig. 4.S10a,b**). This analysis demonstrated that many novel motifs show nearly identical DNase I 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 mouse and human.

Given the conservation of protein occupancy in a distant mammal, we assessed whether the novel motifs are under selection in human populations by analyzing nucleotide diversity across all motif instances found within accessible chromatin. Using high-quality genomic

sequence data from 53 unrelated individuals (Drmanac et al., 2010), we calculated the average nucleotide diversity (Nei and Li, 1979) for each individual motif space (**Fig. 4.S10c**). 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. 4.M4d** and **Fig. 4.S10c**), even after exclusion of motifs containing highly mutable CpG dinucleotides, which underlie the marked increase in nucleotide diversity seen with a subset of known motifs (**Fig. 4.S10c**, right). Collectively, these results demonstrate that DNase I footprints encode an expansive cis-regulatory lexicon encompassing both known transcription factor recognition sequences and novel motifs that are functionally conserved in mouse and bear strong signatures of ongoing selection in humans.

Novel motif occupancy parallels regulators of 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. 4.M5a**), presumably due to the repressive action of the transcriptional regulator NRSF (also called REST) at the VGF promoter in non-neuronal cell types (Schoenherr and Anderson, 1995). Although VGF is expressed only in neuronal cells, its promoter is DNase I-hypersensitive in most cell types. 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 DNase I 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 DNase I footprints and normalized this by the total number of DNase I footprints in that cell type. **Figure 4.M5b** shows a heat-map 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 (also called POU5F1), SOX2, KLF4 and NANOG in human embryonic stem cells (Takahashi et al., 2007); (2) the myogenic factors MEF2A and MYF6 in skeletal myocytes (Yun and Wold, 1996); and (3) the erythrogenic regulators GATA1, STAT1 and STAT5A in erythroid cells (Pevny et al., 1991; Socolovsky et al., 2001; Halupa et al., 2005) (**Fig. 4.M5b**).

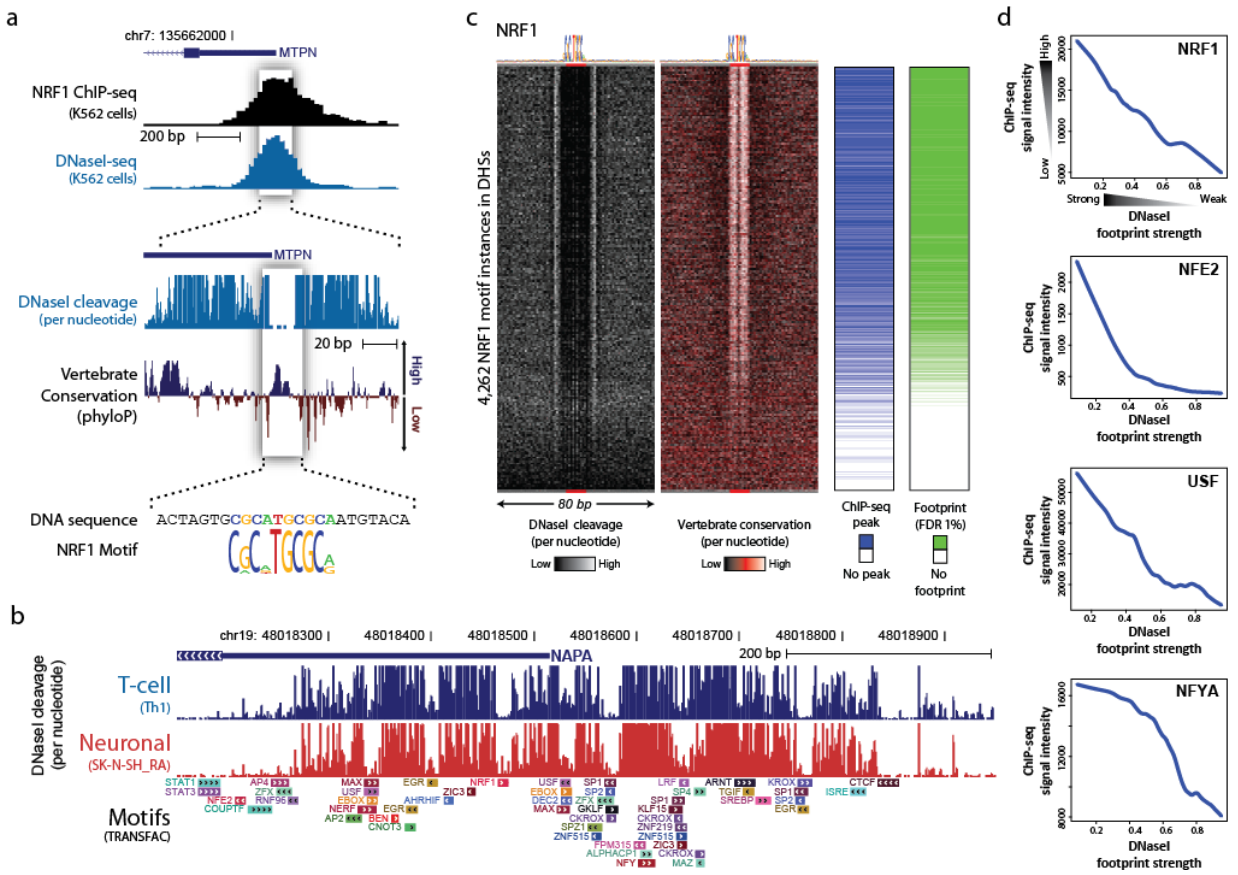
Many of the footprint-derived novel motifs displayed markedly cell-selective occupancy patterns highly similar with the aforementioned well-established regulators. This suggests that many novel motifs correspond to recognition sequences for important but uncharacterized regulators of fundamental biological processes. Notably, both known and novel motifs with high cell-selective occupancy predominantly localized to distal regulatory regions (**Fig. 4.M5c**), further highlighting the role of distal regulation in developmental and cell-selective processes (Treisman and Maniatis, 1985; Grosveld et al., 1987).

4.4 – DISCUSSION

We describe an expansive map of regulatory factor occupancy at millions of precisely demarcated sequence elements across the human genome revealed by genomic DNase I footprinting applied to a wide spectrum of cell types. These elements collectively define a highly information-rich genomic sequence compartment that 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 DNase I cleavage events in most DHSs is not saturating (even in cell types with >500 million mapped unique DNase I cleavages), the present study, although 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, although one that 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 that 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.

4.5 – FIGURES

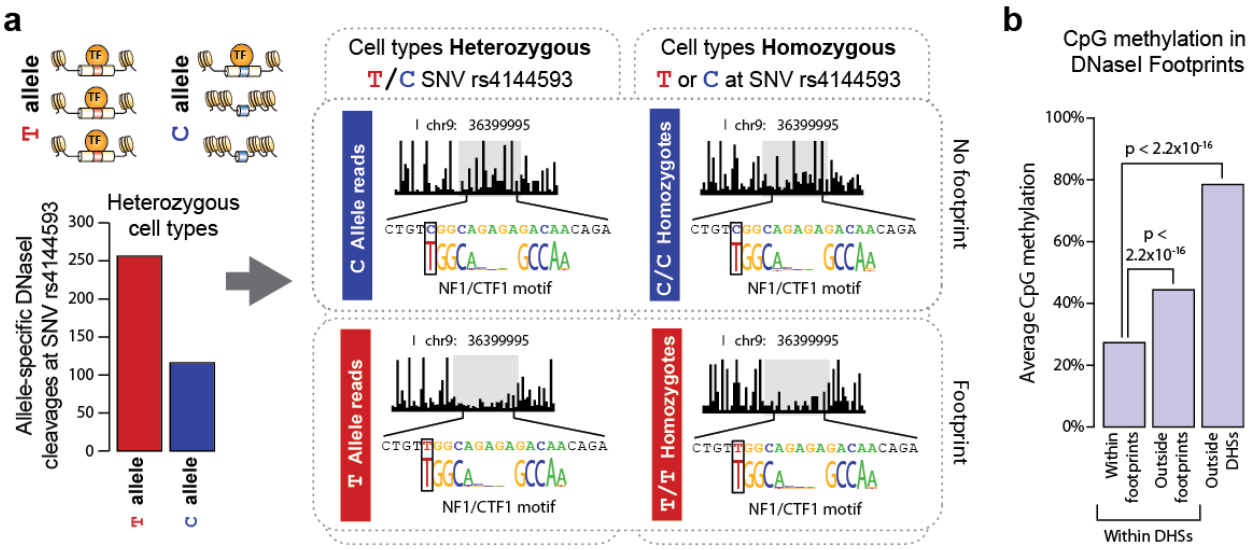
Figure 4.M1. Parallel profiling of genomic regulatory factor occupancy across 41 cell types.



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

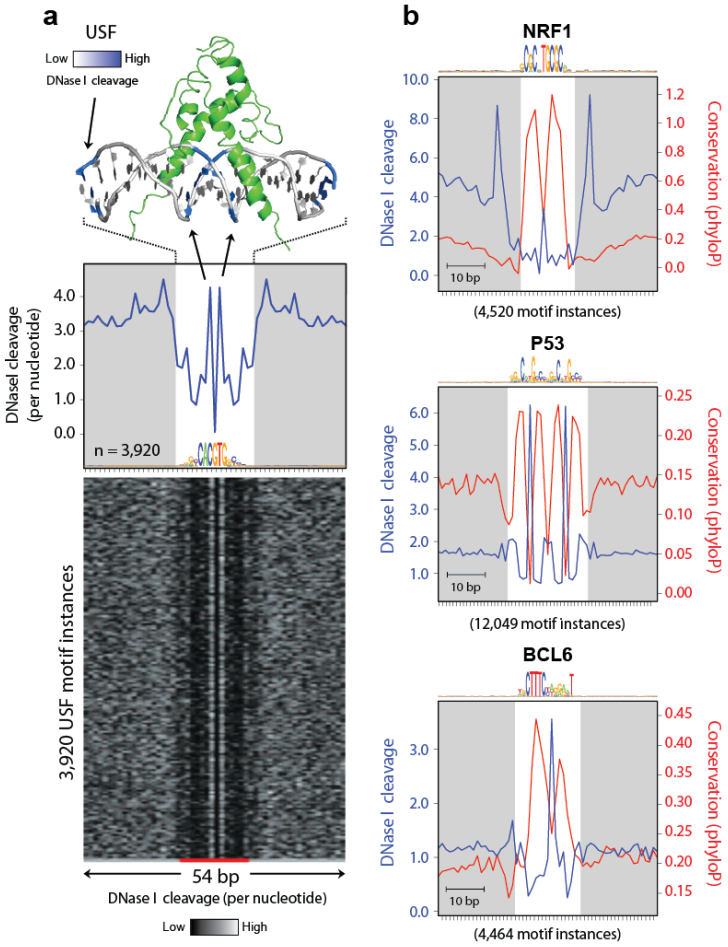
regression of NRF1, USF, NFE2, and NFYA K562 ChIP-seq signal intensities versus DNaseI footprinting occupancy (footprint occupancy score) at K562 DNaseI footprints containing NRF1, USF, NFE2, and NFYA motifs.

Figure 4.M2. 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).

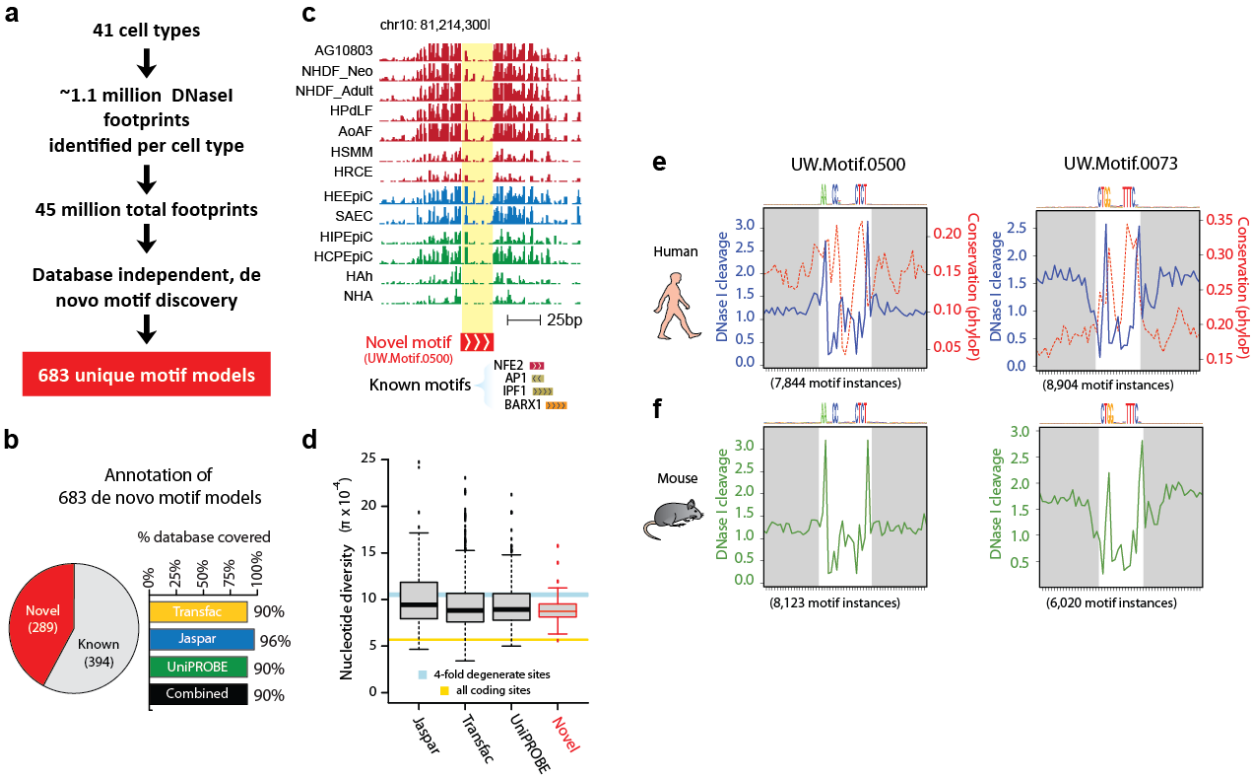
Figure 4.M3. Footprint structure parallels TF structure and is imprinted on the human genome.



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.

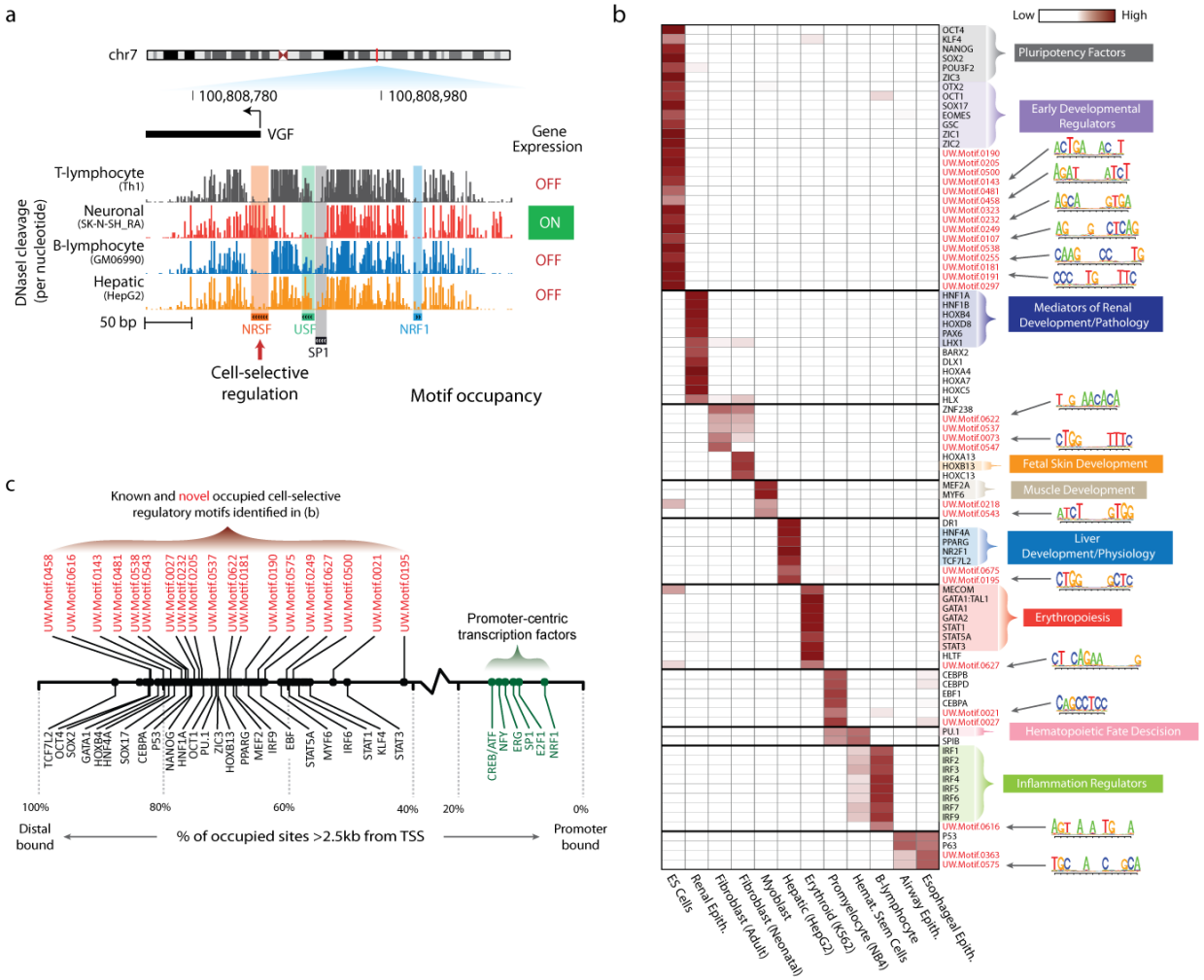
Figure 4.M4. De novo motif discovery expands the human regulatory lexicon.



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.

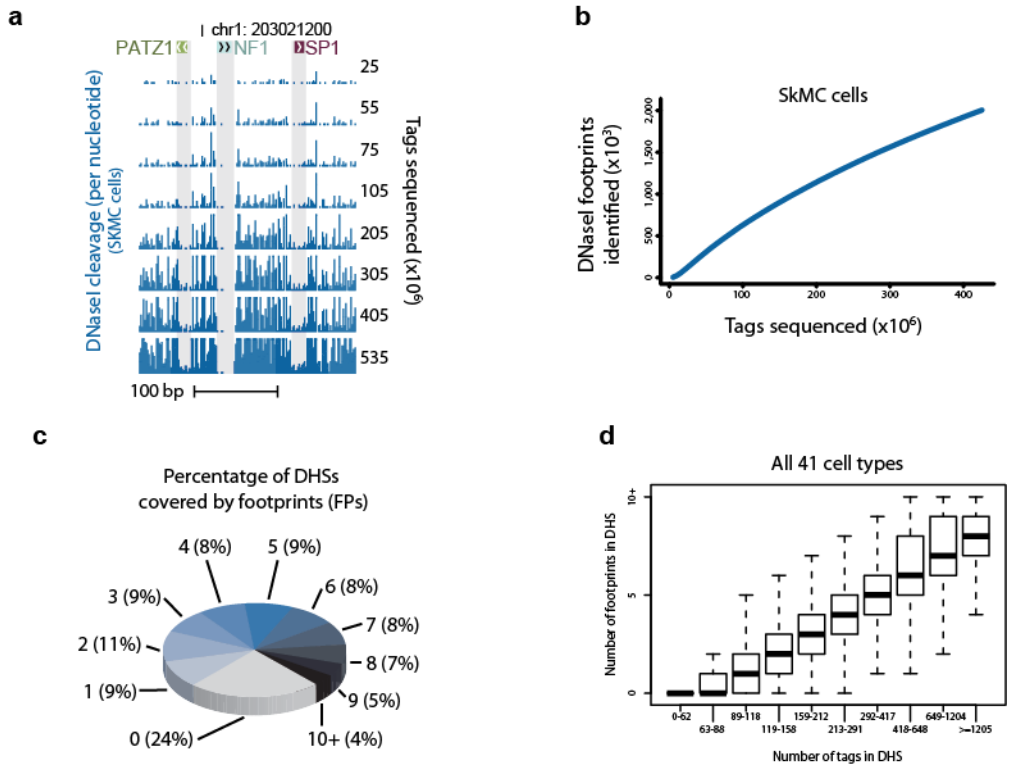
Figure 4.M5. Multi-lineage DNaseI footprinting reveals cell-selective gene regulators.



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

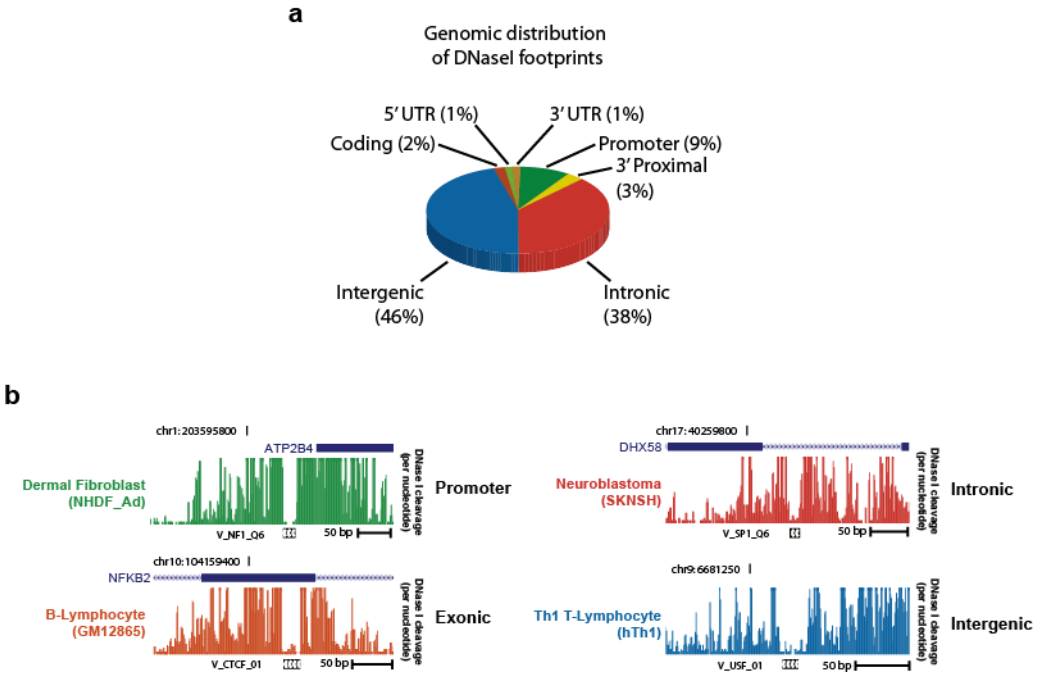
footprints within distal regulatory regions for known (black) and novel (red) cell-type specific regulators in (b) is indicated. Also noted are these values for a small set of known promoter-proximal regulators (green).

Figure 4.S1. Identification and distribution of DNaseI footprints.



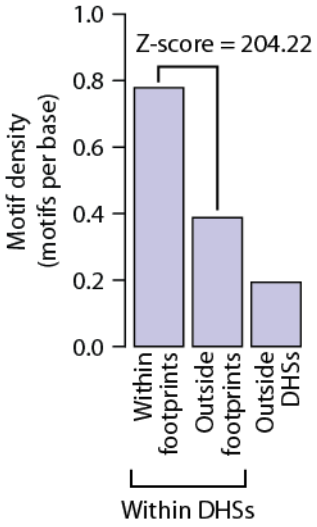
a, As more DNaseI cleavages are sequenced from SKMC cells, individual DNaseI footprints are easier to distinguish. **b,** The number of DNaseI footprints identified in SKMC cells at varying DNaseI cleavage tag sequencing levels. **c-d,** The number of footprints in DHSs is higher for DHSs with more mapped DNaseI cleavages. DHSs from all 41 cell types were broken into deciles based on the sequencing depth of that DHS. The number of mapped DNaseI cleavages for DHSs in each quantile is indicated below the graph. The box-and-wisker plot shows the distribution of the number of footprints within DHSs for each quantile.

Figure 4.S2. Distribution of DNaseI footprints.



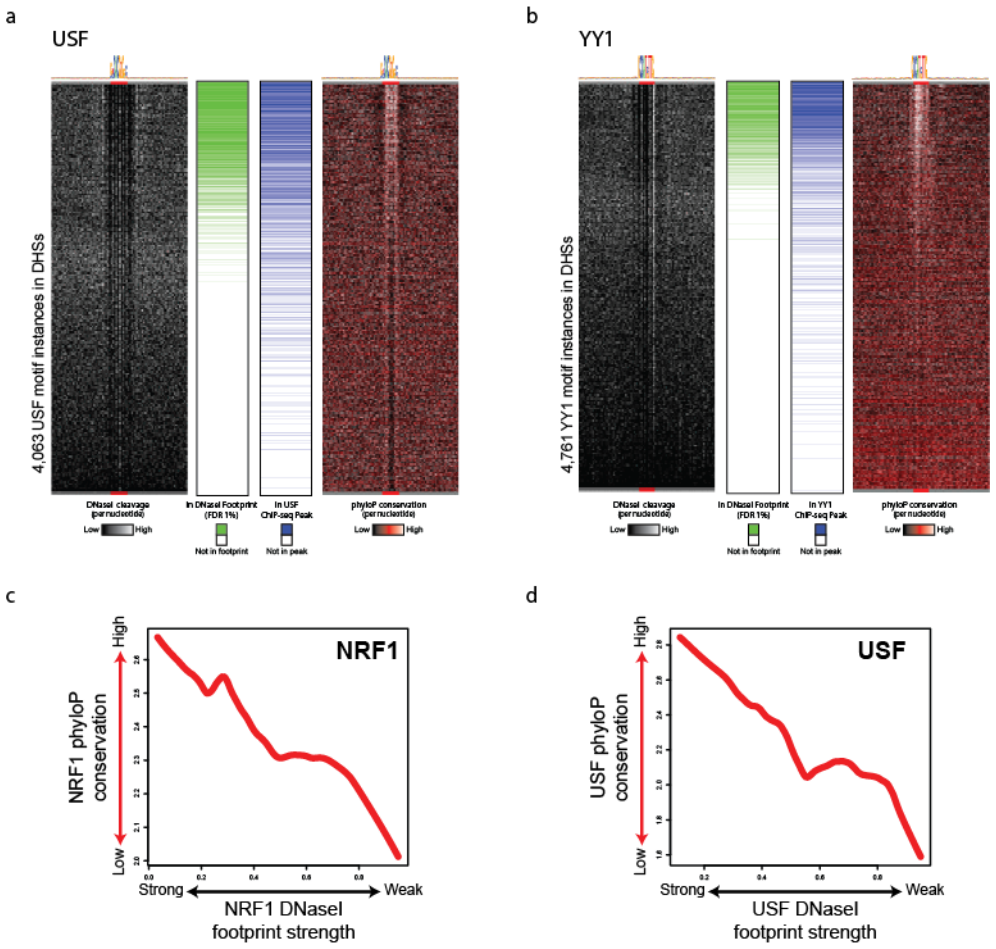
a, The genomic distribution of footprints found in 41 cell types in relation to annotated genomic features. **b,** Examples of DNaseI footprints at different genomic features.

Figure 4.S3. Motif density in DNaseI footprints.



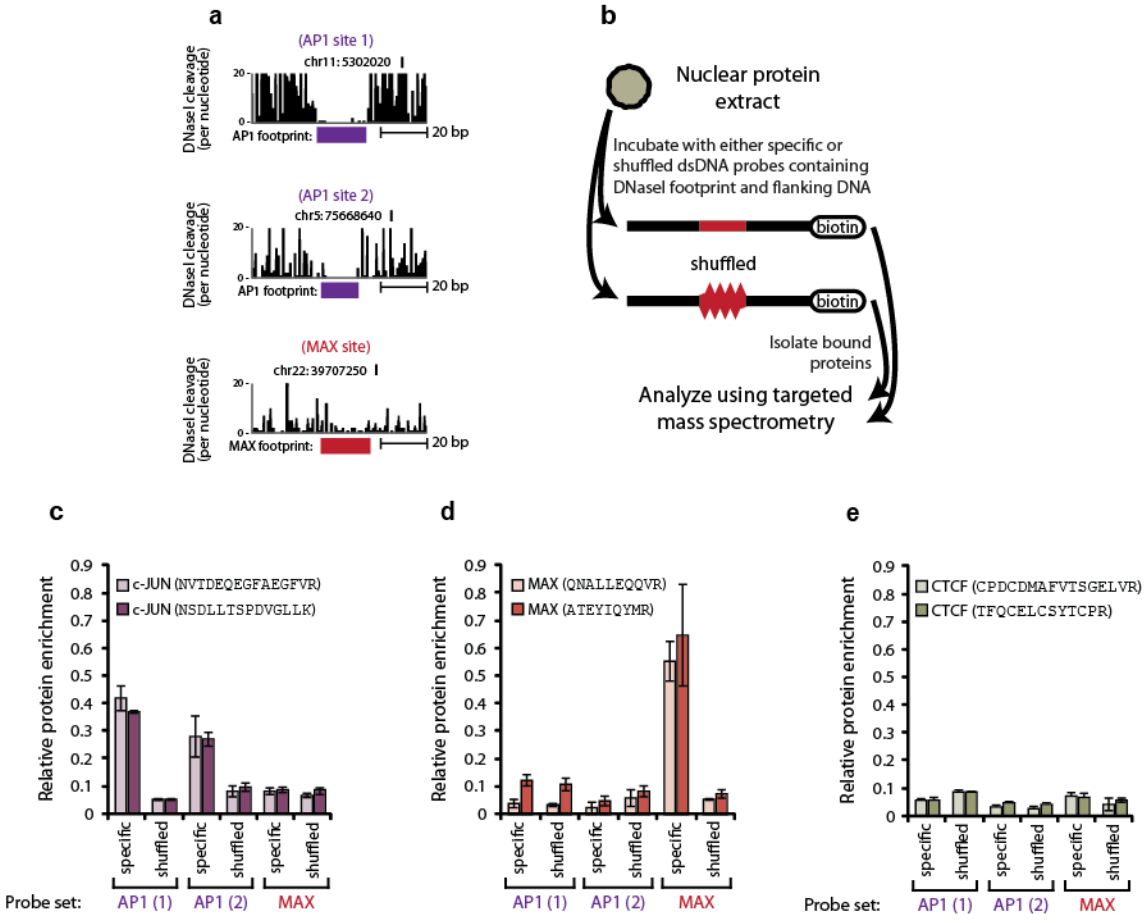
The density of motifs in DNaseI footprints, DHSs (but not in footprints) and non-hypersensitive genomic regions. Motifs are significantly enriched in footpr footprints (Z-score = 204.22, Genome Structure Correction program comparing the locations of TRANSFAC motifs in 1% FDR footprints).

Figure 4.S4. Association of footprint, occupancy and sequence conservation.



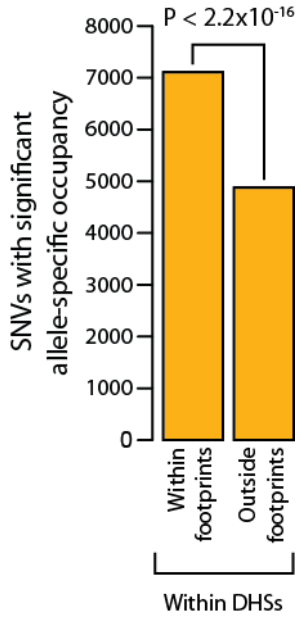
a-b, Heatmaps showing per nucleotide DNaseI cleavage (left) and vertebrate conservation by phyloP (right) for USF (a) and YY1 (b) motifs within K562 DHSs ranked by tag density. Green and blue indicator ticks in the middle indicate the presence of DNaseI footprints and ChIP-seq peaks, respectively, at putative genomic binding sites. **c-d,** Lowess regression of NRF1 (c) and USF (d) maximum phyloP score versus DNaseI footprinting occupancy (footprint occupancy score) at K562 DNaseI footprints marked by NRF1 (c) and USF (d) motifs.

Figure 4.S5. Validation of footprints as potential sites of protein occupancy in vitro.



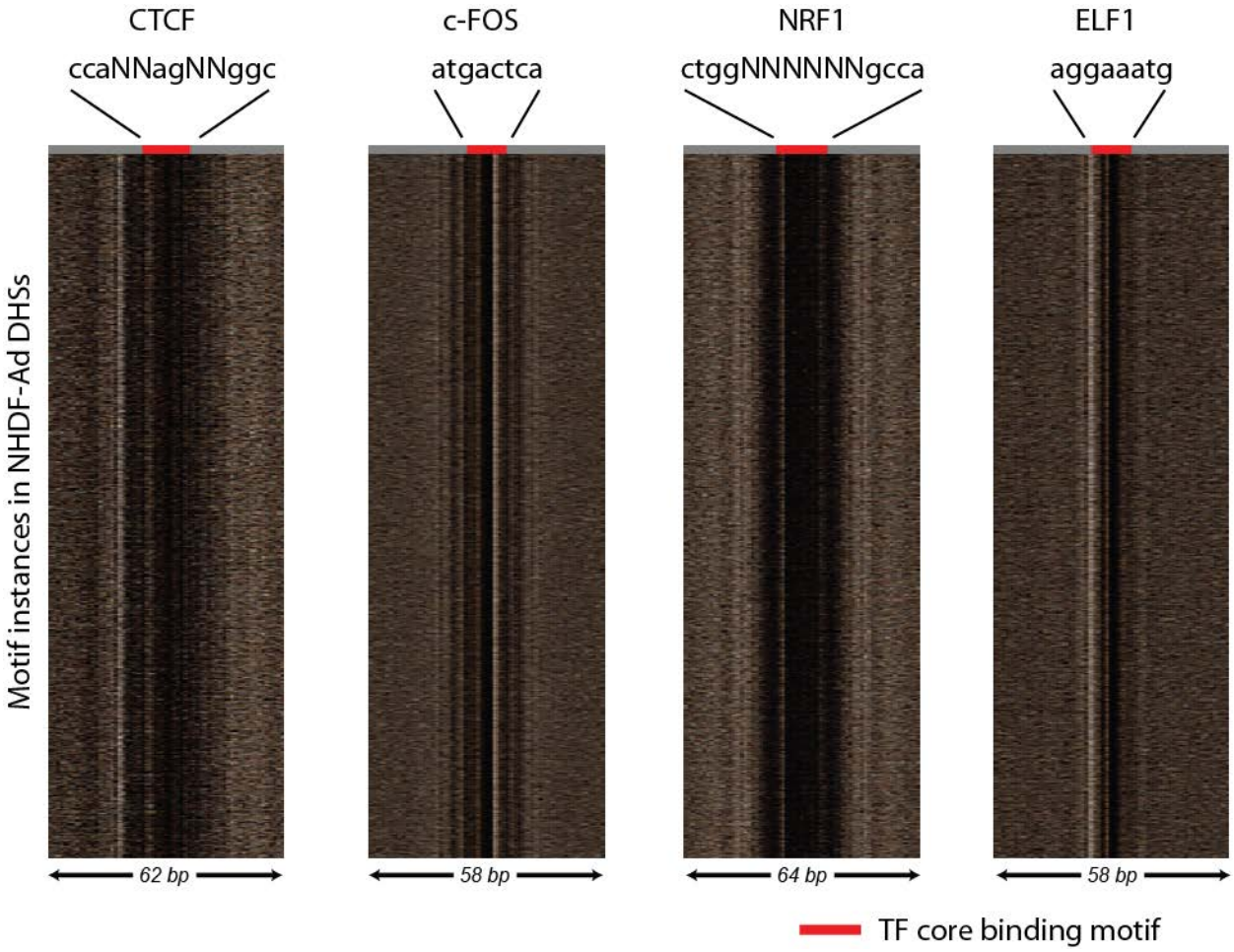
a, Three genomic loci of varying footprint strength targeted using DNA interacting protein precipitation (DIPP). **b,** Schematic overview of the DIPP protocol. **c-d,** Targeted mass spectrometry measurements of the proteins enriched using the different probe sets. The AP1 protein c-Jun was enriched specifically using the AP1 probes (c) and MAX was enriched specifically using the MAX probe (d). **e,** As a negative control for DIPP, we tested for CTCF binding to the six probes. CTCF did not appear to be enriched in any of the pulldowns.

Figure 4.S6. DNaseI footprints mark sites of functional in vivo protein occupancy.



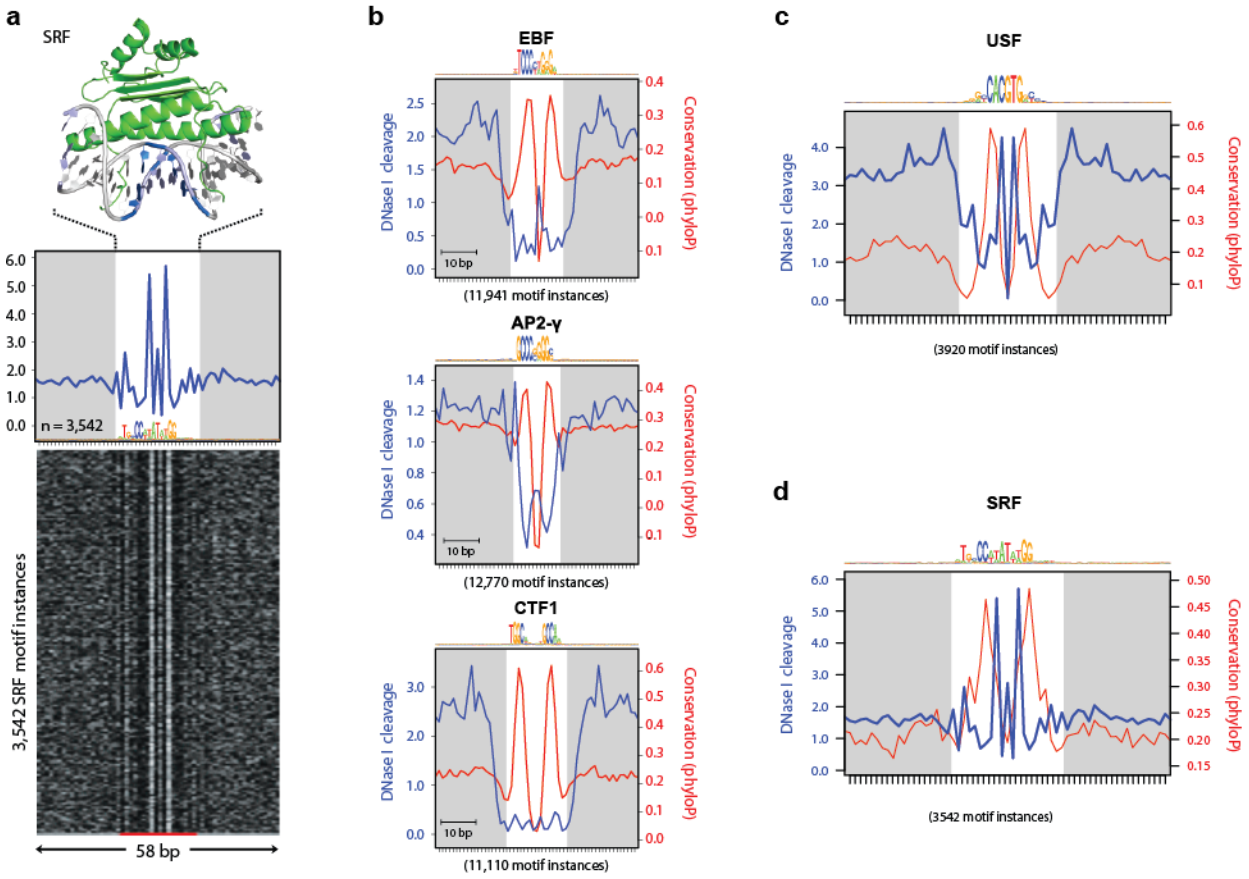
Heterozygous SNVs associated with allele-specific occupancy are significantly enriched inside footprints compared to the rest of the DHS (P < 2.2x10-16, Fisher's exact test).

Figure 4.S7. Stereotyped cleavage patterns for different TFs.



The per-nucleotide DNaseI cleavage patterns at motif instances of 4 different transcription factors in adult dermal fibroblasts (NHDF-Ad). The different motif instances (rows) are randomly ordered.

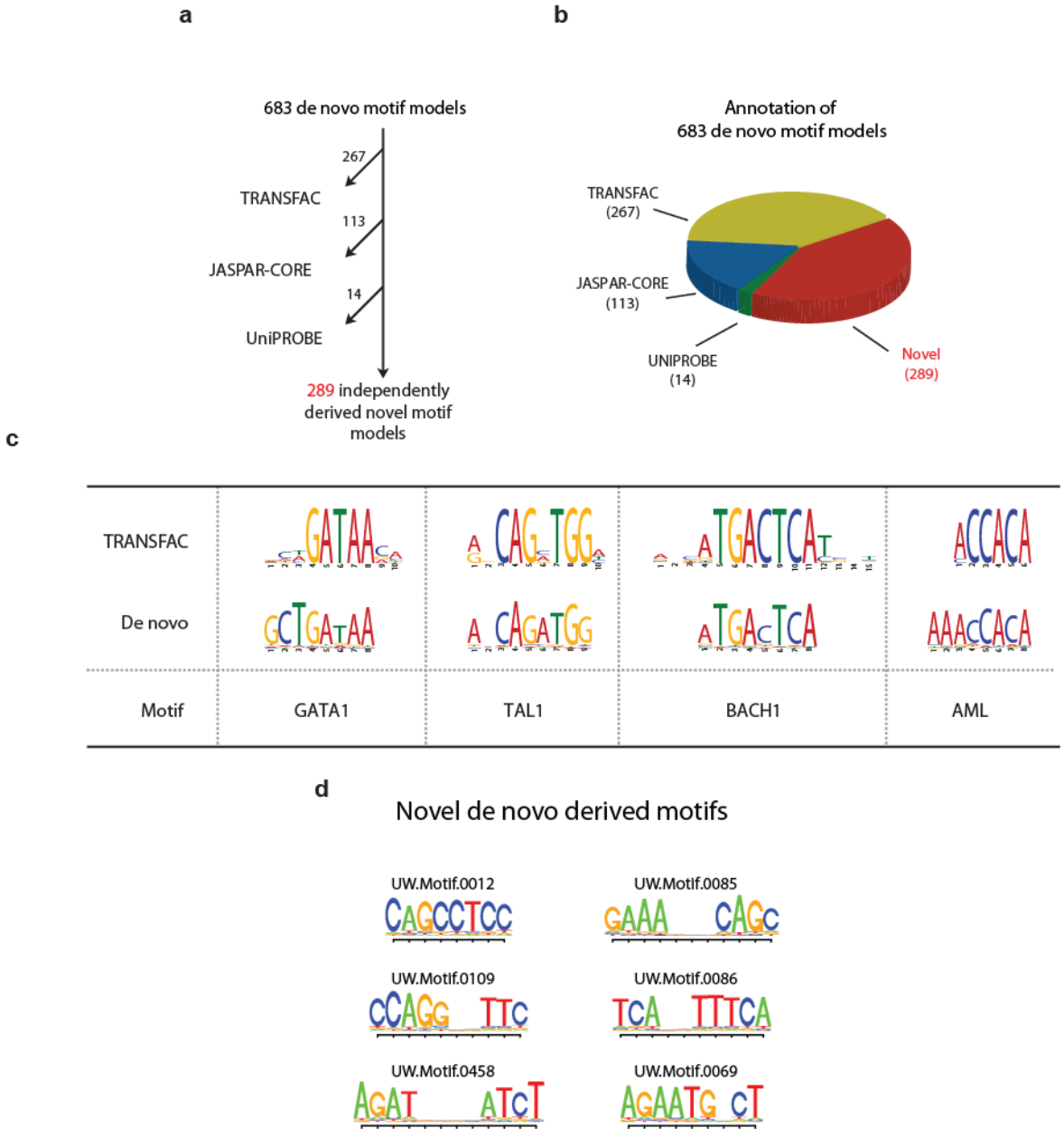
Figure 4.S8. Anti-correlation of conservation and DNaseI cleavage for factors with structural data.



a, Similar to Fig. 3a, the co-crystal structure of Serum Response Factor (SRF) bound to its DNA ligand is juxtaposed above the average nucleotide-level DNaseI cleavage pattern (blue) at motif instances of SRF 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 SRF 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 SRF 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. **c-d,** Cleavage profiles mirror the protein structure and are anti-correlated with vertebrate conservation for USF (c) and SRF (d).

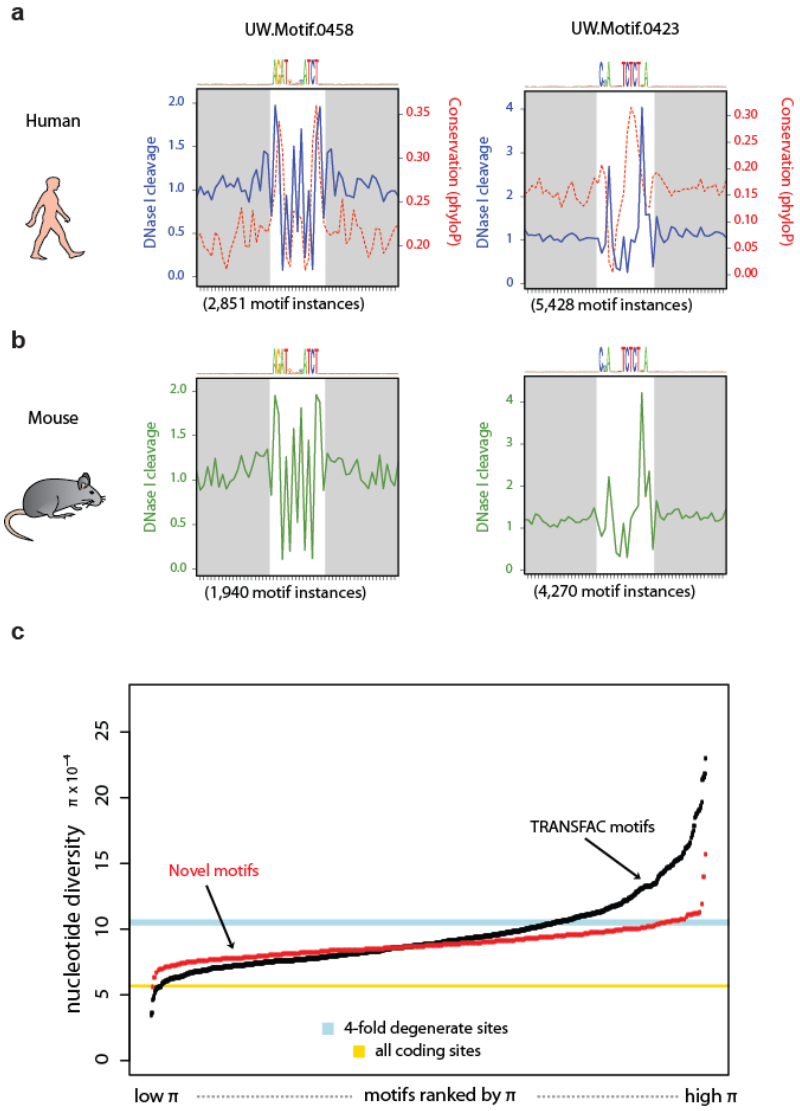
Figure 4.S9. De novo motif discovery in footprints.



a, Diagram of the depletion scheme used to identify novel motifs. 683 motifs were filtered in successive order using TOMTOM with TRANSFAC, JASPAR-CORE and UniPROBE. The numbers on the arrows display the number of de novo motifs matched to the corresponding

database. **b,** Pie chart annotating the partition of de novo motifs into known and novel motifs. **c,** Example consensus logos of de novo derived motifs that match TRANSFAC models. **d,** Example consensus logos of novel de novo derived motifs using DNaseI footprints.

Figure 4.S10. Conservation and selection of DNaseI footprints.



a, 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. **b,** Per-nucleotide mouse liver DNaseI cleavage patterns at occurrences of the motifs in (a) at DNaseI footprints identified in mouse liver. **c,** The average human nucleotide diversity (π, y-axis) across all motif instances within DNaseI

footprints is plotted for each of the motif models in the TRANSFAC database (black, ordered by mean π) and for each of the novel de novo-derived motif models (red, ordered by mean π). Blue bar indicates the average nucleotide diversity (pi) at 4-fold degenerate coding sites (width is equal to 95% confidence interval); gold bar indicates pi at all coding sites (width is equal to 95% confidence interval).

4.6 – METHODS

Cell types used for digital genomic footprinting

The following human cell types were subjected to DNaseI digestion and high-throughput sequencing, following previous methods (John et al., 2011; Hesselberth et al., 2009) 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 (Langmead et al., 2009), 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.

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 (Galas and Schmitz,

1978; Hesselberth et al., 2009). 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 (Hesselberth et al., 2009). 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. 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 (Lazarovici et al., 2013). 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 (Neph et al., 2012a). 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.

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.

FDR 1% DNaseI hypersensitive Sites

We counted the number of footprints falling within every DNaseI hypersensitive sites (DHS, defined as 150 nt in length) (John et al., 2011) 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.

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.

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 (The ENCODE Project Consortium, 2007). 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 (Matys et al., 2006) and JASPAR Core (Bryne et al., 2008), 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.

DNaseI cleavages vs. ChIP-seq

Motif models (from TRANSFAC, version 2011.1, JASPAR Core, and UniPROBE (Newburger and Bulyk, 2009)) 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 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 (Pavlidis and Noble, 2003), version 1.2. A phyloP evolutionary conservation (Pollard et al., 2010) 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. Motif instances overlapping ChIP-seq peaks by at least 1 nt were also annotated.

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.

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.

DNA Interacting Protein Precipitation (DIPP) experiments

Protein extraction for DNA Interacting Protein Precipitation (DIPP)

Nuclei were isolated using a standard protocol previously described (Sabo et al., 2004). 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).

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 (Mittler et al., 2009). 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 (Stergachis et al., 2011). 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 (Maclean et al., 2010). 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

(Stergachis et al., 2011). 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 (MacLean et al., 2010). 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.

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 (Durbin et al., 2010). 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 (Li et al., 2009). 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.

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

IMR90 methylation calls (Lister et al., 2009) 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.

Rendering of DNA-protein complexes

Crystallography data showing DNA-protein complexes for selected factors (Ferré-D'Amaré et al., 1994; Pellegrini et al., 1995) 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 (Pavlidis and Noble, 2003).

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 (Neph et al., 2012a), 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 (Crooks et al., 2004).

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 (Pruitt et al., 2009b), 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.

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 (Gupta et al., 2007), 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.

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.

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 released by Complete Genomics, version 1.10 (Drmanac et al., 2010). Relatedness was determined both by pedigree and with KING (Manichaikul et al., 2010). 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.

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.

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.

Chapter 5 – Circuitry and dynamics of human transcription factor regulatory networks

Note: This work was published in the September 2012 edition of *Cell* as:

*Neph, S., *Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012). **Circuitry and Dynamics of Human Transcription Factor Regulatory Networks.** *Cell* 150, 1274–1286.

** indicates co-first authors*

5.1 – ABSTRACT

The combinatorial cross-regulation of hundreds of sequence-specific transcription factors defines a regulatory network that underlies cellular identity and function. Here we use genome-wide maps of in vivo DNaseI footprints to assemble an extensive core human regulatory network comprising connections among 475 sequence-specific transcription factors, and to analyze the dynamics of these connections across 41 diverse cell and tissue types. We find that human transcription factor networks are highly cell-selective and are driven by cohorts of factors that include regulators with previously unrecognized roles in control of cellular identity. Moreover, we identify many widely expressed factors that impact transcriptional regulatory networks in a cell-selective manner. Strikingly, in spite of their inherent diversity, all cell type regulatory networks independently converge on a common architecture that closely resembles the topology of living neuronal networks. Together, our results provide the first description of the circuitry, dynamics, and organizing principles of the human transcription factor regulatory network.

5.2 – INTRODUCTION

Sequence-specific transcriptional factors (TFs) are the key effectors of eukaryotic gene control. Human TFs regulate hundreds to thousands of downstream genes (Johnson et al., 2007). Of particular interest are interactions in which a given TF regulates other TFs, or itself. Such mutual cross-regulation among groups of TFs defines regulatory sub-networks that underlie major features of cellular identity and complex functions such as pluripotency (Boyer et al., 2005; Kim et al., 2008), development (Davidson et al., 2002a) and differentiation (Yun and Wold, 1996). On a broader level, cross-regulatory interactions among the entire complement of TFs expressed in a given cell type form a core transcriptional regulatory network, endowing the cell with systems-level properties that facilitate the integration of complex cellular signals, while conferring additional nimbleness and robustness (Alon, 2006). However, despite their central biological roles, both the structure of core human regulatory networks and their component sub-networks are largely undefined.

One of the main bottlenecks limiting generation of transcription factor regulatory networks for complex biological systems has been that information is traditionally collected from individual experiments targeting one cell-type and one transcription factor at a time (Davidson et al., 2002a; Yuh et al., 1994; Kim et al., 2008; modENCODE Consortium et al., 2010; Gerstein et al., 2010). For example, the sea urchin endomesoderm regulatory network was constructed by individually perturbing the expression and activity of several dozen transcription factors and analyzing the effect of these perturbations on the expression of transcription factor genes containing putative *cis*-regulatory binding elements for these factors (Davidson et al., 2002b; Yuh et al., 1994). More recently, genome wide analysis combining chromatin immunoprecipitation of individual TFs with high-throughput sequencing (ChIP-seq) has been

used to derive sub-networks of small numbers of TFs, such as those involved in pluripotency (Kim et al., 2008) or larger scale networks combining several dozen TFs (modENCODE Consortium et al., 2010; Gerstein et al., 2010). However, such approaches are limited by three major factors: (i) the availability of suitable affinity reagents; (ii) the difficulty of interrogating the activities of multiple TFs within the same cellular environment; and, perhaps most critically (iii) the sizable number of TFs and cellular states that need to be studied. *De novo* network construction methods based on gene expression correlations partly overcome the limitation of studying one TF at a time, but lack directness and typically require several hundred independent gene expression perturbation studies to build a network for one cell type (Basso et al., 2005; Carro et al., 2010). Similarly, yeast one-hybrid assays offer a high-throughput approach for identifying *cis*-regulatory element binding partners (Walhout, 2006; Reece-Hoyes et al., 2011). However, such assays lack native cellular context, limiting their direct utility for building cell type-specific networks. Given these experimental limitations, only a handful of well-described multi-cellular transcriptional regulatory networks have been defined, and those that do exist are often incomplete despite the numerous experiments and extended time (typically years) needed to construct them (Davidson et al., 2002b; Basso et al., 2005; Boyer et al., 2005; Kim et al., 2008; modENCODE Consortium et al., 2010; Gerstein et al., 2010).

Given that the human genome encodes >1,000 TFs (Vaquerizas et al., 2009), and that human cellular diversity spans hundreds of different cell types and an even greater number of cellular states, we sought to develop an accurate and scalable approach to transcriptional regulatory network analysis suitable for application to any cellular or organismal state. The discovery of DNaseI footprinting over 30 years ago (Galas and Schmitz, 1978) revolutionized the analysis of regulatory sequences in diverse organisms, and directly enabled the discovery of

the first human sequence-specific transcription factors (Dynan and Tjian, 1983). In the context of living nuclear chromatin, DNaseI treatment preferentially cleaves the genome within highly accessible active regulatory DNA regions, creating DNaseI hypersensitive sites (DHSs) (Wu et al., 1979; Kuo et al., 1979; Wu, 1980; Stalder et al., 1980). Within DHSs, DNaseI cleavage is not uniform but is rather punctuated by sequence-specific regulatory factors that occlude bound DNA, leaving 'footprints' that demarcate TF occupancy at nucleotide resolution (Hesselberth et al., 2009; Neph et al., 2012c; Pfeifer and Riggs, 1991). DNaseI footprinting is a well-established method for identifying direct regulatory interactions, and provides a powerful generic approach for assaying the occupancy of specific sequence elements with *cis*-regulatory functions (Karin et al., 1984; Kadonaga et al., 1987).

DNaseI footprinting has been applied widely to study regulatory interactions between TFs, and to identify cell- and lineage-selective transcriptional regulators (Dynan and Tjian, 1983; Karin et al., 1984; Tsai et al., 1989). In the context of the ENCODE Project, we applied digital genomic footprinting to delineate millions of human DNaseI footprints genome-wide in 41 diverse cell types. Combining DNaseI footprints with defined TF recognition sequences accurately and quantitatively recapitulates ChIP-seq data for individual TFs, while simultaneously interrogating potentially all DNA-binding factors in a single experiment (Neph et al., 2012c).

By performing systematic analysis of TF footprints in the proximal regulatory regions of each transcription factor gene, we develop a foundational experimental paradigm for comprehensive, unbiased mapping of the complex network of regulatory interactions between human TFs. In such networks, TFs comprise the network 'nodes', and the regulation of one TF by another the interactions or network 'edges'. Furthermore, iterating this paradigm across

diverse cell types provides a powerful system for analysis of transcription factor network dynamics in a complex organism. Here, we use genome-wide maps of *in vivo* DNaseI footprints to assemble an extensive core human regulatory network comprising connections among 475 sequence-specific transcription factors, and analyze the dynamics of these connections across 41 diverse cell and tissue types.

5.3 – RESULTS

Comprehensive mapping of transcription factor networks in diverse human cell types

To generate transcription factor regulatory networks in human cells, we analyzed genomic DNaseI footprinting data from 41 diverse cell and tissue types (Neph et al., 2012c). Each of these 41 samples was treated with DNaseI, and sites of DNaseI cleavage along the genome were analyzed using high-throughput sequencing. At an average sampling depth of ~500 million DNaseI cleavages per cell type (of which ~ 273 million mapped to unique genomic positions), we identified an average of ~1.1 million high-confidence DNaseI footprints per cell type (range 434,000 to 2.3 million at a False Discovery Rate of 1% (FDR 1%) (Neph et al., 2012c)). Collectively, we detected 45,096,726 footprints, representing cell-selective binding to ~8.4 million distinct 6-40bp genomic sequence elements. We inferred the identity of factors occupying DNaseI footprints using well-annotated databases of transcription factor binding motifs (Wingender et al., 1996; Bryne et al., 2008; Newburger and Bulyk, 2009) (**Methods**), and confirmed that these identifications matched closely and quantitatively with ENCODE ChIP-seq data for the same cognate factors (Neph et al., 2012c; Samstein et al., 2012).

To generate a TF regulatory network for each cell type, we analyzed actively bound DNA elements within the proximal regulatory regions (i.e., all DNaseI hypersensitive sites within a 10kb interval centered on the transcriptional start site) of 475 transcription factor genes with well-annotated recognition motifs (Wingender et al., 1996; Bryne et al., 2008; Newburger and Bulyk, 2009) (**Fig. 5.M1A**). Repeating this process for every cell type disclosed a total of 38,393 unique, directed (i.e., TF-to-TF) regulatory interactions (edges) among the 475 analyzed TFs, with an average of 11,193 TF-to-TF edges per cell type. Given the functional redundancy of a minority of DNA binding motifs (Berger et al., 2008), in certain cases multiple factors could be designated as occupying a single DNaseI footprint. However, most commonly, mappings represented associations between single TFs and a specific DNA element. Because DNaseI hypersensitivity at proximal regulatory sequences closely parallels gene expression (ENCODE Project Consortium et al., 2012), the annotation process we utilized naturally focuses on the expressed TF complement of each cell type, enabling the construction of a comprehensive transcription regulatory network for a given cell type with a single experiment.

De novo-derived networks accurately recapitulate known TF-to-TF circuitry

To assess the accuracy of cellular TF regulatory networks derived from DNaseI footprints, we analyzed several well-annotated mammalian cell type-specific transcriptional regulatory sub-networks (**Fig. 5.M1B-C**). The muscle-specific factors MyoD, Myogenin (MYOG), MEF2A, and MYF6 form a network vital for specification of skeletal muscle fate and control of myogenic development and differentiation, which was uncovered using a combination of genetic and physical studies, including DNaseI footprinting (Naidu et al., 1995; Yun and Wold, 1996; Ramachandran et al., 2008a). Figure 5.M1B juxtaposes the known regulatory

interactions between these factors determined in the aforementioned studies (**Fig. 5.M1B,** top), with the nearly identical interactions derived *de novo* from analysis of the network computed using DNaseI footprints mapped in primary human skeletal myoblasts (HSMM) (**Fig. 5.M1B**, bottom).

OCT4, NANOG, KLF4 and SOX2 together play a defining role in maintaining the pluripotency of embryonic stem (ES) cells (Takahashi and Yamanaka, 2006; Takahashi et al., 2007), and a network comprising the mutual regulatory interactions between these factors has been mapped through systematic studies of factor occupancy by ChIP-seq in mouse ES cells (Kim et al., 2008) (**Fig. 5.M1C**, top). A nearly identical sub-network emerges from analysis of the transcription factor network computed *de novo* from DNaseI footprints in human ES cells (**Fig. 5.M1C**, bottom).

Critically, both the well-annotated muscle and ES sub-networks are best matched by footprint-derived networks computed specifically from skeletal myoblasts and human ES cells, respectively, vs. other cell types (**Fig. 5.M1D-E**). These findings indicate that network relationships between transcription factors derived *de novo* from genomic DNaseI footprinting accurately recapitulate well-described cell type-selective transcriptional regulatory networks generated using multiple experimental approaches.

Transcription factor regulatory networks show marked cell-selectivity

We next analyzed systematically the dynamics of TF regulatory networks across cell types. 475 transcription factors theoretically have the potential for 225,625 combinations of TF-to-TF regulatory interactions or network edges. However, only a fraction of these potential

edges are observed in each cell type (~5%), and most are unique to specific cell types (**Fig. 5.S1A**).

To visualize the global landscape of cell-selective vs. shared regulatory interactions, we first computed the broad landscape of network edges that are either specific to a given cell type, or are found in networks of two or more cell types (**Fig. 5.M2**). This revealed that regulatory interactions were in general highly cell-selective, though the proportion of cell-selective interactions varied from cell type to cell type. Network edges were most frequently restricted to a single cell type, and collectively the majority of edges were restricted to 4 or fewer cell types (**Fig. 5.S1A**). By contrast, only 5% of edges were common to all cell types (**Fig. 5.S1A**). Interestingly, when comparing networks, we found more common edges than common DNaseI footprints (**Fig. 5.S1B,C**), implying that a given transcriptional regulatory interaction can be generated using distinct DNA binding elements in different cell types.

To explore the regulatory interaction dynamics of limited sets of related factors we plotted the regulatory network edges connecting four hematopoietic regulators and four pluripotency regulators in six diverse cell types (**Fig. 5.M3A**). This analysis clearly highlighted the role of cell-type specific factors within their cognate cell types: regulatory interactions between pluripotency factors within the ES cell network, and hematopoietic factors within the network of hematopoietic stem cells (**Fig. 5.M3A**). Next, we plotted the complete set of regulatory interactions amongst all 475 edges between the same six diverse cell types, exposing a high degree of regulatory diversity (**Fig. 5.M3B**).

Edges unique to a cell type typically form a well-connected sub-network (**Fig. 5.S1D-F**), implying that cell type-specific regulatory differences are not driven merely by the independent actions of a few transcription factors, but rather by organized TF sub-networks. In addition, the

density of cell-selective networks varies widely between cell types (e.g., compare ES cells to skeletal myoblasts in **Fig. 5.M3B**). These observations underscore the importance of using cell type-specific regulatory networks when addressing specific biological questions.

Functionally related cell types share similar core transcriptional regulatory networks

We next sought to determine the degree of relatedness between different transcription factor networks. To obtain a quantitative global summary of the factors contributing to each cell type specific network, we computed for each cell type the normalized network degree (NND) – a vector which encapsulates the relative number of interactions observed in that cell type for each of the 475 TFs (Alon, 2006). To capture the degree to which different cell type networks utilize similar transcription factors, we clustered all cell type networks based on their NND vector (**Fig. 5.M4A**). The resulting network clusters – obtained from an unbiased analysis – strikingly parallel both anatomical and functional cell type groupings into epithelial and stromal cells; hematopoietic cells; endothelia; and primitive cells including fetal cells and tissues, ES cells, and malignant cells with a 'de-differentiated' phenotype (**Fig. 5.M4A**; compare the manually curated groupings in **Fig. 5.M2**). This result suggests that transcriptional regulatory networks from functionally similar cell-types are governed by similar factors. Furthermore, this result suggests a framework for understanding how minor perturbations in network composition might enable trans-differentiation among related cell-types (Graf and Enver, 2009).

To identify the individual transcription factors driving the clustering of related cell type networks, we computed the relative NND (i.e., the normalized number of connections) of each TF across the 41 cell types. This approach uncovered numerous specific factors with highly cell-selective interaction patterns, including known regulators of cellular identity important to

functionally related cell types (**Fig. 5.M4B**). For instance, PAX5 is most highly connected in B-cell regulatory networks, agreeing with its function as a major regulator of B-lineage commitment (Nutt et al., 1999). Similarly, the neuronal developmental regulator POU3F4 (Shimazaki et al., 1999) plays a prominent role specifically in hippocampal astrocyte and fetal brain regulatory networks, while the cardiac developmental regulator GATA4 (Molkentin et al., 1997) shows the highest relative network degree in cardiac and great vessel tissue (fetal heart, cardiomyocytes, cardiac fibroblasts, and pulmonary artery fibroblasts).

In addition to these known developmental regulators, the network analysis implicated many regulators with previously unrecognized roles in specification of cell identity. For instance, HOXD9 is highly connected specifically in endothelial regulatory networks, and the early developmental regulator GATA5 (MacNeill et al., 2000) appears to play a predominant role in the fetal lung network (**Fig. 5.M4B**), providing functional insight into the role of GATA5 as a lung tissue biomarker (Xing et al., 2010). In addition to factors with strong cell-selective connectivity, we found a number of TFs with prominent roles in all 41 cell type networks, including several known ubiquitous transcriptional and genomic regulators such as SP1, NFYA, CTCF and MAX (**Fig. 5.S2**).

Together, the above results demonstrate the ability of transcriptional networks derived from genomic DNaseI footprinting to pinpoint known cell-selective and ubiquitous regulators of cellular state, and to implicate analogous yet unanticipated roles for many other factors. It is notable that the aforementioned results were derived independently of gene expression data, highlighting the ability of a single experimental paradigm (genomic DNaseI footprinting) to elucidate multiple intricate transcriptional regulatory relationships.

Network analysis reveals cell-type specific behaviors for widely expressed TFs

Many transcription factors are expressed to varying degrees in a number of different cell types (Vaquerizas et al., 2009). A major question is whether the function of widely expressed factors remains essentially the same in different cells, or whether such factors are capable of exhibiting important cell-selective actions. To explore this question, we sought to characterize the regulatory diversity between different cell types within the same lineage. Hematopoietic lineage cells have been extensively characterized at both the phenotypic and the molecular level, and a cadre of major transcriptional regulators has been defined, including TAL1/SCL, PU.1, ELF1, HES1, MYB, GATA2 and GATA1 (Orkin, 1995; Swiers et al., 2006). Many of these factors are expressed to varying degrees across multiple hematopoietic lineages and their constituent cell types.

We analyzed *de novo*-derived sub-networks comprising the aforementioned seven regulators in five hematopoietic and one non-hematopoietic cell type (**Fig. 5.M5A**). For each cell type sub-network, we also mapped the normalized outdegree (i.e., the number of outgoing connections) for each factor (**Fig. 5.M5A**). This analysis revealed both subtle and stark differences in the organization of the 7-member hematopoietic regulatory sub-network that reflected the biological origin of each cell-type. For example, the early hematopoietic fate decision factor PU.1 appears to play the largest role in the sub-networks generated from hematopoietic stem cells (CD34+) and promyelocytic leukemia (NB4) cells (**Fig. 5.M5A**). The erythroid-specific regulator GATA1 appears as a strong driver of the core TAL1/PU.1/HES1/MYB sub-network specifically within erythroid cells (**Fig. 5.M5A**), consistent with its defining role in erythropoiesis. In both B-cells and T-cells, the sub-network takes on a directional character, with PU.1 in a superior position. By contrast, the network is

largely absent in non-hematopoietic cells (muscle, HSMM) (**Fig. 5.M5A**, bottom right). These findings demonstrate that analysis of the network relationships of major lineage regulators provides a powerful tool for uncovering subtle differences in transcriptional regulation that drive cellular identity between functionally similar cell-types.

We next extended this analysis to determine whether we could identify commonly expressed factors that manifest cell-type specific behaviors. For example, the retinoic acid receptor-alpha (RAR-α) is a constitutively-expressed factor involved in numerous developmental and physiological processes (Sucov et al., 1996). Rather than simply measuring the degree of connectivity of RAR-α to other factors across different cell types, we sought to quantify the behavior of RAR-α within each cellular regulatory network by determining its position within feed-forward loops (FFLs). Feed-forward loops represent one of the most important network motifs in biological and regulatory systems and comprise a three node structure in which information is propagated forward from the top node through the middle to the bottom node, with direct top node-to-bottom node reinforcement (Milo et al., 2002; Alon, 2006). For each cell type, we quantified the number of feed-forward loops containing RAR-α at each of the three different positions (top vs. middle vs. bottom; **Fig. 5.M5B**, top). In most cell-types, RAR-α chiefly participates in feed-forward loops at 'passenger' positions 2 and 3 (**Fig. 5.M5B**). However, within blood and endothelial cells, RAR-α switches from being a passenger to being a driver (top position) of FFLs. Strikingly, in acute promyelocytic leukemia (APL) cells, RAR-α acts as a uniquely potent driver of feed-forward loops, occurring exclusively in the driver position – a feature unique among all cell types (**Fig. 5.M5B**). APL is characterized by an oncogenic t(15;17) chromosomal translocation which results in a RAR-α/PML fusion protein that misregulates RAR-α target sites (Grignani et al., 1993, 1998). Our results suggest that in

APL cells, RAR-α is additionally altering the basic organization of the regulatory network. Critically, we identified the prominent role of RAR-α in APL using DNaseI footprint-driven network analysis without any prior knowledge of its role in APL cells. This suggests that network analysis is capable of deriving vital pathogenic information about specific factors in abnormal cell types, given a sufficient analyzed spectrum of normal cellular networks. On a more general level, the aforementioned results show clearly that marked cell-selective functional specificities of commonly expressed proteins can be exposed by analyzing factors within the context of their peers.

The common 'neural' architecture of human transcription factor regulatory networks

Complex networks from diverse organisms are built from a set of simple building blocks termed network motifs (Milo et al., 2002). Network motifs represent simple regulatory circuits, such as the feed-forward loop described above. The topology of a given network can be reflected quantitatively in the normalized frequencies (normalized z-score) of different network motifs. Specific well-described motifs including FFL, 'clique', 'semi-clique', 'regulated mutual' and 'regulating mutual' are recurrently found at higher than expected frequency within diverse biological networks (Milo et al., 2002, 2004). We therefore sought to analyze the topology of the human transcription factor regulatory network, and to compare it with those of well-annotated multi-cellular biological networks.

We first computed the relative frequency and relative enrichment or depletion of each of the 13 possible three-node network motifs within each cell type regulatory network. Next, we compared the results for each cell type network with the relative enrichment of 3-node networks motifs found in perhaps the best annotated multi-cellular biological network, the *C. elegans*

neuronal connectivity network (White et al., 1986). This comparison revealed striking similarity between the topologies of human TF networks and the *C. elegans* neuronal network (**Fig. 5.M6A**). Remarkably, in spite of their cell-selectivity, the topologies of each TF network were nearly identical. Notably, the human TF regulatory network topology also closely resembles that of other well-described networks including, the sea-urchin endomesoderm specification network (Davidson et al., 2002a), the *Drosophila* developmental transcriptional network (Serov et al., 1998), and the mammalian signal transduction network (Milo et al., 2004) (**Fig. 5.S3A**), consistent with universal principles for multi-cellular biological information processing systems (Milo et al., 2004).

To test the sensitivity of the above findings to the manner in which the human transcriptional regulatory networks were determined, we re-computed this network solely from scanned transcription factor binding sites within the promoter-proximal regions of each TF gene, without considering whether the motifs were localized within DNaseI footprints. Using this approach, the remarkable similarity of the footprint-derived TF networks to the neuronal network was almost completely lost (**Fig. 5.M6B**). This result affirms the criticality of *in vivo* footprints for biologically meaningful network inference.

Next, we sought to determine if the observed similarity to the neuronal network was a collective property of human transcription factor networks. To test this, we computed a transcriptional regulatory network from the combined regulatory interactions of all 41 cell types and determined the enrichment of network motifs within this network. The resulting network topology diverges considerably from that of the neuronal network (**Fig. 5.M6C**), far more so than was observed for any individual cell type (**Fig. 5.M6A**). This result suggests that the

regulatory interactions within each cell type network are independently balanced to achieve a specific architecture, and that pooling multiple cellular networks together degrades this balance.

Finally, we asked whether a common core of regulatory interactions might be driving the conserved network architecture, by comparing feed-forward loops from biologically similar cell types with one another. This comparison revealed marked diversity among different cellular TF networks (**Fig. 5.M6D,E**), considerably exceeding that observed among individual edges (**Fig. 5.S3C,D**). Indeed, only ~0.1% of all observed FFLs across 41 cell types (784 / 558,841) were common to all cell types (**Fig. 5.M6F** and **Fig. 5.S3E**). Moreover, only a minority of the TFs represented within a given cellular network contribute to the enriched network motifs (**Fig. 5.S3F**). These findings indicate that the conserved 'neuronal' network architecture (**Fig. 5.M6A**) of the human TF regulatory network is specified independently in each cell type using a distinct set of balanced regulatory interactions.

5.4 – DISCUSSION

Transcription factor regulatory networks are foundational to biological systems. Collectively, our results highlight the power of regulatory networks derived from genomic DNaseI footprint maps to provide accurate large-scale depictions of regulatory interactions in human cells, and they suggest such interactions are governed by a core set of organizing principles shared with other multicellular information processing systems.

In a classic treatise, Waddington proposed that the epigenetic landscape of a cell is 'buttressed' by complex interactions among multiple regulatory genes (Waddington, 1939, 1957). These genes – now recognized as sequence-specific transcriptional regulators – form an extended 'cognitive' network that enables the simultaneous integration of multiple internal and

external cues, and conveys this information to specific effector genes along the genome. Consequently, transcriptional regulatory networks influence both the current chromatin landscape of a cell, as well as its epigenetic state, imparting a type of 'memory' that may impact subsequent cellular fate decisions (Waddington, 1957; Groudine and Weintraub, 1982). Such characteristics render transcription factor regulatory networks ideal for governing complex processes such as pluripotency (Boyer et al., 2005; Kim et al., 2008), development (Davidson et al., 2002a) and differentiation (Yun and Wold, 1996). However, despite their central role in human pathology and physiology, human transcriptional regulatory networks are presently poorly understood.

The networks we describe here for 41 diverse cell types represent the first genome-scale human transcriptional regulatory networks, and are among the largest described in any organism. The derivation of regulatory networks from genomic DNaseI footprint maps provides a general, scalable solution for mapping and analyzing cell-selective transcriptional regulatory networks in complex multi-cellular organisms. By comparison, generation of networks of this size across 41 cell-types using traditional approaches such as perturbation or ChIP-seq would have required nearly 20,000 individual experiments. By contrast, the approach we describe can readily scale beyond the 475 factors analyzed in the current study, and is constrained only by the availability of accurate TF recognition sequences.

Our analysis of transcriptional regulatory interactions in a network context has uncovered several novel features of human transcriptional regulation, some quite striking.

First, we observed that human transcriptional regulatory networks are markedly cell-type specific, with only ~5% of all regulatory interactions common across the 41 tested cell types.

This finding highlights the regulatory diversity within humans, and underscores the importance of analyzing cell-selective regulatory networks when addressing specific biological questions.

Second, by detecting factors that predominantly contribute to the transcriptional regulatory networks of only one or a few cell types, we identified both known and novel regulators of cellular identity (**Fig. 5.M4B**). Differences between cell types thus encode a surprisingly rich landscape of information concerning differentiation and developmental processes, and this landscape can be systematically mined for regulatory insights.

Third, we found that commonly expressed TFs within a given cell lineage play distinct roles in the governance of regulatory networks of different cells within that lineage. Our analysis discovered that in acute promyelocytic leukemia cells the widely expressed RAR-α shifts from being a passenger of feed-forward loops (FFLs) to being a strong driver of FFLs. This finding provides novel insights into the broader – and more fundamental – regulatory alterations that accompany the RAR-α/PML fusion protein unique to acute promyelocytic leukemia. On a general level, our results show that commonly expressed proteins may display highly cell-selective actions, and that such activities may be brought to light by analyzing transcription factors in the context of their peers.

Finally, in marked contrast to the high regulatory diversity between cell types, we found that all cell type regulatory networks converge on a common network architecture that closely mirrors the topology of the *C. elegans* neuronal connectivity network and those of other multicellular information processing systems (Milo et al., 2004), highlighting a fundamental similarity in the structure and organizing principles of these biological systems. Strikingly, this common architecture is independently fashioned in each cell type and results from the delicate balance of distinct regulatory interactions.

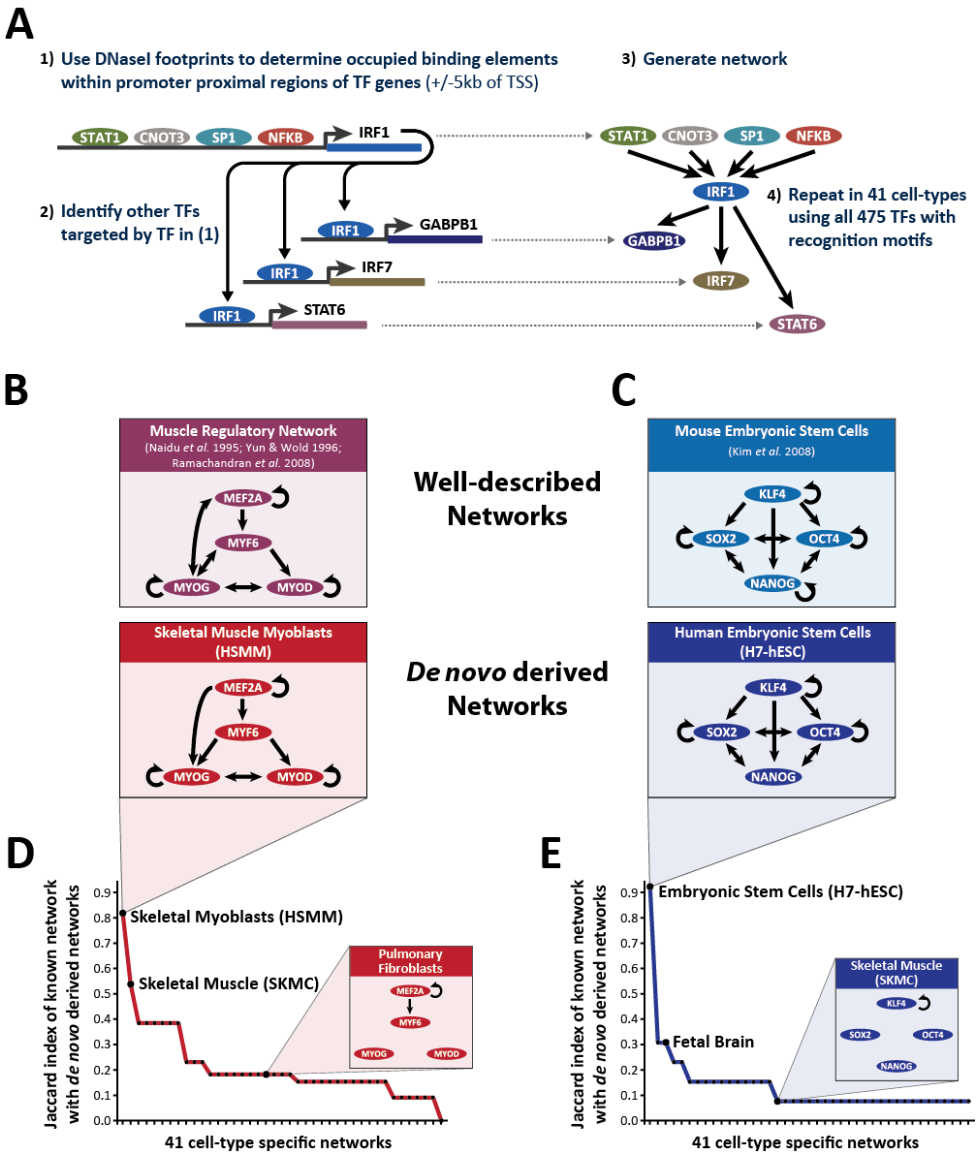
Despite the experimental and computational advantages and successes of our approach, a number of additional steps could be used to refine and improve our regulatory interaction networks. First, as noted above, our approach is limited by the availability of recognition sequences for specific TFs. The pending availability of both more and higher quality recognition sequences through approaches such as Protein Binding Microarrays (Berger et al., 2008; Badis et al., 2009) and SELEX-seq (Jolma et al., 2010, 2013; Slattery et al., 2011) promises to expand considerably the horizons of human transcriptional network analysis. Such refined data may enable differentiation of factors that currently appear to bind similar recognition sequences. Second, the model that we described undervalues the role of distal regulatory elements, which can exert major influences on gene expression. Because enhancers can act over long distances, association of a given distal regulatory element with a specific TF gene is at present difficult. We therefore focused on footprints in DHSs within a 10kb region centered on the transcriptional start site (TSS), in which most regulatory interactions are expected to be directed to the local TSS. Although large numbers of distal regulatory DNA regions marked by DNaseI hypersensitive sites are now available through the ENCODE (ENCODE Project Consortium et al., 2012) and Roadmap Epigenomics (Bernstein et al., 2010) projects, the assignment of distal regulatory elements to their cognate gene(s) has proven to be a formidable challenge. Third, the approach we utilized does not take into account indirect regulatory interactions (e.g., tethering) that may affect the expression of a given TF gene (Davidson et al., 2002a; Rigaud et al., 1991; Biddie et al., 2011). Systematic cross-comparisons between DNaseI footprint and TF ChIP-seq data drawn from the same cell type should enable recognition of such indirect interactions and derivation of rules (e.g., tethering partners) that may enable larger scale modeling of such interactions (Neph et al., 2012c).

In order to interpret human regulatory networks at the organismal level, it will be necessary to analyze cell-selective regulatory networks within the context of surrounding tissues (Barabási and Oltvai, 2004). As initially described by Spemann over 90 years ago, the identity of a given cell can be largely dictated by its surrounding tissue (Spemann, 1918). Consequently, during both normal development and physiological function, the regulatory landscape of one cell type may become intricately dependent upon that of its neighbors (Waddington, 1940). In this context, it is notable that we observed large diversity between the regulatory landscapes of distinct lung cell types (**Fig. 5.M6E**) highlighting the complexity that exists within neighboring tissue from the same organ.

In summary, our results provide the first description of the circuitry, dynamics, and organizing principles of the human transcription factor regulatory network. Systematically applied, the approach we have described has the potential to expand greatly our horizons on the mechanism, architecture, and epistemology of human gene regulation.

5.5 – FIGURES

Figure 5.M1. Construction of comprehensive transcriptional regulatory networks



(A) *Schematic for construction of regulatory networks using DNaseI footprints*. Transcription factor (TF) genes represent network nodes. Each TF node has regulatory inputs (TF footprints within its proximal regulatory regions), and regulatory outputs (footprints of that TF in the regulatory regions of other TF genes). Inputs and outputs comprise the regulatory network interactions 'edges'. For example: *(1)* In Th1 cells, the IRF1 promoter contains DNaseI

footprints matching four regulatory factors (STAT1, CNOT3, SP1 and NFKB). *(2)* In Th1 cells, IRF1 footprints are found upstream of many other genes (for example, GABP1, IRF7, STAT6). *(3)* The same process is iterated for every TF gene in that cell type, enabling compilation of a cell type network comprising nodes (TF genes) and edges (regulatory inputs and outputs of TF genes). *(4)* Network construction is carried out independently using DNaseI footprinting data from each of 41 cell types, resulting in 41 independently-derived cell type networks.

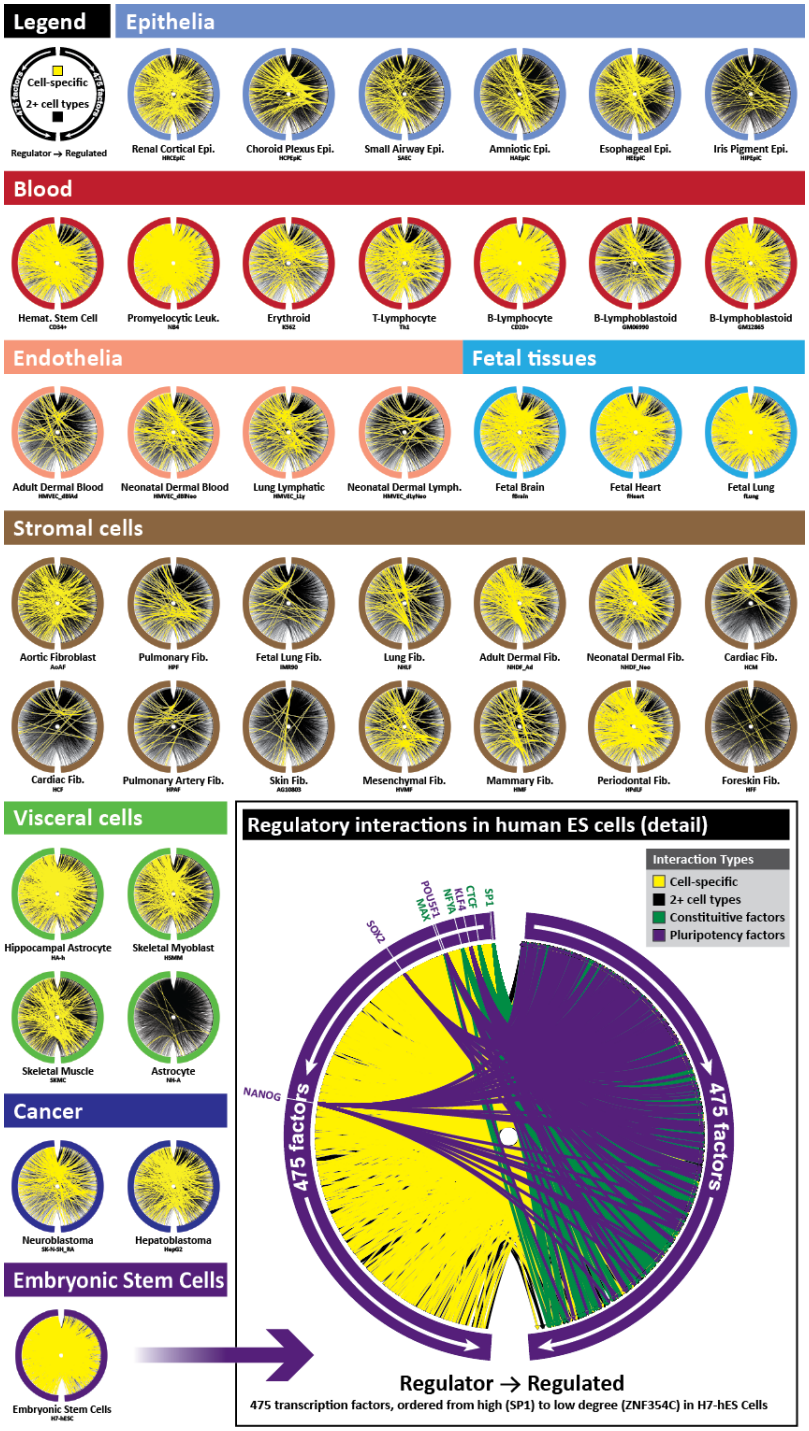
(B and **C)** *Comparison of well-annotated vs. de novo-derived regulatory sub-networks.*

(B) *Muscle sub-network. Top*, experimentally-defined regulatory sub-network for major factors controlling skeletal muscle differentiation and transcription. Arrows indicate direction(s) of regulatory interactions between factors. *Bottom*, regulatory sub-network derived *de novo* from the DNaseI footprint-anchored network of skeletal myoblasts closely matches the experimentally annotated network.

(C) *Pluripotency sub-network. Top*, regulatory sub-network for major pluripotency factors defined experimentally in mouse ES cells (Kim et al. 2008). *Bottom*, regulatory sub-network derived *de novo* from human ES cells is virtually identical to the annotated network.

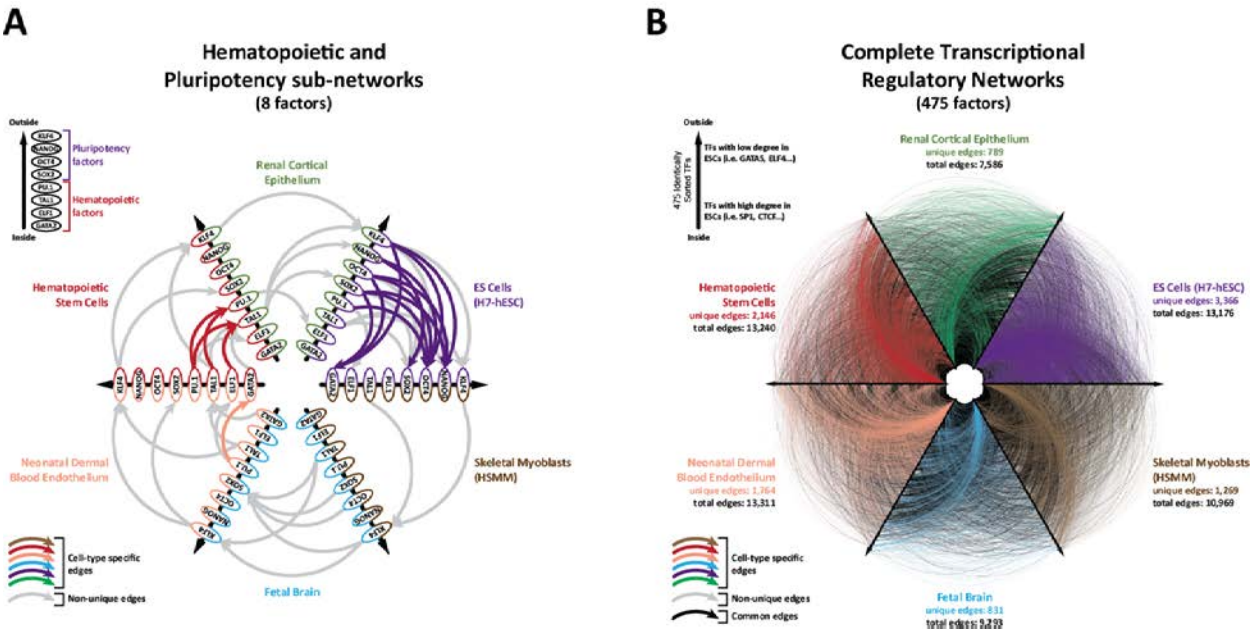
(D,E) *De novo*-derived sub-networks in (B) and (C) match the annotated networks in a cell-specific fashion. *Vertical axes*: Jaccard index, a measure of network similarity, comparing the annotated sub-network with regulatory interactions between the four factors derived *de novo* from each of 41 cell types independently (*horizontal axes*). For the annotated muscle sub-network, the highest similarity is seen in skeletal myoblasts, followed by differentiated skeletal muscle. By contrast, sub-networks computed from fibroblasts are largely devoid of relevant interactions. For the annotated pluripotency sub-network, the highest similarity is seen in human ES cells (H7-ESC).

Figure 5.M2. Cell-specific vs. shared regulatory interactions in TF networks of 41 diverse cell types.



Shown for each of 41 cell types are schematics of cell type-specific vs. non-specific (black) regulatory interactions between 475 TFs. Each half of each circular plot is divided into 475 points (not visible at this scale), one for each transcription factor. Lines connecting the left and right half-circles represent regulatory interactions between each factor and any other factors with which it interacts in the given cell type. *Yellow lines* represent TF-to-TF connections that are specific to the indicated cell type. *Black lines* represent TF-to-TF connections that are seen in two or more cell types. The order of TFs along each half-circular axis is shown in Supplementary Table 1, and represents a sorted list (descending order) of their degree (i.e., number of connections to other TFs) in the ES cell network, from highest degree on top (SP1) to lowest degree on bottom (ZNF354C). Cell-types are grouped based on their developmental and functional properties. Insert on bottom right shows a detailed view of the human ES cell network, and highlights the interactions of four pluripotent (KLF4, NANOG, POU5F1, SOX2) and four constitutive factors (SP1, CTCF, NFYA, MAX) with purple and green edges, respectively.

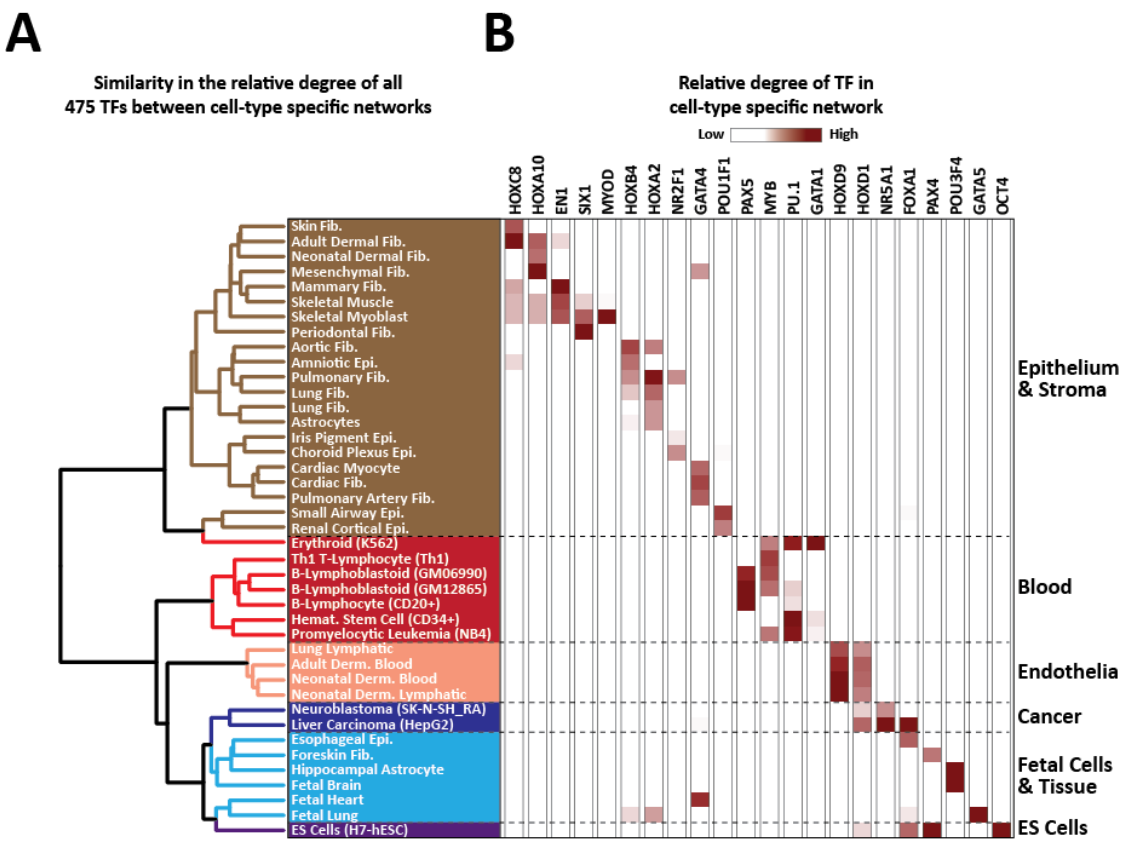
Figure 5.M3. Transcriptional regulatory networks show marked cell-type specificity



(A) *Cross-regulatory interactions between four pluripotency factors and four hematopoietic factors in regulatory networks of six diverse cell types.* All eight factors are arranged in the same order along each axis. Regulatory interactions (i.e., from regulat*or* to regulat*ed*) are shown by arrows in clock-wise orientation. Cell type-specific edges are colored as indicated, whereas regulatory interactions present in two or more cell type networks are shown in grey.

(B) *Cross-regulatory interactions between all 475 TFs in regulatory networks of six diverse cell types.* The 475 TFs are arranged in the same order along each axis, regulatory interactions directed clockwise. Edges unique to a given cell type network are colored as indicated in the legend whereas regulatory interactions present in two or more networks are colored grey. Interactions present in all six cell type networks are colored black. (See also Supplementary Figure S1 and Supplementary Table S2).

Figure 5.M4. Functionally related cell types share similar core transcriptional regulatory networks

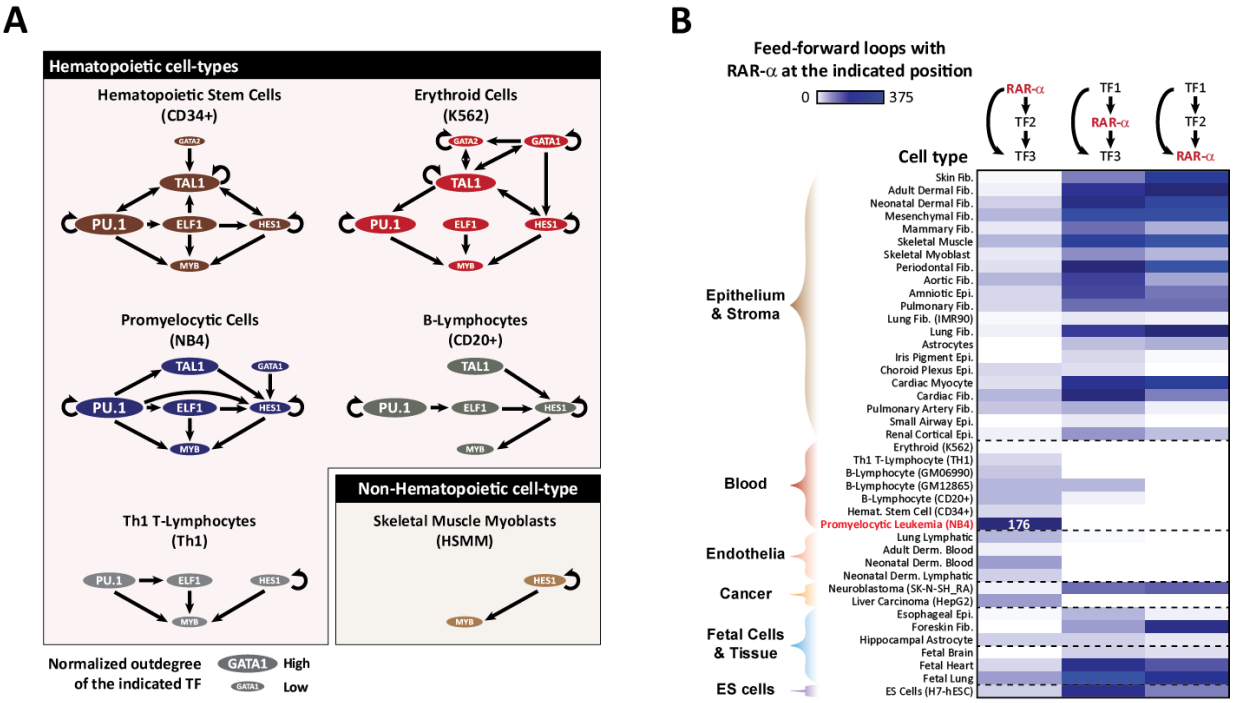


(A) *Clustering of cell type networks by normalized network degree (NND).* For each of 475 TFs within a given cell type network, the relative number of edges was compared between all 41 cell-types using a Euclidean distance metric and Ward clustering. Cell types are colored based on their physiological and/or functional properties.

(B) *Relative degree of master regulatory TFs in cell type networks.* Shown is a heatmap representing the relative normalized degree of the indicated TFs between each of the 41 cell types. For a given TF and cell type, high relative degree indicates high connectivity with other TFs in that cell type. Note that the relative degree of known regulators of cell fate such as

MYOD, OCT4, or MYB is highest in their cognate cell type or lineage. Similar patterns were found for other TFs without previously recognized roles in specification of cell identity.

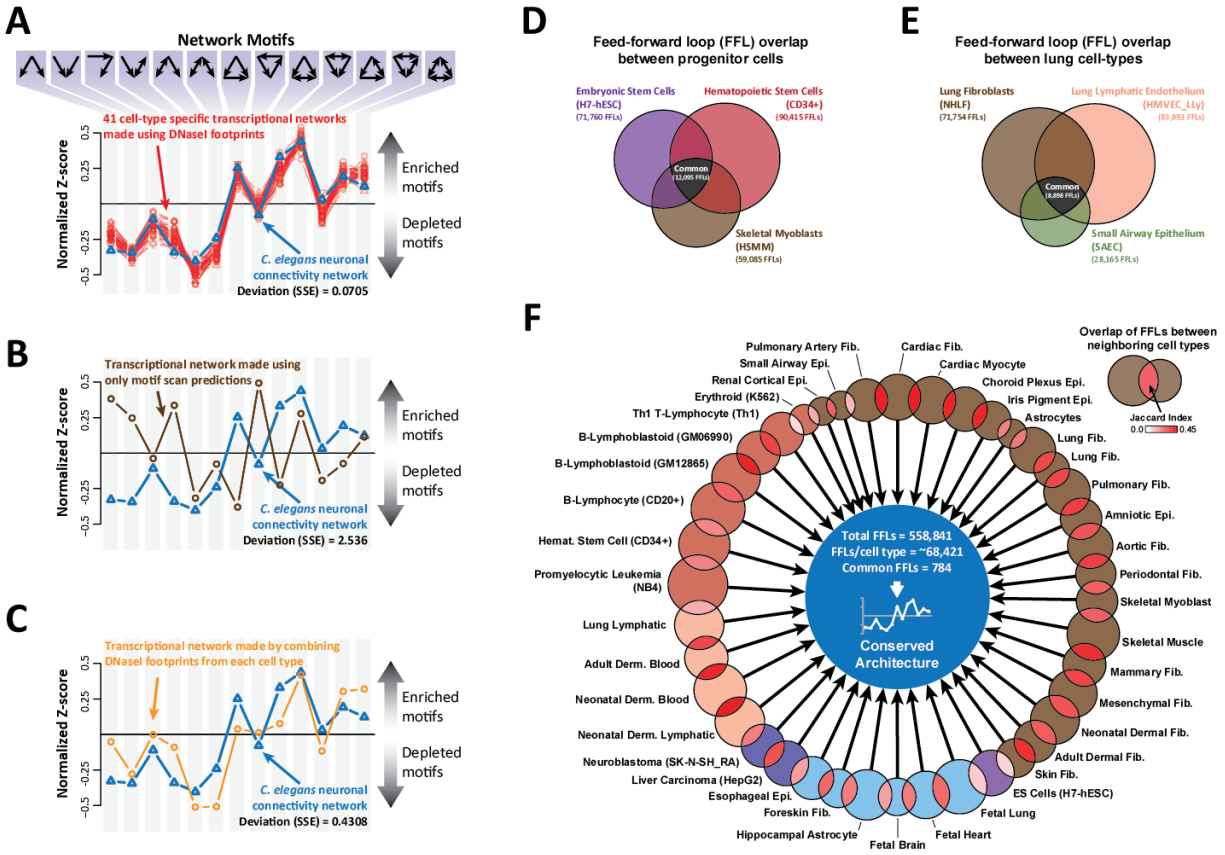
Figure 5.M5. Cell-selective behaviors of widely expressed TFs



(A) Shown are regulatory sub-networks comprising edges (arrows) between seven major hematopoietic regulators in five hematopoietic and one non-hematopoietic cell types. For each TF, the size of the corresponding colored oval is proportional to the normalized out-degree (i.e., out-going regulatory interactions) of that factor within the complete network of each cell type. The early hematopoietic fate decision factor PU.1 appears to play the largest role in hematopoietic stem cells (CD34+) and in promyelocytic leukemia (NB4) cells. The erythroid-specific regulator GATA1 appears as a strong driver of the core TAL1/PU.1/HES1/MYB network specifically within erythroid cells. In both B-cells and T-cells, the sub-network takes on a directional character, with PU.1 in a superior position. By contrast, the network is largely absent in non-hematopoietic cells (muscle, HSMM, bottom right).

(B) Heatmap showing the frequency with which the retinoic acid receptor-alpha (RAR-α) is positioned as a driver (top) or passenger (middle or bottom) within feed-forward loops (FFLs) mapped in 41 cell type regulatory networks. Note that in most cell-types, RAR-α participates in feed-forward loops at 'passenger' positions 2 and 3. However, within blood and endothelial cells, RAR-α switches from being a passenger of FFLs to being a driver (top position) of FFLs. In acute promyelocytic leukemia cells (NB4), RAR-α acts exclusively as a potent driver of feed-forward loops. Cell types are arranged according to the clustered ordering in Figure 4.

Figure 5.M6. Conserved architecture of human transcription factor regulatory networks



(A) Shown is the relative enrichment or depletion of the 13 possible three-node architectural network motifs within the regulatory networks of each cell type (red lines), compared with the relative enrichment of the same motifs in the *C. elegans* neuronal connectivity network. Note that the network architecture of each individual cell type closely mirrors that of the living neuronal network (average summed squared error (SSE) of only 0.0705).

(B) Enrichment of each triad network motifs for a transcription factor network computed using only motif scan predictions within +/- 5kb of TF promoters (brown line). The resulting network bares little resemblance to the *C. elegans* network (blue line) (SSE of 2.536).

(C) The relative enrichment of different triad network motifs is shown for a transcription factor regulatory network generated by pooling DNaseI footprints from all 41 tested cell types into a single archetype (orange line). The resulting topology diverges considerably from that of the neuronal network, far more so than was observed for any individual cell type (SSE of 0.4308).

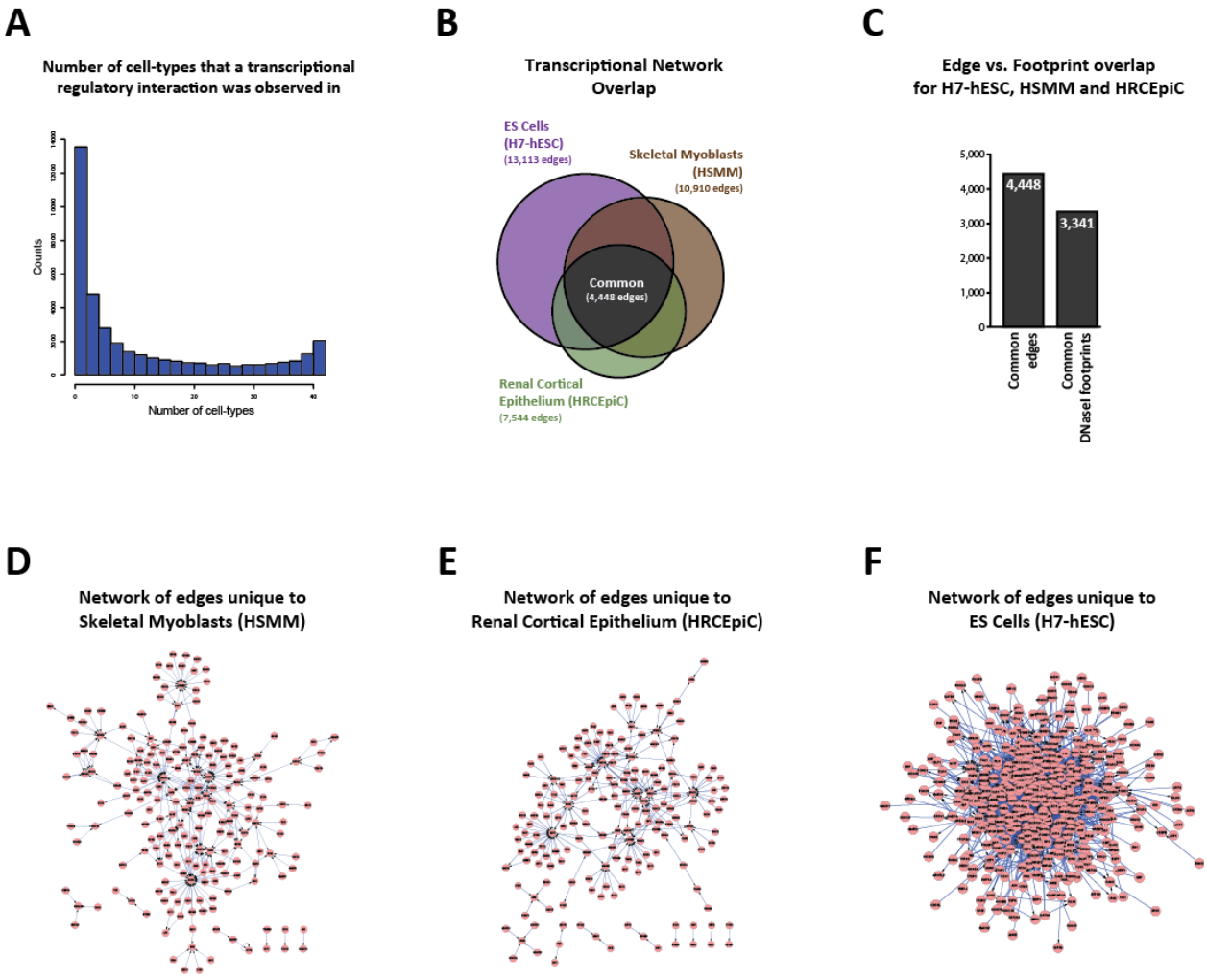
(D - E) *Network architectures are highly cell-specific*

(D) Overlap of feed-forward loops (FFLs) identified in three different progenitor cell types - embryonic stem cells (H7-hESC), hematopoietic stem cells (CD34+) and skeletal muscle myoblasts (HSMM). Note that most FFLs are restricted to an individual cell type.

(E) Overlap of feed-forward loops (FFLs) identified in three pulmonary cell types - lung fibroblasts (NHLF), small airway epithelium (SAEC), and pulmonary lymphatic endothelium (HMVEC_LLy). Highly distinct architectures are present even among cell types from the same organ structure.

(F) Overlap of FFLs from networks of neighboring cell types, following the ordering and coloration shown in Figure 4A. The size of each circle is proportional to the number of FFLs contained within the network of the corresponding cell type. The color of the intersection region between adjacent cell types indicates the Jaccard index between FFLs from those two cell types (see legend in upper right of panel F). The average number of FFLs in each network, the total number of FFLs across all networks and the number of common FFLs across all networks is indicated in the center of the graph. (See also Supplemental Figure S3 and Supplemental Table S3).

Figure 5.S1. Overlap of cell-type specific transcriptional regulatory networks



(A) Histogram showing the number of cell types that each transcriptional regulatory interaction (edge) was observed in.

(B) The overlap of transcriptional regulatory interactions (edges) identified in ES cells (H7-hESC), skeletal muscle myoblasts (HSMM) and renal cortical epithelium (HRCEpiC).

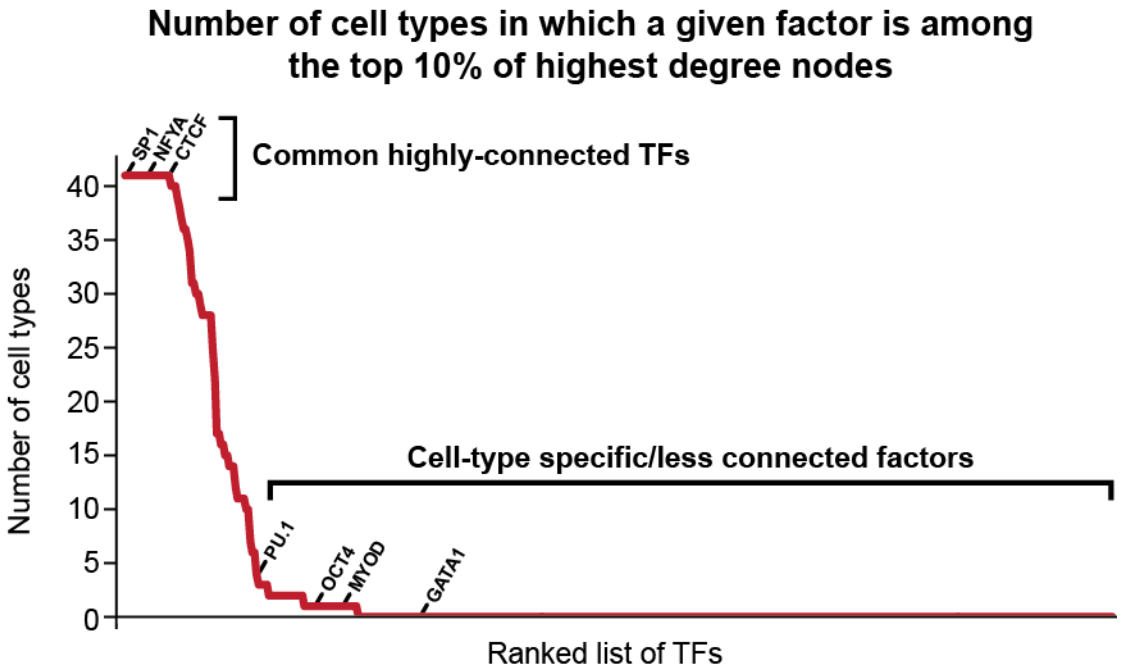
(C) The number of common edges and common DNaseI footprints between the ES cells (H7-hESC), skeletal muscle myoblasts (HSMM) and renal cortical epithelium (HRCEpiC) networks.

(D) Cytoscape derived network showing all edges that are unique to the skeletal muscle myoblasts (HSMM) network.

(E) Cytoscape derived network showing all edges that are unique to the renal cortical epithelium (HRCEpiC) network.

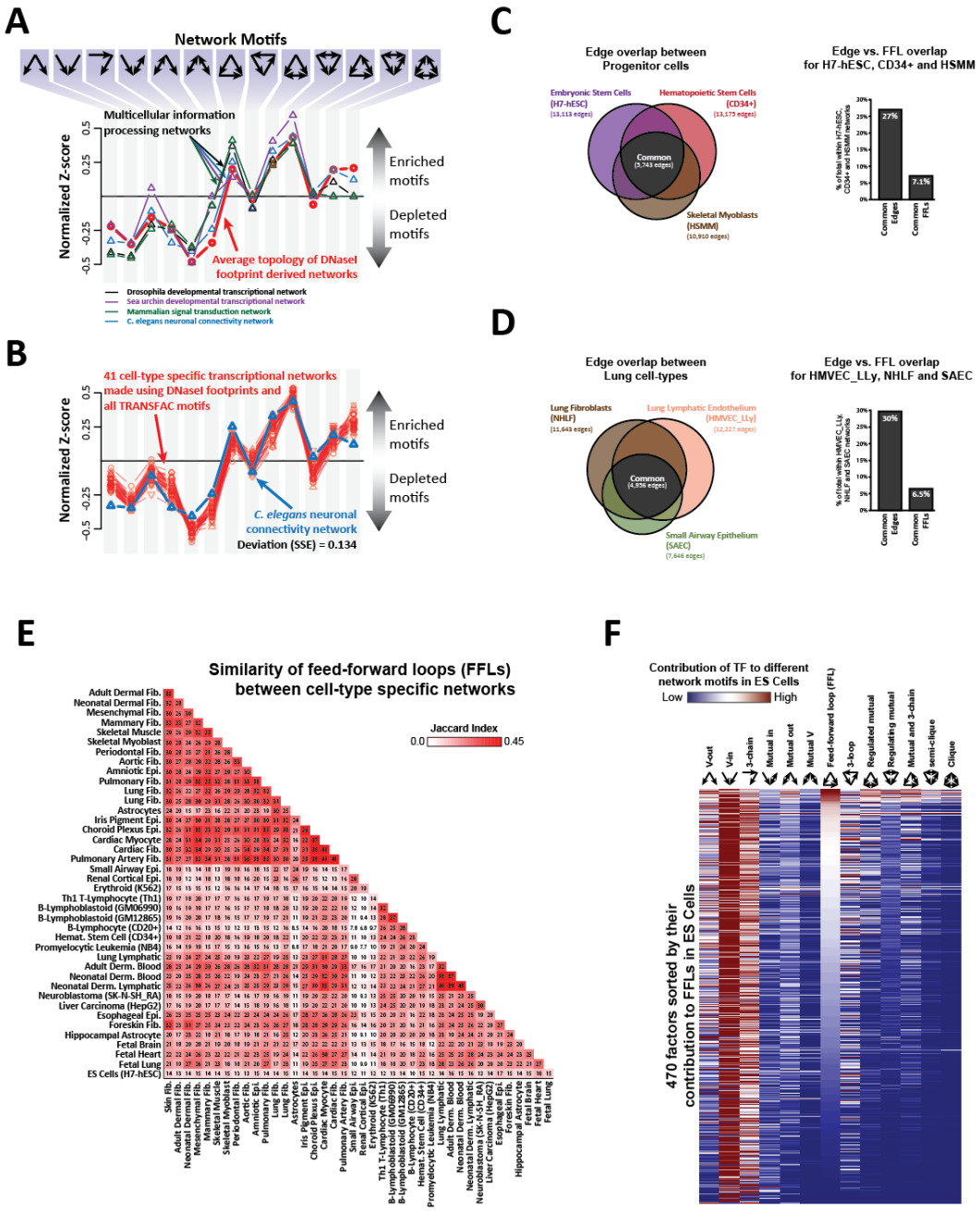
(F) Cytoscape derived network showing all edges that are unique to the ES cell (H7-hESC) network.

Figure 5.S2. Identification of common highly-connected TFs



Shown is the number of cell-type specific networks in which a given factor is among the top 10% of highest degree nodes.

Figure 5.S3. Transcriptional regulatory networks have a conserved network motif architecture



(A) Shown is the average relative enrichment or depletion of the 13 possible three-node architectural network motifs within the regulatory networks of each cell type (red line), compared with the relative enrichment of the same motifs in four previously published multicellular biological networks (Milo 2004); *C. elegans* neuronal connectivity network (blue line), the mammalian signal transduction network (green line) and the sea-urchin (purple line) and *Drosophila* (black line) developmental transcriptional networks.

(B) Shown is the relative enrichment or depletion of the 13 possible three-node architectural network motifs within the regulatory networks of each cell type constructed using all 538 TRANSFAC motifs, including redundant motifs (red lines).

(C) The overlap of edges identified in three progenitor cell types - embryonic stem cells (H7-hESC), hematopoietic stem cells (CD34+) and skeletal muscle myoblasts (HSMM). Shown to the right is the percentage of all edges common to these three cell types, as well as the percentage of all FFLs common to these three cell types.

(D) The overlap of edges identified in three pulmonary cell types - lung fibroblasts (NHLF), lung lymphatic endothelium (HMVEC_LLy) and small airway epithelium (SAEC). Shown to the right is the percentage of all edges common to these three cell types, as well as the percentage of all FFLs common to these three cell types.

(E) Overlap of FFLs from networks of each cell type, following the ordering shown in Figure 4A. The color of each box corresponded to the Jaccard index between FFLs from the two cell-type specific networks contributing to that box.

(F) Heatmap showing the contribution of all 470 TFs with interactions in ES cells (H7-hESC) to 13 possible three-node architectural network motifs in the ES cell-type specific network. The factors are sorted by their contribution to feed-forward loops (FFLs).

5.6 – METHODS

Regulatory network construction

We mapped motif-binding protein information found in TRANSFAC to 538 coding genes, using GeneCards (Rebhan et al., 1997) and UniProt Knowledgebase (Magrane and Consortium, 2011). Some genes were indistinguishable when viewed from a potential motif-binding event perspective, as their respective gene products were annotated as binders to the same set of motif templates by TRANSFAC. In such cases, we chose a single gene, randomly, as a representative and removed others which reduced the number of genes from 538 to 475. Networks built by removing the first redundant motif, alphabetically, or by including all redundant motifs showed very similar properties to the one described in this paper (**Fig. 5.S3B** and data not shown).

We symmetrically padded the transcriptional start sites (TSSs) of the remaining genes by 5,000 nt and scanned for predicted TRANSFAC motif binding sites using FIMO (Bailey et al., 2009), version 4.6.1, with a maximum *p*-value threshold of 1e-5 and defaults for other parameters. For each cell type, we filtered putative motif binding sites to those that overlapped footprints as previously described (Neph et al., 2012c). Each network contained 475 vertices, one per gene. A directed edge was drawn from a gene-vertex to another when a motif instance, potentially bound by the first gene's protein product, was found within a DNaseI footprint contained within 5,000 nt of the second gene's TSS, indicating regulatory potential. Supplemental Table S3 shows the number of such edges in every cell-type specific network.

Network Visualization

We identified interactions which were unique to a single cell type, or cell-specific, and marked those found in two or more of the 41 tested cell types as “common”. Interactions were

rendered with Circos (Krzywinski et al., 2009), version 0.55. Within Circos nomenclature, two pseudo-chromosomes (ideograms) represent identically sorted lists of "regulator" and "regulated" factors, with a directed edge between ideograms indicating that the first factor regulates the second. Ideograms were colored by association of the cell type with tissue category. Unique and common interactions between ideograms were labeled with yellow and black colors, respectively, to visually differentiate cell types by the number and distribution of edges. Transcription factors were oriented along both ideograms by the sort order provided by the H7-hESC cell type, from highest degree (SP1) to lowest (ZNF354C). For the detail view of H7-hESC, we also highlighted the interactions of four pluripotent (KLF4, NANOG, POU5F1, SOX2) and four constitutive factors (SP1, CTCF, NFYA, MAX) with purple and green edges, respectively.

Hive plots

We generated a hive plot (Krzywinski et al., 2011) using the R library HiveR, version 0.2.1, to visualize directed interactions for four hematopoietic (PU.1, TAL1, ELF1, GATA2) and four pluripotent genes (KLF4, NANOG, OCT4, SOX2) among six cell types (H7-hESC, HRCEpiC, CD34+, HMVEC_dBlNeo, fBrain and HSMM). The hive plot was divided into six sections, one for each cell type. Reading the figure in clockwise fashion, a directed edge drawn from one axis to the next indicates the first gene regulating the second. Genes were oriented identically along each axis. Common interactions were defined by an interaction existing in two or more cell types. A second qualitative hive plot was created between the same six cell types and over all 475 TFs.

Unique edge connectedness

We calculated the mean weakly-connected component size using edges unique to a cell type. To identify whether these unique component sub-networks were more connected than would be expected by chance, we randomly subsampled the same number of real edges in the same cell type and recalculated the mean-component size. This process was iterated 100,000 times, and the number of times for a cell type that the mean-component size in random graphs equaled or exceeded that of the unique component graph counterpart was tallied. An empirical p-value was calculated as the tally plus one divided by 100,000. Sub-networks made up of unique edges belonging to each of HSMM, HRCEpiC, and H7-hESC were separately plotted in Cytoscape (Smoot et al., 2011).

Network clustering

We counted the total number edges for every TF gene-vertex (sum of in and out edges) in a cell type and calculated the proportion of edges for that TF relative to all edges in that cell type (normalized network degree (NND)). We computed the pairwise euclidean distances between cell types using the rescaled NND vectors and grouped the cell types using Ward clustering (Ward, 1963). We observed similar cluster patterns when comparing rescaled in-degree, rescaled out-degree, or unscaled total degree (results not shown).

Cell-type specific behaviors

We utilized the mfinder software (Milo et al., 2004), version 1.20, to pull out all feed-forward loop instances in regulatory networks. Prior to using the software, all self-edges, those from a TF gene-vertex to itself, were removed per the requirements of the software. The software parameters were set to *-ospmem <motif-number> -maxmem 1000000 -s 3 -r 250 -z -2000*, where *<motif-number>* was one of 13 possible unique 3-node network motif identifiers.

Triad Significance Profiles

We removed self-edges from every network and used the mfinder software tool for network motif analysis (Milo et al., 2004). A Z-score was calculated over each of 13 network motifs of size 3 (3-node network motifs), using 250 randomized networks of the same size to estimate a null. We vectorized Z-scores from every cell type and normalized each to unit length to create triad significance profiles (TSP) as described in Milo et al., 2004. We computed the average TSP over all cell-type specific regulatory networks and compared to the TSP of the highly-curated multi-cellular information processing networks described in Milo et al., 2004. All sum squared error (SSE) calculations were done by comparing our derived networks against the *Caenorhabditis elegans* profile (White et al., 1986).

To generate a transcriptional network using only motif scan predictions we created a new network, with 86,242 edges, by using all putative motifs within 5,000 nucleotides of the TSSs of each of the 475 TF genes, without conditioning on footprint overlaps. We analyzed this network using the mfinder software as described above, creating a TSP and comparing to the *Caenorhabditis elegans* profile.

To generate a transcriptional network from DNaseI footprints from all cell types we merged footprints across all cell types and filtered motif instances to those overlapping the merged set by at least 3 nt, creating another new network with 38,165 edges. We analyzed this network using the mfinder software as described above, creating a TSP and comparing to the *Caenorhabditis elegans* profile.

Network feature overlaps

We compared cell-type specific networks in greater detail using only feed-forward loops. Summaries of overlaps were made between a small number of cell types using Venn diagrams

and barplots. All pairwise overlaps were computed and summarized using the Jaccard index (number of feed-forward loops in the pairwise set intersection divided by the number in the pairwise set union – **Fig. 5.S3E**). We additionally computed overlaps and differences between entire regulatory networks in terms of shared and unshared edges, as well as footprints (**Figs. 5.S1B,C**).

To identify the contribution of each factor to each network motif, we counted the number of times a factor was present in each of the 13 3-node network motifs within the H7-hESC cell type, in any motif position (**Fig. 5.S3F**). We scaled each column vector to length 100, and then divided each element of a row vector by the maximum value in that row to visualize contributions in heatmap form using the matrix2png program without row normalization (Pavlidis and Noble, 2003).

Chapter 6 – Chromatin environment models the affinity and evolutionary constraint of TF binding elements

6.1 – ABSTRACT

The genomic and chromatin features that model the occupancy landscape of a TF along the genome are poorly understood. To address this, we mapped the genome-wide *in vivo* affinity landscape of a natively expressed human TF using a novel method termed 'salted ChIP-seq'. We show that the *in vivo* affinity landscape of CTCF is modeled by both base-specific and structure-specific protein-DNA interactions. However, the *in vivo* occupancy landscape of CTCF is extensively modeled by surrounding chromatin features, not affinity. The affinity of TF binding elements genome-wide have evolved to match their surrounding chromatin environments, and cell-selective TF binding sites necessitate low-affinity binding elements, which are sensitive to changes in the surrounding chromatin environment. Finally, TF affinity, not occupancy, is predominantly driving the extent of evolutionary sequence constraint at TF binding elements genome-wide. Our results indicate a central role chromatin environment plays in modeling the affinity and evolutionary constraint of TF binding elements genome-wide.

6.2 – INTRODUCTION

The organization of nuclear chromatin is intimately connected with genome function and gene expression (Weintraub and Groudine, 1976). Within a cell, the overwhelming majority of the genome is tightly packaged into chromatin and only a small minority of the genome is actually accessible to sequence-specific transcription factors (TFs) (Thurman et al., 2012). However, as a cell progresses through development, TFs continually model this chromatin

landscape, opening and closing chromatin at cell-type and lineage-specific regulatory elements in a highly reproducible and evolutionarily conserved fashion (Stergachis et al., *in revision*). To accomplish this, TFs negotiate diverse chromatin environments to find and occupy their cognate recognition sequences within regulatory elements. The cohort of regulatory elements occupied by TFs within a cell directly models the identity of that cell (Weintraub and Groudine, 1976), however, our understanding of the genomic and cellular features that model the occupancy landscape of a TF within a cell is woefully inadequate. To tease apart the contribution of DNA sequence and chromatin environment to the occupancy profile of a TF, we sought to obtain *in vivo* genome-wide measurements of TF occupancy, TF affinity and chromatin environment.

Chromatin environment is a composite measurement that encapsulates multiple features of the chromatin surrounding a binding element including: (i) chromatin accessibility; (ii) co-bound TFs; and (iii) location relative to transcriptional start sites (TSSs). The *in vivo* genome-wide localization of distinct chromatin environments can be readily mapped using standard chromatin immunoprecipitation (ChIP-seq) and DNaseI hypersensitivity (DNaseI-seq) methods in combination with gene model annotations (Filion et al., 2010; Ernst et al., 2011; Thurman et al., 2012).

TF occupancy is the frequency that a TF occupies a regulatory DNA element within a population of cells. *In vivo* TF occupancy is widely mapped using standard chromatin immunoprecipitation (ChIP-seq) methods (Johnson et al., 2007; ENCODE Project Consortium et al., 2012). Unfortunately, when describing ChIP-seq data, the terms affinity and occupancy are often incorrectly used interchangeably. However, the premise behind ChIP-seq is to map sites of TF occupancy along the genome, where the height of the ChIP-seq peak corresponds to the number of DNA templates within a population of cells that are occupied by the TF (Biggin,

2011). The actual relationship between this value and the affinity of a TF at a site is currently unknown.

TF affinity is a biochemical attribute of the TF-DNA interface that describes the penchant of a TF to bind to a given regulatory DNA element, independent of the concentration of the TF or DNA ligand. Numerous approaches have been developed over the past several decades to quantify the affinity of a TF towards a single (Galas and Schmitz, 1978; Garner and Revzin, 1981; Fried and Crothers, 1981; Kneale and Wijnaendts van Resandt, 1985; Privalov et al., 1999), or thousands of DNA ligands *in vitro* (Berger et al., 2006; Slattery et al., 2011; Nutiu et al., 2011). These *in vitro* approaches rely on the use of purified proteins and are performed outside of the native genomic or chromatin context of a TF binding element. In contrast, measurements of *in vivo* TF affinity have proven much more difficult, and currently can only be obtained using native footprinting approaches such as DNaseI footprinting (Pfeifer and Riggs, 1991). *In vivo* DNaseI footprinting measures the relative affinity of TFs at bound elements along the genome by quantifying the rate of DNaseI cleavage within and surrounding each TF binding element - the footprint 'depth'. As detailed by Galas and Schmitz in their initial description of DNaseI footprinting (Galas and Schmitz, 1978), the 'depth' of a DNaseI footprint at a given binding element is directly proportional to the relative affinity of a TF for that element since the concentration of a TF should be largely invariant throughout the nucleus and DNaseI is only able to cleave available DNA templates (Thurman et al., 2012). However, measurements of TF affinity derived from *in vivo* DNaseI footprinting can be influenced by the inherent bias DNaseI has towards cleaving certain DNA structures (Galas and Schmitz, 1978; Drew and Travers, 1984, 1985; Lazarovici et al., 2013), as well as potential subtle differences in TF concentration within different nuclear domains (Phair et al., 2004), and when scaled genome-wide is confounded by

the sequencing depth at a given binding element (Neph et al., 2012c). Given the shortcomings of current methods, we sought to develop an alternative approach to mapping the *in vivo* affinity of a TF genome-wide that is readily scalable to any TF and organism.

To map the relative affinity of all binding sites of a TF genome-wide, we leveraged the fact that protein-DNA interactions are largely electrostatic in nature. The binding energy of a protein-DNA complex is comprised of both: (i) electrostatic interactions between positively charged amino acids within the TF and negatively charged phosphates along the DNA backbone that constitute a 'generalized' affinity of a TF towards DNA; and (ii) non-electrostatic interactions between the TF and specific DNA bases that form the basis of 'DNA element-specific' affinity (Privalov et al., 2011). According to the counterion condensation (CC) concept these electrostatic interactions are directly dependent on the salt concentration of the binding reaction (Manning, 1969, 1978; Record et al., 1976, 1991). Based on this model, a previous study of the salt dependencies of protein-DNA interactions (Privalov et al., 2011) found that: (i) the electrostatic component provides the bulk of the energy stabilizing a protein-DNA complex; (ii) the non-electrostatic (salt-independent) interactions result from sequence-specific protein-DNA interactions; and (iii) the salt dependency of the binding constant is linear on the logarithmic scale. Consequently, a protein-DNA complex with many sequence-specific interactions (i.e. a high affinity binding element) would take a higher concentration of salt to disrupt than a protein-DNA complex with few sequence-specific interactions (i.e. a low affinity binding element), as the stability of the low affinity protein-DNA complex is more salt-dependent. As such, the overall stability of a protein-DNA complex can be quantified by measuring the relative concentration of salt needed to disrupt that specific protein-DNA complex.

We have developed and implemented a novel method leveraging this property of protein-DNA interactions to generate genome-wide *in vivo* maps of TF affinity termed '*salted chromatin immunoprecipitations'* or *'salted ChIPs'*. Briefly, salted ChIPs measures the occupancy (using ChIP-seq) of a TF genome-wide after a nucleus has been treated with varying concentrations of salt. Using these occupancy measurements, we can calculate the exact concentration of salt needed to disrupt the protein-DNA complex formed at each individual binding site of a TF along the genome, and thus the relative affinity of a TF for each of its binding sites genome-wide. One of the major benefits of this approach is that it can be readily applied to any TF, cellular system or organism in which ChIP-seq can be applied. To address the physiological and evolutionary interplay between TF affinity, TF occupancy and chromatin environment we sought to apply salted ChIP-seq to the binding landscape of CTCF in human K562 cells. This system provides several unique benefits including: (i) CTCF is a well studied TF that is involved in diverse cellular processes and is known to occupy a multitude of different chromatin environments (Phillips and Corces, 2009); and (ii) the chromatin environments of K562 cells have been extensively mapped as part of the ENCODE consortium (ENCODE Project Consortium et al., 2012).

Our findings, detailed below, suggest five fundamental conclusions. First, the *in vivo* affinity of CTCF at a binding element results from CTCF engaging DNA in both base-specific and DNA structure-specific interactions using its full repertoire of 11 zinc fingers. Second, CTCF is involved in extensive 'accessibility-mediated cooperativity' with other TFs, resulting in a poor association between the affinity and occupancy of CTCF at a binding element. Third, low-affinity CTCF binding elements enable cell-selective CTCF regulation. Fourth, CTCF affinity, not occupancy, is largely dictating the extent of evolutionary sequence constraint at CTCF

binding elements genome-wide. Finally, the DNA sequences of CTCF binding elements genome-wide have evolved such that the affinity of CTCF at these elements is matched to their chromatin environments. Together, these findings indicate a central role chromatin environment plays in modeling the affinity and evolutionary constraint of TF binding elements genome-wide.

6.3 – RESULTS

Genome-wide in vivo affinity map of CTCF

Salted ChIP-seq provides a novel methodology for generating genome-wide *in vivo* maps of TF affinity. By measuring the occupancy (using ChIP-seq) of a TF genome-wide after a nucleus has been treated with varying concentrations of salt, one can calculate the exact concentration of salt needed to disrupt a specific protein-DNA interaction, and thus the relative affinity of a TF for each of its binding sites along the genome. To identify the *in vivo* affinity landscape of a TF, we applied salted ChIP-seq to the human master regulator CTCF. Specifically, we measured the occupancy of CTCF (using ChIP-seq) in intact erythroleukemia nuclei (K562) as well as K562 nuclei that had been treated with 15 different concentrations of salt ranging from 80mM to 600mM NaCl (**Fig. 6.M1A**). Visual inspection of the occupancy patterns of CTCF after varying concentrations of salt revealed that different CTCF binding elements along the genome exhibited markedly distinct salt stabilities, with the salt stability of neighboring binding elements sometimes differing by over 200mM NaCl (**Fig. 6.M1B**). To test the reproducibility of these salted ChIP-seq occupancy patterns, we generated biological replicates for intact nuclei and four of the salt extraction conditions (80mM, 240mM, 320mM and 450mM NaCl). Both a visual and quantitative inspection of these data revealed that salted ChIP-seq occupancy patterns are highly reproducible (**Fig. 6.S1A-B**).

To systematically quantify the exact concentration of salt needed to disrupt each CTCF binding site along the genome, we first identified a set of 21,569 high-quality CTCF ChIP-seq peaks (**Fig. 6.S1C-D**) and then calculated the occupancy of CTCF at each of these peaks after treatment with 15 different concentrations of salt. As the response of CTCF occupancy to salt concentration appeared to closely mirror a dose-response curve (**Fig. 6.M1C**), we applied a four-parameter log-logistic function independently to each CTCF binding site genome-wide and using these fitted curves we calculated the concentration of salt needed to disrupt 50% of the CTCF occupancy (hereafter referred to as the EC_{50}) (**Fig. 6.M1C** and Methods). The extensive number of salt concentrations used in this experiment enabled us to derive precise EC_{50} values for > 97% of all CTCF ChIP-seq peaks (21,015 total). These EC_{50} values provide a quantitative metric of the relative affinity of CTCF at each of its binding sites along the genome, and reveal a previously unrecognized rich diversity in the genome-wide affinity landscape of a transcription factor (**Fig. 6.M1D**).

Extensive validation of salted ChIP affinity measurements

To validate these salted ChIP-seq measurements of *in vivo* CTCF affinity we employed the only other method that enables *in vivo* measurements of TF affinity - *in vivo* DNaseI footprinting. DNaseI footprints are formed within DNaseI hypersensitive regulatory elements (DHSs) by the sequence-specific binding of transcription factors. When bound within DHSs, TFs occlude bound DNA from cleavage by DNaseI, leaving DNaseI footprints that demarcate transcription factor occupancy at nucleotide resolution (Hesselberth et al., 2009; Neph et al., 2012c). The extent to which a binding element is occluded from cleavage by DNaseI is directly

related to the affinity of the TF towards that binding element, with *'deeper'* DNaseI footprints corresponding to higher affinity binding elements (**Fig. 6.M2A**) (Galas and Schmitz, 1978).

Comparison of the average DNaseI cleavage profiles at CTCF binding elements of differing EC_{50} values revealed a strong relationship between the depth of the DNaseI footprint and the affinity of CTCF at that binding element as measured using salted ChIP-seq (**Fig. 6.M2B** and **Fig. 6.S2A-B**). Whereas few DNaseI cleavages were observed within the footprinted region of high affinity CTCF binding elements, the DNaseI cleavage profile within the footprinted region of low affinity CTCF binding elements began to mirror that at unbound DNA (**Fig. 6.S2C**), demonstrating that the proportion of unbound, DNaseI accessible DNA templates at a binding site is directly related to the affinity of CTCF towards that binding site.

Although inherent biases in DNaseI cleavage specificity can influence the quantification of the DNaseI footprint depth at a single binding element (Galas and Schmitz, 1978), a genome-wide comparison of CTCF DNaseI footprint depths and CTCF affinity revealed a strong correlation (**Fig. 6.M2C** and **Fig. 6.S2D**). In contrast, the CTCF DNaseI footprint depth only weakly correlated with CTCF occupancy (**Fig. 6.M2D**) and failed to correlate with DNaseI accessibility (**Fig. 6.M2E**), hinting at a possible complex interplay between CTCF affinity, CTCF occupancy and chromatin accessibility. Overall, these *in vivo* DNaseI footprinting data provide extensive validation of our salted ChIP-seq measurements of *in vivo* CTCF affinity.

CTCF affinity is modeled by base-specific protein-DNA interactions

We next sought to understand whether DNA sequence features might contribute to the diversity of affinity values observed at different CTCF binding elements along the genome. Of note, as a transcription factor, CTCF is largely unique in its ability to engage three distinct DNA

binding elements (**Fig. 6.M3A**) (Rhee and Pugh, 2011; Schmidt et al., 2012)(Stergachis et al., in preparation). At CTCF_M (medium) binding elements CTCF engages in sequence-specific interactions only within a core binding region. In contrast, at CTCF_L (long) and CTCF_XL (extra-long) binding elements CTCF engages in sequence-specific interactions at both an upstream and core binding region, which differ in their spacing at CTCF_L and CTCF_XL elements (**Fig. 6.M3A**). We find that these extended, non-canonical CTCF binding elements (CTCF_L and CTCF_XL) are preferentially found at high affinity CTCF binding sites (**Fig. 6.M3B-C** and **Fig. 6.S3A**), suggesting that CTCF may utilize these different binding modes to modulate its affinity at different binding sites.

We next wanted to determine if CTCF is engaged in more sequence-specific protein-DNA interactions at high affinity binding elements, independent of which binding mode it is using. Of note, we find that the similarity of different CTCF binding elements to the CTCF consensus motif poorly predicts the occupancy of CTCF at these elements (**Fig. 6.M3D** and **Fig. 6.S3B**), similar to what has been observed for other TFs (Carr and Biggin, 1999; Yang et al., 2006; Li et al., 2008; Cuellar-Partida et al., 2012). In contrast, we find that the affinity of CTCF at a binding element is more strongly predicted by how well that binding element matches the consensus CTCF motif (**Fig. 6.M3E** and **Fig. 6.S3C**), suggesting that the affinity of CTCF at a binding element is largely related to the extent of available sequence-specific interactions. To better visualize this, we plotted the base composition at both high affinity and low affinity CTCF_M (**Fig. 6.M3F**), CTCF_L (**Fig. 6.S3D**) and CTCF_XL (**Fig. 6.S3E**) binding elements. This revealed that: (i) high affinity CTCF binding sites utilize sequence-rich binding elements; (ii) certain bases appear to be necessary to establish a binding element, whereas other bases

appear to contribute to the affinity of CTCF at that element; and (iii) CTCF has slight differences in its base preferences at each of the three different binding elements (**Fig. 6.S3F**).

CTCF affinity is modeled by structure-specific protein-DNA interactions

Although CTCF binds DNA with all 11 of its zinc fingers, anywhere between 3 (CTCF_L and CTCF_XL binding modes) to 6 (CTCF_M binding mode) of these zinc fingers do not appear to be involved in base-specific protein-DNA contacts (**Fig. 6.M2** and **Fig. 6.S2**)(Stergachis et al., *in preparation*). Given the ability of DNA shape preferences to guide TF specificity (Joshi et al., 2007; Slattery et al., 2011; Gordân et al., 2013), we wanted to determine whether these DNA base-independent CTCF-DNA contacts were in fact DNA structure-dependent. To test this, we computed the minor groove width at all CTCF binding elements using their primary DNA sequences and well-established DNA prediction tools (Slattery et al., 2011; Lazarovici et al., 2013). Comparison of the DNA minor groove width at high affinity and low affinity CTCF binding elements revealed that: (i) all three binding modes exhibit minor groove width shapes that are preferentially bound by high affinity binding elements and are located outside of the binding region involved in base-specific CTCF-DNA contacts (**Fig. 6.M3H** and **Fig. 6.S3G**); and (ii) the DNA shape at sequences bound by CTCF ZFs 8-11 appears to significantly affect the affinity of CTCF in all three binding modes (**Fig. 6.M3H** and **Fig. 6.S3G**), in spite of the fact that the CTCF_M binding mode has no base-specific contacts formed by these 4 zinc fingers (**Fig. 6.M3F** and **Fig. 6.M3G**). Together these results demonstrate that both DNA base-specific and DNA structure-specific CTCF-DNA interactions contribute to the affinity of CTCF at a binding element *in vivo* (**Fig. 6.M3H**) and that CTCF utilizes its full

repertoire of 11 zinc fingers when interacting with a binding element, independent of the perceived sequence-specific interactions.

Affinity appears to drive the evolutionary sequence constraint at CTCF binding elements

Evolutionary constraint within coding sequence has been well studied, and appears to be largely driven by the selection against variants at functionally important codons. In contrast, the genomic and cellular forces driving evolutionary constraint within non-coding sequence is much less well understood. In particular, it is largely unknown how much the genomic environment of a TF binding element influences the evolutionary constraint at that binding element. To address this, we first wanted to understand the relationship between the evolutionary constraint at a CTCF binding element and the affinity of CTCF at that binding element. Strikingly, for all three CTCF binding modes, the affinity of CTCF at a binding element extensively tracked with the per-nucleotide evolutionary constraint at that binding element (**Fig. 6.M4A**) as well as the total extent of evolutionary constraint at that binding element (**Fig. 6.M4B** and **Figs. 6.S4A-B**).

Alternatively, the occupancy (and hence functional output) of CTCF at a binding element could be driving the evolutionary sequence constraint at that binding element. To test for this, we repeated our analysis of the correlation between evolutionary constraint and CTCF affinity after controlling for CTCF occupancy (**Fig. 6.S4C**). Strikingly, the correlation between sequence constraint and affinity was nearly unchanged even after controlling for CTCF occupancy (**Fig. 6.M4B** and **Fig. 6.S4C**). In contrast, after controlling for CTCF affinity the correlation between evolutionary constraint and CTCF occupancy was largely abrogated (**Fig. 6.S4D**). Together these findings indicate that CTCF affinity, not occupancy, is the predominant driver of evolutionary sequence constraint at CTCF binding elements.

CTCF engages in accessibility-mediated cooperativity with other transcription factors.

In a simple model, the affinity and occupancy of CTCF at different binding elements along the genome would be perfectly correlated. In stark contrast to this simple model, we observed that the affinity of CTCF at a binding site poorly predicts the occupancy of CTCF at that site (**Fig. 6.M5A**). This suggests that certain features of the chromatin environment of a CTCF binding element may be independently affecting the affinity and occupancy of CTCF at that site. To interrogate the effect of chromatin environment on CTCF occupancy we systematically identified CTCF binding sites that were co-occupied by other transcription factors using K562 ChIP-seq data from 38 diverse TFs (ENCODE Project Consortium, 2012). In total, 57% of CTCF binding sites overlapped a ChIP-seq peak from at least one other TF (**Fig. 6.M5B**). Importantly, these co-occupied CTCF binding sites exhibited wider regions of chromatin accessibility (**Fig. 6.M5C**), characteristic of regulatory elements co-occupied by multiple TFs bound along the genome in parallel (Thurman et al., 2012).

Using this TF co-occupancy data, we systematically interrogated the association of TF co-occupancy with CTCF occupancy, CTCF affinity and chromatin accessibility using a linear regression framework (Methods). Strikingly, TF co-occupancy at CTCF binding sites was categorically associated with a significant increase in chromatin accessibility at these CTCF binding sites (average of 259% increase in chromatin accessibility; range 78% to 478% increase) (**Fig. 6.M5D**). In addition, TF co-occupancy at CTCF binding sites was associated with a significant increase in CTCF occupancy for 79% of these TFs (average of 24% increase in CTCF occupancy, range 12% to 41% increase) (**Fig. 6.M5D**). In contrast to the association of TF co-occupancy with chromatin accessibility and CTCF occupancy, TF co-occupancy was never

associated with a significant increase in CTCF affinity and was actually associated with a significant *decrease* in CTCF affinity for 50% of these TFs (average of 6% decrease in CTCF affinity, range 2% to 11% decrease) (**Fig. 6.M5D**). This overall decrease in CTCF affinity at co-occupied CTCF binding sites was further validated using *in vivo* DNaseI footprinting (**Fig. 6.S5A**).

To further investigate the association of TF co-occupancy with CTCF occupancy and affinity we next studied CTCF binding sites which were presumed to be co-occupied by two or more CTCF proteins in parallel (**Fig. 6.S5B**). Consistent with our above findings, CTCF binding sites containing two or more CTCF binding elements have a significantly higher chromatin accessibility (**Fig. 6.S5C**) and CTCF occupancy (**Fig. 6.S5D**), but have no significant difference in CTCF affinity (**Fig. 6.S5E**) when compared to CTCF binding sites with only a single CTCF binding element.

To directly test the effect of co-binding TFs on CTCF occupancy, we determined whether single nucleotide variants (SNVs) neighboring CTCF binding elements have a significant effect on CTCF occupancy at that binding element (**Fig. 6.S5F**) (Maurano et al., 2012b). Since these neighboring SNVs do not alter the CTCF-DNA interface, it is presumed that they are affecting binding elements for co-binding TFs. Overall, 29% of all SNVs that have a significant effect on CTCF occupancy (FDR 5%) are located outside of the CTCF binding element (**Fig. 6.S5G**). Furthermore, these neighboring SNVs tend to be located at CTCF binding sites that are clearly co-occupied by multiple TFs (**Fig. 6.S5H**). Together, these data indicate that genetic variants that affect co-bound TFs can have a significant effect on CTCF occupancy.

Overall, these data suggest that CTCF is able to act in a cooperative manner with neighboring TFs to increase both the accessibility of a CTCF binding element and the occupancy

of CTCF at that binding element. Of note, out of the 38 co-bound TFs studied above, CTCF is only known to form direct protein-protein interactions with YY1 (Donohoe et al., 2007). Consequently, the observed cooperativity is likely not caused by direct protein-protein interactions. In contrast, our data suggests that this cooperativity is mediated through maintaining more of an element within an accessible state, which enables more DNA templates to be bound by CTCF within a population of cells, even though the affinity of these co-occupied sites is not increased (**Fig. 6.M5D** and **5E**). Due to these characteristics, we term this prevailing model of CTCF cooperativity '*accessibility-mediated cooperativity*'.

CTCF affinity compensates for the chromatin environment of a binding site

Our finding that co-occupied CTCF binding sites tended to contain lower affinity CTCF binding elements (**Fig. 6.M5D**) led us to further investigate the association of chromatin environment with CTCF affinity and CTCF occupancy. Strikingly, we found that CTCF binding sites show a complex relationship between chromatin accessibility and CTCF affinity and occupancy (**Fig. 6.M6A**). CTCF binding sites with low to medium chromatin accessibility show a positive association between chromatin accessibility and both CTCF affinity and CTCF occupancy (**Fig. 6.M6A**), and have few co-occupying TFs (**Fig. 6.M6B**), suggesting that the chromatin accessibility at these binding sites is dependent on the affinity of CTCF (affinity-dependent CTCF elements). However, at CTCF sites of medium to high chromatin accessibility, CTCF occupancy appears largely invariant to chromatin accessibility, and chromatin accessibility is *negatively* associated with CTCF affinity (**Fig. 6.M6A**). Furthermore, these binding sites tend to be co-occupied by multiple TFs (**Fig. 6.M6B**), suggesting that the chromatin accessibility at these binding sites may be dependent not on CTCF, but on its co-

binding TFs, and that the chromatin accessibility at these binding sites may somehow limit the affinity of CTCF at these sites (accessibility-limited CTCF elements).

Since promoters are among the most accessible chromatin environments (Thurman et al., 2012), we sought to understand the relationship between the overlap of a CTCF binding site with a promoter and the occupancy and affinity of CTCF at that site (**Fig. 6.M6C**). As expected, the chromatin environment at promoter CTCF binding sites is highly accessible and contains multiple co-bound TFs that appear to bind in parallel to CTCF along the genome (**Figures 6D** and **6E**). Despite the elevated chromatin accessibility at promoter CTCF binding sites, these sites have an identical CTCF occupancy as their poorly accessible distal counterparts (**Fig. 6.M6E**). This appears to result from the fact that promoter CTCF binding sites utilize lower affinity CTCF binding elements than distal CTCF binding sites (**Fig. 6.M6E** and **Fig. 6.S6A**). These findings imply that some type of pressure exists to 'correct' the affinity of CTCF binding elements genome-wide such that the occupancy of CTCF at both highly and less accessible binding sites is maintained at a similar level.

To test whether this pressure might be evolutionary in nature, we analyzed the association of chromatin accessibility and CTCF binding element conservation. Analogous to CTCF affinity, CTCF binding element conservation appeared to be maximal at CTCF binding sites of moderate chromatin accessibility, and was significantly depressed at both poorly and highly accessible CTCF binding sites (**Fig. 6.M6F**). Similarly, CTCF binding elements containing co-bound TFs were significantly less evolutionarily constrained than those lacking co-bound TFs (**Fig. 6.M6G**). Furthermore, although promoters are typically considered to contain highly conserved DNA binding elements (Xie et al., 2005), CTCF binding elements proximal to promoters were significantly less evolutionarily constrained than those distal to promoters (**Fig. 6.M6H** and **Fig.**

6.S6B). Together, these findings reveal that: (i) CTCF occupancy at a binding element is modeled by both the affinity of CTCF towards that binding element, as well as that element's chromatin environment; (ii) highly accessible binding sites utilize lower affinity CTCF binding elements to maintain a similar level of CTCF occupancy as their less accessible peers (**Fig. 6.M6I**); and (iii) Independent of CTCF occupancy, CTCF binding elements within highly accessible chromatin environments are under less evolutionary constraint due to their use of low affinity CTCF binding elements.

Low-affinity CTCF binding elements enable cell-selective CTCF regulation

Although the binding landscape of CTCF was initially considered to be largely invariant (Kim et al., 2007; Cuddapah et al., 2009; Heintzman et al., 2009), recent studies have demonstrated that a large population of CTCF binding sites are in fact regulated across different cell types (Wang et al., 2012), and that these regulated CTCF binding elements play a critical role in cell-selective gene regulation (Sekimata et al., 2009; Monahan et al., 2012). Consequently, we sought to understand whether the affinity of a CTCF binding site may contribute to its ability to become regulated across different cell types. To accomplish this, we compared the occupancy of CTCF across 19 diverse human cell types and identified a set of 3,126 K562 CTCF binding sites that appear to be regulated across cell types (**Fig. 6.M7A** and **7B**). Strikingly, low affinity CTCF binding sites were over 17 times more likely than high affinity CTCF binding sites to be regulated, with ~40% of all low affinity CTCF binding sites regulated across these 19 cell types (**Fig. 6.M7C**).

Overall, regulated CTCF binding sites were significantly lower affinity as opposed to stable CTCF binding sites (**Fig. 6.M7D**), as is seen at the protocadherin-gamma (PCDH-

gamma) locus where regulated CTCF elements are stochastically utilized within a population of cells to determine which starting exon is used for the PCDH-gamma gene (**Fig. 6.M7E**) (Monahan et al., 2012). Furthermore, regulated CTCF binding sites had significantly more co-binding TFs than stable CTCF binding sites (**Fig. 6.M7F**), indicating that co-binding TFs may be contributing to cell-selective changes in chromatin accessibility, and consequently CTCF occupancy, at these low-affinity regulated CTCF binding sites, as is seen at one of the interferon gamma enhancers (**Fig. 6.S7**) (Sekimata et al., 2009).

Of note, despite the critical role regulated CTCF binding elements play in gene regulation, these regulated binding elements are under significantly less evolutionary sequence constraint than stable CTCF binding elements (**Fig. 6.M7G**). However, the lower evolutionary sequence constraint at these elements most likely results from less selective pressure to maintain base-rich high-affinity binding sites, rather than a decrease in functional constraint at these sites.

6.4 – DISCUSSION

By studying the interplay between TF affinity, TF occupancy and chromatin environment we have made the following observations: First, salted ChIP-seq provides a robust and accurate tool for measuring the genome-wide affinity of a natively expressed TF, and application of salted ChIP-seq to the human master regulator CTCF revealed a previously unrecognized rich diversity in its genome-wide affinity landscape. Second, both DNA base-specific and DNA structure-specific CTCF-DNA interactions contribute to the affinity of CTCF at a binding element *in vivo*. Furthermore, CTCF utilizes its full repertoire of 11 zinc fingers when interacting with a binding element, independent of the perceived sequence-specific interactions. Third, CTCF affinity, not occupancy, is driving the evolutionary sequence constraint at CTCF binding elements. Fourth,

CTCF and other TFs are involved in extensive 'accessibility-mediated cooperativity,' which enables the modulation of CTCF occupancy at a binding site through the increase in chromatin accessibility associated with having multiple factors bound at that site. Fifth, the affinity of CTCF binding elements genome-wide have evolved such that the occupancy of CTCF at both highly and less accessible binding sites is maintained at a similar level. Finally, low affinity CTCF binding elements, which appear to be under significantly less evolutionary sequence constraint, play a fundamentally important role in cell-selective gene regulation, as high affinity sites largely lack the ability to be regulated across different cell types.

We interpret the above findings to indicate that chromatin environment plays a principal role in driving the evolutionary constraint of regulatory elements genome-wide. Below we place our findings in a broader perspective.

Determinants of CTCF affinity

We find that the affinity of CTCF at a particular binding element is largely determined by; (i) the binding mode CTCF utilizes when engaging the element: (ii) the base composition of the binding element; and (iii) the DNA structure of the binding element.

First, depending on the DNA sequence at a CTCF binding element, CTCF can engage DNA using one of three biophysically distinct binding modes (Stergachis et al., *in preparation*). We find that the choice of which binding mode CTCF utilizes can have a large impact on the affinity of CTCF at that binding element (**Fig. 6.M3**). This suggests that the 'multivalent' nature of CTCF (Filippova et al., 1996; Ohlsson et al., 2001) may largely function to modulate the affinity of CTCF at different binding sites along the genome.

Second, the base composition at a CTCF binding element plays a critical role in determining the affinity of CTCF at that element. However, closer inspection of the base preferences at low-affinity and high-affinity CTCF binding elements reveals that certain bases within the CTCF binding element appear to be critical for the genesis of a CTCF binding element, whereas other bases appear to play a larger role in modulating the affinity of CTCF at a binding element (**Fig. 6.M3** and **Fig. 6.S3**). Consequently, these results show strong dependencies between the different positions within the CTCF binding element, and provide a possible explanation for why a positional weight matrix (PWM) cannot properly describe the affinity of CTCF at its binding elements genome-wide (**Fig. 6.M3** and **Fig. 6.S3**).

Third, all three CTCF binding modes exhibit minor groove width shapes that are preferentially bound by high affinity binding elements and are located outside of the binding region involved in base-specific CTCF-DNA contacts (**Fig. 6.M3H** and **Fig. 6.S3F**). For example, unlike the CTCF_L and CTCF_XL binding modes, when CTCF engages DNA using the CTCF_M binding mode it does not utilize ZFs 8-11 for base-dependent interactions. However, when using the CTCF_M binding mode, CTCF does utilize ZFs 8-11 for DNA structure-dependent interactions, indicating that CTCF utilizes its full repertoire of 11 zinc fingers when engaging DNA in each of its three binding modes, independent of the perceived sequence-specific interactions.

Of note, due to the substantial length of the CTCF binding element (44-45 bases in length) these affinity-based findings could not have been observed using a standard *in vitro* affinity mapping method such as protein binding microarrays (PBMs) or SELEX-seq (Berger et al., 2006; Jolma et al., 2013). Since these *in vitro* methods typically start from a random pool of oligos, these methods would have required 45 bases of DNA to be independently varied across

different probes, a task that would require over 2,000 moles of dsDNA to represent each unique combination only once.

Accessibility-mediated cooperativity

In stark contrast to what we expected, the affinity of CTCF at a binding site poorly predicted the occupancy of CTCF at that site (**Fig. 6.M5A**). This appears to result from the finding that the chromatin environment of a CTCF binding element can greatly affect the occupancy of CTCF at that site, independent of the affinity of CTCF at that site. Specifically, we observed that when multiple TFs co-occupy a CTCF binding site, these TFs can act in a cooperative manner to increase the proportion of that DNA template that is contained within an accessible state. This increase in the accessibility of a CTCF binding element directly results in an increase in the occupancy of CTCF at that binding element simply through mass action. Due to these characteristics, we term this prevailing model of CTCF cooperativity '*accessibility-mediated cooperativity*'. Of note, this form of cooperativity does not necessitate any direct protein-protein interactions between the cooperative TFs, rather, cooperativity is mediated through the effect co-binding TFs have on chromatin accessibility. In agreement with this model, of the 38 TFs with which CTCF appears to act cooperatively, CTCF only forms direct protein-protein interactions with YY1 (Donohoe et al., 2007).

This form of noninteracting TF cooperativity, or indirect cooperativity, was recently proposed in a theory paper by Leonid Mirny (Mirny, 2010). Specifically, Mirny proposed that cooperativity between noninteracting TFs can be achieved through their competition with nucleosomes. In this model, the presence of multiple neighboring TF binding elements within a regulatory element increases the frequency with which that regulatory element is bound by at

least one of the TFs, and consequently, reduces the chances that a competing nucleosome will occupy that DNA template. Using a similar framework, Voss et al. have proposed that TFs bound at the same regulatory element can undergo noninteracting cooperativity by increasing the chromatin accessibility at that binding site (Voss et al., 2011), and this model has been used to explain the ability of AP1 to potentiate glucocorticoid receptor occupancy (Biddie et al., 2011). When taken together with these previous studies, our findings suggest that accessibility-mediated cooperativity is fundamental force driving TF occupancy genome-wide.

Accessibility mediated cooperativity provides a mechanistic understanding for why the sequence features of a binding element have historically been a poor predictor of the occupancy at that binding element (**Fig. 6.M3D-E**) (Carr and Biggin, 1999; Yang et al., 2006; Li et al., 2008; Cuellar-Partida et al., 2012), and why integrating binding site sequence information with chromatin accessibility does a superior job at predicting TF ChIP-seq signals (Kaplan and Biggin, 2012; Cuellar-Partida et al., 2012). Mainly, although the sequence features are largely deterministic of the affinity of a TF at a binding site (**Fig. 6.M3**), the occupancy of a TF at that site is extensively modulated by that site's chromatin environment (**Fig. 6.M5**). In addition, accessibility mediated cooperativity provides a mechanistic understanding for why a significant proportion of SNPs that affect the occupancy of a TF are often located outside of the TF binding element (McDaniell et al., 2010; Kasowski et al., 2010; Maurano et al., 2012b).

Furthermore, accessibility mediated cooperativity provides a mechanism whereby constitutively expressed TFs may be regulated in a cell-selective manner. For example, co-binding TFs appear to regulate the occupancy of low affinity CTCF binding elements across different cell types (**Fig. 6.M7**), as appears to be the case at an interferon gamma enhancer (Sekimata et al., 2009).

Chromatin environment drives regulatory element evolutionary constraint

Our results indicate that the evolutionary sequence constraint at a binding element is the byproduct of at least three driving pressures: (i) a pressure to create a binding element at that site from random background sequence; (ii) a pressure to maintain that binding element at a certain affinity level given its chromatin environment; and (iii) a pressure to maintain that binding element at a certain occupancy level given its functional importance. Strikingly, the evolutionary constraint to maintain a binding element at a certain affinity level appears dominant over the evolutionary constraint to maintain an element at a certain occupancy level (**Fig. 6.M4** and **Fig. 6.S4**). Since the affinity of CTCF at a binding element is largely dictated by the chromatin environment of that site (**Fig. 6.M6**), these findings indicates that chromatin environment plays a dominant role in modeling the evolutionary sequence constraint of a binding element.

The strong association between evolutionary sequence constraint and CTCF affinity likely results from the fact that DNA sequences with low-affinity binding elements can accommodate more evolutionary 'wiggle room' than sequences with high affinity binding elements, since high affinity binding elements rely on more base-specific protein-DNA interactions (**Fig. 6.M3**). These findings are in stark contrast to evolutionary pressures in coding sequence, where the evolutionary sequence constraint of a codon is thought to be predominantly dependent upon its functional importance, not its genomic environment.

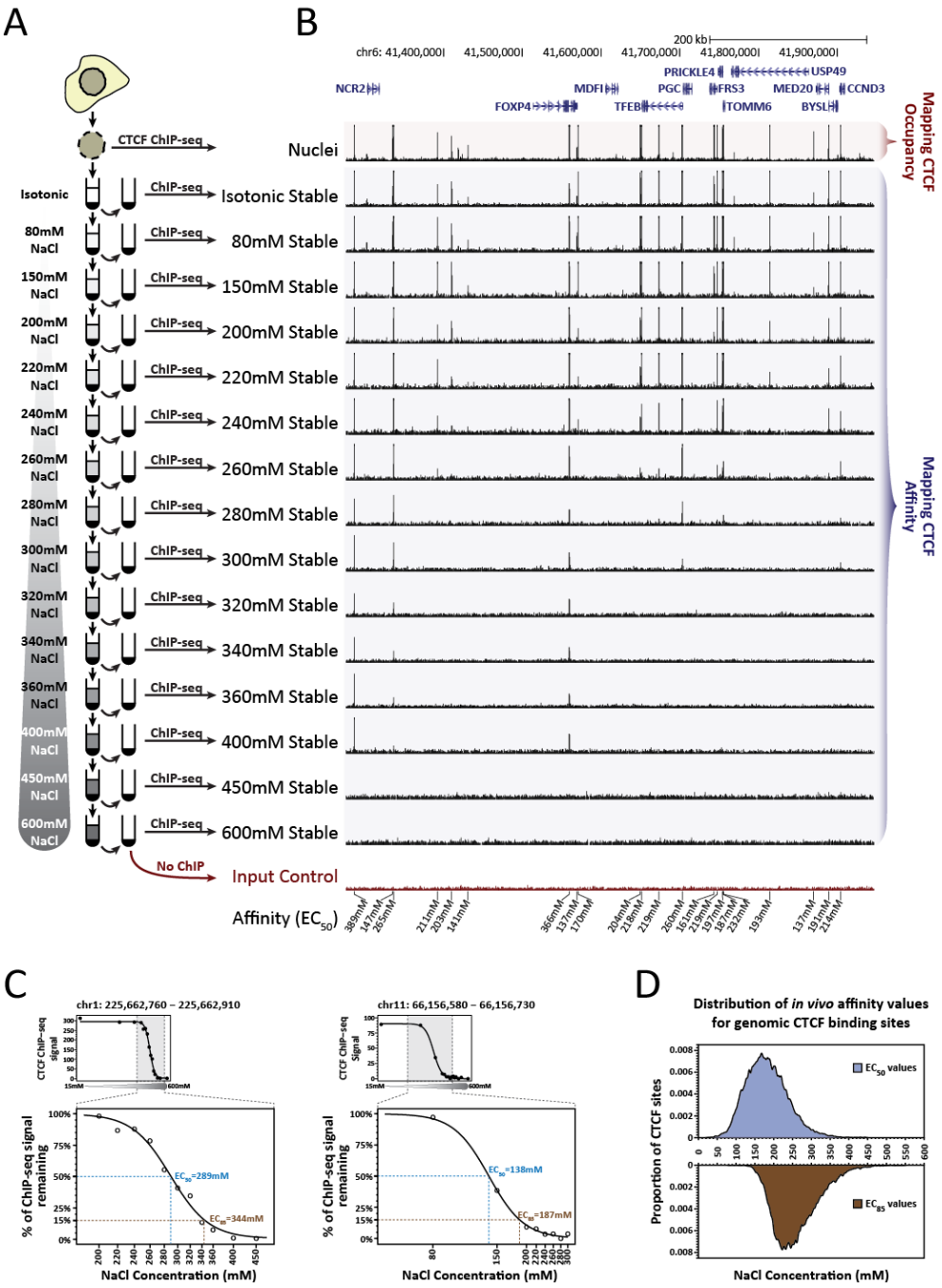
The dependence of a regulatory element's evolutionary constraint on its chromatin environment is most arresting at regulated CTCF binding elements. Since CTCF is a constitutively expressed TF, cross cell-type regulation of a CTCF binding site appears to require the utilization of a low affinity CTCF binding element (**Fig. 6.M7**). Our findings suggest that

these low affinity CTCF binding elements enable cross cell-type regulation through either: (i) stochastic activation of binding elements, as is the case with the PCDH-gamma locus (**Fig. 6.M7E**) (Monahan et al., 2012); or (ii) through accessibility-mediated cooperativity with co-binding TFs, as is the case at one of the interferon gamma enhancers (**Fig. 6.S7**) (Sekimata et al., 2009). However, due to the presence of low affinity CTCF binding elements at these regulated CTCF binding sites, these functionally important regulated CTCF binding elements are inherently under significantly less evolutionary constraint than their stable counterparts.

In summary, our results provide a genome-wide depiction of the *in vivo* affinity landscape of a transcription factor and reveal that the affinity of TF binding elements genome-wide have evolved to match their chromatin environment. These findings support the conclusion that evolutionary sequence constraint within non-coding regulatory elements is predominantly dependent on the chromatin environment at that binding site.

6.5 – FIGURES

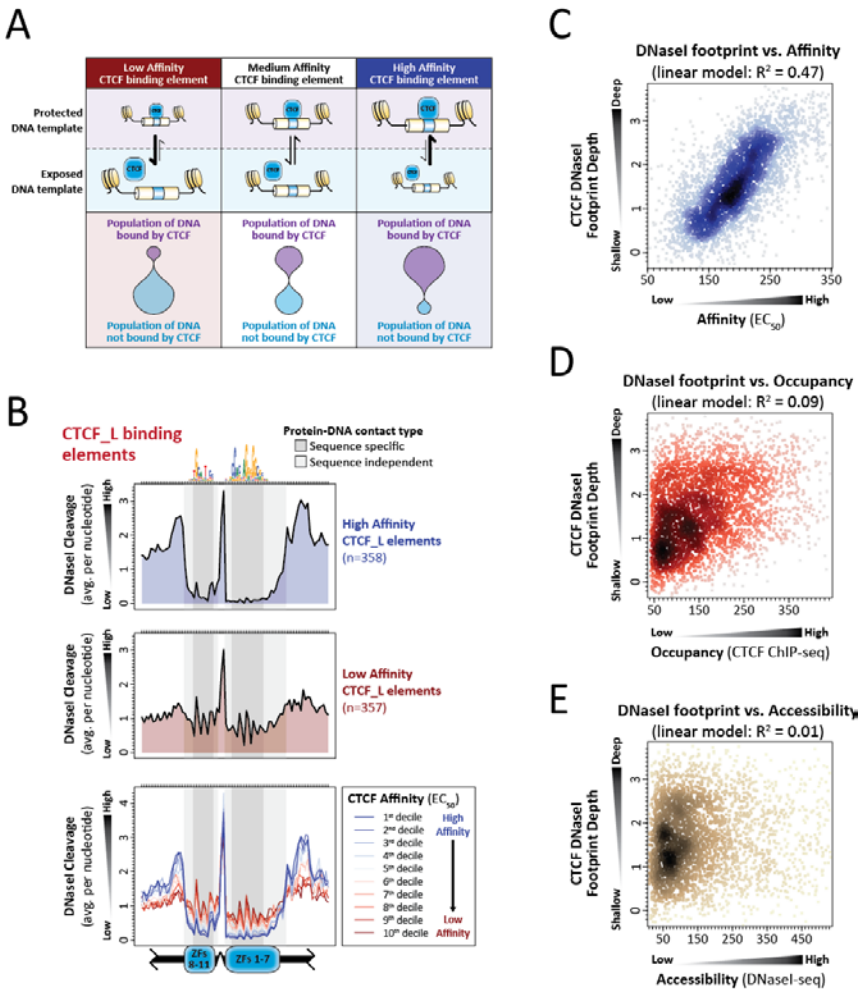
Figure 6.M1. Genome-wide in vivo affinity map of CTCF



(A) *Salted-ChIPs expose the relative affinity of each genomic binding site for a transcription factor.* To perform Salted-ChIPs, nuclei are treated with successively higher concentrations of salt (NaCl) and the protein-DNA interactions stable at these different salt concentrations are interrogated using ChIP-seq.

(B-D) *CTCF genomic binding sites display a diverse stability to salt.* (B) A representative ~600kb genomic locus displaying CTCF binding sites that are stable after treatment with different concentrations of salt. (C) The concentration of salt needed to disrupt a given genomic CTCF binding site was calculated using a 4-parameter log-logistic model fit using the CTCF ChIP-seq intensity at that site after extraction with different salt concentrations. EC50 values are used as a proxy of affinity, and correspond to the concentration of salt needed to disrupt the CTCF ChIP-seq signal at a site by 50%. (D) The distribution of EC50 values (top) and EC85 values (bottom) at each of the 21,015 CTCF binding sites interrogated in this study. Note that the affinity of CTCF at different genomic binding elements widely differs.

Figure 6.M2. Validation of CTCF salted-ChIP affinity measurements using genome-wide in vivo DNaseI footprinting



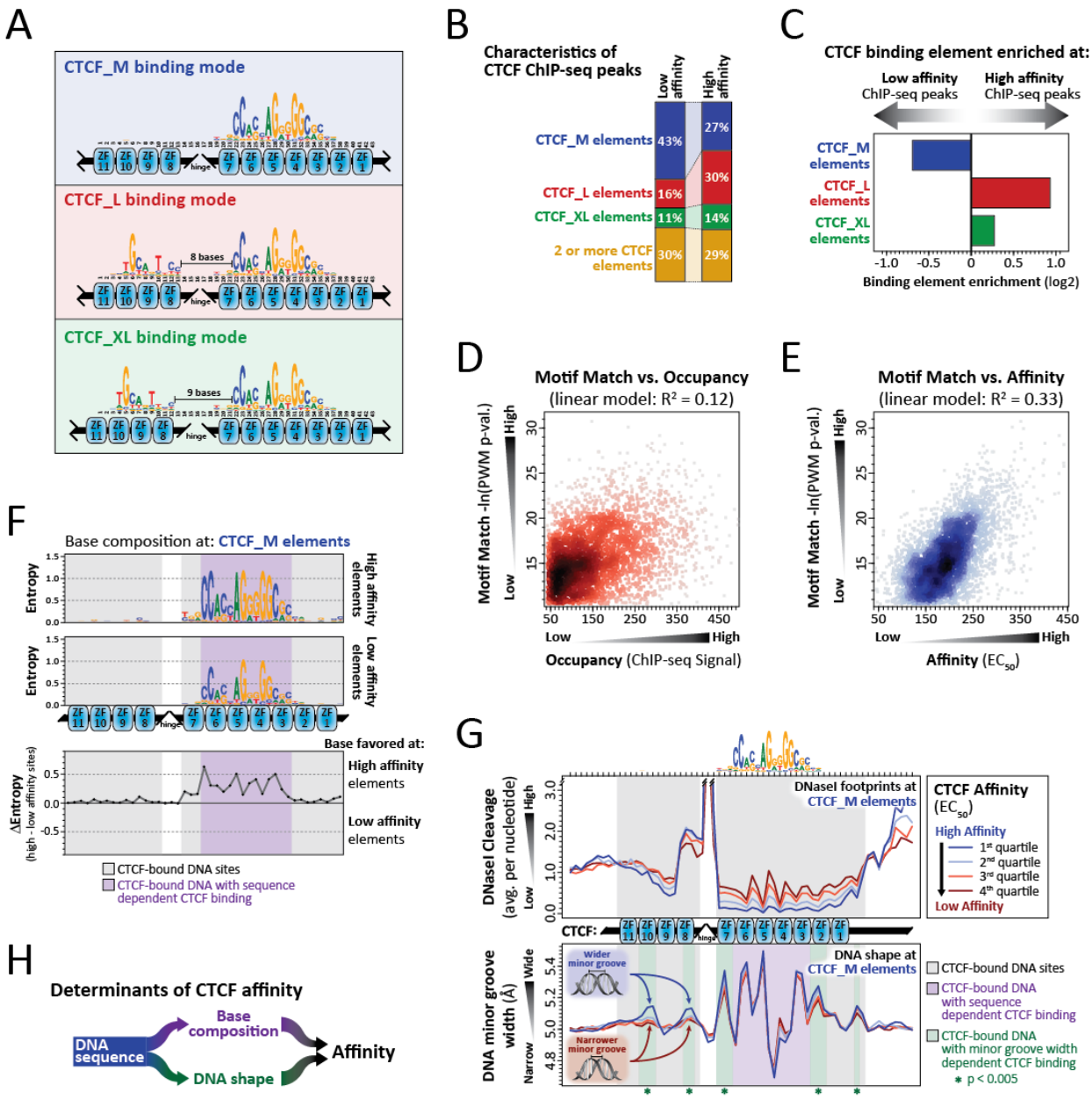
(A) Model of the relationship between the affinity of a CTCF binding element and the proportion of that binding element that is bound by CTCF. Note that low affinity CTCF binding elements should have a relatively larger population of unbound DNA template relative to high affinity CTCF binding elements.

(B-C) *CTCF binding site affinity mirrors DNaseI footprint depth.* (B) Shown are per-nucleotide DNaseI cleavage rates for CTCF binding elements identified as high affinity (top) low affinity

(middle) or intermediate affinity (bottom). Note that nucleotides interacting with CTCF's zinc finger domains have a much higher cleavage rate at low affinity CTCF binding elements indicative of a larger population of unbound DNA template at low-affinity CTCF binding elements. (C) Shown is the relationship between the depth of CTCF's DNaseI footprint at a given binding element and the affinity of CTCF at that element.

(D-E) *CTCF binding site occupancy and chromatin accessibility do not mirror CTCF DNaseI footprint depth.* Shown is the relationship between the depth of CTCF's DNaseI footprint at a given binding element and the occupancy of CTCF (D) or chromatin accessibility (E) at that element.

Figure 6.M3. In vivo CTCF affinity is modeled by both base-specific and structure-specific protein-DNA interactions.



(A-C) *High affinity CTCF binding sites preferentially utilize extended, non-canonical binding elements.* (A) Shown are the sequence preferences for the medium (CTCF_M) long (CTCF_L)

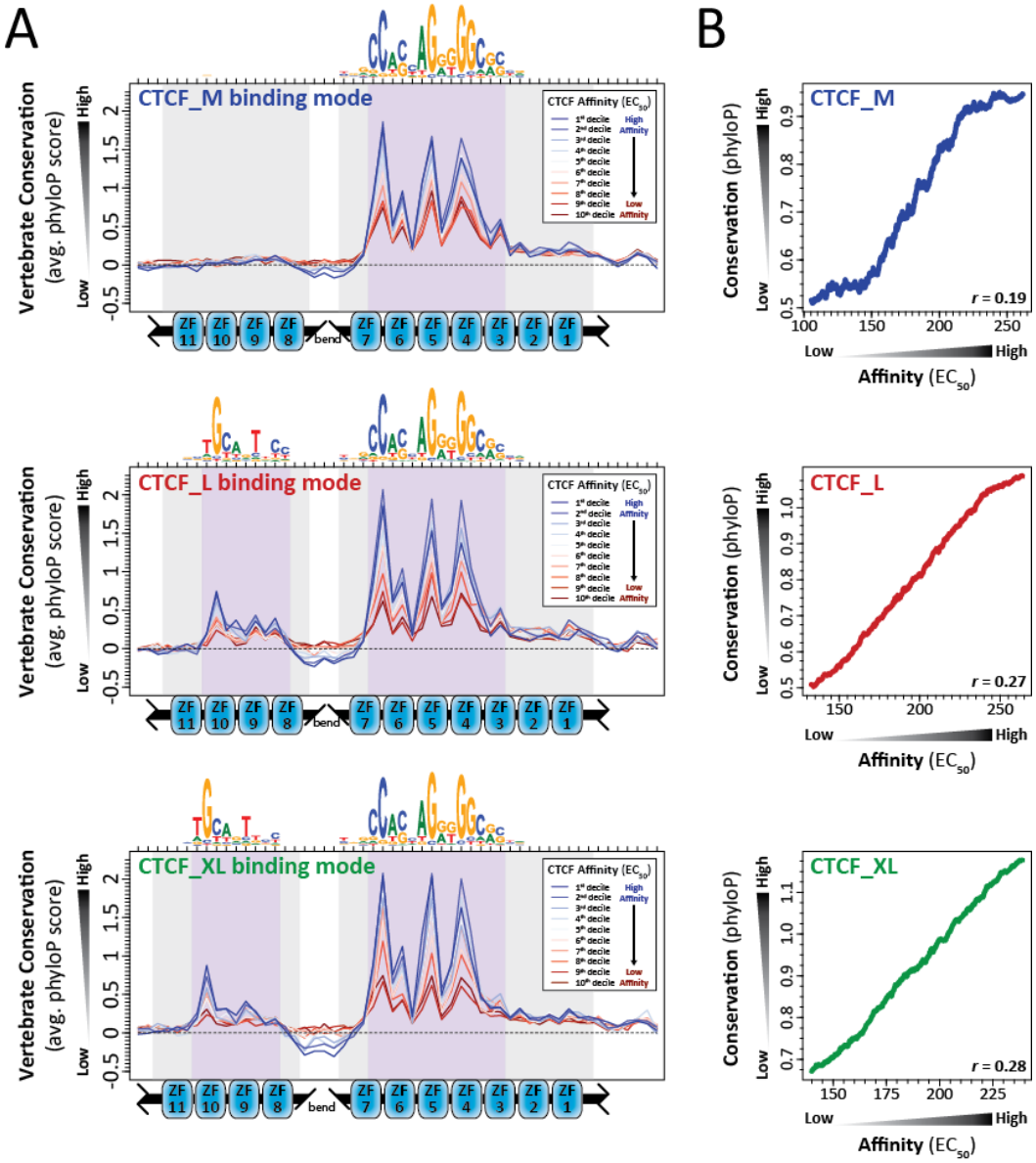
and extra-long (CTCF_XL) CTCF binding modes, as well as the specific zinc fingers utilized at each position along the motif. (B) The distribution of these three CTCF binding elements across low affinity (top) and high affinity (bottom) CTCF binding sites. (C) The enrichment of these three CTCF binding elements at high affinity vs. low affinity binding sites.

(D-F) *High affinity CTCF binding events utilize more DNA sequence-specific protein-DNA interactions.* Shown is the relationship at CTCF_L binding elements between the positional weight matrix (PWM) motif match and CTCF occupancy (D) and CTCF affinity (E). (F) (top) Sequence information at high affinity and low affinity CTCF_M binding elements. (bottom) The difference in sequence information between high affinity and low affinity CTCF_M binding elements. Values above zero indicate bases that are favored at high affinity sites. Note that certain bases are disproportionately favored at high affinity CTCF binding elements.

(G) *High affinity CTCF binding events utilize distinct DNA structure-specific protein-DNA interactions.* (top) Shown are per-nucleotide DNaseI cleavage rates at CTCF_M binding elements identified as having different strengths of *in vivo* CTCF affinity. CTCF-bound DNA bases are highlighted in grey. (bottom) Shown are the predicted DNA minor groove widths at CTCF_M binding elements identified as having different strengths of *in vivo* CTCF affinity. DNA bases that are bound by CTCF in a base-dependent manner are highlighted in purple. Base-independent CTCF-bound nucleotides that significantly differ ($p<0.005$) in minor groove width at high- and low-affinity binding elements are highlighted in green.

(H) Model showing the contribution of DNA structure and base composition to CTCF affinity

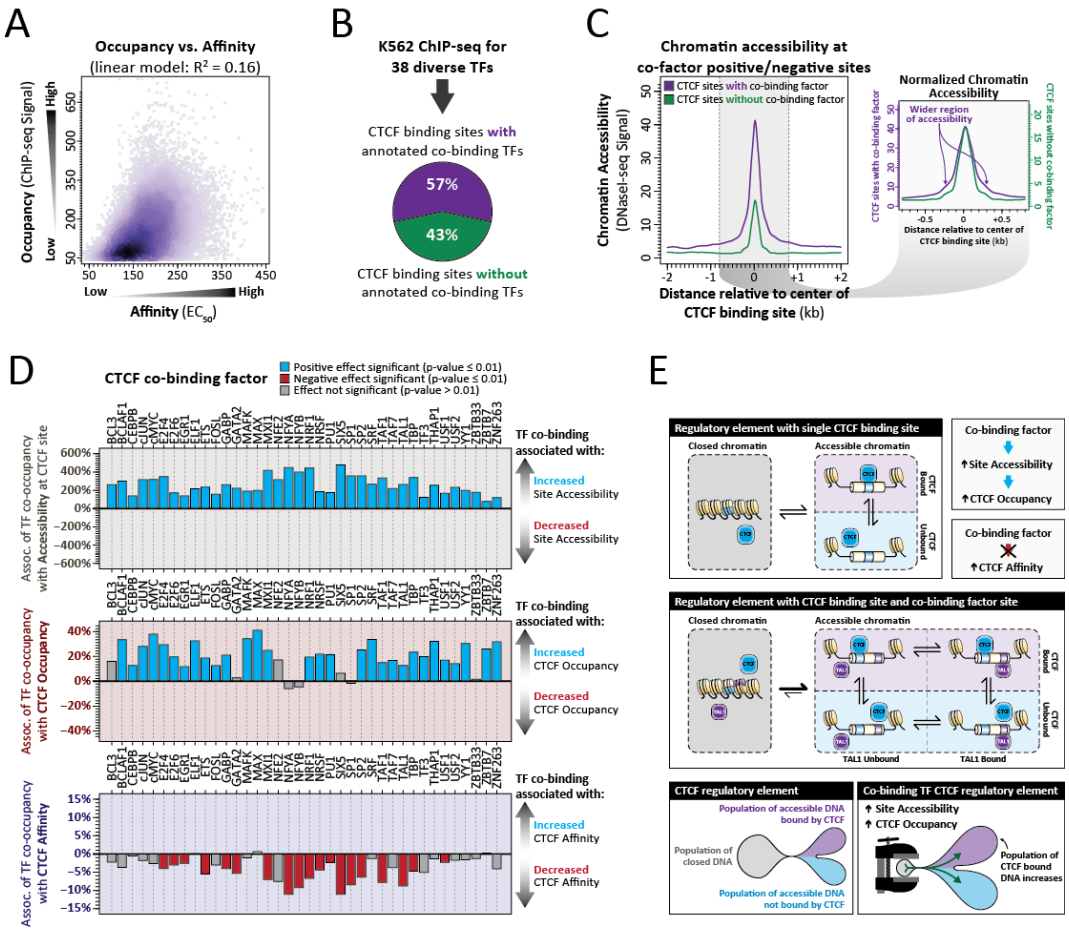
Figure 6.M4. The evolutionary constraint of a CTCF binding element is largely dependent on the affinity of CTCF at the binding element.



(A) Shown are per-nucleotide average vertebrate conservation values for (top) CTCF_M, (middle) CTCF_L and (bottom) CTCF_XL binding elements identified as having different strengths of *in vivo* CTCF affinity.

(B) Shown is a running average plot displaying the association of CTCF affinity and binding element conservation at all non-coding (top) CTCF_M, (middle) CTCF_L and (bottom) CTCF_XL binding elements genome-wide. Pearson correlation value is displayed in the bottom right hand corner.

Figure 6.M5. CTCF engages in accessibility-mediated cooperativity with other transcription factors.



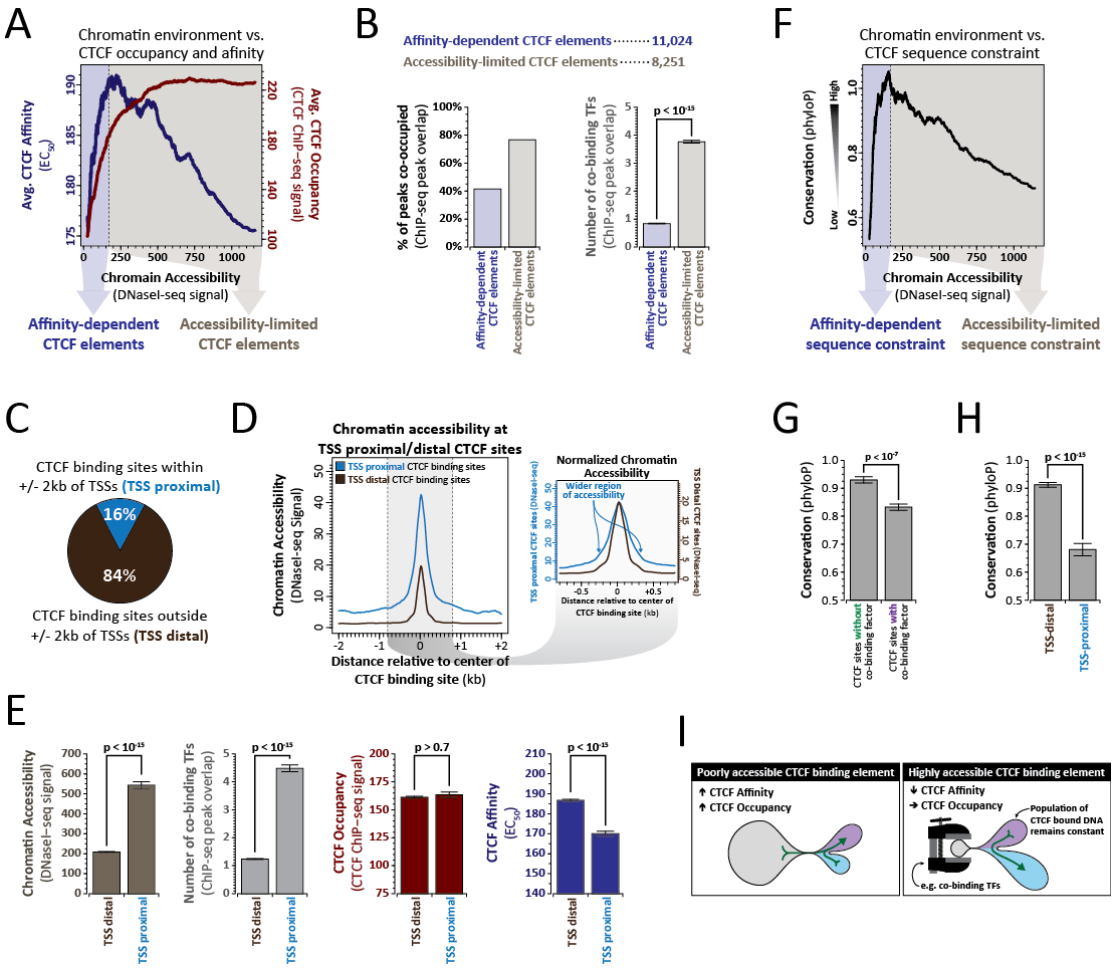
(A) *The affinity of CTCF at a binding site poorly predicts the occupancy of CTCF at that site.* Shown is the relationship between CTCF occupancy and CTCF affinity genome-wide.

(B-D) *Co-binding of CTCF with other factors increases chromatin accessibility and CTCF occupancy, but not CTCF affinity.* (B) Shown is the percentage of CTCF binding sites that overlap a ChIP-seq peak for at least one other transcription factor within K562 cells. (C) Profile plots showing the chromatin accessibility at CTCF binding sites with (purple) and without

(green) co-binding factors. (inset) Note that CTCF regulatory elements with co-binding factors typically are contained within a broader region of accessibility. (D) Shown is the association of a given CTCF co-binding factor with the chromatin accessibility (top) CTCF occupancy (middle) and CTCF affinity (bottom) at CTCF binding sites. Co-binding factors that have a significantly positive effect ($p<0.01$) on these chromatin features are colored in blue, those that have a significantly negative effect are colored in red ($p<0.01$) and those that do not have a significant effect are colored in grey. P-values were calculated based on a linear model fit of the data between co-binding factor positive and co-binding factor negative sites.

(E) Model of the contribution of co-binding factors to chromatin accessibility and CTCF occupancy.

Figure 6.M6. CTCF affinity compensates for the chromatin environment of a binding site



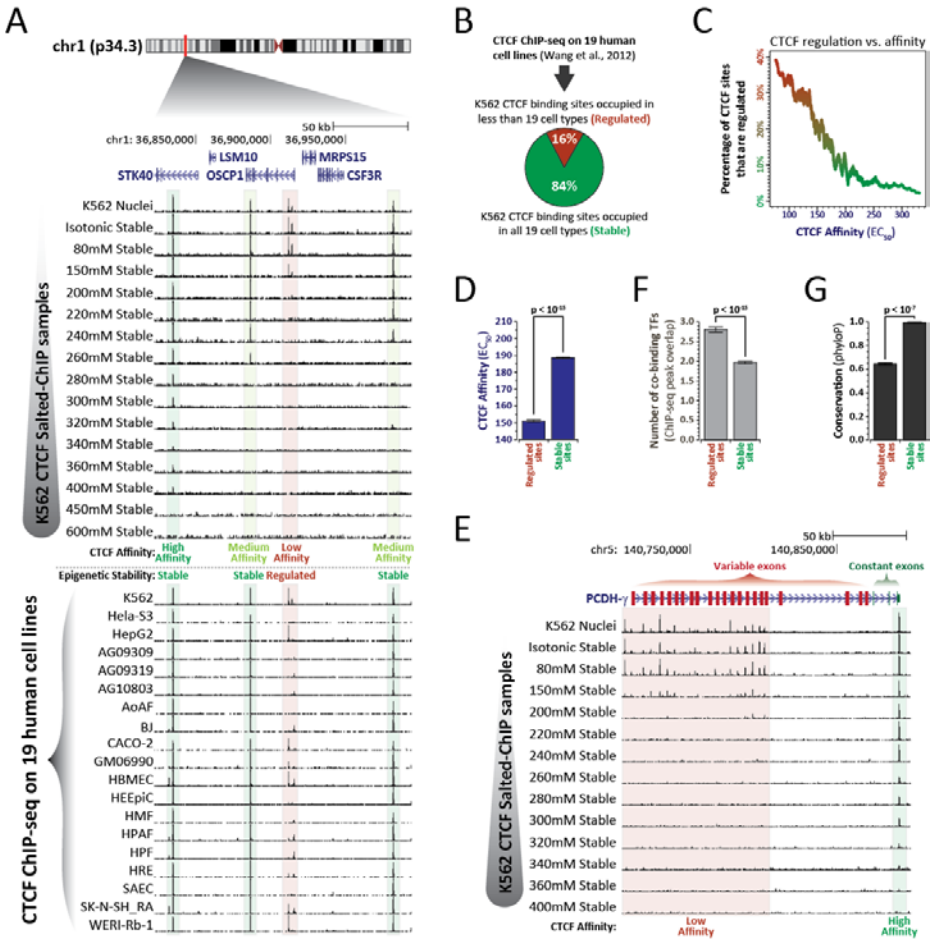
(A-B) *CTCF affinity compensates for chromatin accessibility.* (A) The relationship between chromatin accessibility and CTCF affinity (blue) and CTCF occupancy (red). (B) (top) The number of affinity-dependent and accessibility-limited CTCF elements. (bottom-left) The proportion of affinity-dependent and accessibility-limited CTCF elements that are co-occupied by at least one other TF. (bottom-right) The number of TFs co-occupying affinity-dependent and accessibility-limited CTCF elements.

(C-E) *Distal CTCF binding sites require higher affinity CTCF binding elements to maintain the same amount of CTCF occupancy as promoter CTCF binding sites.* (C) Shown is the percentage of CTCF binding sites that are located within 2kb of an annotated transcriptional start site (TSS). (D) Profile plot showing the chromatin accessibility at CTCF binding sites proximal (blue) or distal (brown) to TSSs. (inset) Note that CTCF regulatory elements proximal to TSSs typically are contained within a broader region of chromatin accessibility. (E) Shown is the chromatin accessibility, number of ChIP-seq verified co-binding TFs, CTCF affinity and CTCF occupancy at CTCF binding sites proximal or distal to TSSs.

(F-H) *CTCF binding elements within highly accessible chromatin environments are under less evolutionary constraint.* (F) Shown is the relationship between chromatin accessibility and evolutionary conservation at CTCF binding elements. Note that CTCF binding elements with moderate accessibility show a positive correlation between chromatin accessibility and conservation, whereas highly accessible CTCF binding elements show a negative correlation between chromatin accessibility and conservation. Shown is the evolutionary conservation at (G) CTCF binding element with and without co-binding TFs as measured using ChIP-seq as well as (H) TSS-distal and TSS-proximal binding elements.

(I) Model for the response of CTCF affinity to highly and poorly accessible chromatin environments.

Figure 6.M7. Regulated CTCF binding sites utilize low-affinity, co-occupied CTCF binding elements.



(A-D) *Low affinity CTCF binding sites are often regulated across different cell types.* (A) A representative ~150kb genomic locus displaying (top) the salt-stability of five CTCF binding sites in K562 cells and (bottom) the occupancy of these CTCF binding sites across 19 human cell lines. (B) Shown is the percentage of CTCF binding sites that are stably occupied across all 19 cell types (green) or regulated in at least one of the cell types (red). (C) Shown is the relationship

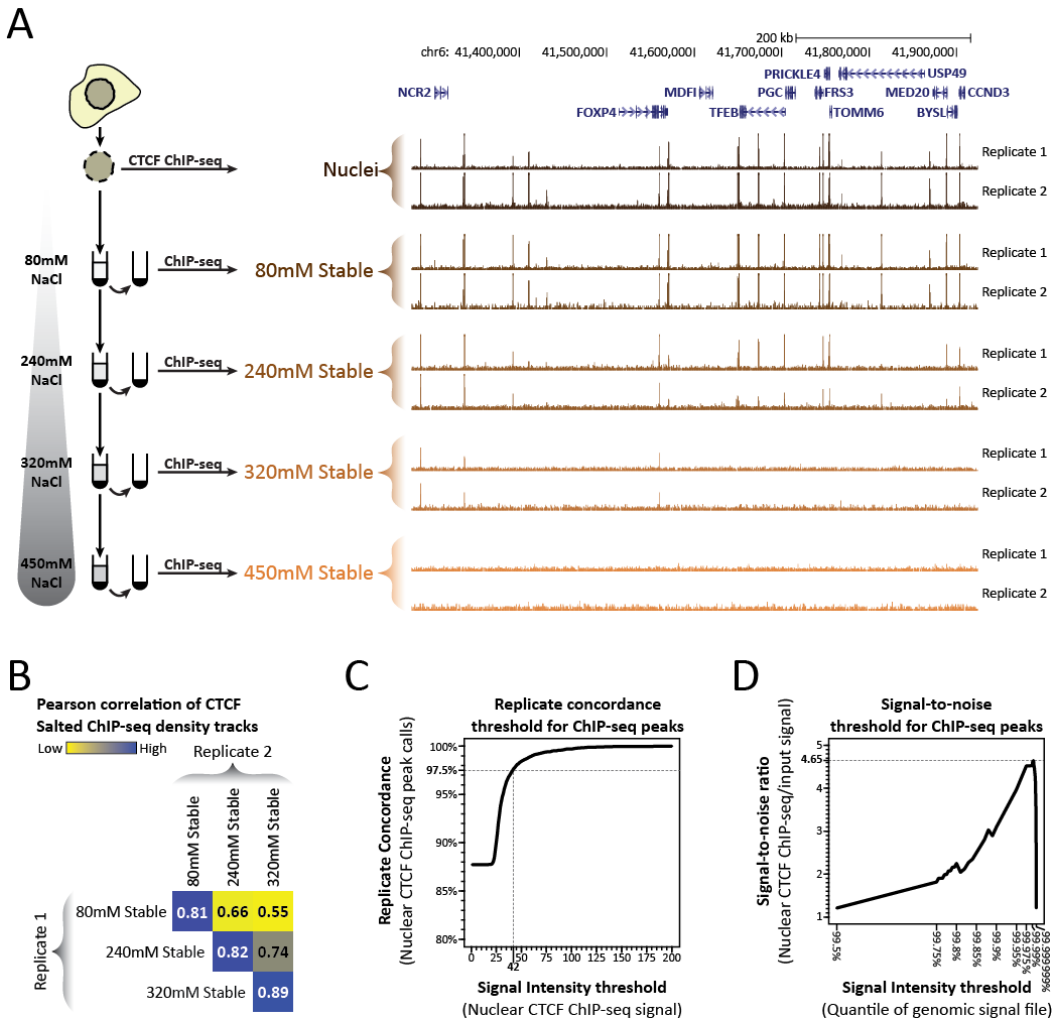
between the affinity of a CTCF site and the probability that it is regulated across these 19 cell types. (D) Shown is the CTCF affinity at regulated and stable CTCF binding elements.

(E) *The protocadherin-gamma gene utilizes low-affinity CTCF binding elements to regulate start codon usage.* Shown is the affinity of CTCF binding elements within the protocadherin-gamma locus. Note that CTCF binding sites near variable exons utilize low-affinity CTCF binding elements, whereas the CTCF binding site in the constant exon that these sites loop to utilizes a high-affinity CTCF binding element.

(F) *Regulated CTCF binding sites utilize co-occupied CTCF binding elements.* Shown is the number of ChIP-seq verified co-binding TFs at regulated and stable CTCF binding elements.

(G) *Regulated CTCF binding sites show diminished evolutionary sequence constraint.* Shown is evolutionary sequence constraint at regulated and stable CTCF binding elements.

Figure 6.S1. CTCF Salted-ChIP binding landscapes are highly reproducible

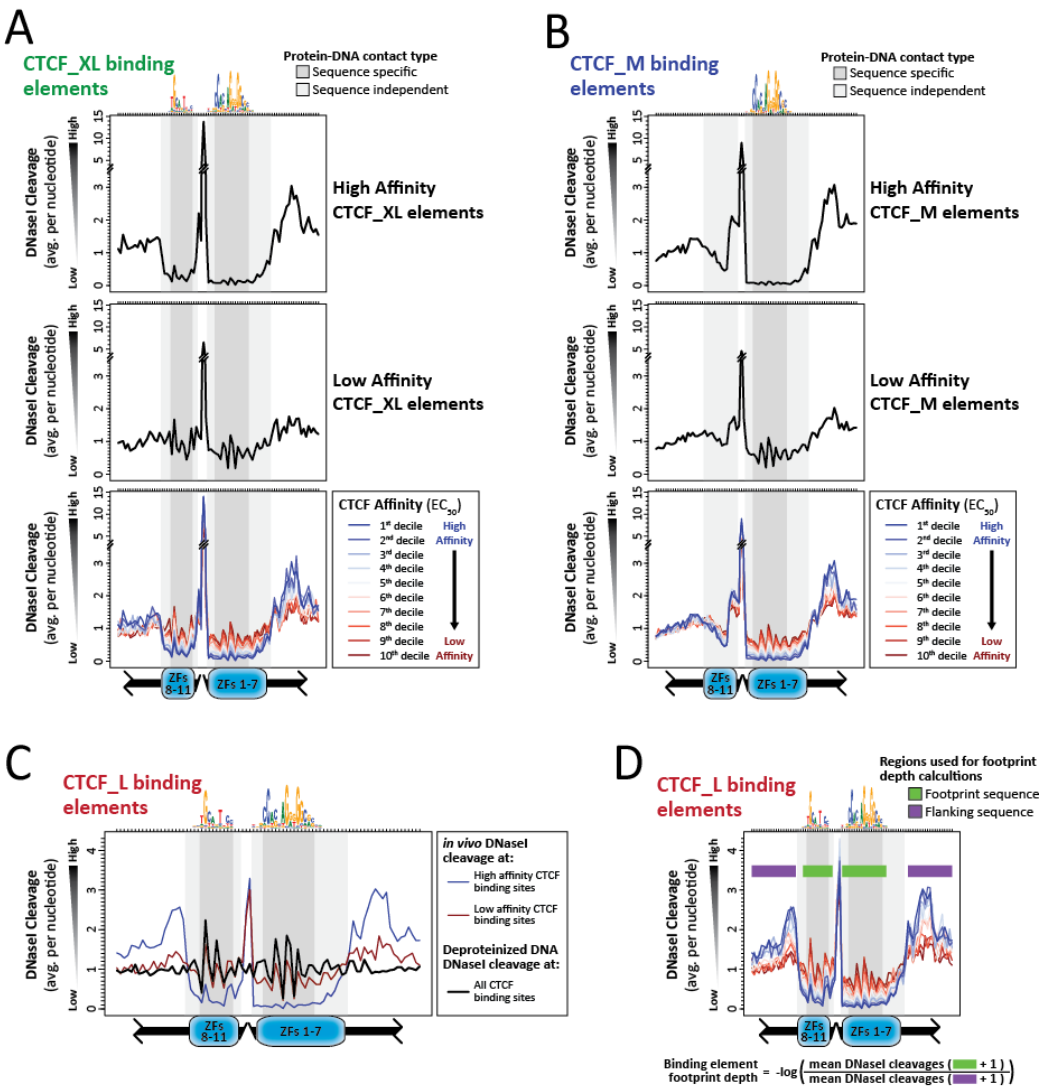


(A) A representative ~600kb genomic locus (same as in Figure 1B) displaying CTCF salted ChIP-seq profiles from two biological replicates that were each treated with four different concentrations of salt.

(B) Pearson correlation of the genome-wide CTCF salted ChIP-seq profiles from both biological replicates.

(C-D) *Thresholds used to identify CTCF binding elements.* (C) The replicate concordance of the two CTCT Nuclei ChIP-seq experiments a different signal intensity thresholds. A threshold of 42 (corresponding to 97.5% replicate concordance) was used to generate a final list of CTCF binding elements. (D) The ratio of CTCF nuclei ChIP-seq to input signal (signal-to-noise ratio) at different quantiles of the two datasets. A signal-to-noise threshold of 4.65 was used to generate a final list of CTCF binding elements that were significantly enriched over input signal.

Figure 6.S2. Validation of CTCF salted-ChIP affinity measurements using genome-wide in vivo DNaseI footprinting



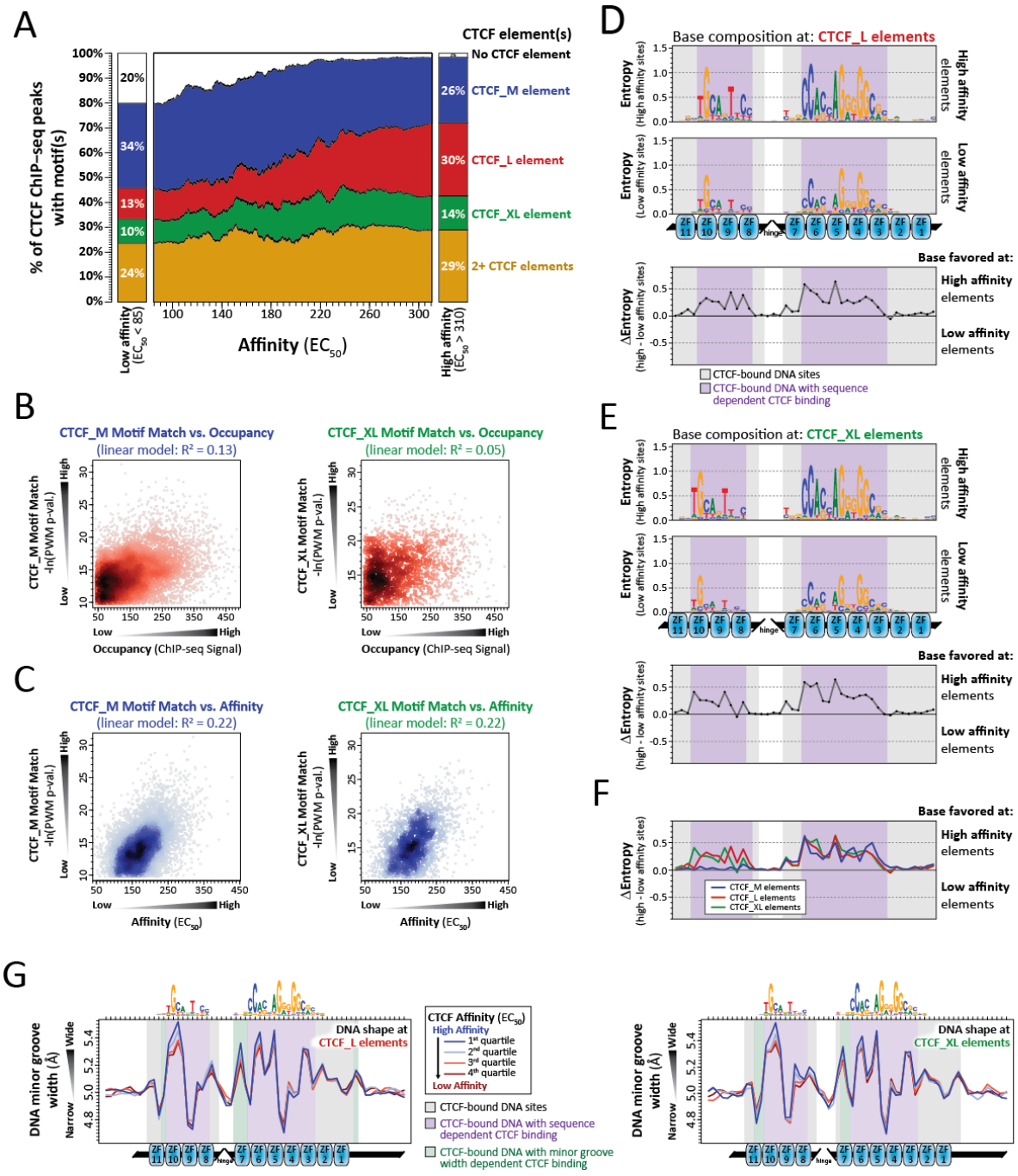
(A-B) Shown are per-nucleotide DNaseI cleavage rates for CTCF_XL (A) and CTCF_M (B) binding elements identified as high affinity (top) low affinity (middle) or intermediate affinity (bottom).

(C-D) *Calculation of CTCF footprint depth.* Shown are the footprinted and flanking sequences (C), as well as the formula (D) used to calculate footprint depth at a given locus.

(E) Shown is the DNaseI cleavage rate within linker regions for CTCF binding sites of different affinities. Note that higher affinity sites show greater linker region accessibility.

(F) *The DNaseI cleavage pattern at low-affinity sites begins to mirror that at unbound naked DNA.* Shown is the per-nucleotide DNaseI cleavage rates for high-affinity (blue) and low-affinity (red) CTCF_L binding elements as well as the deproteinated DNA DNaseI cleavage rate at these binding elements.

Figure 6.S3. DNA sequence and structure determinants of CTCF affinity

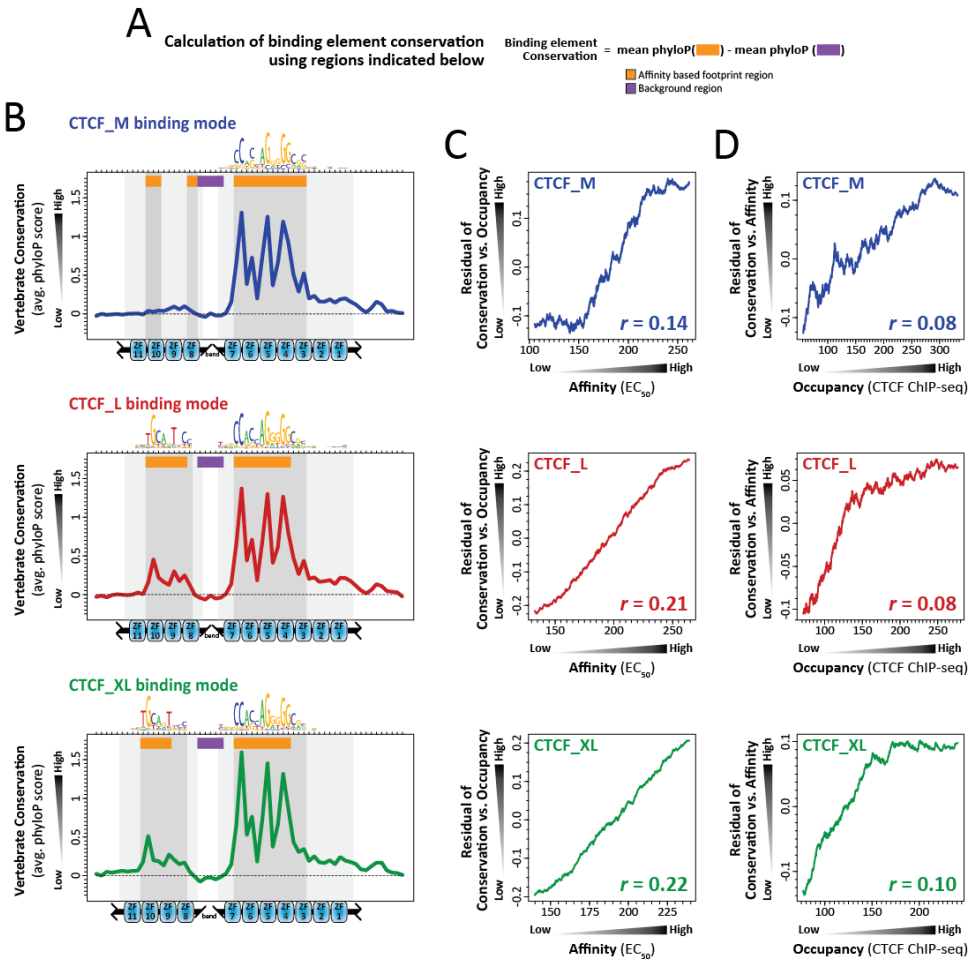


(A) Shown is the proportion of CTCF binding sites with a given affinity that contain one of the three different CTCF binding elements, no CTCF binding element, or multiple CTCF binding elements.

(B-F) *High affinity CTCF binding events utilize more DNA sequence-specific protein-DNA interactions.* Shown is the relationship at CTCF_M (left) and CTCF_XL (right) binding elements between the positional weight matrix (PWM) motif match and CTCF occvupancy (B) and CTCF affinity (C). (D) (top) Sequence information at high affinity and low affinity CTCF_L binding elements. (bottom) The difference in sequence information between high affinity and low affinity CTCF_L binding elements. Values above zero indicate bases that are favored at high affinity sites. (E) Same as (D) but for CTCF_XL binding elements. (F) Comparison of the sequence information plots in Figures 3F Supplemental Figure 3D and Supplemental Figure 3E shows the relative contribution of different bases within the binding element to binding site affinity at CTCF_M (blue), CTCF_L (red) and CTCF_XL (green) binding elements.

(G) *High affinity CTCF binding events utilize distinct DNA structure-specific protein-DNA interactions.* Shown are the predicted DNA minor groove widths at CTCF_L (left) and CTCF_XL (right) binding elements identified as having different strengths of *in vivo* CTCF affinity. DNA bases that are bound by CTCF in a base-dependent manner are highlighted in purple. Base-independent CTCF-bound nucleotides that significantly differ ($p<0.005$) in minor groove width at high- and low-affinity binding elements are highlighted in green. Other CTCF-bound sites are highlighted in grey.

Figure 6.S4. The per-nucleotide evolutionary constraint at CTCF binding elements is largely dependent on the affinity, not occupancy, of CTCF



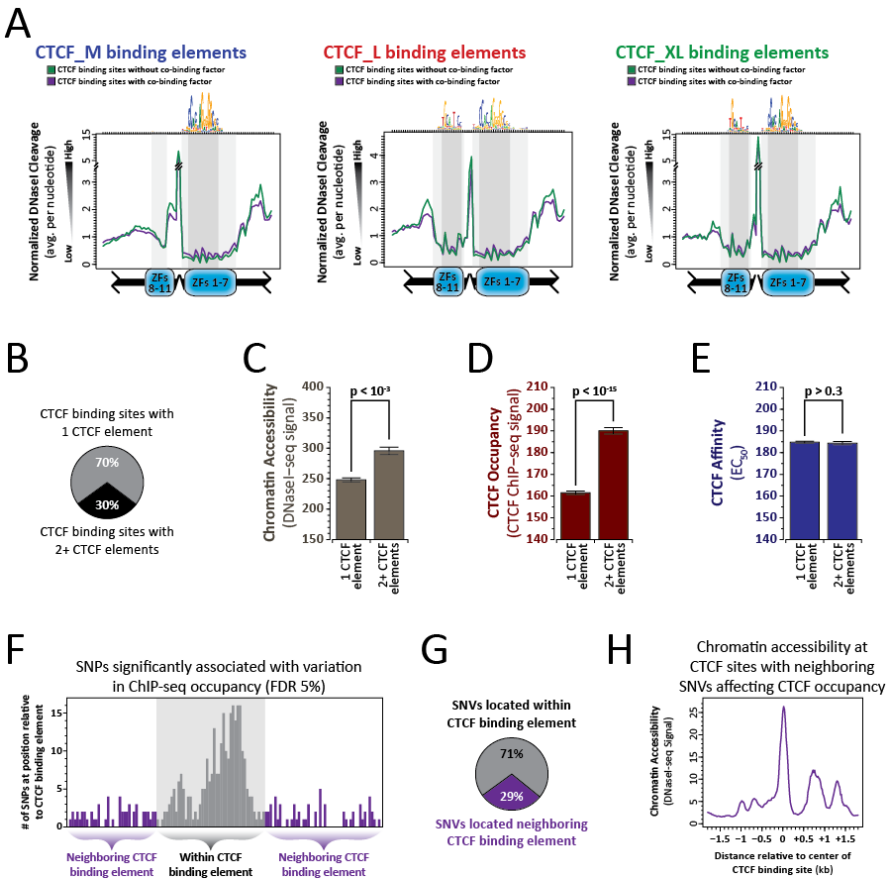
(A) Shown is the formula used to calculate the conservation at a given binding element that results from CTCF binding (i.e. controlling for the background conservation at the binding element).

(B) Shown are per-nucleotide average vertebrate conservation values for CTCF_M (top), CTCF_L (middle) and CTCF_XL (bottom) binding elements, as well as the footprint region (orange) and background region (purple) used to calculate the overall conservation at an element.

(C) For all CTCF_M (top), CTCF_L (middle) or CTCF_XL (bottom) binding elements, shown is a plot of the residual values from a linear regression of CTCF element conservation vs. occupancy versus CTCF affinity. Pearson correlation values are displayed in the bottom right hand corners.

(D) For all CTCF_M (top), CTCF_L (middle) or CTCF_XL (bottom) binding elements, shown is a plot of the residual values from a linear regression of CTCF element conservation vs. affinity versus CTCF occupancy. Pearson correlation values are displayed in the bottom right hand corners. Note that when CTCF affinity is controlled for (D), CTCF occupancy is only minimally associated with binding element conservation.

Figure 6.S5. CTCF engages in accessibility-mediated cooperativity



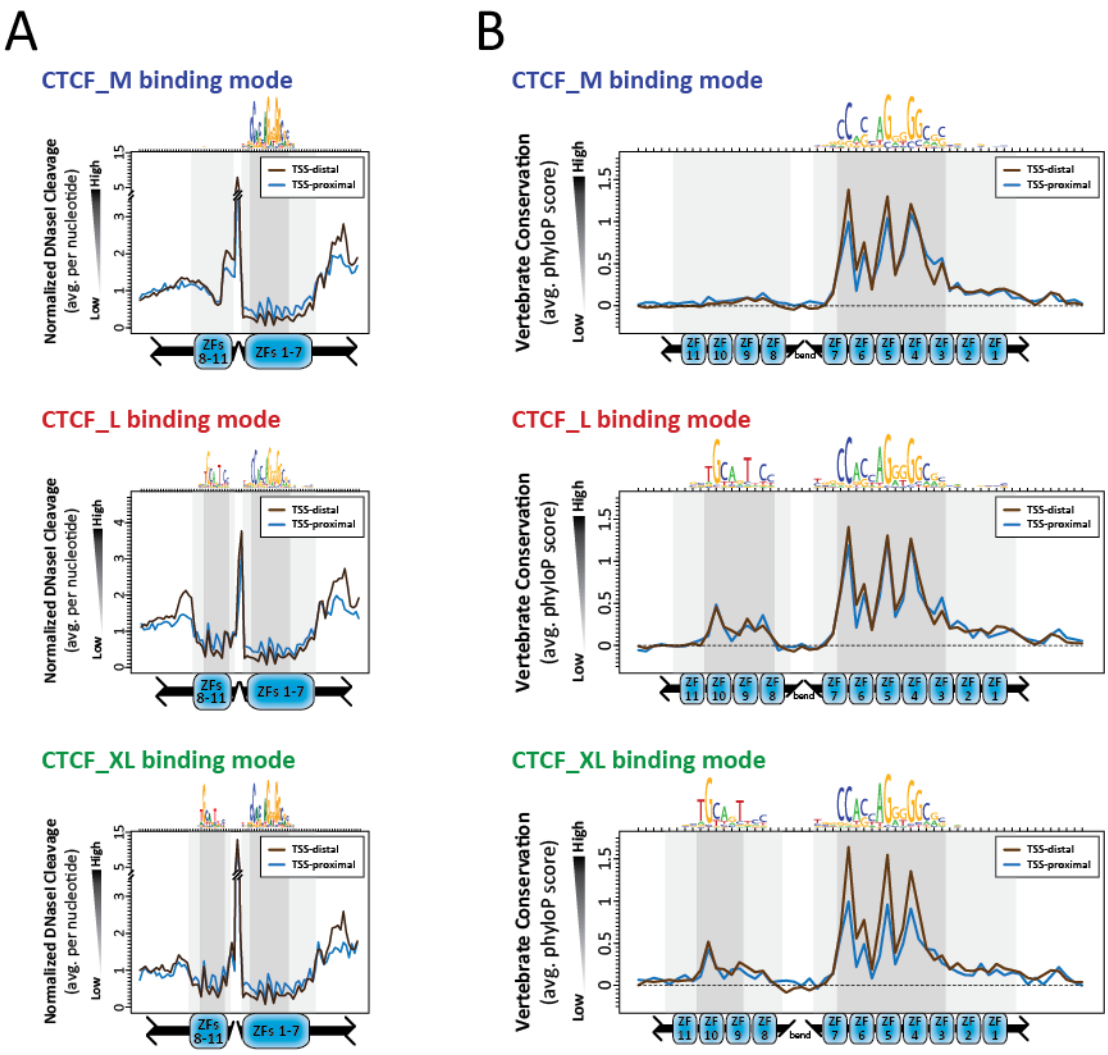
(A) Profile plots showing the per-nucleotide DNaseI cleavage rate at CTCF_M (left), CTCF_L (middle) and CTCF_XL (right) binding sites with (purple) and without (green) co-binding factors. Note that binding elements without co-binding factors have deeper footprints, indicative of higher-affinity CTCF-DNA interactions.

(B-E) *Co-binding of CTCF with itself increases chromatin accessibility and CTCF occupancy, but not CTCF affinity.* (B) Shown is the percentage of CTCF binding sites containing more than one CTCF binding element. Shown is the (C) chromatin accessibility, (D) CTCF occupancy and

(E) CTCF affinity at CTCF binding sites containing one or more than one CTCF binding elements.

(F-H) *CTCF occupancy is affected by SNPs contained in co-bound TF binding elements.* (F) The position, relative to CTCF binding elements, of SNPs that significantly affect the occupancy of CTCF (FDR 5%) (Maurano et al., 2012b). (G) The proportion of SNPs that significantly affect the occupancy of CTCF that are located outside of a CTCF binding element. (H) The average chromatin accessibility at the 112 CTCF binding elements with neighboring SNPs that significantly affect the occupancy of CTCF. The rugged chromatin accessibility landscape at these CTCF elements indicates that CTCF is co-bound by additional TFs.

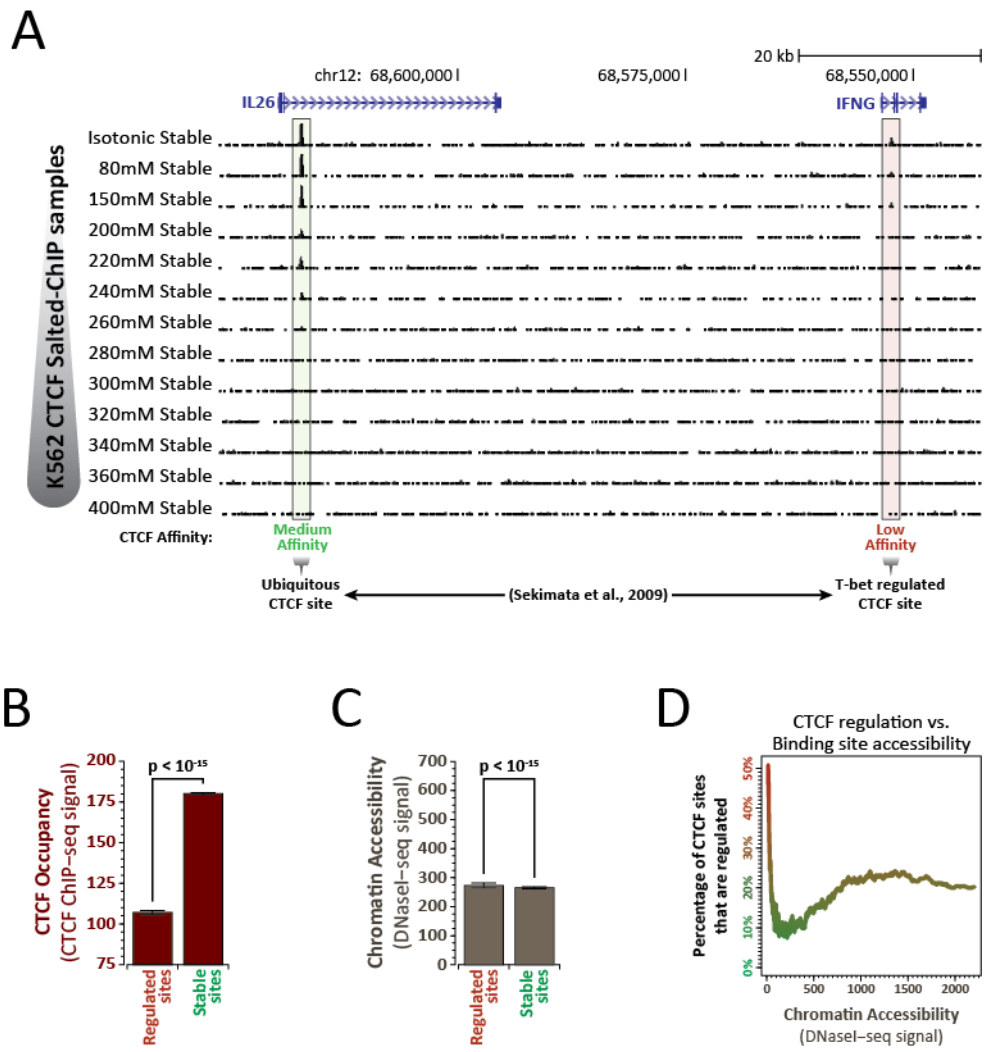
Figure 6.S6. Promoter-proximal CTCF binding elements are low-affinity and under less evolutionary constraint



(A) Profile plots showing the per-nucleotide DNaseI cleavage rate at CTCF_M (top), CTCF_L (middle) and CTCF_XL (bottom) binding sites proximal (blue) or distal (brown) to transcriptional start sites (TSSs).

(B) Profile plots showing the per-nucleotide vertebrate conservation at CTCF_M (top), CTCF_L (middle) and CTCF_XL (bottom) binding sites proximal (blue) or distal (brown) to transcriptional start sites (TSSs).

Figure 6.S7. Highly accessible CTCF binding elements are often regulated



(A) *A cell-selective interferon-gamma (IFNG) enhancer utilizes a low-affinity CTCF binding element.* Shown is the affinity of CTCF binding elements within the *IFNG* locus. Note that the CTCF binding site (left) previously identified as cell-type invariant (Sekimata et al., 2009) is medium affinity, whereas the cell-selective CTCF binding site within the *IFNG* gene is low-affinity.

(B) Shown is the CTCF occupancy at regulated and stable CTCF binding elements.

(C) Shown is the chromatin accessibility at regulated and stable CTCF binding elements.

(D) Shown is the relationship between the chromatin accessibility at a CTCF binding site and the probability that it is regulated across 19 cell types.

6.6 – METHODS

Nuclear extraction and salt-fractionation

For replicate 1, nuclei were isolated using a standard protocol (Dorschner et al., 2004). The nuclear extraction and salt-fractionation protocol is schematized in Figure 1A. Briefly, 5*10^7 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 2.5*10^6 cells/mL in 0.05% NP-40 (Roche) in Buffer A (15mM Tris pH 9.0, 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 400xg for 7 minutes and washed once with Buffer A. Nuclei were then resuspended at $1.2*10^7$ nuclei/mL in either Isotonic Buffer (10mM Tris pH 8.0, 15mM NaCl, 60mM KCl, 1.5mM EDTA pH 8.0, 0.5mM Spermidine) or Buffer B containing a variable amount of NaCl (10mM Tris pH 8.0, [80mM, 150mM, 200mM, 220mM, 240mM, 260mM, 280mM, 300mM, 320mM, 340mM, 360mM, 400mM, 450mM, 600mM NaCl], 1.5mM EDTA pH 8.0, 0.5mM Spermidine). After a 35 minute incubation at 4°C rocking, nuclei were pelleted at 2,000xg for 7 minutes and the supernatant was removed. Salt-extracted nuclei were then resuspended at $1.2*10^7$ nuclei/mL in a fresh batch of the same buffer used for the first extraction and incubated for an additional 35 minutes at 4°C rocking. After this second extraction, nuclei were again pelleted at 2,000xg for 7 minutes and the supernatant was removed. Salt-extracted nuclei were then resuspended in 10mL of the same buffer used for the two extractions, and 1mL of Buffer F (50mM Tris pH 8.0, 100mM NaCl, 1.5mM EDTA pH 8.0, 11% fofrmaldahyde) was added to perform cross-linking. After a 10 minute incubation rocking at room temperature, the cross-linking reaction was quenched by adding 1.57mL of 1.0M

Glycine and incubated for an additional 5 minutes at room temperature while rocking. Cross-linked chromatin samples were then washed 2 times with 10mL of 1xDPBS and then flash frozen in liquid nitrogen.

For replicate 2, nuclei were isolated as described above, but salt extractions were performed sequentially for 35 minutes each at 4°C rocking starting with Isotonic buffer and then progressing to the final concentration of NaCl Buffer B through all buffers in between. For example, the 80mM Buffer B extracted sample was first extracted with Isotonic Buffer for 35 minutes and then extracted with 80mM Buffer B for 35 minutes. After each extraction step, the chromatin was pelleted at 2,000xg for 7 minutes and the supernatant was removed before the next buffer was added. Cross-linking was performed as described above.

Chromatin Immunoprecipitation (ChIP) of salt extracted intact chromatin

For each salt extracted fraction, ChIP assays were performed as described (Welch et al., 2004) with modifications. Briefly, the crosslinked chromatin was lysed in lysis buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA pH 8.0, 1% SDS) supplemented with protease inhibitor cocktail (Roche). Chromatin was sheared by a Diagenode bioruptor. The supernatants were further diluted to a final concentration of 50 mM Tris-HCl pH 8.0, 0.15 M NaCl, 1 mM EDTA pH 8.0, 0.1 % SDS, 1 % Triton X-100, 0.1 % sodium deoxycholate. ~80 µl Dynabeads (M-280, sheep anti-rabbit IgG, Invitrogen) were incubated with 20 µl antibody (#2899 for CTCF, Cell Signal) for 6 hours at 4^{o}C and then incubated overnight with ~100 µg sheared chromatin. The complexes were washed sequentially with IP wash buffer I (50 mM Tris-HCl pH 8.0, 0.15 M NaCl, 1 mM EDTA pH8.0, 0.1 % SDS, 1 % Triton X-100, 0.1 % sodium deoxycholate), high salt buffer (50 mM Tris-HCl pH8.0, 0.5 M NaCl, 1 mM EDTA pH8.0, 0.1 % SDS, 1 % Triton X-100, 0.1 % sodium deoxycholate), IP wash buffer II (50 mM Tris-HCl pH8.0, 1 mM EDTA pH8.0, 1 % NP-

40, 0.7 % sodium deoxycholate, 0.5 M LiCl) and TE buffer (10 mM Tris-HCl pH8.0, 1 mM EDTA pH8.0). The complexes were incubated with elution buffer (10 mM Tris-HCl pH 8.0, 0.3 M NaCl, 5 mM EDTA pH8.0, 0.5 % SDS) at 65°C overnight. The reverse-crosslinked DNA was separated from the beads and treated with RNase A (Ambion) and Proteinase K (Fermentas), and then purified by phenol-chloroform extraction and ethanol precipitation. ChIP libraries were generated and sequenced on an Illumina Genome Analyzer IIx (Illumina, Haywaard, CA) by the High-Throughput Genomics Unit (University of Washington) following a standard protocol.

CTCF ChIP-seq peak calling

Thirty-six base pair Illumina sequence reads were mapped to the human genome (UCSC hg18) allowing up to 2 mismatches using bowtie (Langmead et al., 2009). Reads mapping to more than one location were then discarded. Smoothed density tracks were generated using Bedmap (Neph et al., 2012a) to count the number of tags overlapping a sliding 150-bp window, with a step width of 20-bp. Density tracks were normalized for sequencing depth by dividing the tag count by the number of mapped tags, then multiplying by 10 million. To identify ChIP-seq positive regions genome-wide, we applied the scan statistic algorithm “hotspot” (Sabo et al., 2004; John et al., 2011) to the nuclei ChIP-seq profile and generated a list of CTCF ChIP-seq peaks using a 1% false discovery rate (FDR) threshold. An additional signal intensity threshold corresponding to a signal intensity that resulted in a 97.5% replicate concordance of the two CTCF Nuclei ChIP-seq replicates was generated to screen out low-abundance variable peak calls (**Fig. 6.S1C**). In addition, a signal-to-noise threshold was generated to screen out noisy peaks and peaks found in the input track (**Fig. 6.S1D**). A signal-to-noise threshold was identified by quantifying the ratios of the signal intensity of varying quantiles of both the CTCF nuclei ChIP-seq and input tracks and identifying a ratio that best discriminated the CTCF nuclei ChIP-seq

from the input track. Using these thresholds, we identified 21,569 highly reproducible CTCF binding elements genome-wide. The pairwise Pearson correlations between the two replicates (**Fig. 6.S1B**) were computed using the tag-normalized combined density tracks from each of the ChIP-seq experiments.

Quantifying CTCF affinity using salted ChIP-seq data

For each of the 21,569 CTCF ChIP-seq peak that passed the above filters, CTCF occupancy at that peak within each of the 15 salt extracted nuclear CTCF ChIP-seq samples and the input sample was calculated by quantifying the sum of tags within the the 150bp CTCF ChIP-seq peak. To correct for slight differences in ChIP-seq enrichment efficiency between samples, we normalized the signal from each CTCF ChIP-seq sample using the average values derived in each sample from 8 manually identified CTCF binding sites that were occupied at similar levels in all of these samples. To identify the relative affinity at each binding site, we utilized a general 4 parameter model fitting function in R commonly utilized for concentration/dose/time-response models (*drm*; R package *drc*). For this model, the x-values were set as the concentrations of NaCl used to extract each chromatin sample with the isotonic extraction corresponding to 15mM NaCl (15, 80, 150, 200, 220, 240, 260, 280, 300, 320, 340, 360, 400, 450, 600). At each CTCF binding site genome-wide we independently applied a four-parameter log-logistic function (LL.4) using the corrected CTCF ChIP-seq signal intensity values at this site as y-values and fixed the LL.4 function parameter corresponding to the lower limit of the model as the minimum CTCF ChIP-seq occupancy observed at that site across all 15 samples. Using these models we then calculated the concentration of salt needed to disrupt 50% of the CTCF occupancy (hereafter referred to as the EC50). We were able to calculate EC50

values for XX of the CTCF binding sites in this dataset (XX% of all binding sites). The remainder YY binding sites were too noisy to accurately estimate an EC50 value.

Quantifying CTCF occupancy

For each ChIP-seq peak CTCF occupancy was calculated by quantifying the sum of CTCF nuclei ChIP-seq tags (both replicates combined) within the the 150bp CTCF ChIP-seq peak.

Quantifying binding site chromatin accessibility

Available DNaseI cleavage data from K562 cells was utilized (ENCODE Project Consortium et al., 2012). For each ChIP-seq peak, binding site chromatin accessibility was calculated by quantifying the sum of K562 DNaseI cleavage tags within the the 300bp region containing the CTCF ChIP-seq peak.

Identifying CTCF motif matches contained within CTCF ChIP-seq peaks

The previously identified core CTCF motif model (Kim et al., 2007) was used in our initial analysis. The upstream CTCF motif model was readily identified from a MEME analysis (Bailey and Elkan, 1994) of DNA sequences originating from a subset of the CTCF ChIP-seq sites that were stable at 350mM NaCl. This upstream motif matches previously published upstream CTCF motif descriptions (Filippova et al., 1996; Bowers et al., 2009). Three distinct motif models were constructed using the core and upstream CTCF motifs. One model contained solely the core CTCF motif (CTCF_M). A second model contained the core CTCF motif with the upstream motif spaced 8 bases away (CTCF_L). A third model contained the core CTCF motif with the upstream motif spaced 9 bases away (CTCF_XL). Motif models were padded with background nucleotide frequencies such that all were 35 bases in width.

For each ChIP-seq peak, the 300bp region containing the CTCF ChIP-seq peak was scanned using FIMO (Bailey et al., 2009) with the 3 motif models described above. Given the high similarity of the 3 motif models used and the propensity of CTCF motifs to match genomic locations in both directions, we employed a q-value based scoring system to identify which motif occurrence best matched a given genomic sequence (we set our maximum q-value cutoff at 0.02). In instances where FIMO identified 2 or more overlapping motif occurrences, we took the motif occurrence with the smallest q-value as the best motif. Since all of the motif models are the same width and the FIMO q-values are derived from the same dataset, this FDR based scoring system should allow us to confidently compare motif occurrences (Bailey et al., 2009).

Calculating per-nucleotide DNaseI cleavage profiles

For each CTCF motif occurrence, the DNaseI cleavage per nucleotide was calculated by assigning each nucleotide position an integer score equal to the number of uniquely mappable DNaseI-seq sequence tags with 5' ends mapping there (Hesselberth et al., 2009; Neph et al., 2012c). Aggregate DNaseI plots were generated by averaging over all strand-oriented motif occurrences the number of DNaseI cleavages at the given nucleotide. Furthermore, these profiles were normalized by the average accessibility at these sites to enable the accurate comparison of profiles between sets of sites with differing chromatin accessibility.

Relative DNaseI footprinting calculations

Relative DNaseI footprinting calculations were performed using genome-wide DNaseI cut count data from K562 cells (ENCODE Project Consortium et al., 2012) in a manner analogous to that explained by Galas & Schmitz in their initial description of DNaseI footprinting (Galas and Schmitz, 1978). Briefly, for each genomic binding site a footprinting score was calculated based on the relative depth of a given footprint (**Fig. 6.S2D**). As described

by Galas & Schmitz, this footprinting score is a function of 3 parameters: (i) the concentration of the transcription factor; (ii) the affinity of a transcription factor for a given binding site; and (iii) the specificity of DNaseI for nucleotides within and around a given site. Since these binding sites are all contained in the same nucleus, the observed transcription factor protein concentration should be the same for all of the different genomic binding sites. Additionally, if we aggregate the footprinting signal over multiple genomic sites, we should dilute any effects that the nucleotide specificity of DNaseI has on the footprinting statistic for any given binding site (Drew and Travers, 1984). Consequently, by comparing the average footprinting statistic between groups of genomic binding sites we can determine which group is composed of higher affinity *in vivo* binding sites. Since we are using the relative footprinting depth as opposed to the absolute footprinting depth, any differences in the DNaseI hypersensitivity of the sites should minimally affect our results.

Calculating Deproteinized DNA DNaseI cleavage profiles

Deproteinized DNA DNaseI cleavage profiles were calculated using data from (Lazarovici et al., 2013) using a similar approach as described for per-nucleotide DNaseI cleavage.

Calculating per-nucleotide vertebrate conservation profiles

For each CTCF motif occurrence, the vertebrate conservation per nucleotide was calculated using a similar approach as described for per-nucleotide DNaseI cleavage, except that a nucleotide specific phyloP score was used instead of a nucleotide specific DNaseI cleavage score (Siepel et al., 2006). These scores are negative log-transformed p-values that measure acceleration and conservation at each nucleotide, derived from multiple alignments of 43 vertebrate genomes to the human genome.

Binding element vertebrate conservation calculations

The overall vertebrate conservation of a binding element was calculated as the difference in the average evolutionary constraint at bases involved in direct sequence-specific or structure-specific protein-DNA contacts (**Fig. 6.S4B** - orange box) and the average evolutionary constraint at bases contained within the CTCF binding element, but not involved in direct sequence-specific or structure-specific protein-DNA contacts (**Fig. 6.S4B** - purple box). Binding sites contained within coding sequence or UTRs were removed from these analyses.

Calculating base composition and entropy profiles

High-affinity (top 5%) and low-affinity (bottom 5%) CTCF peaks containing a single CTCF_M, CTCF_L or CTCF_XL motif occurrence were analyzed for the sequence composition of these binding elements. Briefly, strand-oriented genomic DNA sequences from these binding elements were compiled and motif logos and entropies at each base were generated using WebLogo 3.3 (Crooks et al., 2004).

DNA minor groove width calculations

The DNA minor groove at all CTCF binding elements was calculated as previously described (Slattery et al., 2011; Lazarovici et al., 2013). For high affinity (top 25 percentile), medium-high affinity (25-50 percentile), medium-low affinity (50-75 percentile) and low affinity (bottom 25 percentile) CTCF peaks containing a single CTCF_M, CTCF_L or CTCF_XL motif occurrence we calculated the average predicted minor groove width at each position within the CTCF binding element. To identify if the minor groove width at any of these positions significantly correlated ($p<0.005$) with the affinity of CTCF at that site, we used a two-sided K-S test to compare the minor groove width at that base at high affinity (top 25 percentile) and low affinity (bottom 25 percentile) binding elements.

Identification of CTCF peak co-occupancy with other TFs

K562 ChIP-seq peak calls generated by the ENCODE consortium (ENCODE Project Consortium et al., 2012) was used for the following TFs: BCL3, BCLAF1, CEBPB, cJUN, cMYC, E2F4, E2F6, EGR1, ELF1, ETS, FOSL, GABP, GATA2, MAFK, MAX, MXI1, NFE2, NFYA, NFYB, NRF1, NRSF, PU1, ZIX5, SP1, SP2, SRF, TAF1, TAF7, TAL1, TBP, TF3, THAP1, USF1, USF2, YY1, ZBTB33, ZBTB7, ZNF263. A TF was called as co-occupying a CTCF binding site if it contained a ChIP-seq peak overlapping the 150bp CTCF ChIP-seq peak.

Identification of CTCF binding elements coding sequence or located proximal to transcriptional start sites (TSSs)

Coding sequences overlapping CTCF ChIP-seq peaks were identified using refseq annotations. CTCF peaks located within +/- 2kb of a refseq annotated transcriptional start sites (TSS) were labeled as promoter-proximal.

Identifying regulated CTCF binding sites

Regulated CTCF binding sites were identified using previously published CTCF ChIP-seq peak calls from K562, Hela-S3, HepG2, AG09309, AG09319, AG10803, AoAF, BJ, CACO-2, GM06990, HBMEC, HEEpiC, HMF, HPAF, HPF, HRE, SAEC, SK-N-SH_RA and WERI-Rb-1 cells (Wang et al., 2012). CTCF binding sites that were not observed in at least one of these cell types were labeled as ‘regulated’ whereas those observed in all 19 cell types were labeled as ‘Stable’.

Chapter 7 – Developmental fate and cellular maturity encoded in human regulatory DNA landscapes

7.1 – ABSTRACT

How and to what degree information about prior cellular states propagates during development is unknown. Here we show that early developmental fate decisions and enhancer activity states can be derived from the genomic patterning of DNaseI-hypersensitive sites (DHSs) in terminally differentiated cells, distinct from information conveyed by gene expression. Developing cells share a proportion of their DHS landscapes with ES cells (ESCs) that decreases monotonically with the progress of differentiation, providing a quantitative measure of cellular maturity. Developmentally stable DHSs are chiefly occupied by transcription factors that participate in autoregulatory feedback circuits. In contrast to normal cells, cancer regulatory landscapes feature both extensive reactivation of silenced ESC DHSs and ectopic activation of regulatory DNA from alternative developmental programs external to the cell lineage from which the malignancy derives. Our results indicate a central role for regulatory DNA patterning in propagating cellular state and fate information during normal development.

7.2 – INTRODUCTION

Under natural conditions, tissue and cellular differentiation along defined lineages is characterized by an inexorably forward-moving process that terminates in highly specialized cells. Waddington, following Morgan (Morgan, 1901), characterized the process of development as essentially ‘epigenetic’ (from ‘epigenesis’), though imbued with a pre-determined sense of direction (Waddington, 1939). To describe the molecular and genetic forces guiding

development and differentiation, Waddington introduced the metaphor of an 'epigenetic landscape' (Waddington, 1940), which he depicted with a ball rolling down a hill of bifurcating valleys symbolizing the specification of defined cell lineages, and the restriction of possible alternative fates during the progress of differentiation (Waddington, 1939, 1957). It is notable that Waddington's usage of 'epigenetic' to denote the origination and propagation of information about cellular states during differentiation differs considerably from its recent reformulation to mean 'on the genome' and its association with chemical modifications to DNA or chromatin (Ptashne, 2007; Misteli, 2013). Here we employ throughout the classical usage.

The epigenetic landscape paradigm has been widely invoked to explain not only normal developmental processes, but also abnormal or artificial processes such as the creation of pluripotent cells (rolling back up the hill) (Yamanaka, 2009), the 'de-differentiation' of cancer cells (Pujadas and Feinberg, 2012), or 'trans-differentiation' from one cell lineage to another (Graf and Enver, 2009). However, it is presently unknown whether or to what extent development and differentiation in fact manifest predominantly the sequential restriction of pre-specified directives (either encoded in the genome sequence or otherwise inherited from prior cellular states), as opposed to the sequential acquisition of new developmental potential (e.g. through genesis of new cellular control features).

Waddington astutely reasoned that development is a 'historical' process requiring a memory 'faculty' to keep directed lineage programs on track (Waddington, 1939). Indeed, developing cells are frequently exposed to stimuli, whether exogenous (e.g., a morphogen) or endogenous (e.g., a transcription factor), that permanently alter cellular fate, even after the inciting agent is no longer extant. Whether or in what form cells in fact maintain information

concerning prior developmental fate decisions in an organized fashion during epigenesis is currently unknown.

Chromatin structure represents a highly plastic vehicle for specifying cellular regulatory states, and a conceptually attractive template for recording and transmitting such information in a form that might be accessed at subsequent developmental time points. The chromatin template can be modified both chemically and structurally. Chemical modification of chromatin by enzymatic activities that covalently modify histone tails has been intensively studied, and the patterns of many such modifications differ between primitive and definitive cells (Hawkins et al., 2010; Zhu et al., 2013; Bernstein et al., 2006), as well as undergo graded directional transitions during differentiation (Paige et al., 2012; Wamstad et al., 2012). However, the underlying mechanisms responsible for establishing these patters remain unclear.

DNaseI hypersensitive sites (DHSs) represent focal alterations in the primary structure of chromatin that result from engagement of sequence-specific transcriptional factors in place of a canonical nucleosome (Gross and Garrard, 1988; Felsenfeld, 1996; Thurman et al., 2012). In a classic experiment, Groudine and Weintraub demonstrated that induced DHSs could be propagated to, and stably perpetuated by, daughter cells even after the inducing stimulus had been withdrawn (Groudine and Weintraub, 1982). This result suggests that newly arising primary chromatin structures created by transcription factor occupancy of quiescent regulatory DNA have the potential to encode cellular states directly and to perpetuate that information through continued transcription factor (TF) occupancy in daughter cells. Whether, or to what extent, such a mechanism operates during normal development and differentiation, however, is currently unknown.

A connection between development and cancer has long been posited, centering on the potential for both developing and malignant cells to undergo rapid proliferation and self-renewal (Beard, 1902; Waddington, 1935). Based on limited analyses of metabolic (Warburg, 1956), histological (Gleason and Mellinger, 1974) and gene activity phenotypes (Hirszfeld et al., 1932; Tatarinov, 1964) cancer cells are widely described as being 'de-differentiated' compared with their normal counterparts. However, quantifying this concept and generalizing it beyond a few selected markers has proven difficult.

To explore the role of transcription factor-driven chromatin structure at regulatory DNA in normal and transformed cells during epigenesis, we analyzed genome-wide patterns of DNaseI hypersensitive sites across a wide array of cell types and states including definitive adult primary cells; embryonic stem (ES) cells; cells undergoing directed lineage differentiation from ES cells to cardiomyocytes; and diverse cancer cell types. Our findings, detailed below, are interpreted to indicate four fundamental conclusions. First, patterns of DHSs in definitive cells encode 'memory' of early developmental fate decisions that establish lineage hierarchies, and of the tissue activity spectra of primitive enhancers. Second, lineage differentiation couples the extensive activation of novel regulatory DNA compartments with propagation and sequential restriction of the ES DHS landscape as a monotonic function of cellular maturity. Third, developmentally stable DHSs are chiefly populated by self-regulating transcription factors, suggesting a mechanistic role for TF-encoded feedback circuits in propagating developmental information. Finally, malignant transformation is accompanied by retrogression of the regulatory DNA landscape toward a primitive state, but in a disordered fashion that defies normal developmental pathways and departs fundamentally from the paradigm of the epigenetic landscape. Together these findings indicate a central role for patterning and propagation of

regulatory DNA marked by DHSs in the genesis and proper maintenance of developmental programs.

7.3 – RESULTS

Lineage programming of human regulatory DNA

Regulatory DNA landscapes defined by DNaseI hypersensitive sites are both highly cell type-specific and highly stable (Thurman et al., 2012). We first sought to determine how the regulatory landscapes of diverse definitive cells were related to one another and to the regulatory DNA of embryonic stem (ES) cells. To address this, we aggregated genome-wide maps of DHSs from human ES cells plus 38 diverse normal definitive primary cell types from (Thurman et al., 2012) for which anatomical and histological origins could be unambiguously verified. To expand the phenotypic range of cell types and to deepen coverage of the well-characterized hematopoietic lineage, we obtained 9 additional definitive cell samples from adult donors including B-cells (CD19+, CD20+), NK-cells (CD56+), CD34+ hematopoietic progenitors (3 separate donors); and skin keratinocytes (3 donors). The relative representation of different major embryological lineages (mesoderm, ectoderm, endoderm) among these 49 cell types parallels that of recognized cell types (Bard et al., 2005), of which those of mesodermal origin comprise the significant majority (~80%).

We performed DNaseI hypersensitivity cleavage mapping on each of the 49 cell types using a common protocol (**Methods**), and delineated DHSs using a common algorithm that has been extensively validated for both sensitivity and specificity (John et al., 2011; Thurman et al., 2012) (**Methods**), resulting in an average of 161,160 autosomal DHSs per cell type (at FDR 1% threshold, range 91,720 to 257,172 autosomal DHSs per cell type). Preliminary inspection of the

DNaseI profiles suggested systematic commonalities and differences between major cell type groups (**Fig. 7.M1A** and **Fig. 7.S1A-B**).

To visualize quantitatively the relationships between the diverse regulatory DNA landscapes, and to avoid biasing toward promoters, which display higher average DNaseI sensitivity than distal elements (Thurman et al., 2012), we considered each DHS to be either present or absent within a given cell type. Although the magnitude of DNaseI hypersensitivity may convey important information about differential factor occupancy (John et al., 2011; Samstein et al., 2012), it does not appear to be related to the functional significance of a given element, as known enhancers evince a wide range of DNaseI hypersensitivity in their cognate cell types (Thurman et al., 2012). We thus computed the Euclidean distance between all non-redundant pairs of cell types (**Fig. 7.S1C**), and rendered the results using simple unsupervised nearest-neighbor clustering (**Fig. 7.M1B**).

The resulting *ab initio* dendrogram recapitulated known cell lineage relationships with remarkable detail, as well as broader features of embryological origin. On a gross level, ES cells occupied the deepest root, and derivatives of the three germ layers (mesoderm, ectoderm and endoderm) were correctly partitioned into separate high-level clusters (**Fig. 7.M1B**). Mesodermal progeny were further partitioned into paraxial mesoderm, primitive mesoderm, and hemangioblast derivatives. The common embryological origin of endothelia and blood was clearly represented, as was the fine partitioning of the hematopoietic tree into hematopoietic progenitors and lymphoid cells, and the different subtypes of lymphoid tissue including B-cells, T-cells, Natural killer (NK) cells, and more primitive lymphoblastoid cells. While relationships between the derivatives of paraxial mesoderm are less well

understood, we observed subgroups that were organized into anatomical units, such as grouping of the stroma of the heart and great vessels (aorta, pulmonary artery).

The stability of the clustered cell relationships was attested by the strict cohesion of multiple samples of the same cell type that were derived from different individuals at different times, including gingival fibroblasts (n=2), cardiac fibroblasts (n=2), hematopoietic progenitors (n=3), and keratinocytes (n=3). To affirm the distinctiveness of the major cluster groups, we also performed three-dimensional principal coordinate analysis (PCoA) using the aforementioned pairwise Euclidean distance measures **(Fig. 7.M1C)**. This analysis revealed a clear separation between developmentally distinct cell types, with the regulatory DNA landscape of ES cells occupying a central position relative to all other cell types.

To validate the robustness of the clustering, we mapped DHSs in 8 additional cell types of diverse embryological origin (two paraxial mesoderm, two hematopoietic, two endothelial, one ectodermal and one endodermal). We found that all 8 new cell types clustered tightly within their appropriate clades **(Fig. 7.S1C)**, indicating that DHS patterns can correctly identify both developmental origin and specific effector cell type in a prospective fashion (**Fig. 7.S1D-E)**.

Importantly, clustering gene expression patterns for the same cell types failed to recover the fundamental lineage branching relationships exposed by clustering DHS patterns (**Fig. 7.S3A-B),** including rooting of the lineage tree in ES cells; improper high-level segregation of germ layer derivatives; and improper partitioning of mesodermal derivatives. These results demonstrate that the dendogram in **Fig. 7.M1B** is not driven by functional convergence on terminal regulatory states driven by gene expression patterns.

The fact that the aforementioned lineage relationships – including representation of specific primitive commitment events – can be derived from a simple clustering of the DHS

landscapes of terminally differentiated cells suggests that the linear patterning of regulatory DNA along the genome encodes an imprint of prior cellular fate decisions. Given that ES cells represent a common developmental ancestor to the other cell types, the centrality of ES cells within the PCoA plot suggests that significant yet distinct components of the ES regulatory landscape are shared in each of the definitive cell types.

An 'hourglass' pattern of conservation at developmental regulatory DNA

Next, we sought to determine the pattern of evolutionary constraint on regulatory DNA as a function of developmental maturity. To address this we first inferred regulatory DNA stably arising at seven distinct developmental branch points (epiblast, mesoderm, hemangioblast, paraxial mesoderm, endothelia, hematopoietic and lymphoid) by identifying DHSs common to the corresponding dependent branches of the dendogram in Fig. 7.M1 (Methods). We then calculated the mean level of evolutionary constraint for each set of elements using phyloP (Pollard et al., 2010) (Methods). This analysis revealed that regulatory DNA common to mesodermal derivatives (and thus inferred to be stably arising during the onset of the mesodermal lineage) is significantly more evolutionarily constrained than that arising either during early embryogenesis or later lineage differentiation (**Fig. 7.M1D**). This pattern is compatible with the 'hourglass' model of development (Duboule, 1994; Raff, 1996) that has been variably described using cross-species morphology (Von Baer, 1828), gene expression (Kalinka et al., 2010; Irie and Kuratani, 2011; Levin et al., 2012) and gene conservation (Domazet-Lošo and Tautz, 2010; Quint et al., 2012).

Developmental persistence of chromatin accessibility at primitive enhancers

We next asked whether enhancers active during early development could be persistently marked by DHSs in definitive cells. Systematic studies of evolutionarily conserved human DNA elements in transgenic mice have identified >700 early developmental enhancers (Pennacchio et al., 2006; May et al., 2011), each of which displays activity in one or more embryonic tissue (**Fig. 7.M2A**). Of 721 non-promoter human enhancers with reproducible tissue staining patterns in transgenic d11.5 embryos, a surprising proportion – 64% – exhibit DNaseI hypersensitivity in at least one definitive human cell type (**Fig. 7.M2B**). To quantify the tissue activity spectra of these elements, we systematically collated images of enhancer-driven lacZ expression in individual transgenic animals, and related these with cross-cell type patterns of DNaseI hypersensitivity at the same elements in definitive cells (**Fig. 7.M2A**). For example, an enhancer that is selectively active in embryonic heart tissue (**Fig. 7.M2A**, 1st image) is DNaseI hypersensitive selectively within cells derived from human heart and great vessel structures (**Fig. 7.M2C**), and an enhancer that is selectively active in embryonic blood vessels (**Fig. 7.M2A**, 3rd image) is DNaseI hypersensitive selectively within hemangioblast derivatives (endothelia and hematopoietic progenitors; **Fig. 7.M2C**). By contrast, an enhancer with extremely broad tissue activity (**Fig. 7.M2A**, 4th image) is DNaseI hypersensitive in nearly all definitive cell types (**Fig. 7.M2C**).

These findings generalize across the spectrum of enhancers: 100% of enhancers active in embryonic blood vessels are found to be DNaseI hypersensitive in adult endothelial cells, whereas only 30% of all other embryonic enhancers are DNaseI hypersensitive in adult endothelial cells (**Fig. 7.S2A-C**). Similarly, 73% of enhancers that are active in embryonic heart tissue are DNaseI hypersensitive within cells derived from human heart and great vessel

structures, whereas only 27% of all other embryonic enhancers are DNaseI hypersensitive in these cell types (**Fig. 7.S2D**).

We also found striking correlation between the number of primitive tissues in which enhancer activity was detected and the number of definitive cell types in which a DHS was detected at that enhancer (**Fig. 7.M2D**). Together, these results suggest both systematic developmental persistence of DNaseI hypersensitivity at a subset of early developmental enhancers, and persistence of enhancer functional spectra in the form of DHS patterning across different definitive cell types.

Restriction vs. expansion of the chromatin landscape during differentiation

The relationship between ES cells and definitive lineage derivatives disclosed in **Fig. 7.M1**, combined with evidence for developmental persistence of individual DHSs, prompted us to analyze in detail the relative gain and loss of DHSs along specific developmental clines. It is notable that ES cells have the largest DHS complement of all cell types analyzed (n=257,172 excluding ChrX/Y; **Fig. 7.S3C**), of which 58% are shared with at least one definitive cell type (**Fig. 7.S3D**).

Development along the hematopoietic lineage has been extensively characterized at both the cellular and molecular level (Orkin, 1995). During hematopoiesis the accessible chromatin landscape undergoes substantial reorganization that is dominated by the inactivation rather than the *de novo* activation of DHSs (**Fig. 7.M3A,B**). Comparison of the DHS landscape of ES cells to that of hematopoietic progenitors reveals a net loss of 119,032 DHSs, achieved through the silencing of 202,412 ES DHSs and the *de novo* activation of 83,380 DHSs. As hematopoietic progenitors terminally differentiate into B- or T-cells, they preferentially inactivate a common

set of ~72,000 DHSs (**Fig. 7.M3A,C**), while preferentially activating an average of ~52,000 chiefly lineage-restricted DHSs (**Fig. 7.M3A,D**). Development along each hematopoietic lineage results in the generation of a more restricted, specialized accessible chromatin landscape (**Fig. 7.M3B**). Of note, for each of the definitive lymphoid cell types, roughly half of the regulatory DNA landscape is retained from hematopoietic progenitors, and roughly one third from ES cells (**Fig. 7.M3B**).

Fixed contribution of ES regulatory DNA to terminal regulatory landscapes

We next asked how cells differ with respect to the proportion of their regulatory landscapes shared with ES cells. To identify the extent to which ES cell DHSs are retained during a developmental lineage we analyzed the contribution of ES cell DHSs to the regulatory landscape of a variety of diverse definitive cell-types. Strikingly, irrespective of its total number of DHSs (which varies >2-fold among cell types), each definitive cell type retained ~37% of its regulatory DNA landscape from ES cells (**Fig. 7.M3F**). The vast majority (~80%) of these elements are distal, non-promoter DHSs (**Fig. 7.S3E**). However, as the majority of retained ES cell DHSs were found in less than half of the 48 definitive cell types, each cell type retained a distinct mix of ES DHSs (**Fig. 7.M3G**).

Regulatory landscape dynamics during directed differentiation from ES cells

To test the above concepts prospectively, we next analyzed regulatory DNA dynamics during the controlled differentiation of ES cells along a specific developmental cline. Cardiomyocyte differentiation from ES cells represents a well-studied paradigm and is orchestrated by highly timed developmental elements (Paige et al., 2012; Kattman et al., 2011;

Yang et al., 2008). We therefore cultured embryonic stem cells (H7) using a defined set of growth factors and conditions to generate committed cardiac progenitors (appearing at day 5) and immature cardiomyocytes (appearing at day 14), and produced DHS maps for each stage. It is notable that although day 14 cells display synchronized beating in culture, they still exhibit primitive features similar to early fetal stage heart (Paige et al., 2012). For comparison, we therefore also performed DHS mapping in heart tissue obtained from an adult donor.

During directed differentiation we observed large scale reorganization of the DHS landscape (**Fig. 7.M3H**), which was again dominated by the inactivation of early developmental regulatory elements, coupled with the forward propagation of DHSs and the activation of stage–selective elements (**Fig. 7.M3I**,**J**). Strikingly, the inactivation of early developmental elements occurred in a monotonic, clock-like fashion, dropping from 71% at day 5, to 49% at day 14 and 35% in adult heart tissue (**Fig. 7.M3K**). Notably, the proportion of ES DHSs in the terminally differentiated adult heart landscape (35%) matches closely with the average of other terminally differentiated cells (37%; **Fig. 7.M3F**).

Together, the above findings indicate that development along multiple lineages is associated with three basic features: (i) extensive propagation of regulatory DNA marked by DHSs; (ii) pruning of DHSs found in progenitor cells as a monotonic function of developmental maturity; (iii) appearance of a smaller number of lineage-restricted DHSs.

Patterning of regulatory DNA by known and novel lineage regulators

We next sought to determine whether the regulatory DNA dynamics described above reflect the actions of specific lineage-restricted transcription factors (TFs). It is well established that lineage-restricted TFs have the capacity to shape dynamically the accessible chromatin

landscape of a cell during development and differentiation (Weisbrod and Weintraub, 1979; Davis et al., 1987). Consequently, uncovering the TFs that interact with developmentally dynamic (i.e., lost or gained) regulatory DNA should facilitate identification of regulators of cellular identity. Because the accessible regulatory DNA compartment of a given cell type occupies only ~1-2% of the genome, the prior probability that the recognition sequence of a given TF within this small compartment is actually occupied *in vivo* is high, and indeed closely mirrors qualitative factor occupancy measured using ChIP-seq and genomic footprinting (Neph et al., 2012c; Samstein et al., 2012). We focused our analysis on two inferred transitions (ES cells to hematopoietic progenitors, and hematopoietic progenitors to T-, B-, and NK-cell lineages), and two observed transitions during cardiac-directed differentiation from ES cells.

Analysis of the DHSs lost vs. gained during the inferred transition from embryonic stem cells to hematopoietic progenitors (CD34+) revealed loss of DHSs enriched in recognition sequences for pluripotency factors (OCT4, SOX2, NANOG), coupled with gain of DHSs selectively enriched in recognition sequences for major master regulators including PU.1 and ELF1 (Scott et al., 1994; Bassuk et al., 1998) (**Fig. 7.M4A**). Analysis of DHSs lost vs. gained during the inferred transition from hematopoietic progenitors to T- or B-lymphocytes or NK-cells revealed selective loss of DHSs enriched in hematopoietic master regulators, and selective gain of DHSs enriched in major T-cell, B-cell, or NK-cell lineage regulators (Nakajima et al., 1997; Imada et al., 1998; Verbeek et al., 1995; Corcoran et al., 1993; Zandi et al., 2008) (**Fig. 7.M4B** and **Fig. 7.S4A**). Moreover, this analysis indicates previously unrecognized roles for numerous distinct TFs in lineage-defining processes. For example, the binding landscape for RREB1 contracts in every lineage except for hematopoietic progenitors, in which it expands (**Fig. 7.S4B**). This repressor has been implicated in a number of different developmental

processes, but these results indicate that it may play an important but as-yet-uncharacterized role in hematopoiesis. Analogous conclusions may be inferred for many other factors (**Fig. 7.S4B**).

Unlike hematopoietic development, few potent regulators of cardiac differentiation have been characterized. Analysis of DHSs lost vs. gained during the differentiation of ES cells to early cardiac progenitors revealed selective loss of DHSs enriched in recognition sequences for pluripotency factors, coupled with gain of DHSs enriched in recognition sequences for the well-described early cardiac regulator PBX1 (Chang et al., 2008) and the recently-described cardiac regulator MEIS2 (Paige et al., 2012) (**Fig. 7.S4C**, left). Analysis of the subsequent transition from cardiac progenitors to early cardiomyocytes exhibited further pruning of regulatory DNA enriched in pluripotency factors and the early regulator PBX1, and appearance of new DHSs enriched in recognition sites for late cardiac regulators NKX2-5, NKX2-6 and MEF2A (Lyons et al., 1995; Tanaka et al., 2001; Naya et al., 2002) (**Fig. 7.M4C**, right). Together, these findings indicate that regulatory DNA activated along specific lineage pathways directly reflects the actions of major lineage regulating TFs. Moreover, they show that sampling of regulatory DNA patterns across a differentiation gradient can distinguish regulators important for early vs. late differentiation processes.

Coordinated deactivation of alternative regulatory programs during differentiation

The actions of lineage restricted factors were not only reflected in the regulatory elements activated along that lineage, but also in the regulatory elements inactivated along all other lineage paths. For example, the regulatory landscape for the NK-cell master regulator NFIL3 (Gascoyne et al., 2009) remained largely unchanged during NK cell development, but was greatly diminished during T- and B-lymphocyte development (**Fig. 7.M4D**). Similarly the

regulatory landscape for the endothelial regulator SOX17 (Liao et al., 2009) remained largely unchanged during endothelial development, but was greatly diminished during the development of all of the other cell types (**Fig. 7.M4E**). Similar lineage restricted patterns were observed for a number of other factors including PU.1, RREB1 and OCT4 (**Fig. 7.M4E** and **Fig. 7.S4B**). This suggests that (i) the regulatory DNA target landscape for certain lineage restricted factors is largely pre-positioned in progenitor cell types via DHSs that contain their cognate recognition sequences; and (ii) that such DHSs are selectively inactivated along lineage paths in which the lineage-relevant transcription factor is lacking. Consequently, development according to a specific lineage program is associated with both the activation of lineage-restricted regulatory elements, as well as the inactivation of regulatory DNA involved in the specification of alternative lineage fates directed by different TF programs.

'Memory' DHSs are chiefly occupied by TFs that regulate their own expression

As noted above, each adult cell type shares ~37% of its accessible chromatin landscape with ES cells (**Fig. 7.M3F**). Given the role of TFs in shaping regulatory DNA dynamics during development, we next sought to investigate the role of TFs at inferred developmentally stable DHSs. Many diverse biological processes are perpetuated by reinforcing feedback loops. For example, transcription factors involved in autoregulatory feedback loops can stabilize their expression during cell division and development (Alon, 2006; Zheng et al., 2010; Ptashne et al., 1980). We therefore hypothesized that the propagation of developmentally stable DHSs may employ such a mechanism.

We specifically asked whether developmentally stable DHSs were preferentially constructed from autoregulating TFs. To test this, we utilized comprehensive maps of human

transcription factor regulatory networks constructed using genome-wide DNaseI footprinting (Neph et al., 2012b). These maps extensively delineate the regulatory interactions amongst 475 TFs with known recognition sequences, mapped across 23 of the cell types studied here. Using these network maps, we identified, on average, 68 simple autoregulating TFs per cell type (range 48-75) (**Fig. 5A**). Next, we partitioned the accessible chromatin landscape of defined adult cell types into developmentally stable DHSs (DHSs shared with ES cells) and developmentally gained DHSs (DHSs not shared with ES cells) (**Fig. 7.M5A**). We then identified TF occupancy sites within each DHS compartment using a combination of TF recognition sequences and transcription factor footprints (Neph et al., 2012c) and compared the proportion of simple autoregulating TFs targeting epigenetically stable DHSs with those targeting developmentally gained DHSs (**Methods**). In every cell type analyzed, developmentally stable DHSs were chiefly and preferentially bound by simple autoregulating TFs when compared to developmentally gained DHSs (**Fig. 7.M5A** and **Fig. 7.S5A**). Furthermore, DHSs that remained stable throughout the transition from hematopoietic progenitor cells to T- and B-lymphocytes were also preferentially targeted by autoregulating TFs (**Fig. 7.M5B**).

We extended this analysis further to encompass TFs involved in 2-node directed and 3-node directed loop network architectures, which enable indirect autoregulatory behavior. Within each cell type, we identified ~223 TFs involved in 2-node directed loops and ~336 TFs involved in 3-node directed loops. In every cell type analyzed, developmentally stable DHSs were preferentially and chiefly bound by 2-node- and 3-node directed loop TFs (**Fig. 7.M5C-F** and **Fig. 7.S5A**). These features were found for both developmentally stable distal DHSs and developmentally stable promoter-associated DHSs (**Fig. 7.S5B-C**).

Retrograde remodeling of the regulatory DNA landscape during oncogenesis

We next sought to explore how the epigenetic landscape reorganizes during a major pathological deviation. To address this, we produced genome-wide DNaseI hypersensitive maps from 21 diverse cancer cell lines and purified sorted cells from two primary malignancies (two subtypes of acute myelogenous leukemia arising in different unrelated individuals) (**Fig. 7.M6A**), defining between 74,292 and 209,903 autosomal DHSs per cancer cell type. To compare the accessible chromatin landscapes of cancerous cells with those of normal cells we performed principal coordinate analysis of DNaseI hypersensitivity landscapes. Whereas normal developmentally distinct cell types are clearly separated (**Fig. 7.M1C** and **Fig. 7.M6B**), the accessible chromatin landscapes of cancer cells converged on that of ES cells (**Fig. 7.M6B**). Hematological malignancies were a notable exception, forming a distinct group toward the ES-facing pole of the hematopoietic lineage cluster (**Fig. 7.S6A**). This result is consistent with results showing gene expression profiles of leukemic cells are highly similar to their normal counterparts (Krivtsov et al., 2006).

To quantify retrograde remodeling of the regulatory DNA landscape during oncogenesis, we focused our analysis on four cancer types for which DHS maps were available for the presumed corresponding normal precursor cell type (mammary epithelium for breast cancer (2 types); T cells for T cell leukemia; melanocytes for melanoma). Compared with normal precursors, all four cancer cell types exhibited substantial reorganization of the accessible chromatin landscape in a manner largely specific to each cancer type (**Fig. 7.M6C,D** and **Figs. 7.S6B,C**). This reorganization has three major components: (i) reactivation of silenced ES cell DHSs; (ii) ectopic activation of DHSs from early programs, chiefly of cell lineages different than the cognate normal cell from which malignancy arises; and (iii) appearance of novel DHSs not

detected in other cell types. Of these, the first two categories accounted for the vast majority of the observed alterations, with ~90% of the DHSs activated during oncogenesis observed at some other point during normal development other than in the specific lineage giving rise to the cancer precursor cell type (**Fig. 7.M6D**). Moreover, we noted the appearance of a significant number of DHSs shared with ES cells that were different from the ES-shared sites present in the normal precursor cells (**Fig. 7.M6E**), in stark contrast to what was observed during normal differentiation (**Fig. 7.M3E**). Notably, this reactivated ES DHS compartment differed substantially between different cancer cell types (**Fig. 7.S6C**), analogous to the differences in the ES-shared DHS compartment between normal cells. Together these results indicate that oncogenesis is characterized by both retrograde remodeling of the regulatory DNA landscape by reactivation of silenced ES DHSs, as well as the aberrant co-option of regulatory DNA from alternative lineage paths, with each cancer cell type accomplishing this end using a largely distinct set of elements.

Transcription factor drivers of cancer regulatory landscapes

The ability of some transcription factors to function as oncogenes or tumor suppressors is well known (Persson and Leder, 1984). We therefore sought to identify TFs that potentiate reorganization of the accessible regulatory DNA landscape during oncogenesis by focusing on DHSs arising or disappearing compared with a presumed normal cell precursor. This analysis revealed significant enrichment of recognition sequences for known oncogenic TFs in cancer-arising regulatory DNA. In parallel we observed enrichment of known tumor suppressor TF binding sites in DHSs from normal cell counterparts that were not shared with cancer cells. For example, the recognition sequence landscape of the breast cancer oncogene FOXA1 is

specifically and significantly expanded compared with normal breast epithelium (**Fig. 7.M7A**), consistent with its role in mediating estrogen receptor-dependent chromatin remodeling in breast cancer (Carroll et al., 2005; Hurtado et al., 2011). Furthermore, the target landscape of the melanocyte factor SOX9 within regulatory DNA is specifically reduced/contracted in melanoma cell compared with melanocytes, consistent with its role as a potent melanoma tumor suppressor (**Fig. 7.M7A**) (Cook et al., 2005; Passeron et al., 2007, 2009). This analysis revealed a variety of TFs with similar patterns of expansion or contraction of regulatory recognition space in normal vs. tumor cells (**Fig. 7.M7A** and **Fig. 7.S7A**), potentially exposing novel cell-selective oncogenic and tumor suppressor roles for such factors.

Functionally distorted memory of normal lineage programs in malignant cells

Next we asked whether the transformed regulatory landscapes of cancer cells maintained systematic memory of earlier developmental fate decisions, akin to normal cells (**Fig. 7.M1**). Clustering 23 cancer cell types based on their patterning of DHSs resulted in well-defined clusters that were predominantly typified by the functional characteristics of the cancers, rather than their developmental origin (**Fig. 7.M7B**). For example, hormone responsive cancers (LNCap, T-47D and MCF-7) formed a tight cluster, distinct from those of other adult and pediatric solid cancers, germ cell neoplasms, and hematological malignancies. These findings suggest that the oncogenic transformation of the regulatory DNA landscape is accompanied by the loss of developmental information and is dominated by selective activation of regulatory elements associated with the derived phenotype of the cancer.

Reduced evolutionary pressure on cancer regulatory landscapes

To investigate whether regulatory DNA activated during oncogenesis is under similar selective constraint within human populations with that of normal developmentally-patterned elements, we compared the extent of human nucleotide diversity (pi) between these classes of DHSs. Nucleotide diversity calculated using high-quality genomic sequence data from multiple unrelated individuals provides a quantitative assessment of the extent of ongoing purifying selection at DHSs within the human population (Vernot et al., 2012; Thurman et al., 2012; Neph et al., 2012c). Analysis of DHSs gained during oncogenesis revealed significantly higher nucleotide diversity compared with those retained from normal development (**Fig. 7.M7C**). Furthermore, regulatory DNA arising during oncogenesis had significantly higher nucleotide diversity than normal developmental elements (**Fig. 7.S7B-C**). Of note, cancer cells do not typically have more DHSs than their normal precursor cell types (**Fig. 7.M6D**). As such, this finding cannot be explained by positing a hyperactive regulatory state in which new unselected elements with e.g. low-affinity TF binding sites are being activated in large numbers. Rather, the results suggest that cancer cells selectively recruit regulatory elements that are active at other developmental stages, yet are under less selective pressure in the human population. Such reactivation events are likely generated through the mis-regulation of key developmental TFs (**Fig. 7.M7A**), and can have a large effect on the expression of neighboring genes (Akhtar-Zaidi et al., 2012). Consequently, our results suggest that oncogenesis favors the recruitment of functionally advantageous regulatory DNA that likely plays a secondary role in normal developmental processes.

7.4 – DISCUSSION

The salient findings can be recapitulated as follows: First, early developmental fate decisions and enhancer activity states can be derived from the genomic patterning of DNaseI-hypersensitive sites (DHSs) in definitive cells, distinct from information conveyed by gene expression. Second, development along multiple lineages is associated with three features: (i) extensive propagation of DNaseI hypersensitivity at regulatory DNA; (ii) monotonic pruning of DHSs shared with ES cells; and (iii) blossoming of a smaller number of lineage-restricted regulatory elements – all resulting in a more restricted, specialized accessible chromatin landscape. Third, the regulatory DNA landscapes of terminally differentiated cells retain a nearly constant proportion of DHSs shared with ES cells. Fourth, the strength of evolutionary conservation at regulatory DNA is maximal at elements shared among all derivatives of a major lineage (mesoderm) relative to both earlier (ES) and later branches. Fifth, developmentally stable DHSs are chiefly occupied by self-regulating transcription factors. Finally, in contrast to normal cells, cancer regulatory landscapes feature both extensive reactivation of silenced ES cell DHSs and ectopic activation of regulatory DNA from alternative developmental programs external to the cell lineage from which the malignancy derives.

We interpret the above findings to signify a central role for DHS patterning in propagating cellular state and fate information during development that is abrogated by oncogenesis. Below we place the findings in historical context, and consider both the features of differentiation that Waddington presaged, as well as a number of novel and telling insights our results afford into basic developmental mechanisms and strategies.

From epigenesis to epigenetics

The generation of consistent body plans by the sequential differentiation of totipotential material is a foundational conceptual paradigm for development. First articulated by Aristotle (*De generatione animalium 739a*), this concept was termed 'epigenesis' (literally, 'moving toward coming into being') by Harvey to distinguish it from preformationism (Harvey, 1651). By the early 20th century, the concept that animal development was epigenetic in character was widely accepted (Patten, 1920). Waddington's enduring epigenetic landscape paradigm crystalized this concept and added two important features (Waddington, 1940, 1957). The first synthesized a series of observations by many investigators concerning the stability of cellular phenotypes. As a cell negotiates the epigenetic landscape, it passes through various valleys and pauses at "states of competence", representing branching points within the landscape. The valleys act to guide the ball down a particular "pathway of change which is equilibrated in the sense that the systems tends to return to it after disturbance" (Waddington, 1957). The second feature Waddington introduced was mechanistic: the proposition that the epigenetic landscape itself is controlled by a complex system of "regulatory genes" that interact with one another in a combinatorial and temporally-coordinated fashion to shape the topography of epigenesis (Waddington, 1957).

In view of our results, Waddington's paradigm appears to provide a remarkably prescient schematization the transformation of the regulatory DNA landscape during development and differentiation. The analysis of systematic alterations in regulatory DNA accessibility afforded by global maps of DHSs has now introduced a missing quantitative dimension for major facets of epigenesis.

DHSs as epigenetic signposts

The fact that our results were derived exclusively by analysis of the patterns of DNaseI hypersensitive sites without reference to other data types has important implications for understanding both central mechanisms of differentiation and the roles of other accompanying changes in the chromatin landscape. The fact that proper historical lineage branching relationships can be recovered from DHS data but not from gene expression data alone suggests that the DHS compartment of both differentiating and definitive cells contains both developmentally dynamic sites involved in active expression of unique lineage features, and 'marker' or ‘memory’ elements that preserve information about prior developmental states (e.g. in the form of persistent DNaseI hypersensitivity at tissue-selective early developmental enhancers). The direct role of TF binding in underpinning DHSs thus indicates a defining role for TFs in both phenomena. The DHS landscapes of *in vivo* differentiated cells are highly reproducible and thus extremely stable. While the precise molecular mechanism underlying the simple mitotic propagation of DHSs has not been definitively elucidated, we have every expectation that the same mechanism should be operative during the mitotic cascade of development. TFs rapidly re-associate with daughter DNA strands following replication and can bookmark regulatory DNA through mitosis (Egli et al., 2008; Kadauke et al., 2012). As such, multiplex TF binding within regulatory DNA marked by DNaseI hypersensitivity is mechanistically well-positioned to convey information to daughter cells without invoking other features of the chromatin template.

Developmental transformation of the regulatory DNA landscape

Persistence of DNaseI hypersensitivity at regulatory DNA appears to be a pervasive feature of developing chromatin landscapes. A consistent finding in our analysis was that

development and differentiation, irrespective of lineage, balance the propagation, extinction, and *de novo* activation of chromatin accessibility at regulatory DNA in a highly formulaic manner. Of these forces operating on the regulatory landscape, programmed extinction of DHSs appears to play the largest role. As a cell progresses through development, its accessible chromatin landscape invariably becomes progressively smaller and more restricted, driven by the pruning of early developmental regulatory sites and the blossoming of a smaller number of lineage-restricted DNA elements – metaphorically, a narrowing of the epigenetic landscape's valley floors (**Fig. 7.M8A**). Critically, this pruning process results in passive yet systematic repression of alternative lineage programs through the wholesale loss of DHSs associated with alternative fates (**Fig. 7.M4**), cementing or "canalizing" a particular developmental pathway. Our data thus suggest that during normal development the processes of differentiation and lineage commitment occur in lockstep.

It is notable that the same basic process unfolding across the regulatory landscape during the transition away from ES cells – restriction of the DHS landscape, persistence of memorized DHSs, acquisition of novel DHSs – is successively repeated during subsequent transitions from intermediate stages (e.g., hematopoietic stem cells, cardiac progenitors) on the way to fully differentiated cells. The directional evolution of the regulatory landscape during differentiation thus appears to employ an essentially recursive process, reminiscent of those modeled by classical finite state automata (Turing, 1937).

It has been observed that the proportion of the genome evincing repressive polycomb-associated histone modification (H3K27me3) increases during directed differentiation from ES cells to neuronal progenitors (Zhu et al., 2013), or more generally from primitive to definitive cells (Hawkins et al., 2010), chiefly in the form of large modified domains. However, the

developmental restriction of DHSs we report is far more extensive, and also affects elements within active regions of the genome that contain other DHSs in close proximity.

Relation to the developmental 'hourglass'

The concept that developmental biology can provide insights into evolution was first recognized by Darwin (Darwin, 1859), who expanded on Von Baer's observation that, following convergence on a common form during the pharyngula stage of mid-embryogenesis, embryos of different species successively diverge in a species-specific manner (Von Baer, 1828). This point of maximal convergence has been termed the 'phyletic' or 'phylotypic' stage (Cohen, 1963; Sander K., 1982), and coincides with the activation of Hox genes and other major transcriptional regulators of development (Duboule, 1994). These observations have given rise to an 'hourglass' model of development (Duboule, 1994; Raff, 1996), now supported further by analyses of embryonic gene expression patterns in flies (Kalinka et al., 2010; Domazet-Lošo and Tautz, 2010), worms (Levin et al., 2012), fish (Domazet-Lošo and Tautz, 2010), frogs, birds, and mice (Irie and Kuratani, 2011), all of which show maximal similarities at the time of appearance of pharyngeal arches.

Our finding that the evolutionary constraint of regulatory DNA shared among mesodermal lineages is higher than for elements arising either earlier or later during development (**Fig. 7.M1D**) accords with this model. As such, the results raise the possibility that the 'hourglass' phenomenon may be grounded within discrete sets of regulatory DNA regions, the study of which may yield further insight into this process. Although shown for mesoderm, the same pattern presumably exists in endo- and ectodermal lineages, awaiting only the fleshing out of these branches with additional definitive lineage exemplars.

Temporal 'memory' of cellular state and fate

As a cell differentiates, it retains a large subset of DHSs from its immediate progenitors many of which can be traced back to totipotential (ES) cells (**Fig. 7.M8B**), and are populated chiefly by autoregulatory transcription factors (**Fig. 7.M5**). This suggests that perpetuation of regulatory states embodied in DHSs is largely achieved through feedback circuits such as positive feedback loops, which are found in a variety of biological contexts involving propagation of regulatory information (Iliopoulos et al., 2009). Surprisingly, the proportion of developmentally stable sites in the terminal (definitive) regulatory DNA landscapes of all different cell types remains remarkably constant, irrespective of the size of that landscape: each definitive cell type retains roughly 1/3rd of its regulatory DNA landscape from ES cells (**Fig. 7.M3F-I**). The particular combination of such developmentally stable sites is unique for each definitive cell type, and effectively encodes a history of prior developmental fate decisions that gave rise to it.

A further remarkable feature of the developmentally stable DHS compartment is its almost linear association with developmental time or cellular maturity. As cells become more differentiated, they prune ES cell-originated DHSs from their landscapes in a progressive, clock-like fashion. As such, for a given cell population, simply measuring the proportion of DHSs within its regulatory landscape that are shared with a fixed reference from ES cells provides a quantitative measure of developmental maturity (with ~1/3rd comprising the lower bound, marking a fully differentiated cell).

As such, in view of the above, the widespread persistence of DNaseI hypersensitivity at regulatory DNA may serve two intertwined purposes, providing both a type of memory of earlier cell states as well as a template for marking cellular maturity and thus developmental ‘time’.

Role of TFs in propagation of active vs. repressive chromatin states

The fact that proper developmental lineage relationships can be recovered by analysis of DHS patterns but not gene expression patterns suggests that developmentally persistent DHSs are not actively involved in expression regulation in terminal cells, even though such elements may function actively as enhancers in earlier stages. Whether or how the cell utilizes the information maintained in developmentally stable DHSs remains an open question. Irrespective of whether they maintain active function in controlling gene expression, the prevalence of developmentally stable DHSs appears to generalize the 'active' chromatin perpetuation phenomenon described by Weintraub (Groudine and Weintraub, 1982; Burch and Weintraub, 1983; Weintraub, 1985).

A key feature of the active, DHS-centric process we document is its direct links to transcription factors acting through specific cis-regulatory elements. Our data strongly indicate that changes in accessible regulatory DNA during development and differentiation are orchestrated by specific combinations of lineage restricted transcription factor genes that act differentially on distinct cohorts of DHSs (**Fig. 7.M4**). In combination with the finding that developmentally stable DHSs are occupied chiefly by autoregulatory transcription factors (**Fig. 7.M5**), this result emphasizes the central influence of the cellular transcription factor regulatory network on modeling the accessible chromatin landscape across developmental time (**Fig. 7.M8C-D**).

In this context it is notable that other examples of 'epigenetic memory' which have come to light feature propagation of chromatin states through primarily repressive mechanisms such as CpG methylation (Bird, 2002; Kim et al., 2010; Lister et al., 2011; Hackett et al., 2013), histone H3K9 trimethylation (Hathaway et al., 2012) or polycomb (Cavalli and Paro, 1998). In an

elegant experiment conceptually analogous with (Groudine and Weintraub, 1982), Hathaway et al. recently reported the first clear evidence for epigenetic mitotic perpetuation of a repressive histone modification (H3K9me3) (Hathaway et al., 2012). However, differences in genomic H3K9me3 patterns provide relatively limited discrimination of cell types compared with active chromatin features (Zhu et al., 2013), and cannot be connected directly with our observations concerning the remodeling of the DHS landscape.

Connecting epigenesis and oncogenesis

Our results indicate that oncogenesis is attended by drastic remodeling of the accessible chromatin landscape, resulting in a loss of developmental information, and a reversion to a 'pseudo-primitive' state that combines regulatory DNA features of ES cells with those of other developing lineages (**Fig. 7.M6** and **Fig. 7.M7**). This state is not truly *de*-differentiated – which implies walking back along a path previously taken – but rather *dys*differentiated. Importantly, we find no regulatory elements in common among cancer cells that are not found in normal cells. Rather, malignant cells are mainly characterized by novel combinations of re-activated and cross-activated (ie, extra-lineage) primitive and definitive DHSs that differ for each cancer type. As such, cancer cells encompass a multidimensional deviation from normal development, and can no longer be placed on Waddington's landscape.

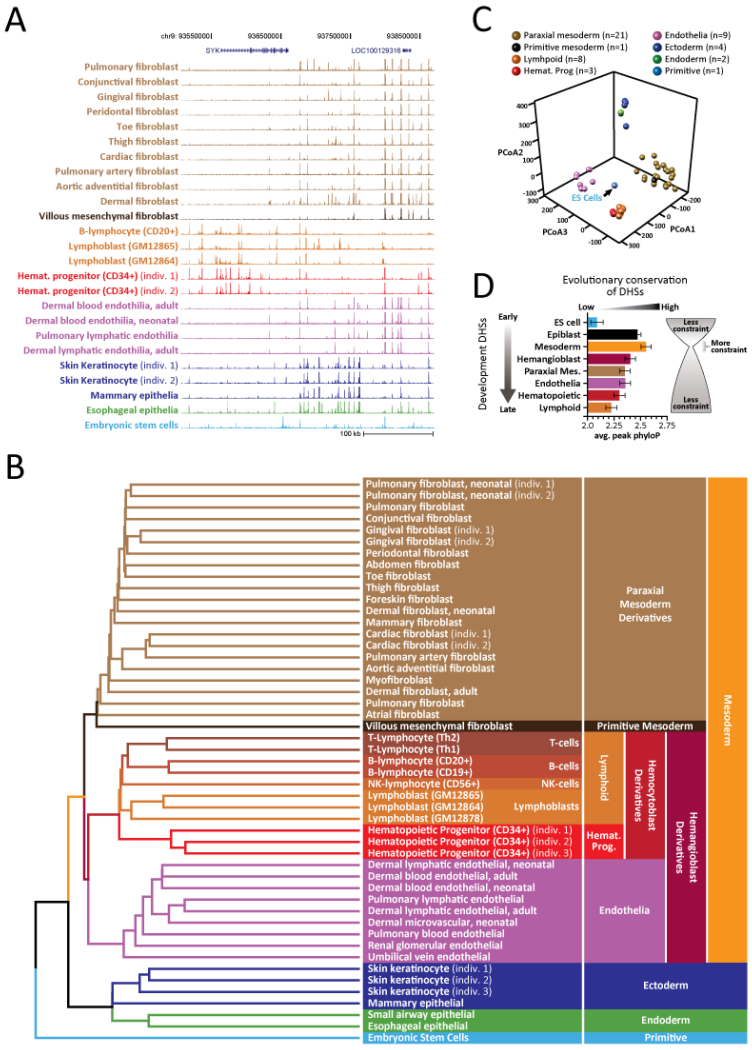
The acquisition of regulatory DNA that is specifically beneficial for the oncogenic phenotype appears to be achieved mainly through the aberrant co-option of 'normal', regulatory elements from alternative lineage paths (**Fig. 7.M6D**). This feature may explain the long-standing observation that oncogenesis is accompanied by the reappearance of fetal antigens (Hirszfeld et al., 1932; Tatarinov, 1964). Our results are also difficult to reconcile with models

of oncogenesis that posit cancer origins from developmental remnants – i.e., rare primitive cells distributed within otherwise differentiated tissue beds (Goldstein et al., 2010). If this were the case, we not expect to find the observed significant sharing of DHSs between cancer cells and their normal, terminally differentiated precursors. Alternatively, if cancer cells simply arose from uncontrolled proliferation of a more primitive cell remnant, we would expect cancer cells to retain strong lineage signatures; however, with the exception of hematological malignancies, this is decidedly not the case.

The actions of transcription factor oncogenes and tumor suppressors are directly represented in the DHS landscapes of cancer cells (**Fig. 7.M7A**). Importantly, this phenomenon should enable the direct discrimination of cell-selective tumor suppressors from oncogenes. For example, our analysis correctly implicates SOX9 as a melanoma tumor suppressor (**Fig. 7.M7A**), though it was initially falsely labeled a melanoma oncogene (Tani et al., 1997; Cook et al., 2005; Passeron et al., 2007, 2009). Despite the growth and phenotypic advantages bestowed by the transformed regulatory DNA landscape, malignant cells have likely lost many of the beneficial regulatory redundancies and feedback mechanisms formed during normal development that maintain epigenetic stability in the face of environmental and genetic stress (**Fig.7.M 8D**). It is tempting to speculate that this patchwork reorganization of the chromatin landscape during oncogenesis may expose exploitable vulnerabilities of the malignant state.

7.5 – FIGURES

Figure 7.M1. Lineage programming of human regulatory DNA



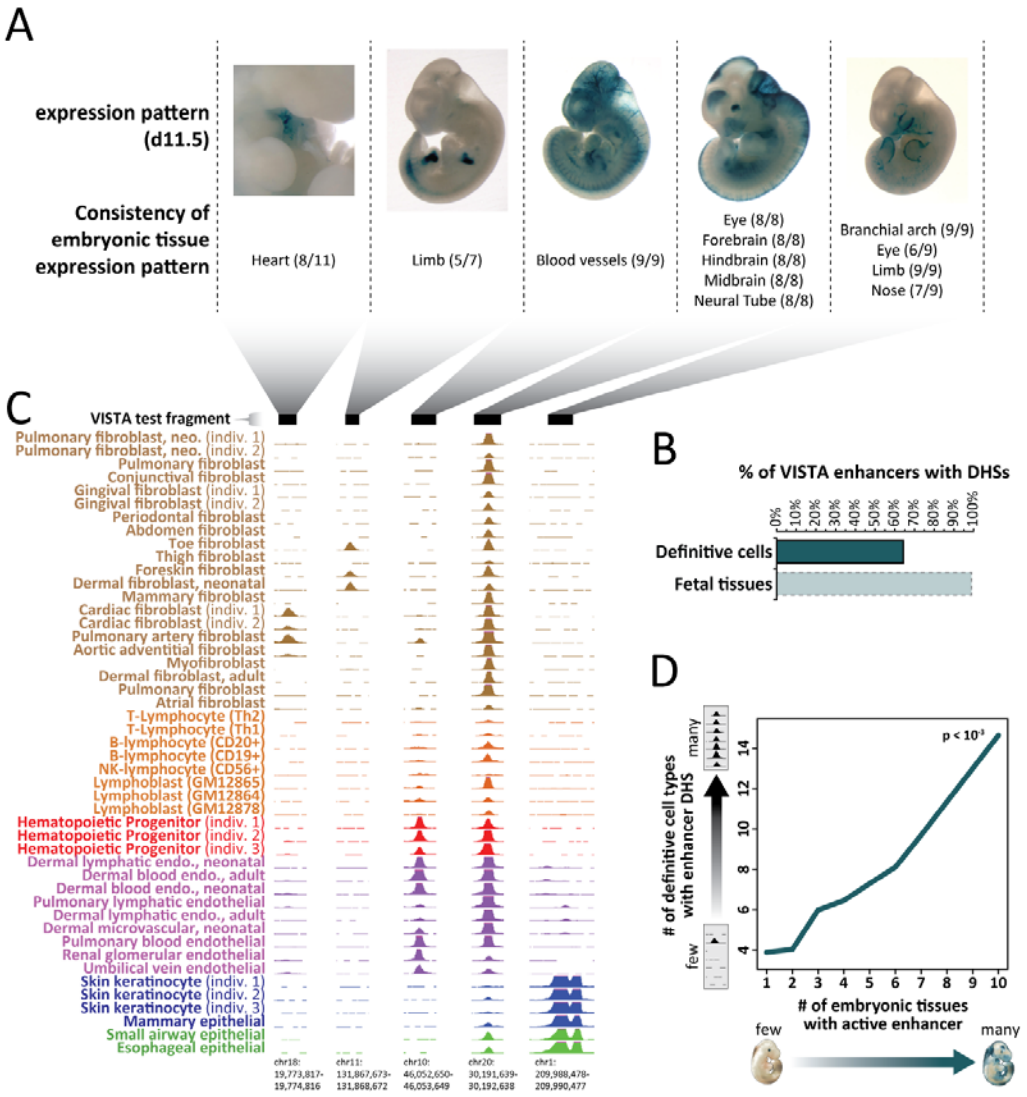
(A) *Evidence of lineage patterning in primary DHS data*. DNaseI cleavage density profiles for 24 exemplary primary human cell types and ES cells across a ~350kb region along chromosome 9. Cell types are colored according to their embryological derivation as indicated in (B).

(B) *Clustering DHS profiles recovers precise embryological relationships.* Unbiased clustering (nearest-neighbor) of the linear patterning of DHSs (present/absent) from 48 diverse, definitive cell types plus embryonic stem cells. Branches and cell types are colored according to their embryological origin, with embryological ancestors common to multiple cell types indicated on the right. Note the rooting of the tree by ES cells and the partitioning of major branches corresponding to the trilaminar embryo. Note also the demarcation of early fate decisions such as partitioning of hemangioblast derivatives into endothelia and blood. Further note, the grouping of anatomically related cell types with common origins such as the heart and great vessel derivatives within paraxial mesoderm.

(C) *Principal coordinate analysis of cell type relationships.* Principal coordinate (PCo) analysis of DHS relationships for each cell type. Shown is each cell type from B projected into a three-dimensional principal coordinates space (PCo1-PCo3). Cell type coloring is indicated above. Note the centrality of ES cells, the grouping of cells by embryological orgin, and the spatial separation of major lineage groups.

(D) *'Hourglass' pattern of regulatory DNA conservation across the developmental spectrum.* Shown is the mean evolutionary conservation (phyloP, x-axis) of DHSs common to the indicated lineage branches. Error bars represent 95% confidence intervals.

Figure 7.M2. Developmental persistence of chromatin accessibility at primitive enhancers



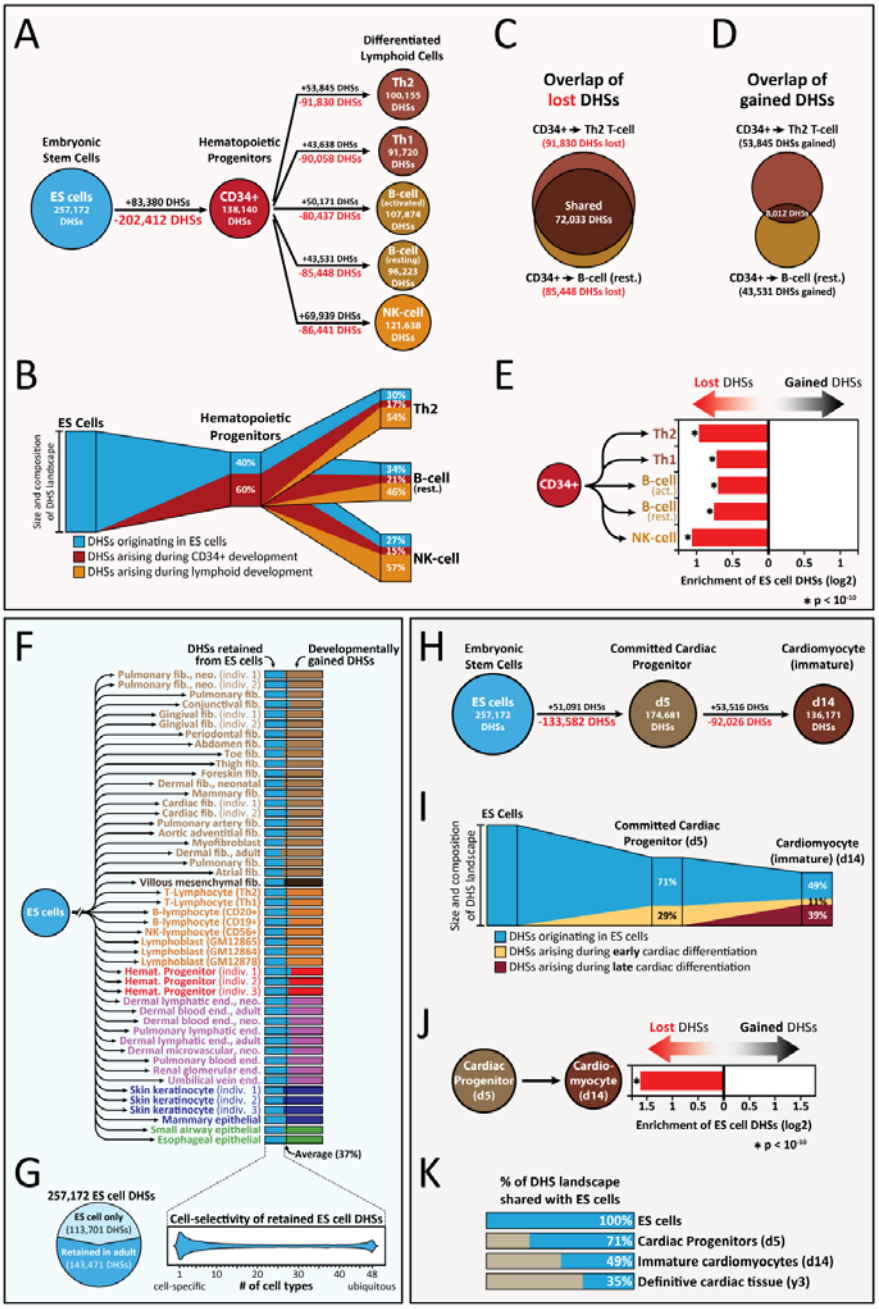
(A) *Enhancer tissue activity spectra in developing embryos (e11.5).* Embryonic tissue activity spectra (LacZ staining) of 5 representative transgenic human enhancer elements from the VISTA database assayed in whole day 11.5 embryos (Visel et al., 2007 NAR). Shown below each image are numbers of individual embryos with enhancer activity (staining) in the indicated anatomical structure.

(B) *Persistence of DNaseI hypersensitivity at embryonic enhancers.* Shown is the percentage of validated embryonic enhancers from the VISTA database (n=721) that are marked with DHSs in one or more definitive cell-types (top). Substantially all embryonic enhancers show DNaseI hypersensitivity in early human fetal tissues (~ day 70-150) (bottom).

(C) *Cross-cell type DHS patterns of primitive enhancers in definitive cells.* Shown is DNaseI hypersensitivity at five enhancer elements corresponding to (A) across 47 definitive cell types from figure 1. Note the relationship between the anatomical staining patterns in (A) and the cellular restriction (or lack thereof) of DNaseI hypersensitivity.

(D) *Embryonic enhancer tissue spectrum parallels DHS spectrum in definitive cells.* The number of embryonic tissues with an active enhancer by LacZ staining (x-axis) is almost linearly proportional (linear regression p-value<10e-3) to the number of definitive cell types showing DNaseI hypersensitivity at the same enhancer (y-axis) (n=721).

Figure 7.M3. Developmental extinction, maintenance and *de novo* activation of chromatin accessibility at regulatory DNA



(A) *Extinction vs. activation of DHSs during hematopoietic differentiation.* Shown for each developmental transition are the number of acquired (black) vs. extinguished (red) DHSs relative

to the prior cell state. Note the progressive contraction of the size of the regulatory DNA landscape from primitive to definitive cells.

(B) *Composition of developing hematopoietic regulatory landscapes.* Schematic illustrating the inherited vs. acquired compositions of the regulatory DNA landscapes in (A). The lymphoid DHS compartment colored blue comprises a strict subset of the hematopoietic progenitor subset colored in blue.

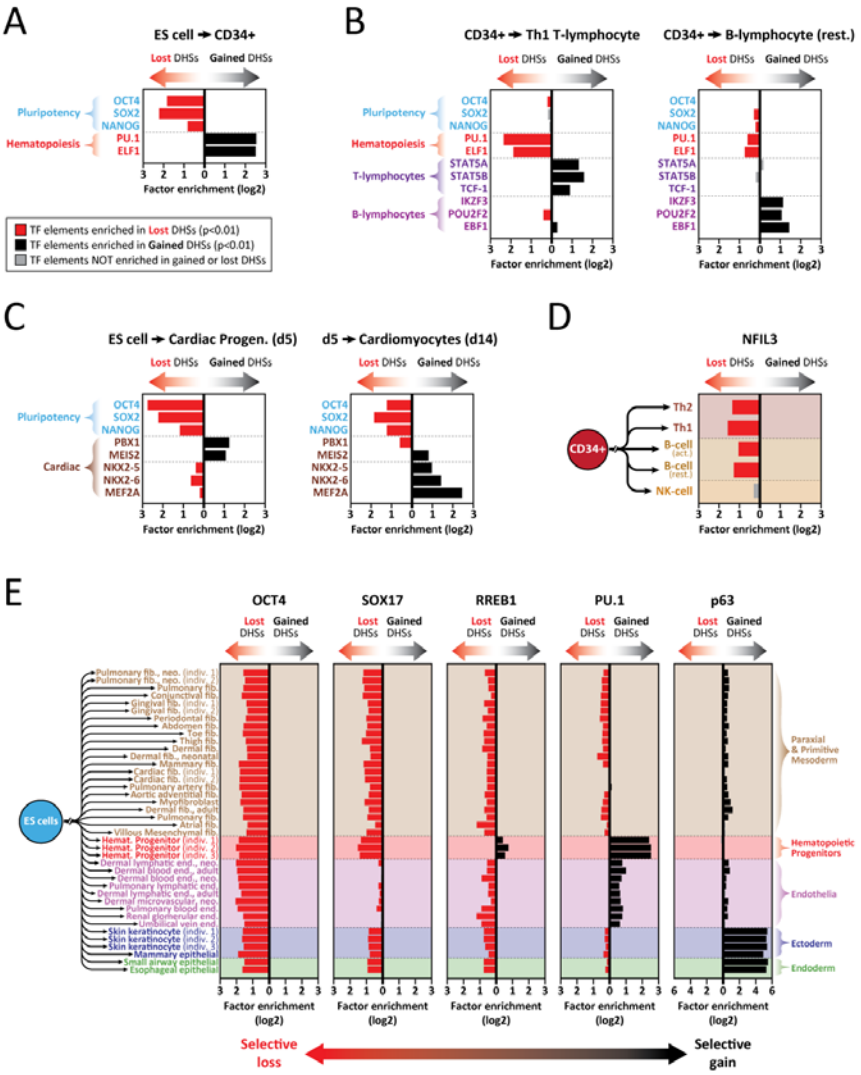
(C-D) *Common DHS extinction vs. lineage selective DHS activation.* DHSs lost (C) during transition of hematopoietic progenitors are highly similar, irrespective of lineage. By contrast, acquired DHSs (D) are highly cell type specific.

(E) *Preferential extinction of ES cell DHSs during later developmental transitions.* Shown is the enrichment of ES cell DHSs within lost vs. gained DHS populations. p-values were calculated using the hypergeometric test.

(F-G) *Formulaic composition of terminal regulatory DNA landscapes.* (F) Shown are the proportions of the DHS landscapes accounted for by DHSs shared by ES cells, vs. those acquired during development. In spite of the fact that the sizes of the regulatory DNA landscapes vary by greater than 2-fold (Supplemental Table 1), the proportion of DHSs shared with ES cells remains nearly constant at ~37%. However, each cell type inherits a different cohort of ES cell DHSs (G). ~60% of the ES DHS landscape persists in varying combinations in definitive cells (G).

(H-K) *Regulatory DNA landscape of cardiac differentiation.* (H) Acquired (black) vs. extinguished (red) DHSs relative to the prior cell state. (I) Schematic illustrating the inherited vs. acquired compositions of the regulatory DNA landscapes in (H). (J) Enrichment of ES cell DHSs within lost vs. gained DHS populations. p-value was calculated using the hypergeometric test. (K) Clock-like extinction of ES cell DHSs during cardiac differentiation.

Figure 7.M4. Selective loss vs. gain of DHSs targeted by lineage regulators



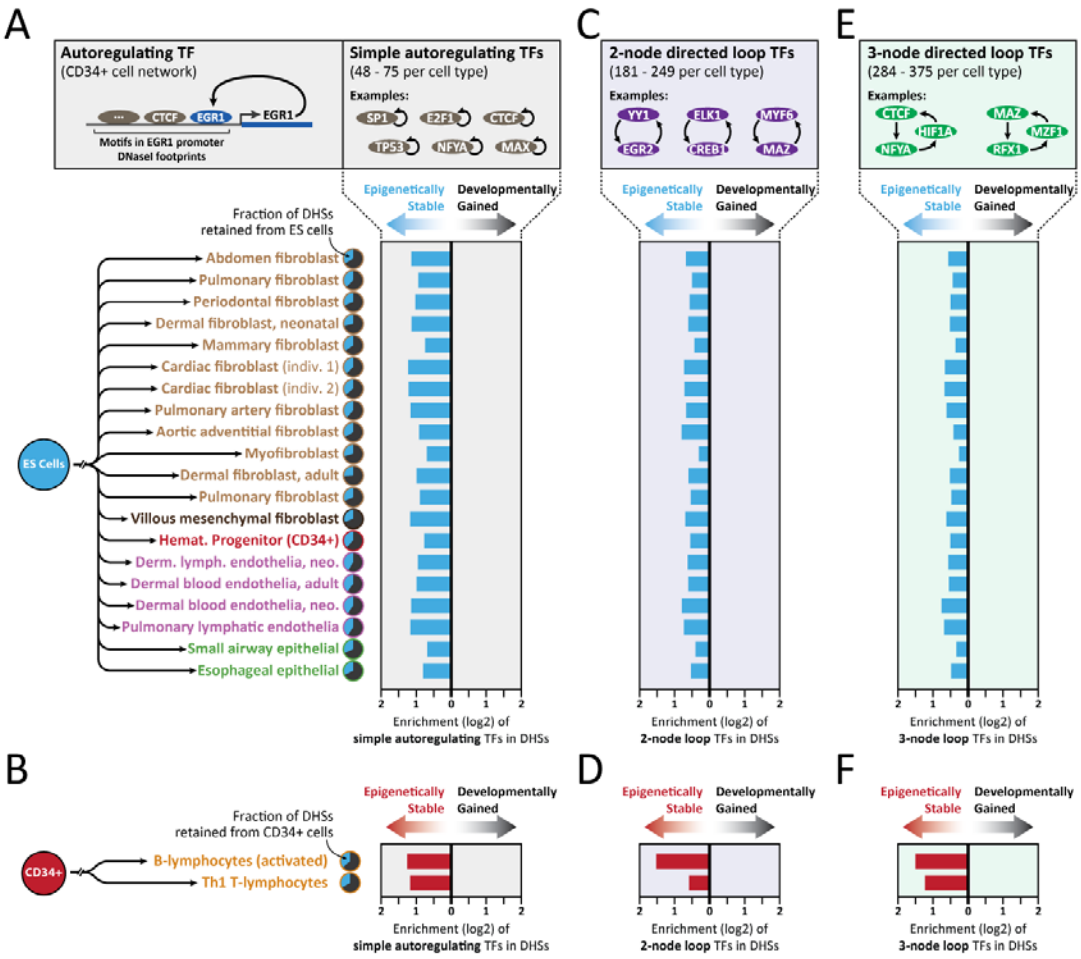
(A-C) *Enrichment of pluripotency vs. hematopoietic and cardiac lineage regulators in lost vs. gained DHSs.* Enrichment of recognition sequences for 3 pluripotency factors (blue), 8 hematopoietic lineage and sub-lineage regulators (red/purple), and 5 cardiac lineage regulators (brown) in DHSs lost vs. gained during (A) the differentiation of ES cells into CD34+ cells, (B)

the differentiation of CD34+ cells into Th1 T-cells or B-cells, and (C) the differentiation of ES cells into cardiac progenitors and immature cardiomyocytes.

(D) *Selective retention of regulatory DNA with lineage regulator recognition sites.* DHSs containing recognition sites for the NK-cell master regulator NFIL3 are selectively retained by NK cells and lost in other lineages.

(E) *Loss vs. gain of lineage regulator selective lineage compartments.* Shown is enrichment of recognition sequences for 5 pluripotency/lineage regulators in the DHSs gained vs. lost during lineage differentiation from ES cells. Lost DHSs are uniformly enriched in OCT4. DHSs containing recognition sequences for the endothelial master regulator SOX17 are selectively retained in endothelial cells. Recognition sites for RREB1, a regulator of the transition from primitive to definitive hematopoiesis, are enriched in DHSs lost from all lineages except hematopoietic, where they are selectively enriched in gained DHSs. DHSs gained in hemangioblast derivatives (blood/endothelia) are selectively enriched in recognition sites for the hemangioblast master regulator PU.1. Selective enrichment of p63 recognition sites in DHSs of ectodermal and endodermal derivatives.

Figure 7.M5. Epigenetically stable DHSs are potentiated by TFs that regulate their own expression

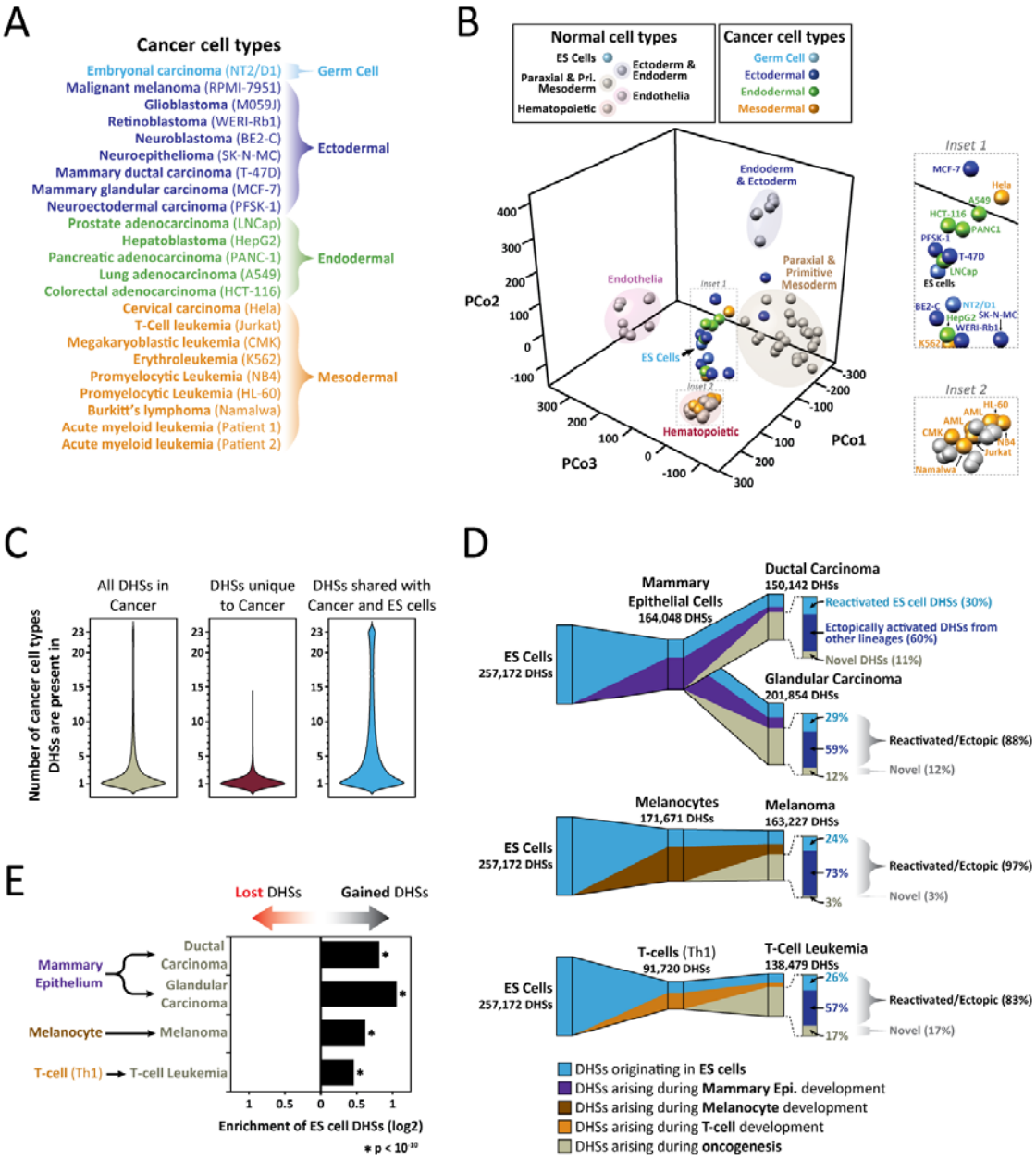


(A) Enrichment of simple autoregulatory TFs (computed from (Neph et al., 2012b) in epigenetically stable (i.e. inherited from ES cells) vs. developmentally gained DHSs in 20 cell types with annotated regulatory networks.

(B) Enrichment of simple autoregulatory TFs in epigenetically stable DHSs retained from CD34+ hematopoietic progenitors, excluding sites retained from ES cells.

(C-F) Enrichment of TFs involved in 2-node (C-D) or 3-node (E-F) autoregulatory loops in epigenetically stable vs. developmentally gained DHSs. All enrichments shown are significant ($p<10e-10$), hypergeometric distribution.

Figure 7.M6. Retrograde remodeling of the regulatory DNA landscape during oncogenesis



(A) Developmental origins of 21 cancer cell lines and two primary cancers analyzed using genome-wide DNaseI mapping.

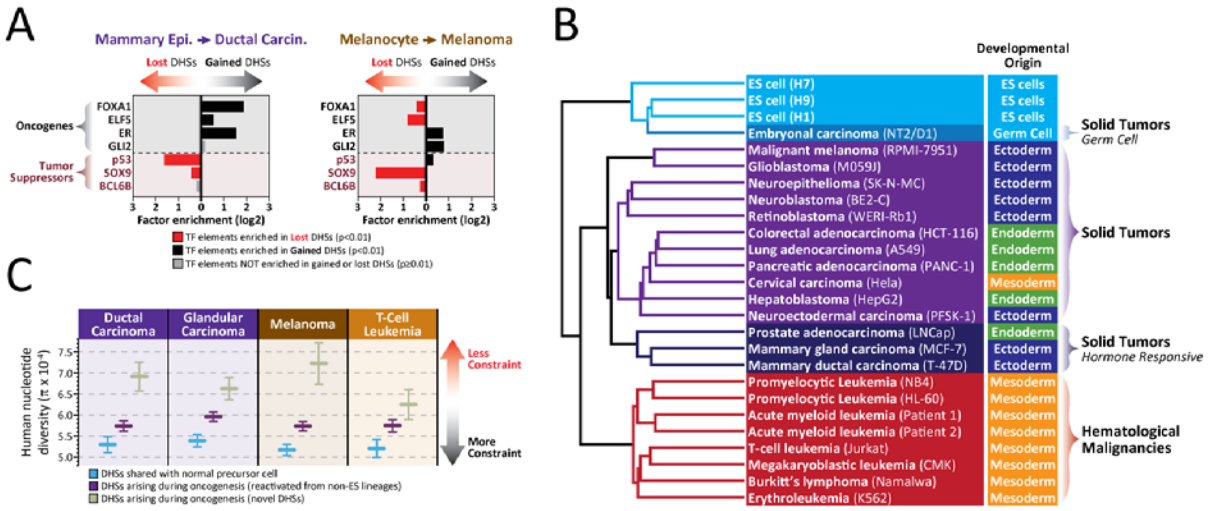
(B) *Principal coordinate analysis of normal vs. malignant cell type relationships.* Mapping of 23 cancer cell types within the principal coordinate framework of Figure 1C. Normal cell types are colored in solid grey and shaded according to developmental orgin. Cancer cell types are colored as indicated in (A). Note the prominent clustering of cancer cell lines around ES cells (inset 1).

(C) *Cell selectivity of cancer cell DHSs.* Distribution of the number of cancer cell types in which; (left) a DHS is observed; (middle) a DHS that is unique to cancer cell types, and not found in any of the normal cell types is observed; and (right), a DHS shared between a cancer cell type and ES cells in observed.

(D) *Oncogenic transformation is accompanied by extensive reactivation of ES cell DHSs.* Enrichment of ES origin DHSs in lost vs. gained DHSs during the oncogenic transformation of; (top) mammary epithelium into ductal and glandular mammary carcinoma; (second from bottom); melanocytes into melanoma; (bottom) Th1 T-cells into T-cell leukemia.

(E) *Disordered reactivation of ES DHSs and those from other lineages during oncogenesis.* (left) Shown is the contribution of non-malignant predecessors to the DHS landscape of each of 4 cancer cell types. Cancer DHS landscapes are partitioned into: DHSs originating in ES cells (blue); DHSs shared with their cognate normal predecessor, but not ES cells (purple/brown/orange); and DHSs arising during oncogenesis (grey). (right) Proportion of DHSs arising during oncogenesis that are reactivated (i.e. not inherited) ES cell DHSs (light blue), reactivated DHSs from other lineages (dark blue), or novel DHSs unique to each cancer cell type (grey).

Figure 7.M7. TF drivers, functional organization and evolutionary pressures on cancer regulatory landscapes



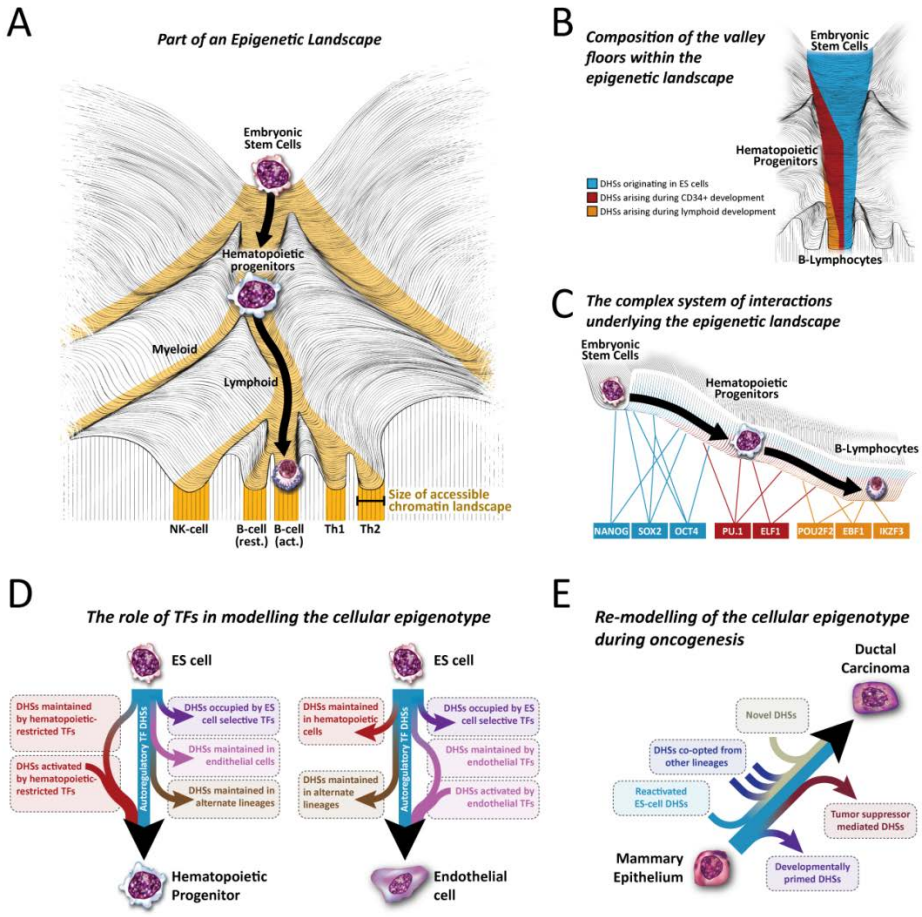
(A) *Oncogenic vs. tumor suppressor TF targets in DHSs gained vs. lost during oncogenesis.* Shown is enrichment of recognition sequences for 4 TF oncogenes and 3 tumor suppressor TFs in the DHSs lost vs. gained during the oncogenic transformation of mammary epithelium and melanocytes.

(B) *Functional cooption of the malignant regulatory DNA landscape.* Clustering of DHSs from 23 cancer cell types and three different ES lines (Euclidean distances, Ward clustering). Note the predominance of functional or phenotypic features over embryological origins.

(C) *DHSs arising during oncogenesis show relaxed evolutionary constraint.* Shown, for different cancer types, is human nucleotide diversity (π, y-axis) at DHSs that are either retained from normal precursors (blue), DHSs gained during oncogenesis but seen in some other normal cell type (purple), and DHSs gained during oncogenesis that are unique to cancer cells (grey). DHSs

shared with ES cells were removed from this analysis to control for the retrograde remodeling. Error bars represent 95% confidence intervals (obtained by bootstrap resampling).

Figure 7.M8. The epigenetic landscapes of differentiation and oncogenesis

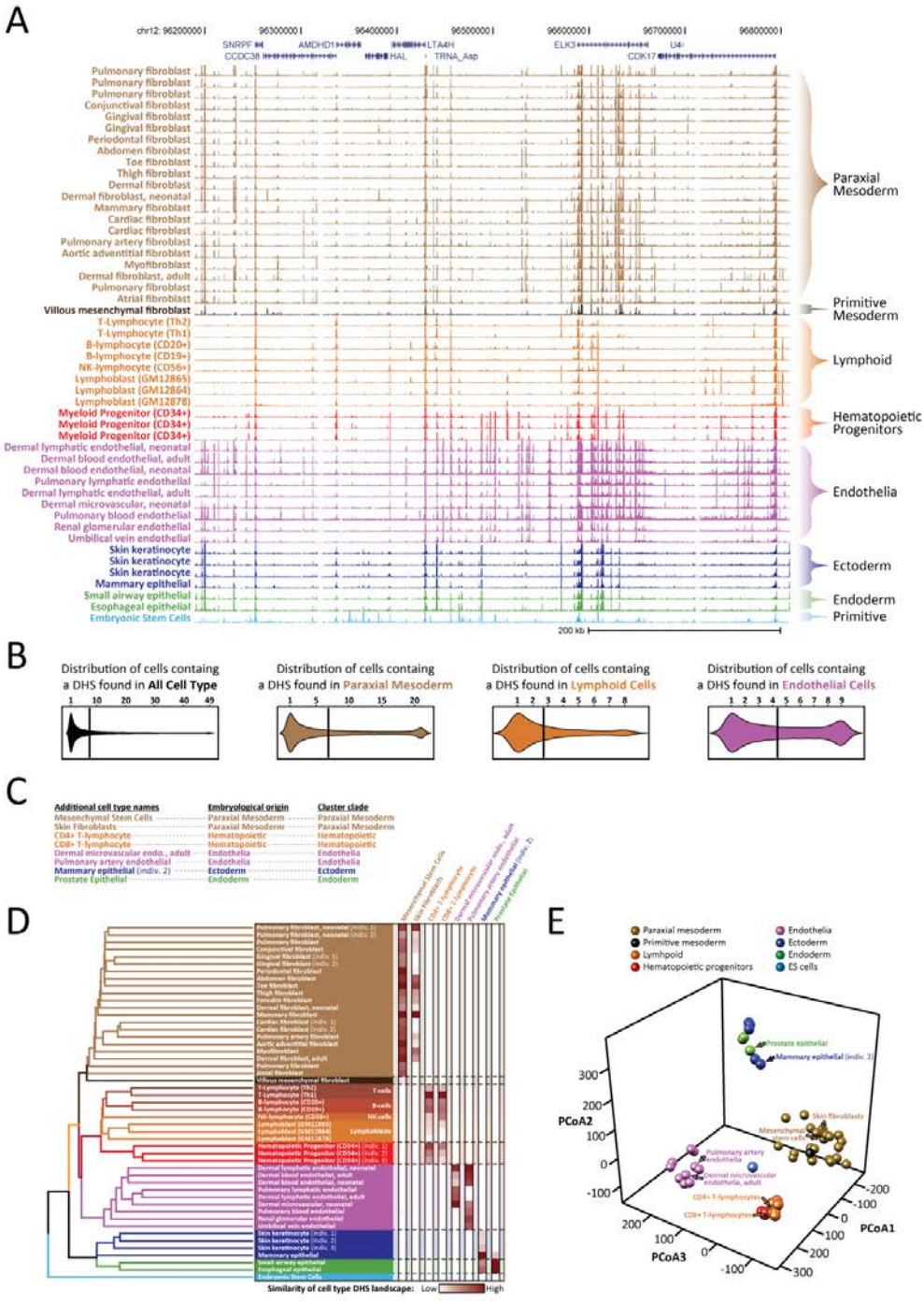


(A-C) *Waddington's epigenetic landscape revisited.* (A) Shown is part of an epigenetic landscape encompassing the differentiation of embryonic stem cells into defined hematopoietic lineages. Differentiation is accompanied by the progressive restriction of the size of the regulatory DNA landscape (gold). (B) Schematic showing restriction, perpetuation, and *de novo* activation that accompanies contraction in the overall size of the landscape during differentiation. (C) Sequential activation of transcriptional regulators underpin the topology and

trajectory of the epigenetic landscape (paralleling Waddington's regulatory genes and 'guy-ropes' (Waddington, 1956)).

(D-E) *Models of normal differentiation vs. oncogenesis.* (D) Shown are the actions of different TF classes on the developing regulatory DNA landscape of hematopoietic progenitors and endothelial cells. Each lineage is built around a distinct core of epigenetically stable DHSs driven by autoregulatory TFs. Gain of DHSs is driven by lineage selective TFs. Restriction of alternative cell fates is governed by selective loss of DHSs targeted by regulators of alternative lineage programs. (E) Remodeling of the regulatory DNA landscape during oncogenesis is dominated by the ectopic reactivation of ES cell DHSs and DHSs co-opted from diverse other lineage programs. DHSs lost during oncogenic transformation are populated by tumor suppressor TFs, while those gained are enriched in targets of oncogenic TFs.

Figure 7.S1. Lineage committed cells share a similar patterning of DHSs

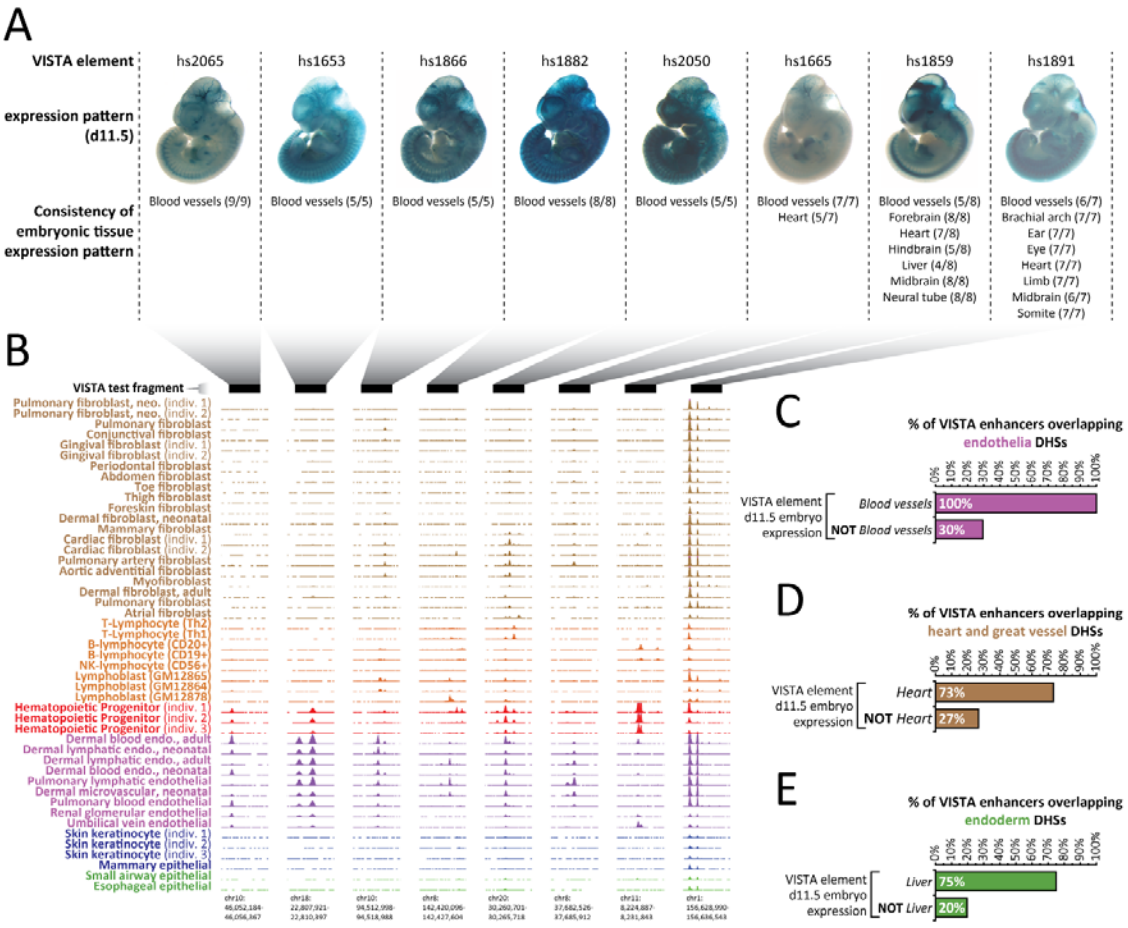


(A) DNaseI cleavage density profiles for 48 diverse, definitive cell types plus embryonic stem cells across a ~600kb region along chromosome 12. Cell types are colored according to their embryological derivation.

(B) Distributions of the number of cell types in which a given DHS is observed for; (left) all cell types (n=49); (middle-left) paraxial mesoderm cell types (n=21); (middle-right) lymphoid cell types (n=8); and (right) endothelial cell types (n=9). Width of each shape at a given y-value shows the relative frequency of DHSs present in that number of cell types.

(C-E) *Testing lineage robustness using 8 new cell types of diverse embryological origin.* (C) Shown are the names of 8 additional cell types used for this robustness analysis, as well as their embryological origin. Each cell type was individually clustered with the original 49 cell types, and the clade that the cell type fell within is indicated on the right. (D) The euclidian distance of the accessible chromatin landscape of the 8 additional cell types versus all 49 original cell types as measured by overlapping DHSs. (E) Principal coordinate (PCo) analysis of DHS relationships for each of the 8 additional cell types and the 49 original cell types. Cell type coloring is indicated above. Each of the 8 additional cell types are labeled with arrows and names.

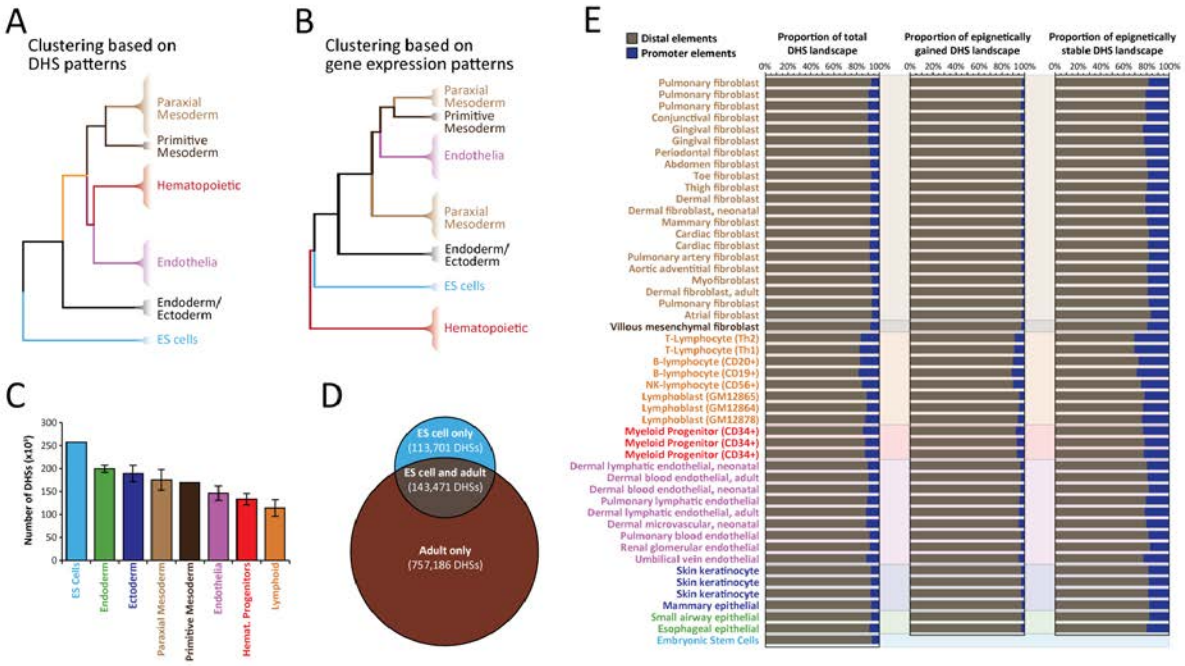
Figure 7.S2. Developmental persistence of chromatin accessibility at primitive enhancers



(A-B) *Blood vessel enhancer tissue activity in developing embryos (e11.5) foretells endothelial DHS activity in adult humans.* (A) Embryonic tissue activity spectra (LacZ staining) from all 8 transgenic human enhancer elements with a blood vessel staining pattern. Shown below each image are numbers of individual embryos with enhancer activity (staining) in the indicated anatomical structure. (B) Shown is DNaseI hypersensitivity at all 8 enhancer elements corresponding to (A) across 47 definitive cell types from figure 1.

(C-E) *Persistence of DNaseI hypersensitivity at embryonic enhancers.* (C) Shown is the percentage of validated embryonic enhancers from the VISTA database with blood vessel staining (*Blood vessels*), and without blood vessel staining (*NOT Blood vessels*) that overlap a DHS in human endothelial cells. (D) Shown is the percentage of validated embryonic enhancers from the VISTA database with heart staining (*Heart*), and without heart staining (*NOT Heart*) that overlap a DHS in human paraxial mesoderm cells. (E) Shown is the percentage of validated embryonic enhancers from the VISTA database with liver staining (*Liver*), and without liver staining (*NOT Liver*) that overlap a DHS in human endoderm cells.

Figure 7.S3. The accessible chromatin landscape of a cell encodes a history of prior developmental states



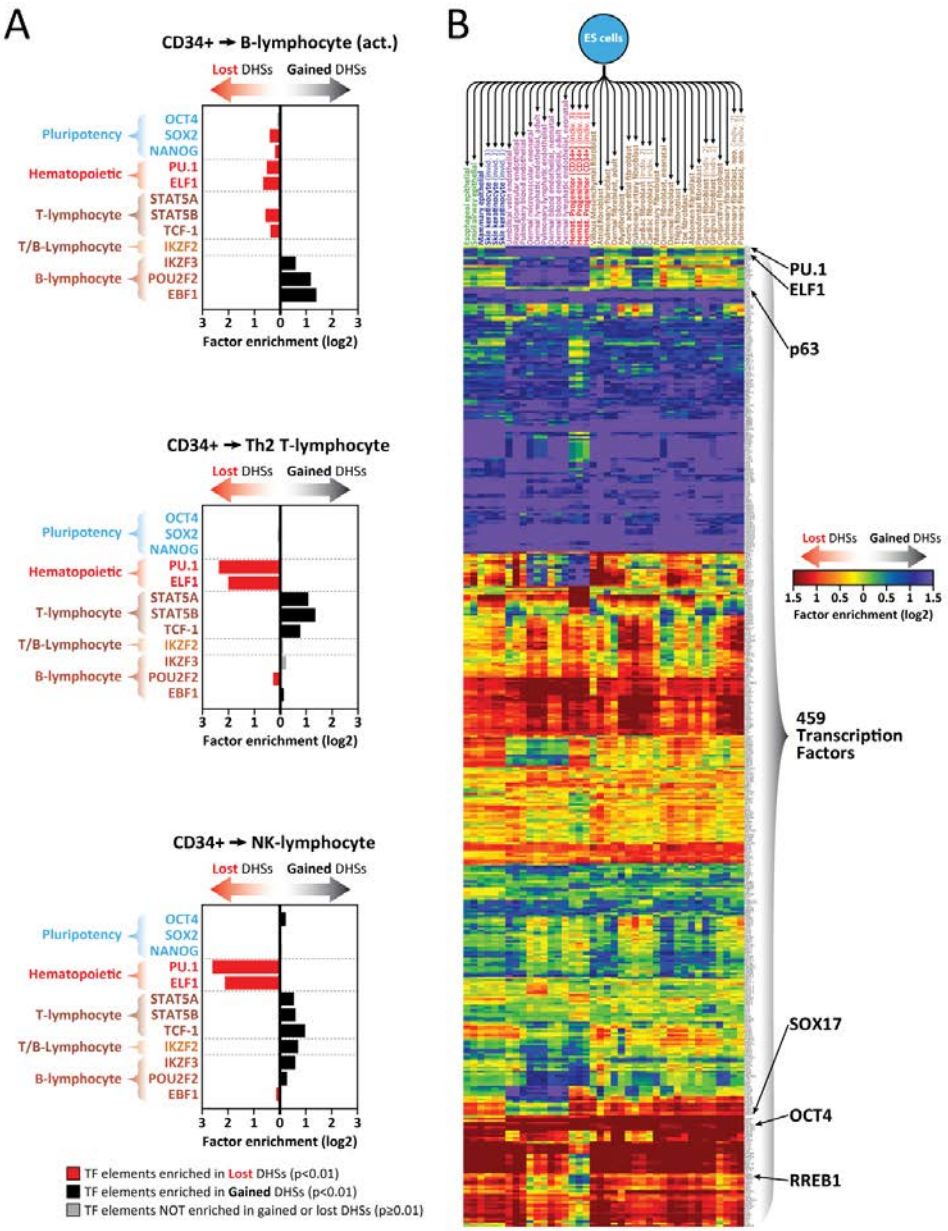
(A-B) *Clustering of gene expression patterns fails to recover correct early lineage branching relationships.* (A) Unbiased clustering (nearest-neighbor) of the linear patterning of DHSs (present/absent) from 48 diverse, definitive cell types plus embryonic stem cells as shown in Figure 1B. Individual cell types clustering together were collapsed according to their embryological origin. (B) Unbiased clustering (nearest-neighbor) of the gene expression patterns from 45 of the cell types shown in Figure 1B for which gene expression data were available. Individual cell types clustering together were collapsed according to their embryological origin. Note that the tree is no longer rooted in ES cells and the developmental lineage relationships are distorted, broken or lost.

(C) The total number of DHSs observed in cell types of different embryological orgin. Bars represent the mean +/- the standard error of the mean (SEM).

(D) Comparison of the accessible chromatin landscape of embryonic stem cells and adult cell types (49 cell types). Note that 59% of ES cell DHSs are seen in at least one adult cell type.

(E) *Composition of a cell types regulatory landscape.* The proportion of total (left), epigenetically gained (middle) and epigenetically stable (right) DHSs that are either proximal or distal to a transcriptional start site.

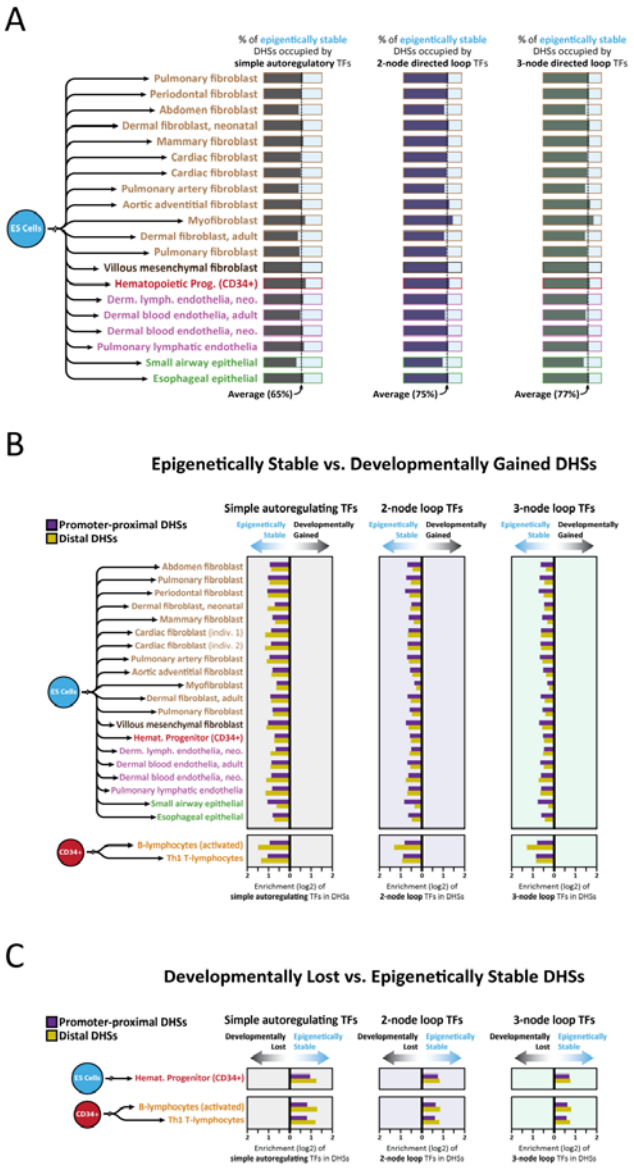
Figure 7.S4. Selective loss vs. gain of DHSs targeted by lineage regulators



(A) Enrichment of recognition sequences for 3 pluripotency factors (blue), 2 hematopoietic lineage regulators (red) and 7 hematopoietic sub-lineage regulators (brown) in DHSs lost vs.

gained during (top) the differentiation of CD34+ cells into B-cells, (middle) the differentiation of CD34+ cells into T-cells (Th2), and (bottom) the differentiation of CD34+ cells into NK-cells. **(B)** Enrichment of recognition sequences for 459 transcription factors in DHSs lost vs. gained during the development of ES cells into 41 different definitive cell types.

Figure 7.S5. Epigenetically stable DHSs are preferentially and chiefly populated by TFs that regulate their own expression

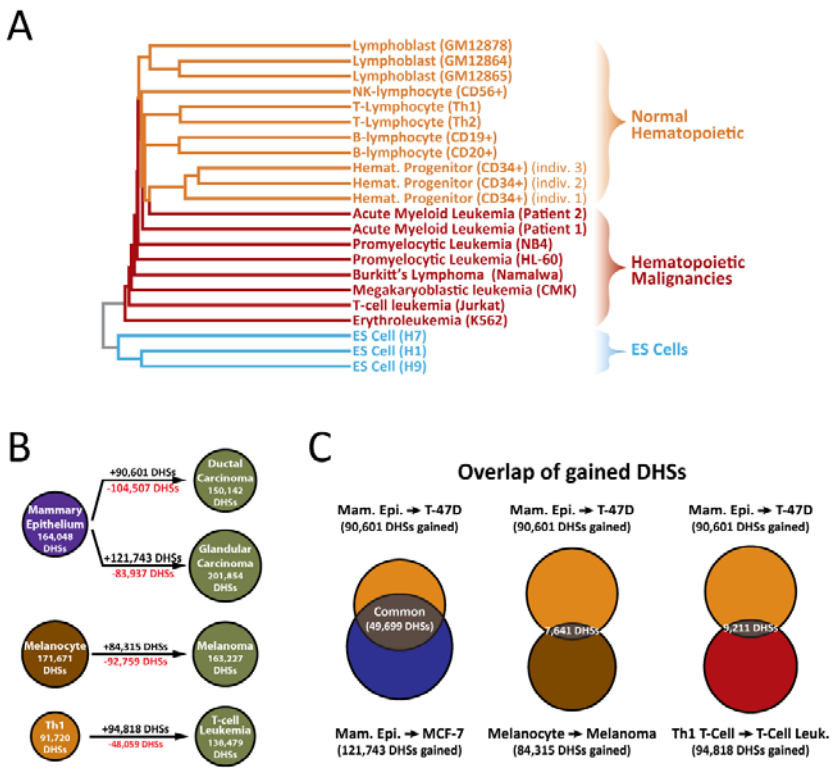


(A) Proportion of epigenetically stable DHSs that are populated by (left) simple autoregulatory TFs, (middle) 2-node directed loop autoregulatory TFs, and (right) 3-node directed loop autoregulatory TFs.

(B) Enrichment of (left) simple autoregulatory TFs, (middle) 2-node directed loop autoregulatory TFs, and (right) 3-node directed loop autoregulatory TFs in either transcriptional start site (TSS)-proximal or TSS-distal epigenetically stable vs. developmentally gained DHSs. All enrichments shown are significant ($p<10e-10$), hypergeometric distribution.

(C) Enrichment of autoregulatory TFs in either transcriptional start site (TSS)-proximal or TSS-distal epigenetically stable vs. developmentally lost DHSs. All enrichments shown are significant ($p<10e-10$), hypergeometric distribution.

Figure 7.S6. Remodeling of the regulatory landscape during oncogenesis

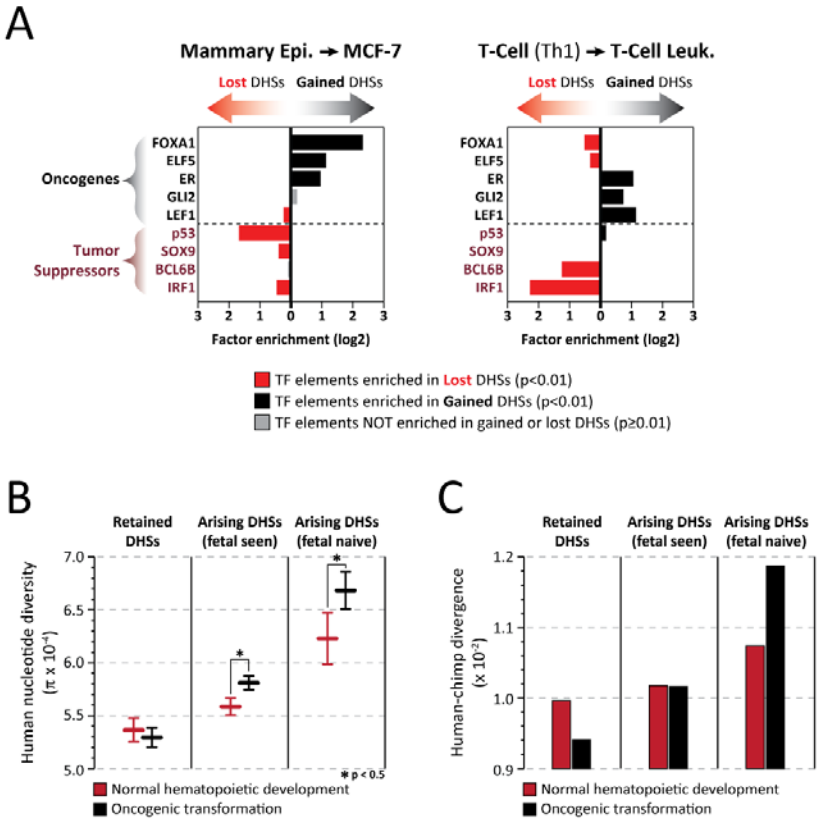


(A) Unbiased clustering (nearest-neighbor) of the linear patterning of DHSs (present/absent) from 11 normal hematopoietic cell types (orange), 3 ES cell types (blue) and 8 hematopoietic malignancies (red).

(B) Shown for each of four oncogenic transformations are the number of acquired (black) vs. extinguished (red) DHSs relative to the precursor cell state.

(C) Shown is the pairwise overlap of sites gained during the oncogenic transformation of the cell types in (B).

Figure 7.S7. DHSs arising during oncogenesis show relaxed evolutionary constraint



(A) *Oncogenic vs. tumor suppressor TF targets in DHSs gained vs. lost during oncogenesis.* Shown is enrichment of recognition sequences for 5 TF oncogenes and 4 tumor suppressor TFs in the DHSs lost vs. gained during the oncogenic transformation of mammary epithelium and T-cells.

(B) For the normal development of CD34+ hematopoietic progenitors (red), and the oncogenic transformation of mammary epithelium, melanocytes and T-cells (black), shown is the nucleotide diversity (π, y-axis) at DHSs that are either retained from their normal precursor (left), DHSs that are gained during development/oncogenesis but seen at some other point during development (middle), and DHSs that are gained during development/oncogenesis and not active

at any other point during development (right). DHSs shared with ES cells were removed from this analysis to control for the retrograde remodeling. Error bars represent 95% confidence intervals (obtained by bootstrap resampling).

(C) Shown is the human-chimp divergence at the DHSs described in (B).

7.6 – METHODS

Cell types

Human cell types IMR90, CD19, CD56, CD34 (indiv. 2), CD34 (indiv. 3), Skin Keratinocyte (indiv. 1), Skin Keratinocyte (indiv. 2), Skin Keratinocyte (indiv. 3), vHMEC, RPMI_7951, NT2_D1, PFSK-1, AML (patient 1), AML (patient 2), Namalwa, Mesenchymal Stem Cells, Skin fibroblasts, CD4+ T-lymphocytes, CD8+ T-lymphocytes, Cardiac_d5, Cardiac_d14, Cardiac_y3, Melanocyte, EScell_H1 and EScell_H9 were subjected to DNaseI digestion and high-throughput sequencing, following previous methods (Thurman et al., 2012; John et al., 2011; Hesselberth et al., 2009). Tags were aligned to the reference genome, build GRCh37/hg19 using Bowtie (Langmead et al., 2009), 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. Data from additional cell types were utilized from Thurman et al. 2012. Reads mapping to the X and Y chromosomes were removed from all analyses.

DHS clustering of non-cancer cell types

DNaseI hypersensitive sites (DHSs), each 150 bp in genomic length, were computed across every cell type at a false discovery rate of 1% using established methods (John et al., 2011). We computed the multiset union of DHSs across all cell types using the BEDOPS suite (Neph et al., 2012a), version 1.2 (using *bedops -u*). For each element of the union, we collected all DHSs in the union with 25% or more of their bases in common with the element. The genomic coordinates of a DHS were redefined to the minimum and maximum coordinates from its overlap set, which always included the original DHS itself (using *bedmap --echo-map-range --fraction-map 0.25*). Unique DHSs at this point comprised a final reference set. The Euclidean distance between two cell types was calculated using vectors of binary values with sizes equal to

the number of elements in the reference set. For each element of the reference set, a cell type received a '1' if it contained a DHS enveloped by the reference element or '0' otherwise (using *bedmap --fraction-map 1 --indicator*). We computed pairwise Euclidean distances and packaged results into matrix form. We then clustered the cell types using the nearest-neighbor algorithm as implemented by the *hclust* and *dendrogram* functions available in the R statistical package (http://www.r-project.org). Pairwise overlap calculations (same procedure as described above) were plotted as a heatmap (**Fig. 7.S1D**).

DHS clustering of cancer cell lines

We used the same procedure to derive a Euclidean distance matrix over several cancer cell lines. We applied Ward clustering (Ward, 1963) using the *hclust* and *dendrogram* functions in R.

Exon clustering of non-cancer cell types

We performed experiments using Affymetrix GeneChip Human Exon 1.0 ST Arrays and extracted results with Affymetrix Expression console (Ver1.1) across a subset of available cell types (45). Extraction parameters were set to the extended annotation confidence level, median polish summarization method as used in Robust Multichip Average, and the sketch-quantile batch normalization method. Probe measurements were mapped onto overlapping coding and noncoding exons annotated by the Reference Sequence project (Pruitt et al., 2009a) in a strand sensitive manner (using *bedmap --max --fraction-either 1*), and these results were converted to percentiles for each cell type. As with the DHS clustering, we computed pairwise Euclidean distances and clustered cell types using the nearest-neighbor algorithm.

The 45 cell types used in the exon clustering were: AG04449, AG04450, AG09309, AG09319, AG10803, AoAF, BJ, CD19, CD20+, CD34+_Mobilized, GM12864, GM12865,

GM12878, H7-hESC, HCFaa, HCF, HCM, HConF, HEEpiC, HGF, HMEC, HMF, HMVEC-dBl-Ad, HMVEC-dBl-Neo, HMVEC-dLy-Ad, HMVEC-dLy-Neo, HMVEC-dNeo, HMVEC-LBl, HMVEC-LLy, HPAF, HPdLF, HPF, HRGEC, Th1, Th2, HUVEC, HVMF, IMR90, NHDF-Ad, NHDF-neo, NHLF, SAEC, and SKMC

Principal coordinate analyses

We applied multidimensional scaling to the Euclidean distance matrix created to cluster cell types (above). We produced 3-dimensional plots using the *cmdscale* function in R.

Evolutionary conservation of DHSs arising in embryological ancestors

For this analysis only replicate concordant peaks were used for each cell type. Replicate concordant peaks were defined as 1% FDR peaks identified in both biological replicates of the same cell type (using *bedops -e -25%*). BJ and GM12864 cells were excluded from this analysis as only one biological replicate existed for these two cell types. DHSs common to a lineage group (i.e. paraxial mesoderm, lymphoid, hematopoietic progenitors, endothelia, ectoderm and endoderm) were identified by first making a master list of all replicate concordant DHSs present within that lineage group and then identifying DHSs within that master list that were identified as DHSs in at least 50% of the cell types from that lineage group. Ancestral lineage groups were inferred using these core lineage groups. Hematopoietic common sites were defined as sites common in both the lymphoid and hematopoietic progenitor groups. Hemangioblast common sites were defined as sites common in both the endothelia and hematopoietic lineage groups. Mesoderm common sites were defined as sites common in both the paraxial mesoderm and hemangioblast lineage groups. Epiblast common sites were defined as sites common in mesoderm, ectoderm and endoderm lineage groups. Evolutionary conservation was calculated using ES cell DHSs, and DHSs arising in specific lineage groups. DHSs arising in the epiblast

lineage group were defined as epiblast common sites not found in ES cells. DHSs arising in the mesoderm lineage group were defined as mesoderm common sites that were not common to the epiblast lineage group. DHSs arising in the paraxial mesoderm lineage group were defined as paraxial mesoderm common sites that were not common to the mesoderm lineage group. DHSs arising in the hemangioblast lineage group were defined as hemangioblast common sites that were not common to the mesoderm lineage group. DHSs arising in the endothelia lineage group were defined as endothelia common sites that were not common to the hemangioblast lineage group. DHSs arising in the hematopoietic lineage group were defined as hematopoietic common sites that were not common to the hemangioblast lineage group. DHSs arising in the lymphoid lineage group were defined as lymphoid common sites that were not common to the hematopoietic lineage group. Ectodermal and Endodermal arising sites were not calculated due to the paucity of cell types within these branches. Evolutionary conservation was calculated using a 45 vertebrate species measurement of evolutionary basewise conservation (phyloP). For each DHS within a set (e.g. ES cell DHSs, epiblast arising DHSs, etc.) the maximum phyloP score was identified. For each set we sampled 1,000 values with replacement 1,000 times to calculate the average evolutionary constraint and 95% confidence intervals. Performing this analysis using only distal DHSs yielded nearly identical results (not shown).

Comparison of adult versus embryonic enhancer activity

Data for tests of human enhancers in a mouse developmental model (Pennacchio et al., 2006) were downloaded from http://enhancer.lbl.gov/**.** For each developmental enhancer, the number of tissues scored positive in the mouse developmental model was recorded, as was the number of ES, adult, and fetal tissues with an overlapping DHS (false discovery rate 1%, (John et al., 2011). These values were plotted using LOWESS curve fitting with the default parameters

for the R *lowess* function. We applied a linear regression model to collect the associated F-test p-value. Embryonic mouse images were downloaded from the VISTA enhancer browser (http://enhancer.lbl.gov/).

Development and differentiation along specific lineages

DHSs 'lost' along a known lineage differentiation path were calculated as those DHSs belonging to a progenitor cell type and not found within a more differentiated cell type (using *bedops -n -1*). Conversely, DHSs 'gained' along a known lineage differentiation path were calculated based on DHSs belonging to a more differentiated cell type but not its progenitor. DHSs 'shared' along a known lineage differentiation path were calculated based on DHSs belonging to a more differentiated cell type and its progenitor (using *bedops -e -25%*). To generate the data for Figures 3E, 3J and 6E, we first calculated the proportions of lost and gained DHSs that were shared with ES cells. We then took the ratio of these two proportions and plotted the log2 values to show the enrichment of ES cell DHSs in either the lost or gained sites.

Motif analysis

We determined the relative predicted motif enrichments for select transcription factors in the DHSs that dynamically change along known lineage differentiation paths (**Figs. 7.M4A,B,C**, **Fig. 7.M6C** and **Figs. 7.S3A,B**). We identified DHSs lost and gained as described above, and scanned for motif instances in each set separately, using FIMO (Bailey et al., 2009) with $P < 1e-5$ and defaults for other parameters. We counted the number of motif instances for each template and normalized to the total number of bases scanned. Log2 ratios of the normalized results were computed between the DHS sets.

Autoregulating transcription factors

To map transcription factor occupancy within DHSs genome-wide, we identified predicted transcription factor (TF) binding sites overlapping DNaseI footprints (FDR 1%) by at least 3 nt similarly to as previously described with the addition of motif models from the JASPAR and Uniprobe databases (Neph et al., 2012b). We applied this analysis to 23 cell types in this study containing genome-wide DNaseI footprint maps (Neph et al., 2012b). TF-coding genes with a DNaseI footprint located within 5,000 nt of the gene's transcriptional start site, predicted to be bound by the gene's own protein product, were labeled as autoregulatory transcription factors. For each cell type, we used a cell-type specific transcription factor occupancy map to identify simple autoregulatory TFs, as well as TFs contained in 2-node and 3-node directed loop network motifs. DHSs retained along a known lineage differentiation path were calculated based on DHSs belonging to a progenitor cell type and also found within a more differentiated cell type (using *bedops -e -25%*). DHSs 'gained' along a known lineage differentiation path were calculated based on DHSs belonging to a more differentiated cell type but not its progenitor (using *bedops -n -25%*). We determined the proportions of shared and gained DHSs harboring footprint-positive binding sites potentially occupied by simple autoregulatory, 2-node directed loop and 3-node directed loop TFs for each differentiation path. We then calculated the log2 ratio between these proportions to obtain an enrichment value. This analysis was also repeated using DHSs within +/- 1kb of transcriptional start sites (proximal DHSs) and DHSs outside of +/-1kb of transcriptional start sites (distal DHSs).

Distribution of cell types containing a class of regulatory elements

For Fig. 7.S1B, the distribution of cell types containing a DHS was calculated separately for DHSs observed in any of the 49 cell types, DHSs observed in any of the 21 paraxial

mesoderm cell types, DHSs observed in any of the 8 lymphoid cell types, and DHSs observed in any of the 9 endothelial cell types. The number of cell types that a specific DHS was active in was calculated by first creating a master list of DHS positions that spanned the cell types of interest. This master list was created by taking the union of DHSs from all cell types and selecting those from one cell type that didn't overlap a DHS from any other cell type. If a DHS from the union overlapped two or more cell-types, the DHS with the highest z-score of all overlapping DHSs was also selected for the master list. Each peak from the master list so constructed was then annotated with the number of cell types having a DHS overlapping at that position. The distribution of these cell-type numbers was then plotted using the *beanplot* package in R. Figure 5C (left) was calculated as above, but using DHSs observed in any of the 19 cancer cell lines. Figure 5C (middle) was calculated using DHSs present in any of these 19 cancer cell lines, but DHS present in any of the 49 normal cell types were removed from this analysis. The distribution in Figure 4D was determined by calculating the number of normal cell types (48 cell types) each ES cell DHS was also present in (using *bedops -e -25%* to calculate the overlap of each individual cell type with ES cells). ES cell DHSs not present in any of these 48 cell types were not displayed. Figure 5C (right) was calculated using the same approach, but for the overlap of the 19 cancer cell lines with ES cells.

For Fig. 7.S3E, epigenetically gained and epigenetically stable DHSs were characterized based on their proximity to transcriptional start sites (TSSs) – sites within +/-1kb of TSSs were labeled as “promoter elements” and sites outside of +/-1kb of TSSs were labeled as “distal elements”.

Human nucleotide diversity measurements

Human nucleotide diversity measurements (π) were calculated using whole genome sequences from 53 unrelated, publicly available human genomes released by Complete Genomics (version 1.1034) as previously described (Vernot et al., 2012). To obtain a per nucleotide estimate of π, we normalized π by the total number of bases considered for a particular analysis (Vernot et al., 2012). All π calculations were performed on replicate concordant peaks that did not overlap ES cell DHSs. Repeats, exonic regions and CpGs were removed from all π calculations. 95% confidence intervals were obtained by bootstrap resampling of genomic regions.

Human-chimp divergence measurements

An estimate of human-chimp divergence (relative mutation rate) within the different classes of DHSs was calculated as previously described (Thurman et al., 2012). Human/chimpanzee alignments were downloaded from the UCSC Genome Browser (hg19/panTro2) and the number of nucleotide differences between human and chimpanzee (d) and the number of bases aligned (n) were used to calculate the relative mutation rates μ per site per generation ($\mu = (d / n) / (2 \times 6 \text{ my} / 25 \text{ years/generation})$, with 6 million years being the approximate age of the human/chimp divergence).

Chapter 8 – Transcription factors extensively occupy coding sequence and constrain both protein evolution and codon usage

8.1 – ABSTRACT

Codons are the fundamental building block of genes and underpin our current understanding of genome function and evolution. However, whether codons carry additional information beyond protein sequence is currently unknown. To address the magnitude and ramifications of transcription factor (TF) binding elements within coding sequence, we mapped 11.6 million TF occupancy events at nucleotide resolution across 81 diverse human cell types using genome-wide DNaseI footprinting. We identified 216,304 TF footprints populating coding sequence, with 14% of coding bases, and 87% of all genes occupied by a TF in at least one cell type. Overall, coding TF occupancy has constrained amino acid divergence throughout human evolution and has systematically shaped codon choice - resulting in its appearance as the major determinant of codon bias in mammalian genomes. Furthermore, TF recognition preferences have co-evolved with coding sequence, resulting in the exaptation by TFs of certain common coding features as well as the avoidance by TFs of generating aberrant stop codons within coding sequence. Finally, we identify that at least 17% of coding variants impact the occupancy of superimposed TFs, independent of the variant's effect on protein sequence and structure. Our results indicate that coding variation can significantly alter the wiring of TF binding elements and that coding TF binding elements play a central role in both protein evolution and codon usage biases.

8.2 – INTRODUCTION

Our understanding of the magnitude and ramification of TF binding elements within coding sequence is largely limited by the technical difficulty associated with identifying specific coding bases occupied by TFs within an organism. Current approaches either offer too low of a resolution to identify individual codons bound by TFs (i.e. ChIP-seq), or require an exorbitant number of experiments to fully interrogate the sizable number of TFs (>1,500) (Vaquerizas et al., 2009) and cellular states (>350) (Bard et al., 2005) contained within the human body. Due to these limitations, only a handful of transcriptional regulatory elements have been identified within coding sequence (Hyder et al., 1995; Lang et al., 2005; Barthel and Liu, 2008; Ritter et al., 2012; Khan et al., 2012).

In contrast, genome-wide DNaseI footprinting provides a reasonable alternative for identifying TF binding elements embedded within coding sequence. First, genome-wide DNaseI footprinting can simultaneously map the genome-wide regulatory interactions of every TF expressed within a cell-type (Hesselberth et al., 2009; Neph et al., 2012c). Furthermore, DNaseI footprints demarcate TF occupancy at nucleotide resolution, enabling the precise determination of bound and unbound genomic elements (Galas and Schmitz, 1978).

8.3 – RESULTS

Transcriptional regulatory elements extensively populate genomic coding sequence

To systematically identify TF binding elements embedded within coding sequence, we analyzed high-resolution DNaseI footprint maps across 81 diverse human cell types generated at an average depth of ~484 million sequencing reads (~302 million uniquely mapping) per cell type (Neph et al., 2012c). In total, these maps define 11,598,043 TF binding elements genome-wide (~1,018,514 per cell-type), 216,304 of which overlap coding sequence (~24,842 per cell-type) (**Fig. 8.M1A-B** and **Fig. 8.S1A-B**). In total, ~14% of all coding bases were contained within a transcription factor footprint in at least one cell type (avg. 1.1% of coding sequence per cell type; **Fig. 8.M1C** and **Fig. 8.S1C**) and 86.9% of genes contained coding TF footprints (avg. 33% of genes per cell type) (**Fig. 8.M1D** and **Fig. 8.S1D**). Together, these results demonstrate that transcription factor binding elements extensively populate genomic coding sequence, and imply that the genetic code is highly capable of encoding TF binding elements (Itzkovitz and Alon, 2007).

Although widespread, the coding TFs detected using this approach likely represent a conservative estimate of the total number of human coding bases with overlapping TF binding elements since only ~25% of the >350 distinct human cell types were sampled, and the ability to detect TF footprints was limited by the extent of mapped DNaseI cleavages within coding sequences (Neph et al., 2012c). Each of the 81 cell types contributed incrementally to coding TF footprints, with little evidence of saturation (**Fig. 8.S2**).

To ascertain coding footprints within a cell type more completely, we developed an approach for targeted footprinting of exons via solution-phase capture of DNaseI cleavage fragment libraries (**Methods**). We performed this approach on DNaseI libraries from two cell

types (abdominal skin fibroblasts and mammary stromal fibroblasts), resulting in 8- to 10-fold increase in DNaseI cleavage density over exons, equivalent to the coverage from sequencing nearly ~1-2 billion genomic reads of conventional footprinting (**Fig. 8.S3A**). This provided superior quantification of DNaseI footprints, both sharpening previously delineated elements, and exposing additional sites not yet evident at a read depth of ~400 million uniquely mapping genomic reads (**Fig. 8.S3B-D**). Overall, we identified an additional ~150,000 coding DNaseI footprints in each cell type (**Fig. 8.M1E**), a 7-12-fold increase. Collectively, the above results indicate that coding sequences are densely populated with transcription factor footprints.

Genetic variation at coding DNaseI footprints significantly alters chromatin state.

Dense sequencing coverage of DNaseI hypersensitive exons (hundreds- to thousands-fold) enabled systematic identification of single nucleotide variants (SNVs; total 592,867 heterozygous variants across 81 cell types each obtained from a distinct individual). In total we identified SNVs within 3% of coding footprints (**Fig. 8.M1F**). Allelic variation within TF footprints may result in abrogation of TF occupancy, and is directly detectable through analysis of the allelic origin of DNaseI cleavage fragments. We detected allelic imbalance in chromatin accessibility at 17.4% of all heterozygous coding SNVs contained within DNaseI footprints (**Fig. 8.M1G**), a significant enrichment over non-footprinted regions ($P < 1 \times 10^{-8}$; Fishers exact test) (**Fig. 8.S4**). Importantly, this encompassed both synonymous and non-synonymous variants. For example, a common nonsynonymous G-to-A SNV (rs8110393) occurs at a high-information position within a DNaseI footprint matching as SP1 occupancy site, and disrupts TF occupancy at that site without affecting neighboring CTCF and SP1 occupancy sites (**Fig. 8.M1H**).

The ability of a coding SNV to cause allelic coding imbalance is not dependent on whether it is a synonymous or nonsynonymous variant (**Fig. 8.M1I**), or whether the variant is predicted to be deleterious to protein function (**Fig. 8.M1J-K**). Notably, 13.5% of all coding GWAS lead SNPs overlap DNaseI footprints (**Fig. 8.S5A**), implicating that these disease associations may result in alterations in gene regulation as opposed to protein sequence (Maurano et al., 2012a), as is likely the case with the psoriasis associated synonymous coding variant rs495337 (**Fig. 8.S5B-F**).

TFs are influenced by and can exploit coding features of exons

We next sought to determine whether certain coding regions are preferentially occupied by TFs. To accomplish this we first mapped the density of DNaseI footprints at different positions within a gene (**Fig. 8.M2A** and **Fig. 8.S6A**). These findings revealed that: (i) for many genes promoter-proximal transcriptional regulatory elements appear to extend well into the first exon (**Fig. 8.M2A**); (ii) internal coding exons are just as likely as their neighboring DNA sequence to harbor TF binding elements (**Fig. 8.M2A**); and (iii) genes that are longer, or more highly expressed within a cell type, have more coding DNaseI footprints (**Fig. 8.S6B-D**).

To identify the occupancy patterns of specific TFs within coding sequence, we used well-annotated databases of TF-binding motifs to infer which TFs are occupying which DNaseI footprints, an approach that closely and quantitatively mirrors ChIP-seq data (Neph et al., 2012c; Samstein et al., 2012). This analysis revealed that certain TFs, such as CTCF, SP1 and NFYA, preferentially avoid occupying coding binding elements (**Fig. 8.M2B**). Of note, the preference of a TF for occupying coding binding elements largely parallels the extent of CpG methylation at these elements (**Fig. 8.S7**), suggesting that gene body methylation may serve as a mechanism to

model the occupancy of certain TFs within coding sequence (Hellman and Chess, 2007; Zilberman et al., 2007).

For example, TFs involved in positioning the transcriptional pre-initiation complex, such as NFYA and SP1 (McKnight and Tjian, 1986), preferentially avoid occupying the first coding exon (**Fig. 8.M2B**), and typically occupy elements just upstream of the methionine start codon (**Fig. 8.M2C** and **Fig. 8.S8A**). In contrast, TFs involved in modulating the expression at a given promoter, such as YY1 and NRSF (Shi et al., 1991; Schoenherr and Anderson, 1995), preferentially occupy the first coding exon (**Fig. 8.M2B**). This appears to be partially accomplished due to the exaptation of certain coding sequences for DNA binding elements. For example, both YY1 and NRSF preferentially occupy start codons, as these TFs prefer binding ATG containing sequences (**Fig. 8.M2D** and **Fig. 8.S8B**). Furthermore, NRSF preferentially occupies the coding sequence for leucine rich protein domains (such as signal peptide domains and transmembrane domains), as NSRF prefers binding CTG containing sequences, which codes for leucine (**Fig. 8.M2E** and **Fig. 8.S8B**). In addition, certain TFs, such as CTCF and SREBP1, have exapted splice sites (**Fig. 8.S9C-D**), which are generally depleted of DNaseI footprints (**Fig. 8.S9A**).

In contrast to the diverse TF occupancy patterns at promoter-proximal regulatory elements, TFs in general avoid occupying stop codons (**Fig. 8.S9A**). Furthermore, TFs also avoid occupying the stop codon TAG, TAA and TGA trinucleotides even in non-coding sequences (**Fig. 8.S9B**), suggesting that the binding preferences of TFs may have evolved to avoid occupying, and thus generating, stop codons throughout the genome and coding sequence.

Transcription factors constrain both protein evolution and codon usage

Given the abundance of TF binding elements within coding sequence, we next sought to understand the impact of these binding elements on coding sequence evolution and codon choice. 4-fold degenerate coding bases are often used as a neutral model of evolution. However, we observed that 4-fold degenerate bases contained within DNaseI footprints are under significant evolutionary constraint (**Fig. 8.M3A**), indicating that these bases are not neutrally evolving. To investigate whether bases at coding DNaseI footprints are still under constraint within the human population, we quantified the age of coding mutations arising within or outside of DNaseI footprints using exome sequencing data from 4,298 individuals of European ancestry (**Fig. 8.M3B**) and 2,217 individuals of African American ancestry (**Fig. 8.S10A**) (Fu et al., 2013). Both synonymous and nonsynonymous coding mutations within DNaseI footprints are significantly younger than those outside of DNaseI footprints (average of ~3,312 years younger) (**Fig. 8.M3C** and **Fig. 8.S10B**), indicating that TF binding elements are constraining both degenerate and nondegenerate coding bases within the human population.

To identify the influence of specific TFs on coding sequence evolution we leveraged the fact that TF regulatory DNA sequences genome-wide have evolved to fit the continuous morphology of the transcription factor-DNA binding interface (Neph et al., 2012c) (**Fig. 8.M3D**). Consequently, by comparing the conservation profile of a TF at non-coding as well as 4-fold degenerate and nondegenerate coding bases, we can identify whether that TF is influencing codon and amino acid choice at its binding elements. For example, the evolutionary conservation profile at 4-fold degenerate and non-degenerate bases for the transcription factors KLF4 and NFIC closely mirrors that at non-coding binding elements (**Fig. 8.M3E**), indicating that these two TFs are constraining which codons (4-fold degenerate bases), and even amino

acids (non-degenerate bases), are used at their binding elements. Overall, of the 63 TFs with ample coding binding elements to calculate their 4-fold degenerate and nondegenerate conservation profile (Methods) we found that 73% appeared to constrain 4-fold degenerate bases, and 51% appeared to constrain nondegenerate bases (**Fig. 8.M3F** and **Fig. 8.S11** and **Fig. 8.S12**), suggesting a widespread influence of TF binding elements on codon and amino acid choice.

Transcription factors influence global codon usage bias.

Given the widespread influence of TF binding elements on codon choice, we next sought to understand whether TF binding elements might be contributing to a global bias in which codons are used in humans. Global codon usage biases have been observed in every organism studying to date (Shabalina et al., 2013). In organisms with a relatively short life span and a large effective population size such as bacteria, yeast, Caenorhabditis and Drosophila, codon usage biases are thought to result from the preferential use of codons that form efficient mRNA structures (Bulmer, 1988; Eyre-Walker and Bulmer, 1993; Gray and Hentze, 1994), and appear to be perpetuated by selection for the use of favored codons are protein sequences that require a high translational efficiency (Ikemura, 1981a, 1981b; Grantham et al., 1981; Gouy and Gautier, 1982; Carlini and Stephan, 2003; dos Reis et al., 2004). However, the basis for codon usage biases within mammals is much less understood. Selection appears to be affecting codon usage biases at >75% of human genes (Yang and Nielsen, 2008), yet codon usage bias within humans is only weakly associated with translational efficiency (Dos Reis et al., 2004), or other selective pressures such as exonic splicing regulatory elements (ESEs) (Parmley and Hurst, 2007) and 5' mRNA stability (Gu et al., 2010).

To understand whether TF binding elements might be causing a global bias in which codons are used in humans, we first determined whether TFs tend to occupy preferred codons. For example, genome-wide asparagine is preferentially encoded by the codon AAC. Whereas 50.8% of Asparagine codons outside of DNaseI footprints are AAC, 60.4% of asparagine codons overlapping DNaseI footprints are AAC, indicating that TFs show a 9.6% preference towards occupying the preferred codon for asparagine (**Fig. 8.M4A**). Strikingly, for every amino acid encoded by two or more codons (except Arginine) the codon that is preferentially used genome-wide is also preferentially occupied by TFs (**Fig. 8.M4B**).

To determine whether TF bias towards occupying preferred codons results from an inherent sequence preference of human TFs, we analyzed the frequency of different trinucleotides within non-coding DNaseI footprints. Strikingly, the codon trinucleotides favored by TFs within coding sequence were the same as those favored by TFs in non-coding sequence (**Fig. 8.M4C**), indicating that genome-wide human TFs preferentially occupy certain trinucleotides, and that these global TF preferences are reflected in the frequency of different codons. Importantly, the baseline tri-nucleotide frequencies within non-coding sequence and coding sequence are largely independent of each other, indicating that TFs preferences have not merely evolved to fit a set of trinucleotides favored throughout the genome.

If TF sequence preferences are generating codon usage biases, then TFs must be imparting excess evolutionary constraint at these preferred codons. Indeed, the third position of preferred codons overlapping DNaseI footprints is under excess evolutionary constraint (**Fig. 8.M4D**), indicating that TFs are selectively preserving these preferred codons within coding sequence.

8.4 – DISCUSSION

While nearly all codon biases reflect TF sequence preferences genome-wide, arginine was a notable exception. Of note, arginine is one of the 5 amino acids encoded by codons containing CpGs, and is the only amino acid where the majority of its codons contain a CpG (4 out of 6 codons). CpGs extensively participate in regulatory elements genome-wide (Bird, 1986), consistent with the genome-wide preference of TFs for CpG dinucleotides, and exhibit a markedly elevated mutational rate (Coulondre et al., 1978). Consequently, although TFs favor these CpG codons (**Fig. 8.M4E**), and are imparting excess constraint at them, the mutational rate at these codons is likely too high for them to be preferentially utilized within coding sequence.

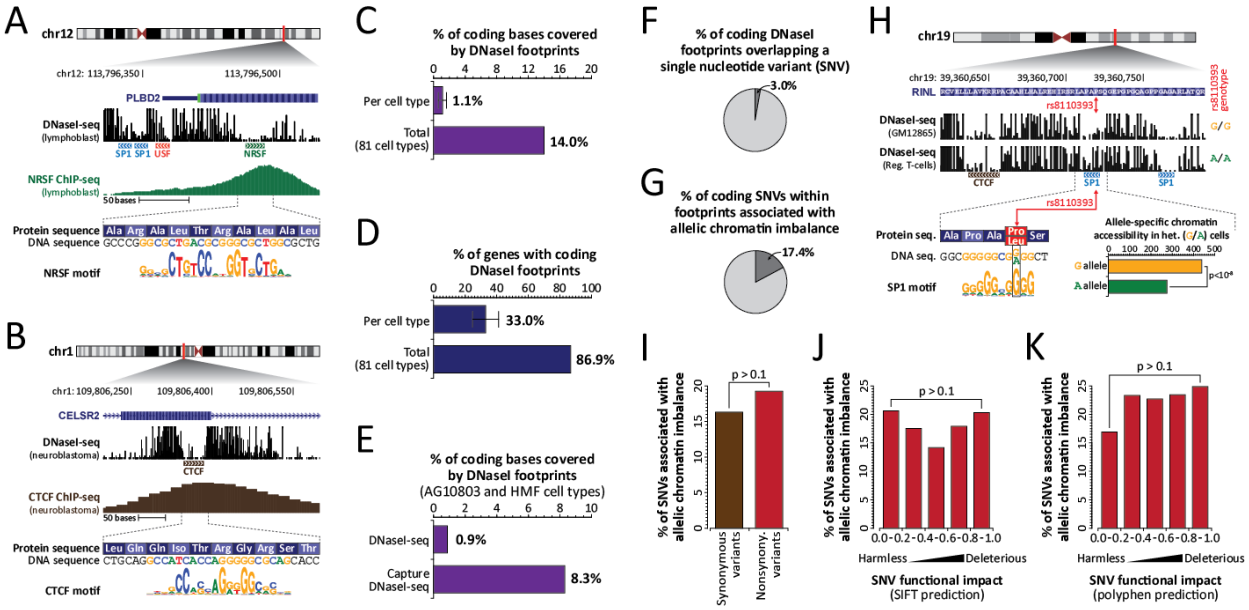
Although codons not overlapping DNaseI footprints still exhibit codon usage biases (**Fig. 8.M4A**), it is likely that these codons also reflect the direct or indirect actions of TF binding elements for several reasons. First, as noted above, our conclusions are drawn from a conservative annotation of which codons are TF binding elements. Consequently, as more human cell types are analyzed at deeper depths, we should have a better estimate of the true extent of TF occupied codons in humans. Second, codon usage biases are generated over evolutionary time scales. Consequently, the codon choice at any given codon may result from a TF binding element that existed in some ancestral species to human. This is likely the case given that TF trinucleotide preferences and codon biases have not changed much since the divergence of humans and mice (**Fig. 8.S13**). Third, codon usage bias can be exaggerated due to positive-feedback loops between codon usage biases and other cellular factors such as tRNA abundances (Bulmer, 1987, 1991). Consequently, codon preferences caused by TF binding elements can 'nucleate' a positive-feedback mechanism that exacerbates the intrinsic TF preference biases and creates both abundant and rare codons and tRNAs. These differences in tRNA abundance can

then affect protein synthesis and stability (Duan et al., 2003; zur Megede et al., 2000; Graf et al., 2000; Leder et al., 2001; Coleman et al., 2008).

Overall, these results have broad implications for the interpretation of the functional consequences of genetic variation within coding sequence. Furthermore, these findings expose an unexpected role of TFs in generating and perpetuating codon usage biases in mammals.

8.5 – FIGURES

Figure 8.M1. Transcriptional regulatory elements extensively populate genomic coding sequence



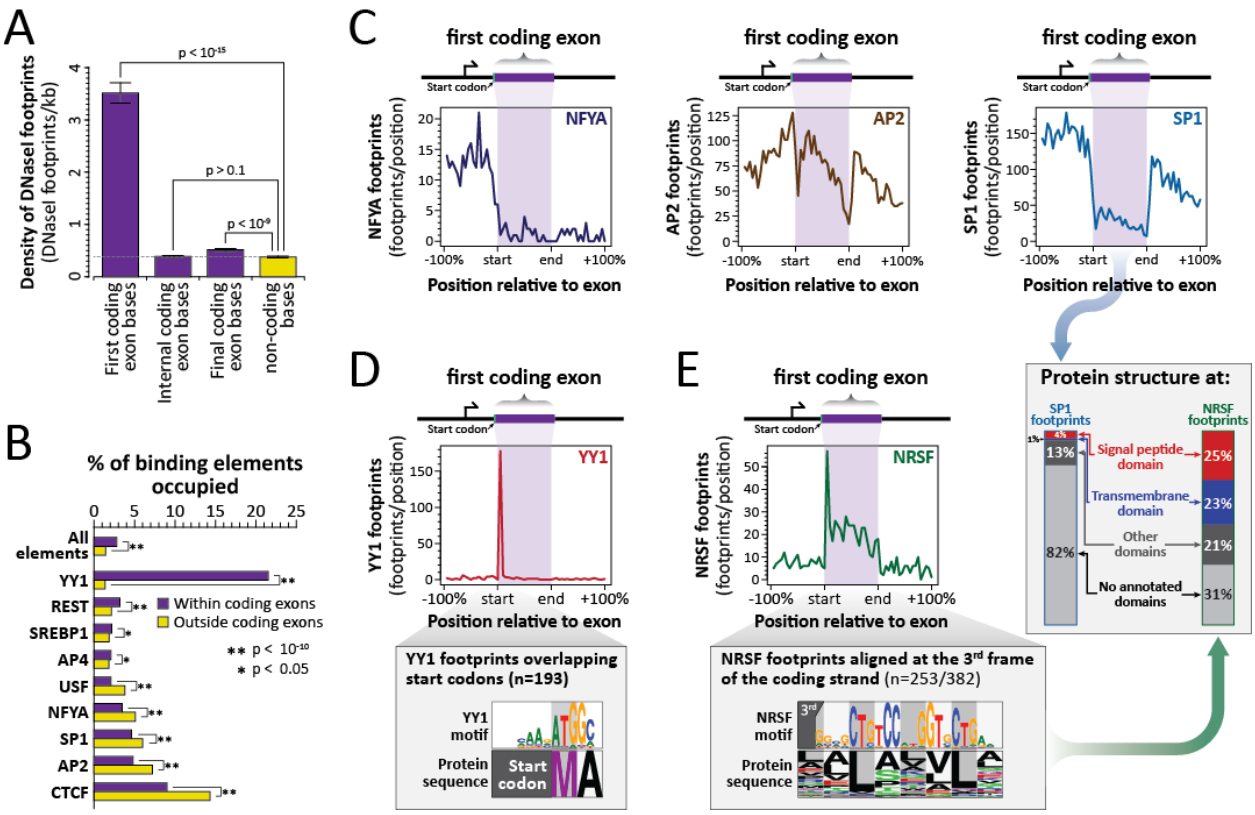
(A-E) *Regulatory factor footprints extensively populate genomic coding sequence.* DNaseI cleavage pattern and ChIP-seq signal for a coding (A) NRSF and (B) CTCF binding element. (C) The proportion of coding bases occupied by DNaseI footprints in each of the 81 cell types studied (top), or in any of the 81 cell types studied (bottom). (D) The proportion of genes containing a coding DNaseI footprint in each of the 81 cell types studied (top), or in any of the 81 cell types studied (bottom). (E) The percentage of coding bases identified as being occupied by DNaseI footprints using DNaseI-seq and Capture DNaseI-seq data.

(F-H) *Genetic variation at coding DNaseI footprints significantly alters chromatin state.* (F) The proportion of coding DNaseI footprints overlapping a SNV called in any of the 81 cell-

types. (G) The proportion of coding SNVs located within DNaseI footprints are associated with an allelically imbalanced chromatin state. (H) DNaseI cleavage pattern at an SP1 binding element overlapping the nonsynonymous SNV rs8110393 in a cell homozygous for the G or A allele. (bottom right) Allele-specific chromatin state at rs8110393 in cells heterozygous at this allele.

(I-K) *Coding variation can simultaneously alter both protein structure and chromatin state.* (I) Percentage of synonymous and non-synonymous coding variants overlapping DNaseI footprints that are associated with allele specific chromatin states. (J-K) Non-synonymous variants are grouped by the predictive functional impact of the variant based on the (J) SIFT or (K) polyphen-2 prediction algorithms ($P > 0.1$, Fisher's exact test using tag normalized datasets).

Figure 8.M2. TFs are influenced by and can exploit coding features of exons



(A) *First and final exons are preferentially occupied by TFs.* The average density of DNaseI footprints within first, internal, or final coding exons. Shown is the mean value for each of the 81 cell types studied plus or minus the standard error of the mean. p-value was computed using a paired t-test across the 81 cell types.

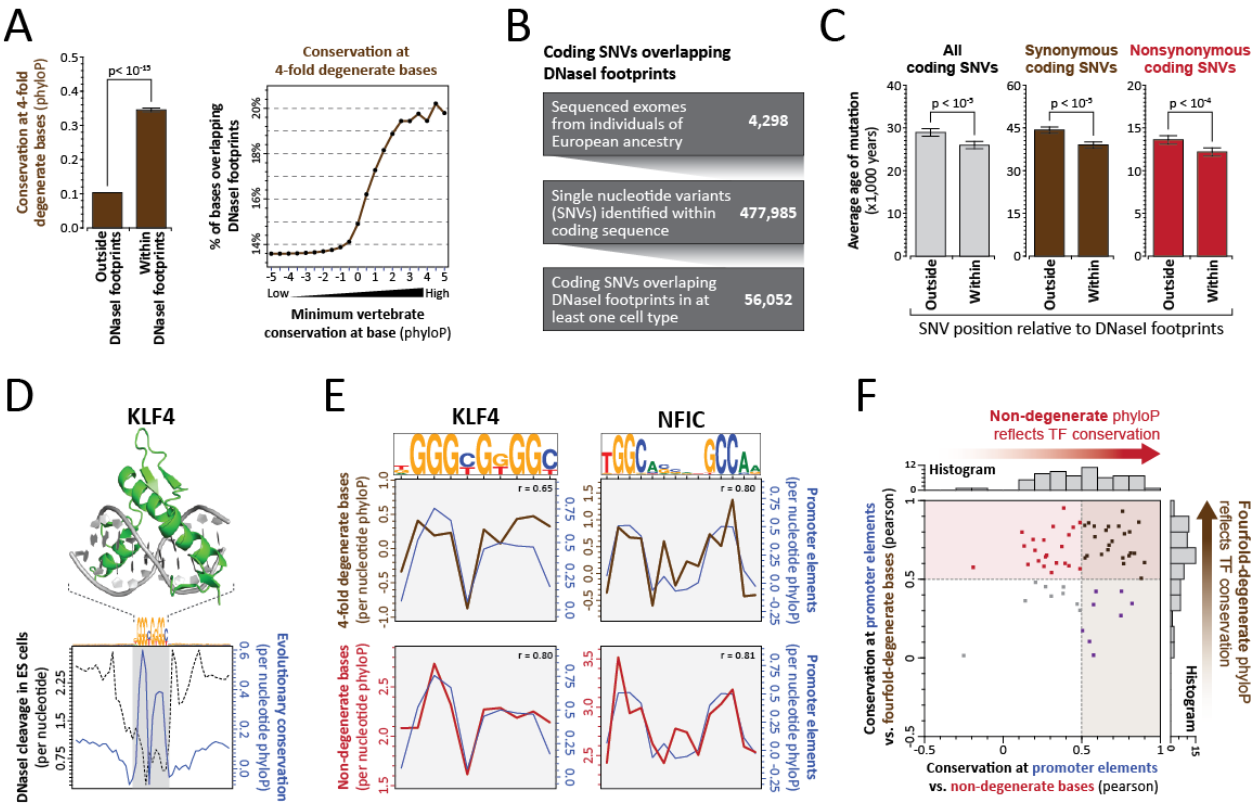
(B) *Individual TFs exhibit unique occupancy profiles within exonic DNA.* The percentage of all annotated coding and non-coding transcription factor binding elements that are occupied in each cell type, as well as the percentage for individual transcription factors. p-values were computed using the paired t-test across the 81 cell types.

(C) *NFYA, AP2 and SP1 preferentially avoid binding within coding sequence, start codons and splice junctions.* The density of (left) NFYA (middle), AP2 and (right) SP1 DNaseI footprints relative to first coding exons. Coding sequence is colored in purple.

(D) *YY1 binding elements exploit start codons.* (top) The density of YY1 DNaseI footprints relative to first coding exons. (bottom) Shown is the YY1 motif model as well as a logo of the amino acid sequence at all occupied YY1 binding elements that overlap a start codon.

(E) *NRSF exploits leucine-rich protein domains.* (left) (top) The density of NRSF DNaseI footprints relative to first coding exons. (bottom) Shown is the NRSF motif model as well as a logo of the amino acid sequence at all occupied coding strand NRSF binding elements that begin in the third frame of the codon. (right) Shown is the protein domain annotation at occupied first exon coding strand NRSF binding elements that begin in the third frame of the codon and occupied SP1 binding elements in the first exon.

Figure 8.M3. Transcription factors influence codon choice.



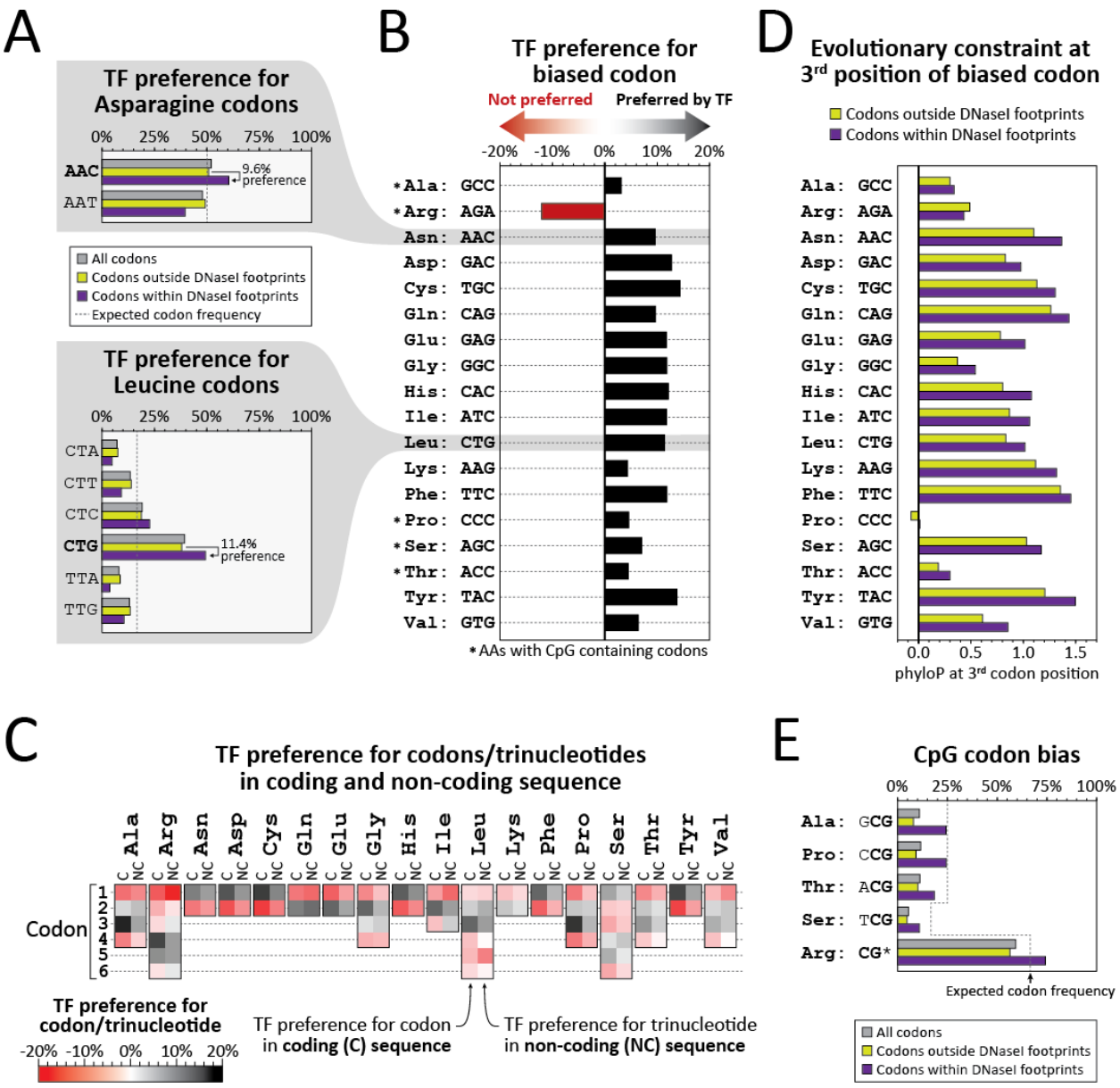
(A) *TF binding elements impart evolutionary constraint on coding sequence.* (left) phyloP conservation at 4-fold degenerate exonic bases within and outside of DNaseI footprints found in each of the 81 cell types. (right) The percentage of 4-fold degenerate bases above a minimum phyloP conservation level that overlap DNaseI footprints found in any of the 81 cell types analyzed. Note that 20% of highly conserved 4-fold degenerate bases overlap a DNaseI footprint in at least one of the 81 cell types analyzed in this study.

(B-C) *TF binding elements impart populational constraint on coding sequence.* (B) 4,298 sequenced exomes from individuals of European ancestry were utilized to identify SNVs overlapping DNaseI footprints in any of the 81 cell types. (C) The average mutational age at all

(grey), synonymous (brown) and non-synonymous (red) European coding SNVs identified within and outside of DNaseI footprints. Mutational ages and p-values were calculated as before (Fu et al., 2013).

(D-F) *Coding TF binding elements constrain both codon and amino acid choice.* (D) The co-crystal structure of KLF4 bound to its DNA ligand is juxtaposed above the average nucleotide-level DNase I cleavage pattern (dashed-black) and nucleotide-level evolutionary constraint (blue) at motif instances of KLF4 in DNase I footprints. Note how KLF4 imparts a stereotyped pattern of evolutionary constraint at its binding elements, which reflects the protein-DNA interface of KLF4. (E) Average per-nucleotide conservation profile at 4-fold degenerate bases (brown) and non-degenerate bases (red) overlapping (left) KLF4 and (right) NFIC footprinted binding elements. Pearson correlation values (r) between conservation profiles at promoter bases and 4-fold degenerate bases (top) or non-degenerate bases (bottom) are shown in the upper right corner of each plot. Note that KLF4 and NFIC impart a nearly identical pattern of evolutionary constraint at both coding and non-coding binding elements. (F) Pearson correlation values comparing the evolutionary constraint profiles imparted by 63 transcription factors at promoter elements, 4-fold degenerate bases and non-degenerate bases.

Figure 8.M4. Transcription factors influence global codon bias.



(A-B) *TFs preferentially occupy and constrain biased codons.* (A) Displayed is the proportion of all codons (grey), codons outside of DNaseI footprints (yellow), and codons within DNaseI footprints (purple) that overlap each of the two codons that encode asparagine (top) or each of

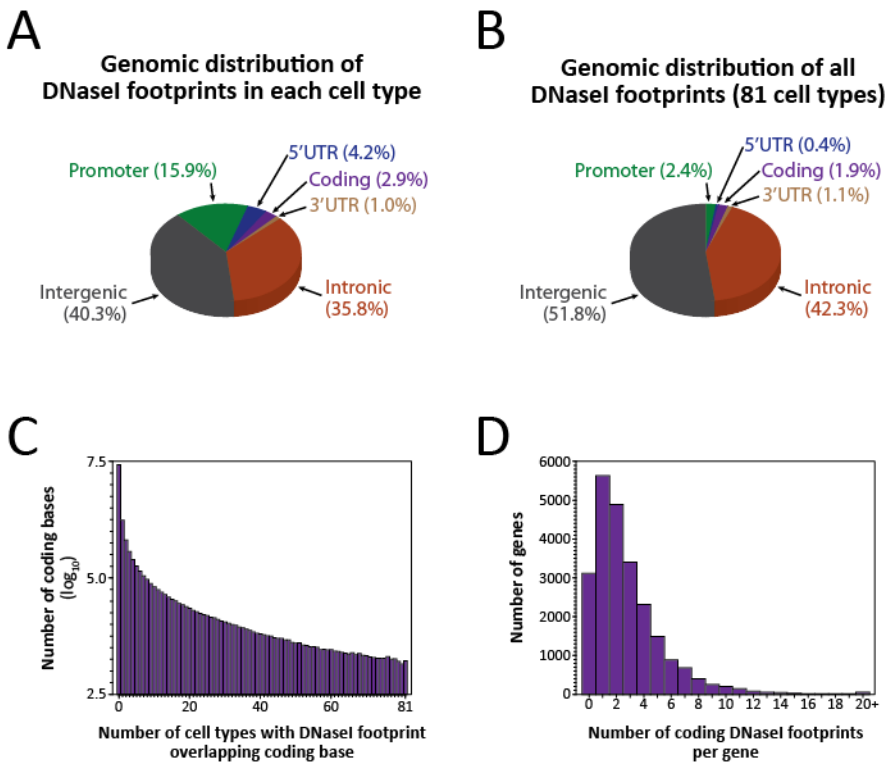
the 6 codons for leucine (bottom). Note that the codons that are biased genome-wide (AAC for asparagine and CTG for Leucine) are preferentially occupied by DNaseI footprints. (B) For each amino acid codon that is biased genome-wide, displayed is the preference for that codon to be occupied by DNaseI footprints, calculated as in (A). Note that nearly all biased codons show a strong preference towards being occupied by DNaseI footprints.

(C) *TF codon preferences reflect a genome-wide preference of TFs for certain tri-nucleotides.* For each codon encoding a specific amino acid, displayed is the preference for that codon to be occupied by DNaseI footprints (C - coding). Also displayed is the preference for that codon trinucleotide to be occupied by DNaseI footprints in non-coding regions of the genome (NC - non-coding).

(D) *TFs impart excess evolutionary constraint at biased codons.* Displayed is the average evolutionary constraint at 3rd codon positions for biased codons outside of DNaseI footprints (yellow), and within DNaseI footprints (purple). Note that TF binding elements impart excess evolutionary constraint on biased codons.

(E) *CpG containing codons are favored by TFs, yet disfavored genome-wide.* For each of the five amino acids that are encoded by a CpG containing codon, displayed is the proportion of all codons (grey), codons outside of DNaseI footprints (yellow), and codons within DNaseI footprints (purple) that overlap the CpG containing codon(s). Note that CpG containing codons are disfavored globally, but preferentially occupied by TFs.

Figure 8.S1. Distribution of Transcription Factor DNaseI footprints



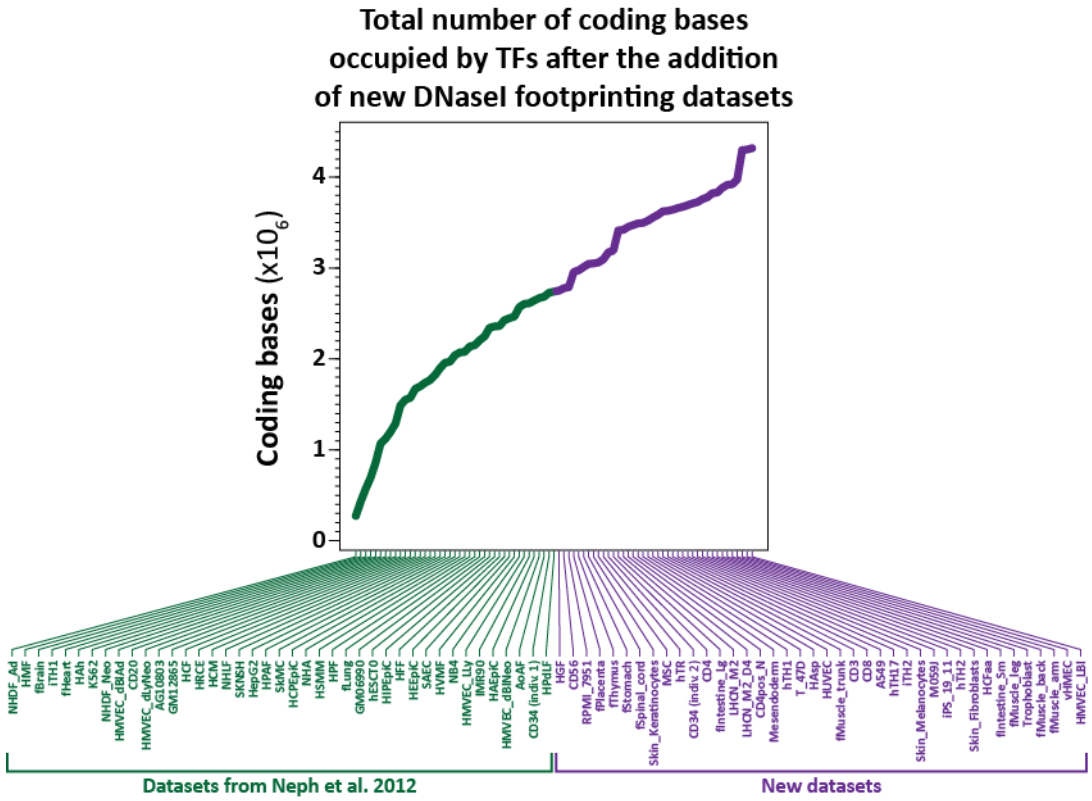
(A) Shown is the average genomic distribution of DNaseI footprints across all 81 cell types

(B) Shown is the genomic distribution of DNaseI footprints identified in any of the 81 cell types

(C) Histogram of the number of DNaseI footprints overlapping each coding base along the genome. Y-axis is log-10 scale.

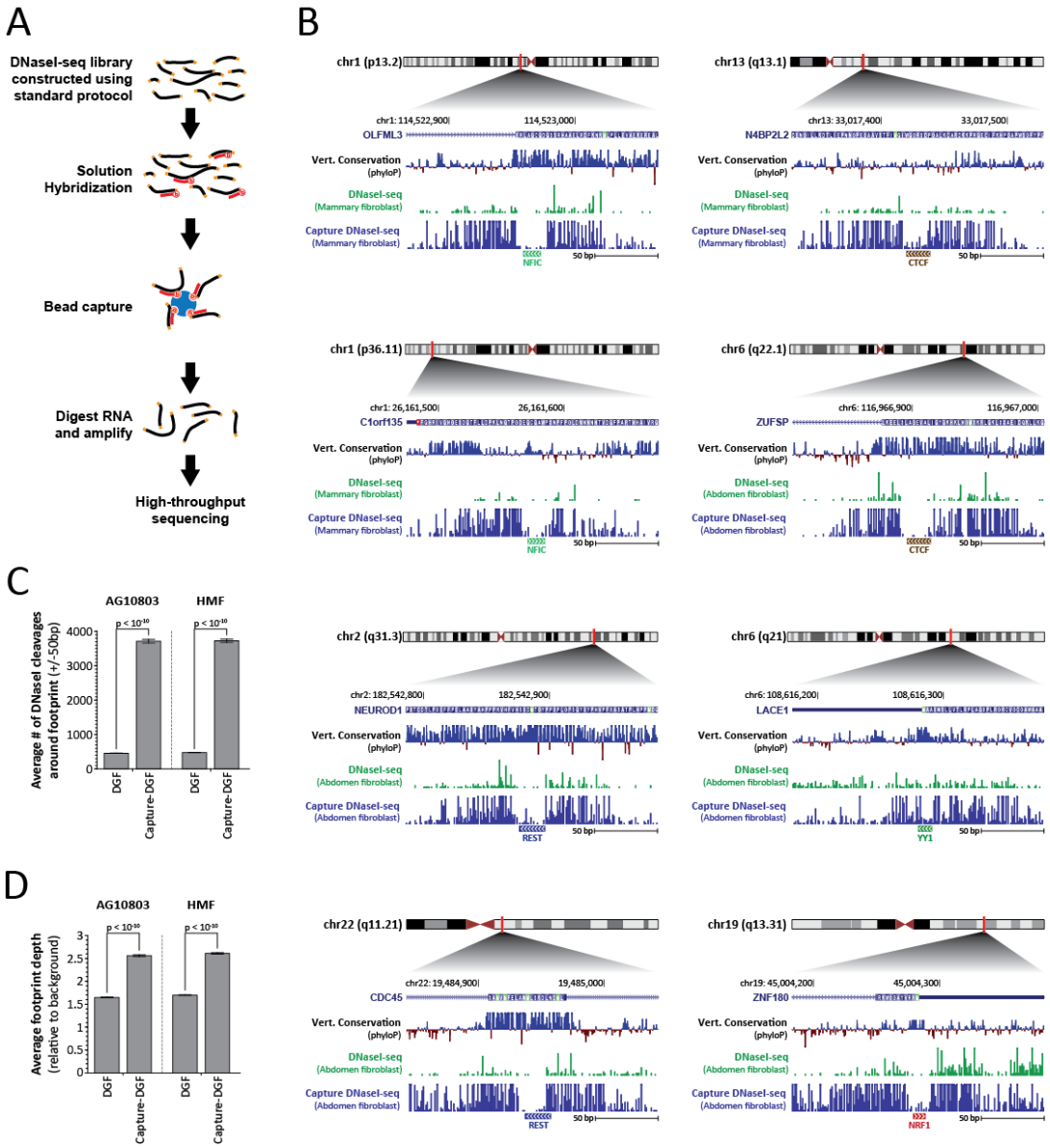
(D) Histogram showing the number of coding DNaseI footprints per gene.

Figure 8.S2. DNaseI footprints identified using additional cell types



Total number of coding bases overlapping DNaseI footprints identified after adding additional cell types.

Figure 8.S3. Sensitivity of coding DNaseI footprints using capture DNaseI-seq.

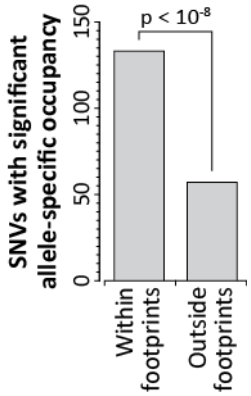


(A) Summary of Capture DNaseI-seq method

(B) Per-nucleotide vertebrate conservation as well as per-nucleotide DNaseI-seq and capture DNaseI-seq cleavage patterns at coding binding elements for NFIC, CTCF, REST, YY1 and NRF1.

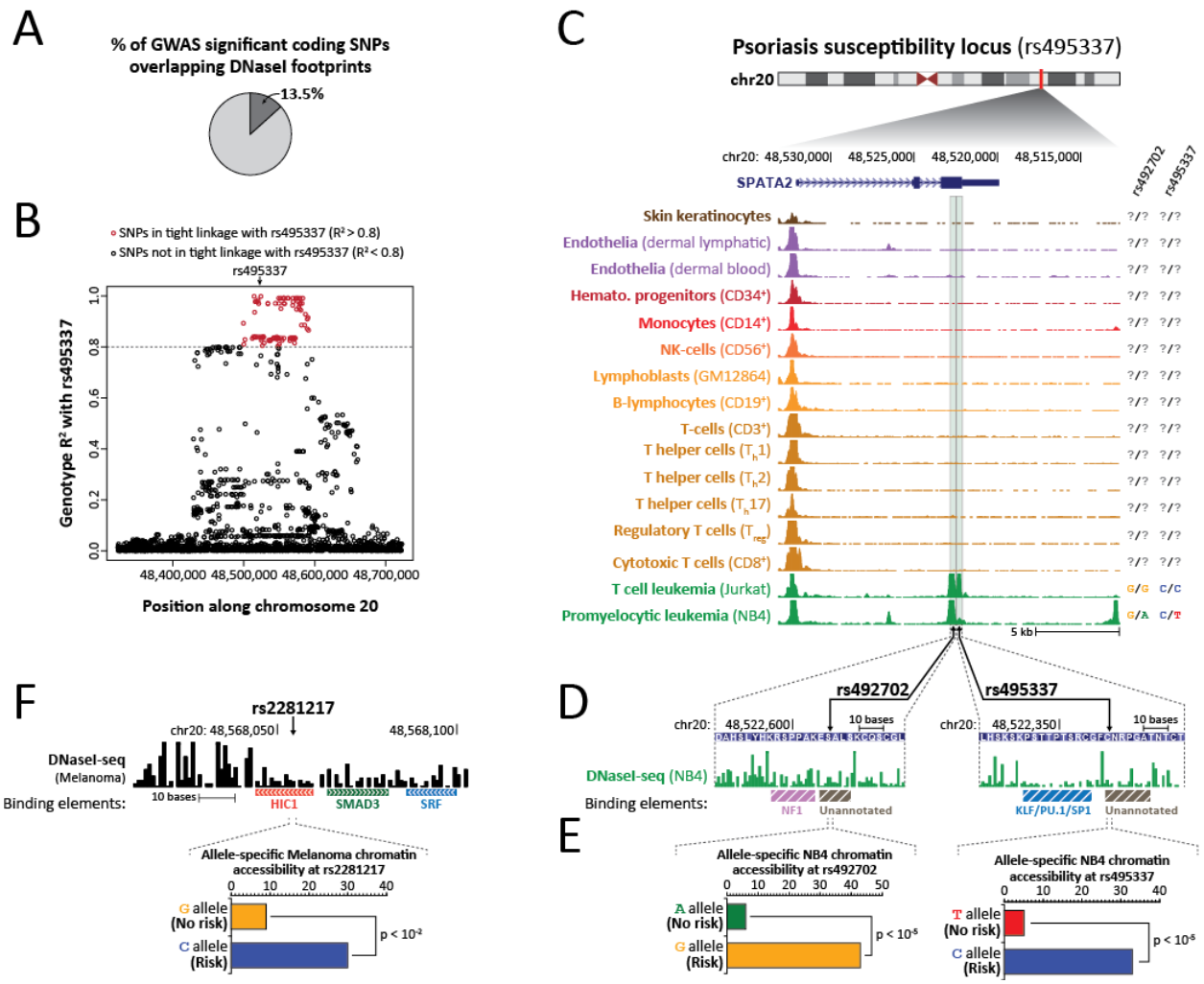
(C-E) *Capture DNaseI-seq enables extensive DNAseI footprint identification and superior quantification.* (C) The average number of sequenced DNaseI cleavages surrounding DNaseI footprints using DNaseI-seq and Capture DNaseI-seq data. (D) The average depth of DNaseI footprints using DNaseI-seq and Capture DNaseI-seq data. Note that Capture DNaseI-seq enables the more precise quantification of DNaseI footprints.

Figure 8.S4. Coding DNaseI footprints are enriched in variants associated with allele-specific chromatin states



Heterozygous coding SNVs associated with allele-specific occupancy are significantly enriched inside DNaseI footprints (P < 1x10-8, Fisher's exact test using tag normalized datasets).

Figure 8.S5. Coding variants linked to disease susceptibility can also influence chromatin state

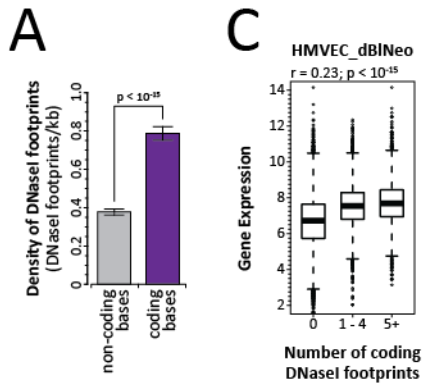


(A) Shown is the proportion of coding GWAS variants linked to disease susceptibility that overlap DNaseI footprints in one of the 81 tested cell types.

(B-E) *Synonymous coding variants linked to psoriasis susceptibility are associated with chromatin state changes selectively within transformed hematopoietic cells.* (B) SNPs in tight linkage with the psoriasis associated coding SNP rs495337 were identified using whole genome

sequencing data from 267 individuals of Northern and Western European (CEPH), Finnish (FIN) and English and Scottish (GBR) ancestry, corresponding to the geographic regions used in the original GWAS studies. Of the 98 SNPs with a genotype R-squared greater than 0.8 (red points), 14 overlap DNaseI footprints in at least 1 cell type and 3 are associated with allelically imbalanced chromatin state, including the initial lead SNP (rs495337, rs492702 and rs2281217). (C) DNaseI cleavage density profiles surrounding two of the psoriasis linked variants that are associated with allelically imbalanced chromatin state (rs495337 and rs492702) for 16 human cells potentially involved in psoriasis pathogenesis. The genotypes of each cell type at rs495337 and rs492702 are indicated to the right of the plot. (D) DNaseI cleavage pattern surrounding the synonymous SNP rs492702 (left) and the synonymous SNP rs495337 (right) in NB4 cells. Binding elements overlapping DNaseI footprints are indicated below. (E) Shown is the chromatin accessibility associated with either the psoriasis risk or non-risk allele of rs492702 (left) and rs495337 (right) within NB4 cells ($P < 1x10-5$, Fisher's exact test). (F) (top) DNaseI cleavage pattern surrounding the psoriasis-linked non-coding SNP rs2281217 in Melanoma cells (RPMI_7951). (bottom) Shown is the chromatin accessibility associated with either the psoriasis risk or non-risk allele of rs2281217 within Melanoma cells (RPMI_7951) ($P < 1x10-2$, Fisher's exact test).

Figure 8.S6. TFs preferentially occupy coding bases from expressed genes.



D

Correlation of exonic footprint count with gene expression (per cell-type)

Cell Type	Pearson (r)	p-value
A549	0.149	7.05E-105
AG10803	0.191	2.50E-171
AoAF	0.158	7.37E-117
CD34	0.172	1.79E-139
CD3	0.223	4.89E-234
CD8	0.238	7.66E-268
fHeart	0.214	6.07E-216
fIntestine_Lg	0.202	5.75E-192
fIntestine_Sm	0.202	3.56E-191
fLung	0.152	3.56E-109
GM06990	0.233	5.09E-257
GM12865	0.154	1.30E-110
HAEpiC	0.144	6.31E-97
HAh	0.211	2.24E-209
HAsp	0.204	5.31E-196
HCFaa	0.187	1.02E-164
HCF	0.165	2.15E-128
HCM	0.168	3.39E-132
HCPEpiC	0.180	1.76E-152
HEEpiC	0.141	6.35E-94
HepG2	0.228	7.07E-246
hESCT0	0.092	6.56E-41
HFF	0.164	4.97E-126
HGF	0.180	3.08E-152
HMF	0.160	2.35E-120
HMVEC_dBlAd	0.185	1.86E-161
HMVEC_dBlNeo	0.232	4.52E-255
HMVEC_dLyNeo	0.208	2.64E-203
HMVEC_LBl	0.226	1.33E-240
HMVEC_LLy	0.225	2.39E-239
HPAF	0.163	2.77E-125
HPdLF	0.139	2.45E-90
HPF	0.160	1.30E-119
HRCE	0.083	2.30E-33
HSMM	0.161	9.48E-122
HUVEC	0.179	1.51E-150
HVMF	0.168	7.96E-132
IMR90	0.149	2.26E-104
iPS_19_11	0.097	1.18E-44
NB4	0.234	8.25E-259
NHA	0.117	8.13E-65
NHDF_Ad	0.140	2.15E-92
NHDF_Neo	0.157	1.17E-115
NHLF	0.154	1.00E-111
SAEC	0.076	5.59E-28
SkMC	0.119	8.03E-67
SKNSH	0.196	5.33E-181

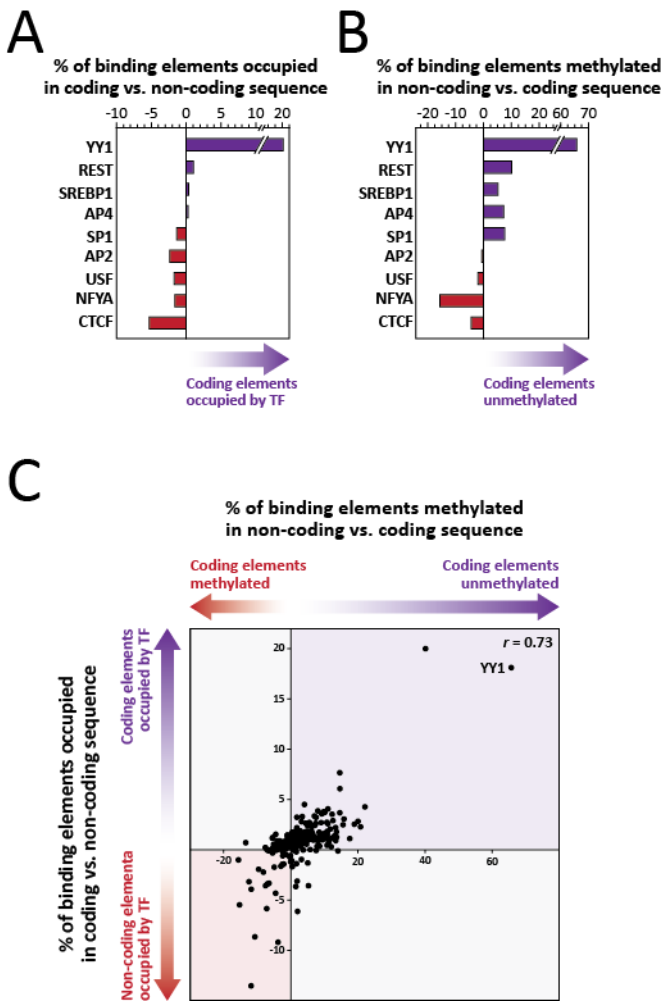
(A) The average density of DNaseI footprints within coding sequence and outside of coding sequence.

(B) *Long genes contain more TF footprints than short genes.* Box-and-whisker plots showing the association of coding gene length with the number of DNaseI footprints within that coding sequence. R-squared and p-values are from a linear regression of coding gene length vs. the number of DNaseI footprints within that coding sequence.

(C-D) *Transcription factors preferentially populate highly expressed genes.* (C) Shown is a box-and-whiskers plot of the gene expression in HMVEC_dBlNeo cells for genes with 0, 1-4 and 5+

coding DNaseI footprints. (D) Shown is the correlation of exonic footprints count with gene expression in 47 cell types with DNaseI footprint calls and gene expression data.

Figure 8.S7. TF occupancy within coding sequence is modelled by CpG methylation



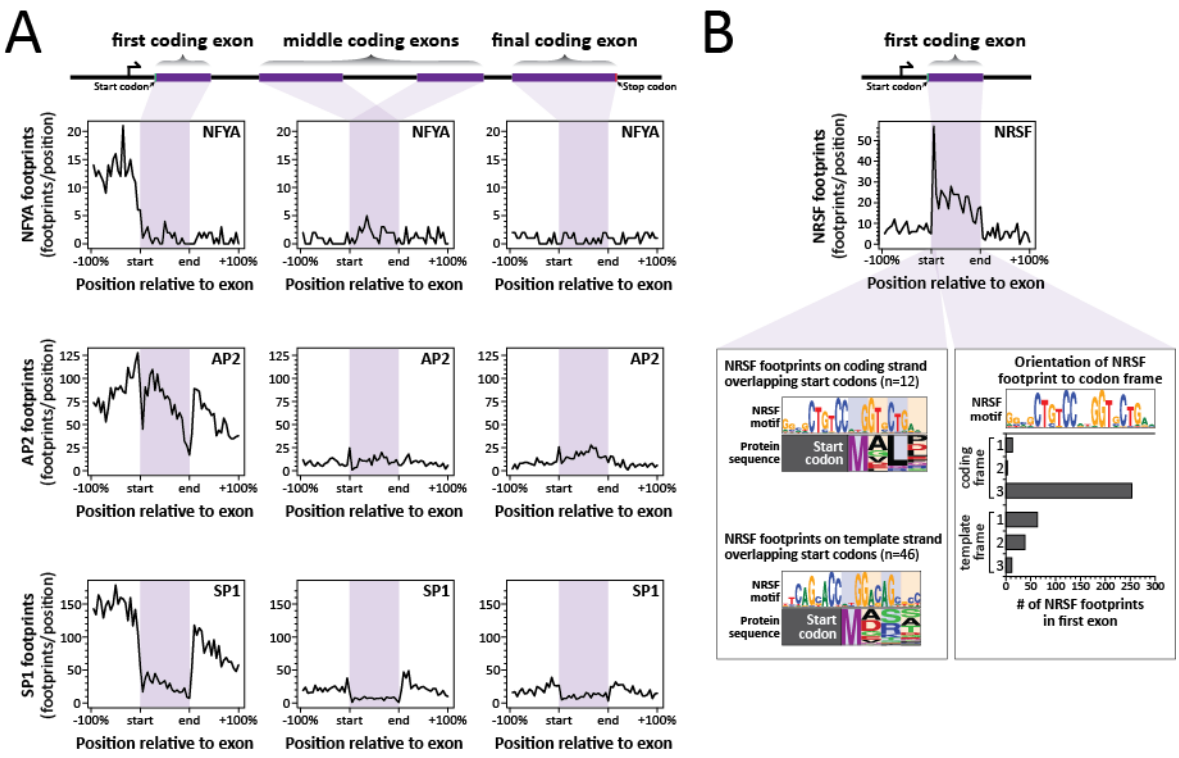
(A) The difference in the percentage of annotated coding and non-coding transcription factor binding elements overlapping a DNaseI footprint in each cell type. Positive values indicate that a greater fraction of coding binding elements are occupied as compared to non-coding elements.

(B) The difference in the percentage of CpGs within annotated coding and non-coding transcription factor binding elements methylated in each cell type. Positive values indicate that a

greater fraction of CpGs in coding binding elements are unmethylated as compared to CpGs in non-coding elements.

(C) Shown is a scatter plot of the preference of 232 TFs for occupying coding vs. non-coding binding elements (y-axis) and being CpG methylated at coding vs. non-coding binding elements (x-axis). Pearson correlation is shown in the upper right corner (r=0.73).

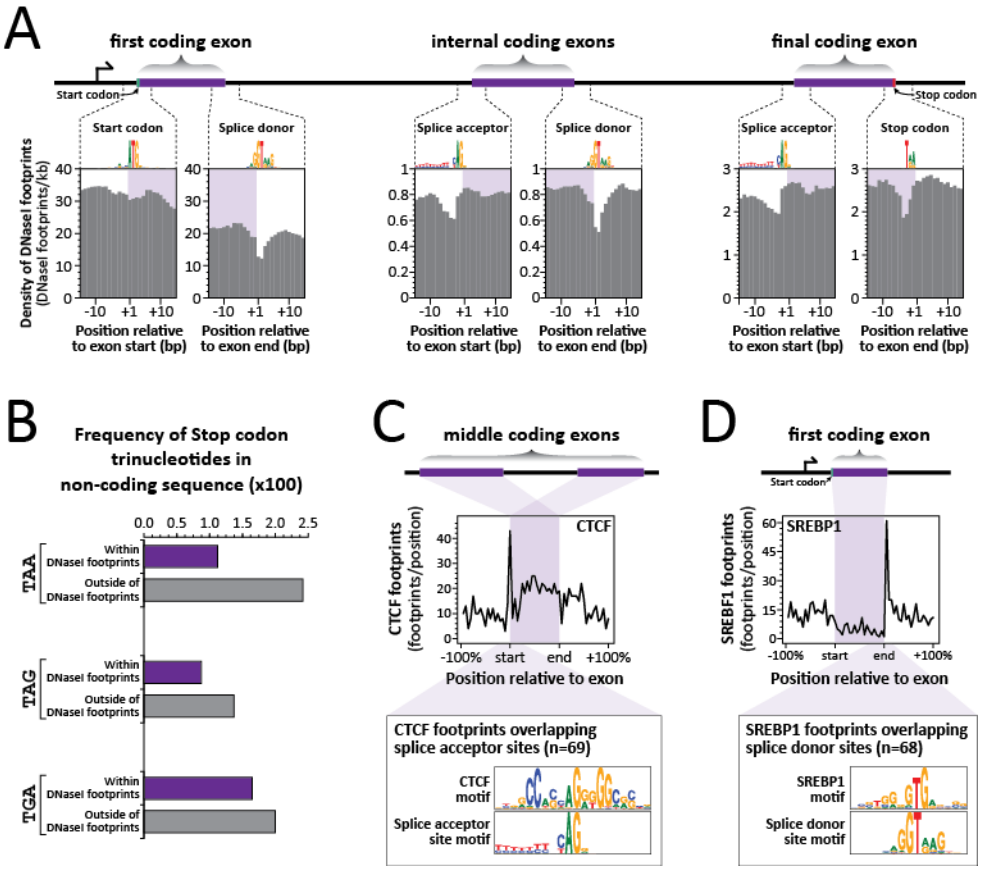
Figure 8.S8. TFs are influenced by and can exploit coding features of exons



(A) *NFYA, AP2 and SP1 preferentially avoid binding within coding sequence, start codons and splice junctions.* The density of (top) NFYA, (middle) AP2 and (bottom) SP1 DNaseI footprints relative to first, middle and final coding exons. Coding sequence is colored in purple.

(B) *NRSF binding elements preferentially align to the coding strand at the third frame of the codon and exploit start codons.* (top) The density of NRSF DNaseI footprints relative to first coding exons. (bottom-left) Shown is the NRSF motif model as well as a logo of the amino acid sequence at all occupied coding strand NRSF binding elements that overlap a start codon. (bottom-right) Shown is the number of NSRF binding elements within first coding exons that align to the three different coding positions along either the coding or template strand.

Figure 8.S9. TF occupancy at stop codons and splice sites reflects global evolution in TF preferences

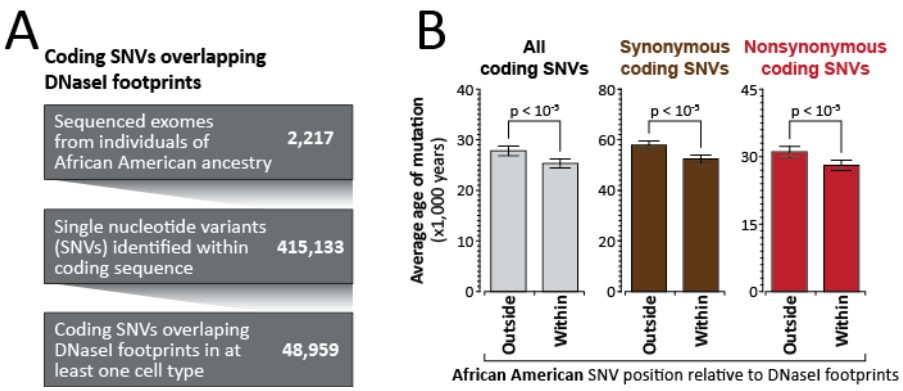


(A-B) *Transcription factors preferentially avoid occupying splice sites and stop codons.* (A) Shown is the density of DNaseI footprints surrounding start codons, splice donor sites, splice acceptor sites and stop codons. Sequence features of these elements are displayed as motif models and coding sequence is colored in purple. (B) The frequency of the stop codon trinucleotides TAA, TAG and TGA within and outside of non-coding DNaseI footprints.

(C) *CTCF binding elements exploit splice acceptor sites.* (top) The density of CTCF DNaseI footprints relative to middle coding exons. (bottom) Shown is the CTCF motif model in comparison with the splice acceptor site motif model.

(D) *SREBP1 binding elements exploit splice donor sites.* (top) The density of SREBP1 DNaseI footprints relative to first coding exons. (bottom) Shown is the SREBP1 motif model in comparison with the splice acceptor site motif model.

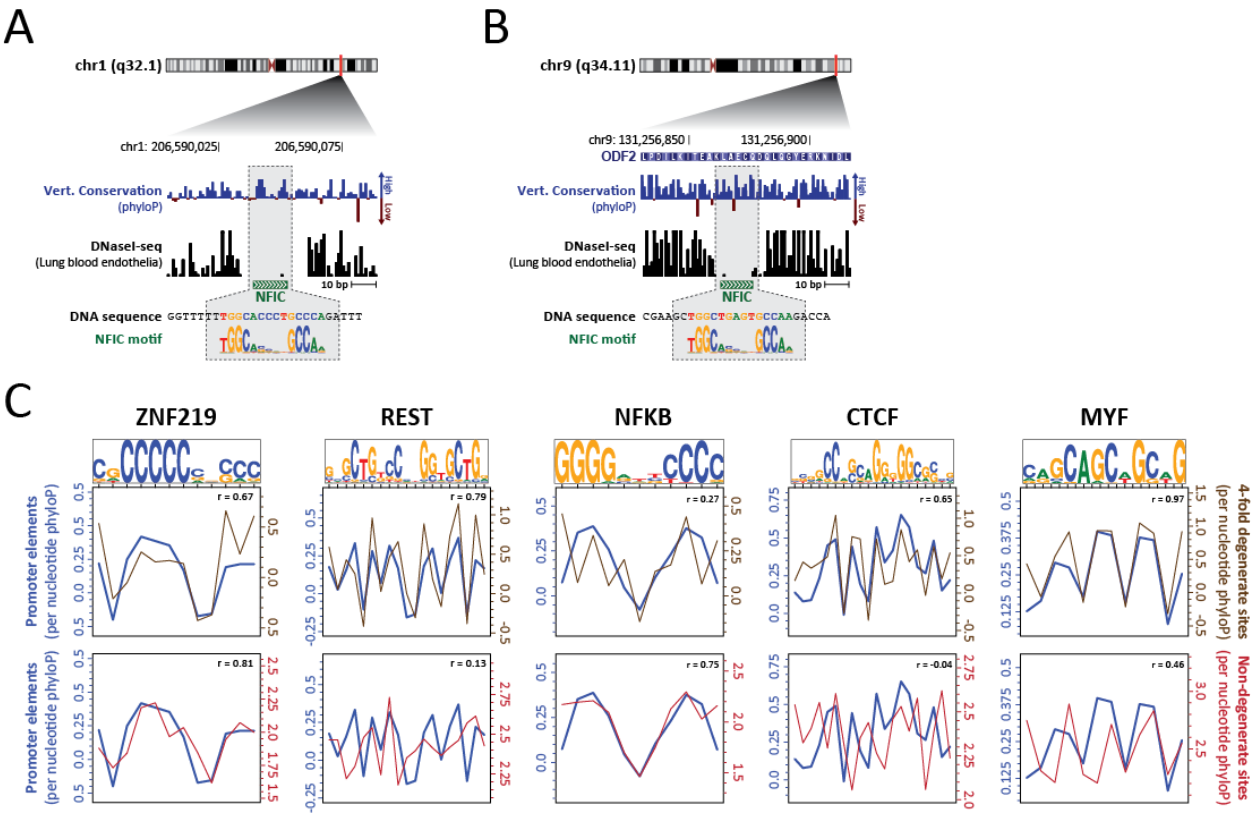
Figure 8.S10. TF binding elements impart populational constraint on coding sequence.



(A) 2,217 sequenced exomes from individuals of African American ancestry were utilized to identify SNVs overlapping DNaseI footprints in any of the 81 cell types.

(B) The average mutational age at all (grey), synonymous (brown) and non-synonymous (red) European coding SNVs identified within and outside of DNaseI footprints. Mutational ages and p-values were calculated as before (Fu et al., 2013).

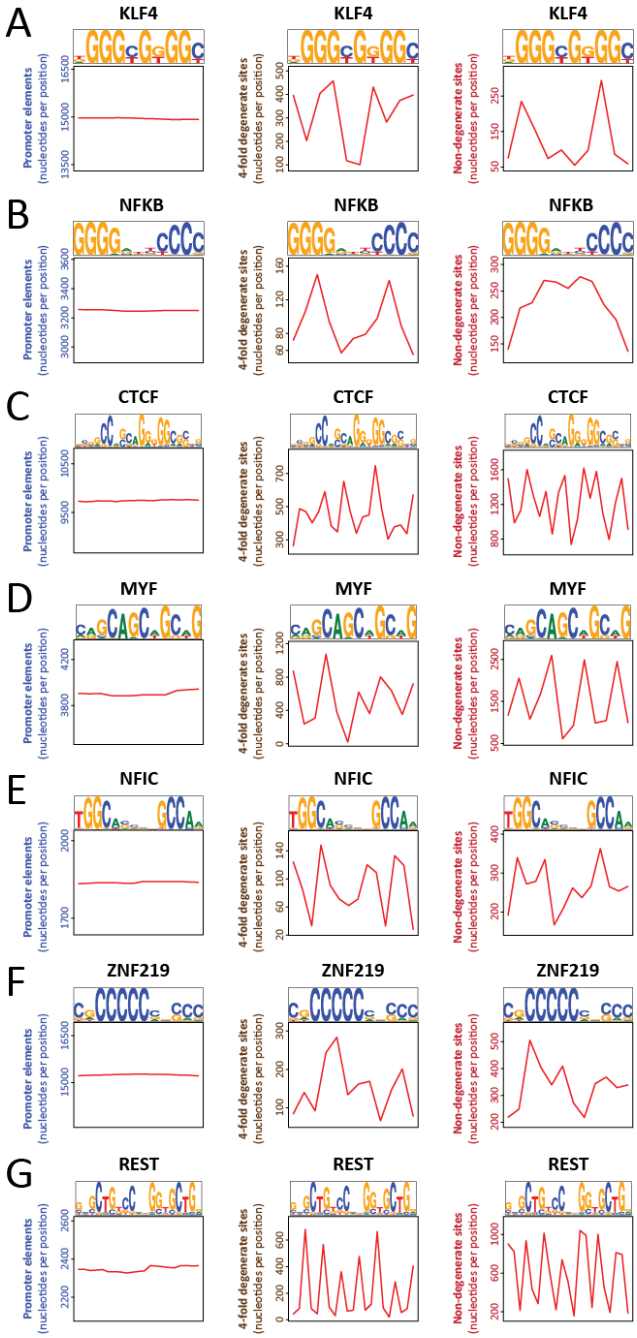
Figure 8.S11. Transcription factors influence codon choice.



(A-B) Per-nucleotide phyloP vertebrate conservation and DNaseI cleavage plots at (A) a non-coding and (B) a coding NFIC regulatory element. Note that NFIC imparts a stereotyped pattern of evolutionary constraint when bound at a non-coding regulatory elements.

(C) Average per-nucleotide conservation profile at footprinted binding elements for ZNF219 (left), REST (second), NFKB (third), CTCF (fourth) and MYF (right) overlapping non-coding bases within promoters (blue), 4-fold degenerate coding bases (brown) and non-degenerate coding bases (red). Pearson correlation values (r) between conservation profiles at promoter bases and 4-fold degenerate bases (top) or non-degenerate bases (bottom) are shown in the upper right corner of each plot.

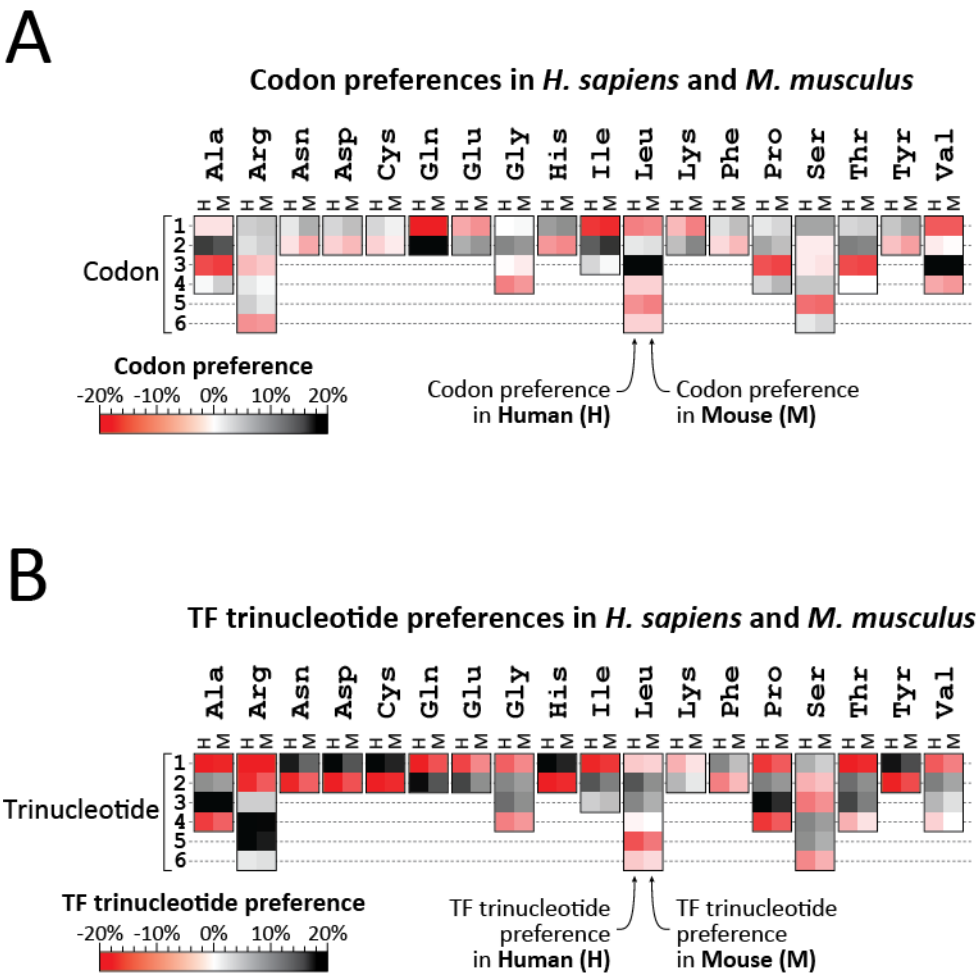
Figure 8.S12. 4-fold degenerate and non-degenerate bases overlapping TF elements.



(A-F) Shown are the number of bases overlapping different positions within footprinted (A) KLF4, (B) NFKB, (C) CTCF, (D) MYF, (E) NFIC, (F) ZNF291, and (G) REST binding

elements in any of the 81 cell types. Bases overlapping binding elements are broken into; (left/blue) promoter element bases; (middle/brown) 4-fold degenerate bases; and (right/red) non-degenerate bases.

Figure 8.S13. TFs sequence preferences and codon usage biases in M. musculus.



(A) Comparison of codon preferences in *H. sapiens* and *M. musculus*.

(B) Comparison of the TF trinucleotide preferences in *H. sapiens* and *M. musculus*. *H. sapiens* tri nucleotide preferences are derived from trinucleotides preferentially localized within non-coding DNaseI footprints in *H. sapiens* B-cells. *M. musculus* tri nucleotide preferences are derived from trinucleotides preferentially localized within non-coding DNaseI footprints in *M. musculus* B-cells.

8.6 – METHODS

DNaseI Footprinting

40 diverse human cell types were subjected to DNaseI digestion and high-throughput sequencing, following previous methods (Thurman et al., 2012; John et al., 2011; Hesselberth et al., 2009). To generate genome-wide per-nucleotide DNaseI cleavage profiles tags were aligned to the reference genome, build GRCh37/hg19 using Bowtie (Langmead et al., 2009), 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 and the 5' ends of the aligned sequencing tags at each position along the genome were summed. Data from additional cell types were utilized from Neph et al. 2012 (Neph et al., 2012c). FDR 1% DNase I footprints were identified in each cell type as previously described (Neph et al., 2012c). DNaseI footprints found in any of the 81 cell types were identified using the BEDOPS command *bedops -m* on each of the individual cell-type DNaseI footprint files (Neph et al., 2012a).

Capture-DGF

The two cell types HMF and AG10803 were DNaseI digested and proceeded on to Illumina PE library construction following the previous methods described above for DNaseI footprinting. The DNaseI libraries was amplified by PCR following the Exon Capture SureSelect protocol recommendations with minimal amount of PCR cycles and then purified using Agencourt AMPure XP beads (Beckman Coulter Genomics). Five hundred nanograms of each library was hybridized to Agilent SureSelect Human All Exon Kit (50 Mb) for 24 h at 65 C. The biotinylated probe/target hybrids were captured on Dynal MyOne Streptavidin T1 (Invitrogen), washed, eluted, and desalted and purified on a MinElute PCR column (Qiagen) as described in the SureSelect protocol. The eluted captured library was amplified by PCR with

minimal amount of PCR cycles. Amplified exon captured libraries were purified using Angencourt AMPure XP beads the samples were then quantified by Qubit dsDNA assay (Invitrogen). Samples were diluted to a working concentration of 10 nM. Cluster generation was performed for each sample and loaded on to single lane of an Illumina HiSeq flowcell. Paired end sequencing was performed for 36 cycles according to manufacturer's instructions.

Overlap of DNaseI footprints with coding sequence

Genomic regions annotated as coding were identified using the Consensus CDS (CCDS) database (Release 6). Each transcript within this database having two or more exons was utilized to identify first, internal and final coding exons, with first coding exons, by definition, always containing the methionine start codon, and final coding exons, by definition, always containing the stop codon. The density of footprints within exons was computed by first calculating the number of DNaseI footprints within a cell type that overlap coding sequence by at least 50% and dividing that by the total number of mappable bases within the coding sequence. The density of DNaseI footprints surrounding splice acceptor sites, splice donor sites, start codons and stop codons was calculated by summing the number of DNaseI footprints that overlap each base surrounding these genomic features in each of the 81 cell types.

Allelic chromatin imbalance measurements

Allelic chromatin imbalance was calculated as previously described (Neph et al., 2012c). Briefly, known autosomal single nucleotide variants (SNVs) were downloaded from the 1000 Genomes Project and SNVs mapping within 36 bases of each other were removed. DNaseI-seq reads overlapping a SNV were counted only if they contained no more than one mismatch, excluding the SNV, and duplicate DNaseI-seq reads were removed from this analysis. To test for allelic chromatin imbalance we first combined DNaseI cleavage reads overlapping a particular

SNV in all cell types heterozygous at that SNV. The difference in DNaseI cleavage reads containing each of the two alleles was tested using a two-sided binomial test, with SNVs having $P < 0.01$ considered as showing significant chromatin allelic imbalance. To test whether two sets of SNVs significantly differed in the proportion showing allelic imbalance, 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 onesided Fisher's exact test. The SIFT (Kumar et al., 2009) and PolyPhen-2 (Adzhubei et al., 2010) scores at each non-synonymous SNV were determined using Ensembl's Variant Effect Predictor (McLaren et al., 2010).

Transcription factor occupancy at SNPs in linkage with rs495337

Genome wide association study (GWAS) lead SNPs were downloaded from Maurano et al., 2012 (Maurano et al., 2012a). The synonymous coding variant rs495337, which is associated with Psoriasis (Capon et al., 2008; Stuart et al., 2010), was identified as overlapping DNaseI footprints in NB4 cells using BEDOPS (Neph et al., 2012a). To identify other SNPs linked with rs495337 that may also be contributing to the psoriasis association signal we utilized whole genome sequencing data from 267 individuals of Northern and Western European (CEPH), Finnish (FIN) and English and Scottish (GBR) ancestry (Durbin et al., 2010), corresponding to the geographic regions used in the original GWAS studies. The 98 SNPs with a genotype R-squared greater than 0.8 were used for further analysis. The allele at each of these SNPs in linkage with the rs495337 psoriasis risk allele (G allele) was also determined using the whole

genome sequencing data from these 267 individuals. Of these 99 SNPs (including rs495337), 14 overlapped a DNaseI footprint in at least one of the 81 cell types. In total three of these SNPs were associated with allelic chromatin imbalance at the overlapping DNaseI footprints (rs495337 in NB4 cells; rs492702 in NB4 cells; and rs2281217 in RPMI cells) and two of these were synonymous coding variants (rs495337 and rs492702).

Distribution of DNaseI footprinted TF recognition sequences

Human genome build hg19 was scanned for predicted TRANSFAC (Matys et al., 2006), JASPAR Core (Bryne et al., 2008) and UniPROBE (Newburger and Bulyk, 2009) motif-binding sites using FIMO (Bailey et al., 2009), version 4.6.1, with a maximum p value threshold of 10^-5 and defaults for other parameters. We marked a putative binding site as being occupied within a cell type if it overlapped a DNaseI footprint within that cell types by at least 3 nt, as previously described in (Neph et al., 2012c). The distribution of DNaseI footprinted TF binding elements within and surrounding coding exons was computed by calculating the distance of the mid-point of the TF binding element to the beginning of the overlapping or neighboring coding exon, relative to the length of that coding exon. TF recognition sequence motif logos were generated using Weblogo 3 and all of the exonic binding sites for that TF (Crooks et al., 2004). Start codon, stop codon, splice acceptor site and splice donor site motifs were generated by aligning all relevant genomic elements using CCDS coding sequence annotations and calculating base enrichments using Weblogo 3 (Crooks et al., 2004).

Protein domain architecture overlapping TF binding sites

The position of NRSF binding sites (JASPAR Core model MA0138.2) relative to the codon frame within the first exon was analyzed using Consensus CDS (CCDS) database gene models (Release 6). Possible alignments included: (1) the first frame of the coding strand; (2) the

second frame of the coding strand; (3) the third frame of the coding strand; (4) the first frame of the template strand; (5) the second frame of the template strand; and (6) the third frame of the template strand. Of the 382 footprinted NRSF recognition sequences within the first exon, 253 were found aligned with the third frame of the coding strand. The protein domain architecture overlapping these binding elements were analyzed using: (1) signal peptide domain predictions from the SignalP 4.1 Server (Petersen et al., 2011); (2) transmembrane domain predictions from the TMHMM Server v. 2.0 (Krogh et al., 2001); (3) leucine-rich nuclear export signal predictions from the NetNES 1.1 Server (La Cour et al., 2004); and (4) the Superfamily database of structural and functional protein domain annotations (Gough et al., 2001). Logos of the frequency of amino acids overlapping TF recognition sequences were generated using Weblogo 3 (Crooks et al., 2004).

Evolutionary constraint at footprinted coding sequences

Evolutionary constraint at TF binding sites was calculated using phyloP evolutionary conservation scores (Pollard et al., 2010). 4-fold degenerate bases were identified based on the sequence features of each codon (e.g. the third position of the following codons: CTA, CTT, CTG, CTC, GTA, GTT, GTG, GTC, TCA, TCT, TCG, TCC, CCA, CCT, CCG, CCC, ACA, ACT, ACG, ACC, GCA, GCT, GCG, GCC, CGA, CGT, CGG, CGC, GGA, GGT, GGG, GGC). Non-degenerate bases were identified based on the sequence features of each codon (e.g. the first and second position of every codon except TTA, TTG, CTA, CTT, CTG, CTC, AGT, AGC, TCA, TCT, TCG, TCC, AGA, AGG, CGA, CGT, CGG, CGC. And the second position of TTA, TTG, CTA, CTT, CTG, CTC, AGA, AGG, CGA, CGT, CGG, CGC). To generate the conservation profile of a TF within exons, we calculated the average phyloP at all 4-fold degenerate bases, or non-degenerate bases, overlapping each position within the TF binding

element. Only TFs with 20 or more data points contributing to each position within the binding element were used for further analysis. The number of bases contributing to each position within the binding element is shown in **Fig. 8.S5**. To generate the conservation profile of a TF within promoters, a similar process was performed for all bases overlapping TF binding elements within non-coding promoter regions. Pearson correlations were calculated to determine the similarity of the conservation profile of a TF at promoter elements and coding 4-fold degenerate and coding non-degenerate sites.

Mutational age at footprinted coding sequences

Exome sequences were obtained from 6,515 individuals (4,298 of European American ancestry and 2,217 of African American ancestry) (Fu et al., 2013). Coding variants, specifically synonymous variants and nonsynonymous variants, were classified according to whether they overlapped a DNaseI footprint in any of the 81 tested cell types. Average mutation age for each category was calculated as previously described (Fu et al., 2013). Briefly, mutation age was estimated based on a derivation of Griffiths and Tavare (Griffiths and Tavare, 1998) by generating a series of coalescent trees under a specified demographic model for European and African American populations (Tennessen et al., 2012). Average mutation age across variants for each category was defined as a weighted average of mutation age, where the weights are calculated according to the site-frequency-spectrum (SFS) in this category. Average mutation age in different categories was compared through permutations to identify significant differences (Fu et al., 2013).

Codon usage biases and TF footprint trinucleotide frequencies

Coding usage biases were obtained using CCDS gene annotations downloaded from the UCSC genome browser, corresponding to human build GRCh37/hg19 or mouse build

NCBI37/mm9. For Drosophila, UCSC refGene was used, and for Arabidopsis the TAIR10_GFF3_genes.gff file from TAIR was used, excluding codons containing an 'N' in the reference. Individual codon locations were parsed into BED format, excluding start codons and any codons overlapping a splice site or that were ambiguous due to overlapping annotations in different reading frames. Coding annotations containing one or more internal stops in the reference sequence were also excluded. Overlaps of codon locations with footprint calls were determined using BEDOPS codons that partially overlapped a footprint were excluded. Non-coding trinucleotide frequencies were obtained using the genomic space uniquely mappable by 36-mer sequencing tags. CCDS coding exons, as well as RepeatMasker annotations also downloaded from the UCSC genome browser, were then subtracted from this space using BEDOPS, and the remaining regions divided by overlap with footprint calls. Finally, all reference-strand genomic 3-mers in the footprint or non-footprint space were tabulated separately.

Chapter 9 – Future directions

Although much progress has been made in the field of gene regulation since Jacob and Monod's initial findings nearly 50 years ago (Jacob and Monod, 1961), the hope of understanding how gene regulation generates the human body plan is still a distant goal. Below I have highlighted some of the methodological and theoretical challenges that face our field and need to be addressed in the near future.

First, targeted proteomics needs to move beyond the quantification of a handful of proteins into the simultaneous quantification of tens of thousands of proteins. Targeted proteomics offers superior quantification and detection of proteins, but is constrained by the need to know the physiochemical properties of a peptide before searching for it. In well-studied organisms such as human, mouse, *D. melanogaster*, *C. elegans*, *S. Cerevisiae* and *E. coli*, these physiochemical properties should be developed and made available for every protein isoform and post-translational modification. Although this task seems daunting, the utility of such a library to the scientific community would be tremendous. Thankfully, recent advances in instrument design and data independent acquisition (DIA) workflows (Marx, 2012) suggest that the utilization of such libraries to quantify thousands of peptides simultaneously is just around the corner.

Second, in this thesis I have demonstrated several methods to obtain global maps of TF occupancy. However, these experiments all required between 1-50 million cells as input. This feature limited the experiments and conclusions presented in this thesis, as critical developmental tissues are often scarce and hard to isolate, and all of these measurements represent population averages. Consequently, in order to gain a fuller understanding of chromatin biology and the developmental patterning of gene regulation, we need to develop new methods for generating

global maps of TF occupancy and chromatin accessibility on a single cell level. Although some success has been achieved in mapping *in vivo* single cell occupancy in bacteria (Elf et al., 2007; Hammar et al., 2012) and on select loci in mammals (McNally et al., 2000; Voss et al., 2011; Gebhardt et al., 2013), we need entirely new technologies to move to genome-wide single cell measures of TF occupancy in human nuclei.

Finally, our 'network-level' understanding of gene regulation is severely limited. We currently have the methodology to map the occupancy patterns of TFs genome-wide, yet our ability to interpret these maps is woefully inadequate. Although much has been made over the past decade regarding chromatin conformation technologies (Dekker et al., 2002; Lieberman-Aiden et al., 2009), we currently do not know how to efficiently identify which distal regulatory elements act on a specific gene. Similarly, while a single gene can be regulated by multiple TFs bound at genomically distinct regulatory elements (Thurman et al., 2012; Neph et al., 2012b), we do not yet know how to properly integrate the qualitative and quantitative aspects of TF co-occupancy and TF co-regulation. Furthermore, nearly all studies of human gene regulation focus on isolated cell types. However, human tissues and organs are comprised of numerous functionally distinct cell types and the regulatory landscape of each of these cell types can be intricately dependent upon that of its neighboring cell types (Spemann, 1918; Waddington, 1940). Consequently, if we are going to gain a 'network-level' understanding of gene regulation during human development, we need to be able to analyze cell-selective regulatory networks within the context of their surrounding tissues (Barabási and Oltvai, 2004).

Finally, as we move into the era where global maps of TF occupancy are readily available, we hope to better understand the ultimate question: how our genome generates complex human physiology and wards against disease.

ACKNOWLEDGMENTS

Science is a collaborative process. The research that I have described in this thesis was only possible thanks to the collaborative and supportive environment of the Stamatoyannopoulos lab and the Department of Genome Sciences. I give thanks to everyone who contributed to the Encyclopedia of DNA Elements (ENCODE) Consortium, NIH Roadmap Epigenomics Mapping Consortium and Hao Wang (Stamatoyannopoulos lab). This dissertation would not have been possible without their tireless efforts, striving for the highest data quality standards.

Within the Department of Genome Sciences, I have been fortunate enough to work with many excellent researchers. Specifically, my efforts studying CTCF and its many biological intricacies would not have been possible without the mentorship of John Stamatoyannopoulos and the countless discussions with Matt Maurano and Hao Wang. Hao's tireless efforts perfecting CTCF ChIP-seq turned 'salted ChIPs' from a funny name into a viable experiment. Furthermore, Matt and Hao's work studying the tissue selectivity and sequence determinants of CTCF occupancy provided countless insights into the role of chromatin in modeling TF occupancy patterns.

Similarly, my efforts studying DNaseI footprinting patterns would not have been possible without the tremendous data resources that the Stamatoyannopoulos lab produced as part of the ENCODE and Roadmap Epigenomics consortiums, as well as strong collaborations with different members of the Stamatoyannopoulos lab, including: Shane Neph, Alex Reynolds, Eric Haugen, Bob Thurman, Richard Sandstrom, Eric Rynes, Rich Humbert, Eunice Choi, Jeff Vierstra, Tony Shafer, Sam John, Kristen Lee and Tanya Kutyavin. Shane's computational acumen was critical for developing the core DNaseI footprint identification algorithm used

throughout these studies, and the ‘Lineages’ and ‘Networks’ projects would not have been possible without our close collaboration. Furthermore, Eric Haugen’s help was indispensable for turning the observation of coding DNaseI footprints into an actual story.

In addition to these members of the Stamatoyannopoulos lab, I would also like to thank other members of the UW Genome Sciences department who were instrumental to several of these analyses. First, I thank Elhanan Borenstein for providing numerous insights into how to approach transcriptional regulation from a networks perspective. Second, I thank Josh Akey and Ben Vernot for teaching me how to think about transcriptional regulation from an evolutionary biology and population genetics perspective.

My efforts developing targeted proteomic assays for human transcription factors would not have been possible without the mentorship of Mike MacCoss and the many members of the MacCoss lab and the UW Proteomics Resource Center. Specifically, I thank Greg Finney, Genn Merrihew, Daniela Tomazela, Michael Bereman and Priska von Haller for their assistance teaching me mass spectrometry. I also thank Brendan MacLean and Vagisha Sharma and the rest of the Skyline and Panorama team—without their outstanding software tools, many of the analyses described in this thesis would have been impossible.

I also sincerely thank Jim Thomas, Michael MacCoss, Stan Fields, Steve Henikoff and Rachel Klevit, who as part of my thesis committee have provided outstanding feedback and guidance during this process.

I would also like to thank the many researchers and teachers who helped cultivate my scientific interests over the past decade. I will always be indebted to George Stamatoyannopoulos and Pat Navas, who let me volunteer with them while I was in high school,

and patiently taught me bench science. Similarly, I thank my college research mentors Piers Nash and Bernard Liu, who taught me how to develop a research project.

I am grateful and inspired by my mentor, John Stamatoyannopoulos, for all the opportunities he provided during my graduate training. John's scientific vision and encyclopedic knowledge has been indispensable for my graduate education, and I am forever grateful for his untiring efforts to teach me and help me grow as a scientist. John's passion for science has continually motivated me to leap into multiple new areas of research, enabling me to approach a broader set of questions than I thought possible when I started my thesis.

I owe deep gratitude to the late Tamara Stevens. Tamara's tireless fight against cancer showed me the power of biomedical medicine and the importance of translating basic science discoveries into clinical practice.

Finally, I would also like to thank my family for their unwavering support. I would not be in science if not for my dad, who demonstrated how exciting academic research can be. I will always strive to fill your shoes (hopefully not with soda this time…). I thank my mom for teaching me to never be afraid and my sister for teaching me to enjoy more than just science. Lastly, this work would not have been possible without the love and dedication of my college sweetheart and wife Emily.

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. Nature methods *7*, 248–249.

Akhtar-Zaidi, B., Cowper-Sal-lari, R., Corradin, O., Saiakhova, A., Bartels, C. F., Balasubramanian, D., Myeroff, L., Lutterbaugh, J., Jarrar, A., Kalady, M. F., et al. (2012). Epigenomic enhancer profiling defines a signature of colon cancer. Science *336*, 736–739.

Alon, U. (2006). An Introduction to Systems Biology: Design Principles of Biological Circuits 1st ed. (Chapman and Hall/CRC).

Andrews, N. C., Erdjument-Bromage, H., Davidson, M. B., Tempst, P., and Orkin, S. H. (1993). Erythroid transcription factor NF-E2 is a haematopoietic-specific basic-leucine zipper protein. Nature *362*, 722–728.

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and Complexity in DNA Recognition by Transcription Factors. Science *324*, 1720–1723.

Von Baer, K. E. (1828). Über Entwicklungsgeschichte der Thiere. Beobachtung und Reflexion. (Königsberg: Gebrüder Bornträger).

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research *37*, W202–W208.

Bailey, T. L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology *2*, 28–36.

Barabási, A.-L., and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. Nature Reviews Genetics *5*, 101–113.

Bard, J., Rhee, S. Y., and Ashburner, M. (2005). An ontology for cell types. Genome biology *6*, R21.

Barthel, K. K. B., and Liu, X. (2008). A transcriptional enhancer from the coding region of ADAMTS5. PloS one *3*, e2184.

Basso, K., Margolin, A. A., Stolovitzky, G., Klein, U., Dalla-Favera, R., and Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. Nature Genetics *37*, 382–390.

Bassuk, A. G., Barton, K. P., Anandappa, R. T., Lu, M. M., and Leiden, J. M. (1998). Expression pattern of the Ets-related transcription factor Elf-1. Molecular medicine (Cambridge, Mass.) *4*, 392–401.

Beard, J. (1902). Embryological Aspects and Etiology of Carcinoma. Lancet *159*, 1758–1761.

Becker, M., Baumann, C., John, S., Walker, D. A., Vigneron, M., McNally, J. G., and Hager, G. L. (2002). Dynamic behavior of transcription factors on a natural promoter in living cells. EMBO Reports *3*, 1188–1194.

Berezney, R., and Coffey, D. S. (1974). Identification of a nuclear protein matrix. Biochemical and bBophysical Research Communications *60*, 1410–1417.

Berger, M. F., Badis, G., Gehrke, A. R., Talukder, S., Philippakis, A. A., Peña-Castillo, L., Alleyne, T. M., Mnaimneh, S., Botvinnik, O. B., Chan, E. T., et al. (2008). Variation in Homeodomain DNA Binding Revealed by High-Resolution Analysis of Sequence Preferences. Cell *133*, 1266–1276.

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nature Biotechnology *24*, 1429–1435.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

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

Biddie, S. C., John, S., Sabo, P. J., Thurman, R. E., Johnson, T. A., Schiltz, R. L., Miranda, T. B., Sung, M.-H., Trump, S., Lightman, S. L., et al. (2011). Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. Molecular cell *43*, 145–155.

Biggin, M. D. (2011). Animal transcription networks as highly connected, quantitative continua. Developmental cell *21*, 611–626.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes & Development *16*, 6–21.

Bloom, K. S., and Anderson, J. N. (1978). Fractionation of hen oviduct chromatin into transcriptionally active and inactive regions after selective micrococcal nuclease digestion. Cell *15*, 141–150.

Bowers, S. R., Mirabella, F., Calero-Nieto, F. J., Valeaux, S., Hadjur, S., Baxter, E. W., Merkenschlager, M., and Cockerill, P. N. (2009). A Conserved Insulator That Recruits CTCF and Cohesin Exists between the Closely Related but Divergently Regulated Interleukin-3 and Granulocyte-Macrophage Colony-Stimulating Factor Genes. Molecular and Cellular Biology *29*, 1682–1693.

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947–956.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. Cell *132*, 311–322.

Bryne, J. C., Valen, E., Tang, M.-H. E., Marstrand, T., Winther, O., Da Piedade, I., Krogh, A., Lenhard, B., and Sandelin, A. (2008). JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. Nucleic Acids Research *36*, D102–D106.

Bulmer, M. (1988). Codon usage and intragenic position. Journal of theoretical biology *133*, 67–71.

Bulmer, M. (1987). Coevolution of codon usage and transfer RNA abundance. Nature *325*, 728–730.

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. Genetics *129*, 897–907.

Burch, J. B., and Weintraub, H. (1983). Temporal order of chromatin structural changes associated with activation of the major chicken vitellogenin gene. Cell *33*, 65–76.

Burton, D. R., Butler, M. J., Hyde, J. E., Phillips, D., Skidmore, C. J., and Walker, I. O. (1978). The interaction of core histones with DNA: equilibrium binding studies. Nucleic Acids Research *5*, 3643–3664.

Capon, F., Bijlmakers, M.-J., Wolf, N., Quaranta, M., Huffmeier, U., Allen, M., Timms, K., Abkevich, V., Gutin, A., Smith, R., et al. (2008). Identification of ZNF313/RNF114 as a novel psoriasis susceptibility gene. Human molecular genetics *17*, 1938–1945.

Carlini, D. B., and Stephan, W. (2003). In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein. Genetics *163*, 239–243.

Carr, A., and Biggin, M. D. (1999). A comparison of in vivo and in vitro DNA-binding specificities suggests a new model for homeoprotein DNA binding in Drosophila embryos. The EMBO journal *18*, 1598–1608.

Carr, S. A., and Anderson, L. (2008). Protein Quantitation through Targeted Mass Spectrometry: The Way Out of Biomarker Purgatory? Clin Chem *54*, 1749–1752.

Carro, M. S., Lim, W. K., Alvarez, M. J., Bollo, R. J., Zhao, X., Snyder, E. Y., Sulman, E. P., Anne, S. L., Doetsch, F., Colman, H., et al. (2010). The transcriptional network for mesenchymal transformation of brain tumours. Nature *463*, 318–325.

Carroll, J. S., Liu, X. S., Brodsky, A. S., Li, W., Meyer, C. A., Szary, A. J., Eeckhoute, J., Shao, W., Hestermann, E. V, Geistlinger, T. R., et al. (2005). Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. Cell *122*, 33–43.

Cavalli, G., and Paro, R. (1998). The Drosophila Fab-7 chromosomal element conveys epigenetic inheritance during mitosis and meiosis. Cell *93*, 505–518.

Chang, C.-P., Stankunas, K., Shang, C., Kao, S.-C., Twu, K. Y., and Cleary, M. L. (2008). Pbx1 functions in distinct regulatory networks to pattern the great arteries and cardiac outflow tract. Development *135*, 3577–3586.

Chudinova, E. M., Ivanov, P. A., and Nadezhdina, E. S. (2004). [Large subunit of translation initiation factor--3 p170 contains potentially functional nuclear localization signals]. Molekuliarnaia biologiia *38*, 684–691.

Clausell, J., Happel, N., Hale, T. K., Doenecke, D., and Beato, M. (2009). Histone H1 subtypes differentially modulate chromatin condensation without preventing ATP-dependent remodeling by SWI/SNF or NURF. PloS one *4*.

Cohen, J. (1963). Living embryos: An introduction to the study of animal development (New York: Pergamon Press).

Coleman, J. R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., and Mueller, S. (2008). Virus attenuation by genome-scale changes in codon pair bias. Science (New York, N.Y.) *320*, 1784–1787.

Cook, A. L., Smith, A. G., Smit, D. J., Leonard, J. H., and Sturm, R. A. (2005). Co-expression of SOX9 and SOX10 during melanocytic differentiation in vitro. Experimental cell research *308*, 222–235.

Corcoran, L. M., Karvelas, M., Nossal, G. J., Ye, Z. S., Jacks, T., and Baltimore, D. (1993). Oct-2, although not required for early B-cell development, is critical for later B-cell maturation and for postnatal survival. Genes & development *7*, 570–582.

Coulondre, C., Miller, J. H., Farabaugh, P. J., and Gilbert, W. (1978). Molecular basis of base substitution hotspots in Escherichia coli. Nature *274*, 775–780.

La Cour, T., Kiemer, L., Mølgaard, A., Gupta, R., Skriver, K., and Brunak, S. (2004). Analysis and prediction of leucine-rich nuclear export signals. Protein engineering, design & selection : PEDS *17*, 527–536.

Crooks, G. E., Hon, G., Chandonia, J.-M. M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. Genome Research *14*, 1188–1190.

Cuddapah, S., Jothi, R., Schones, D. E., Roh, T.-Y., Cui, K., and Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Research *19*, 24–32.

Cuellar-Partida, G., Buske, F. A., McLeay, R. C., Whitington, T., Noble, W. S., and Bailey, T. L. (2012). Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics (Oxford, England) *28*, 56–62.

Darwin, C. (1859). On the Origin of Species (London: John Murray).

Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., et al. (2002a). A genomic regulatory network for development. Science *295*, 1669–1678.

Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C., et al. (2002b). A provisional regulatory gene network for specification of endomesoderm in the sea urchin embryo. Developmental Biology *246*, 162–190.

Davis, R. L., Weintraub, H., and Lassar, A. B. (1987). Expression of a single transfected cDNA converts fibroblasts to myoblasts. Cell *51*, 987–1000.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing Chromosome Conformation. Science *295*, 1306–1311.

Dignam, J. D., Lebovitz, R. M., and Roeder, R. G. (1983). Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. Nucleic Acids Research *11*, 1475–1489.

Domazet-Lošo, T., and Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature *468*, 815–818.

Donohoe, M. E., Zhang, L.-F., Xu, N., Shi, Y., and Lee, J. T. (2007). Identification of a Ctcf cofactor, Yy1, for the X chromosome binary switch. Molecular cell *25*, 43–56.

Dorschner, M. O., Hawrylycz, M., Humbert, R., Wallace, J. C., Shafer, A., Kawamoto, J., Mack, J., Hall, R., Goldy, J., Sabo, P. J., et al. (2004). High-throughput localization of functional elements by quantitative chromatin profiling. Nature Methods *1*, 219–225.

Drew, H. R., and Travers, A. A. (1984). DNA structural variations in the E. coli tyrT promoter. Cell *37*, 491–502.

Drew, H. R., and Travers, A. A. (1985). Structural junctions in DNA: the influence of flanking sequence on nuclease digestion specificities. Nucleic Acids Research *13*, 4445–4467.

Drmanac, R., Sparks, A. B., Callow, M. J., Halpern, A. L., Burns, N. L., Kermani, B. G., Carnevali, P., Nazarenko, I., Nilsen, G. B., Yeung, G., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science *327*, 78–81.

Duan, J., Wainwright, M. S., Comeron, J. M., Saitou, N., Sanders, A. R., Gelernter, J., and Gejman, P. V (2003). Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Human molecular genetics *12*, 205–216.

Duboule, D. (1994). Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. Development, 135–142.

Duma, D., Jewell, C. M., and Cidlowski, J. A. (2006). Multiple glucocorticoid receptor isoforms and mechanisms of post-translational modification. The Journal of Steroid Biochemistry and Molecular Biology *102*, 11–21.

Durbin, R. M., Altshuler, D. L., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., et al. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

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

Egli, D., Birkhoff, G., and Eggan, K. (2008). Mediators of reprogramming: transcription factors and transitions through mitosis. Nature reviews. Molecular cell biology *9*, 505–516.

Elf, J., Li, G.-W., and Xie, X. S. (2007). Probing transcription factor dynamics at the single-molecule level in a living cell. Science (New York, N.Y.) *316*, 1191–1194.

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

Eng, J. K., McCormack, A. L., and Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry *5*, 976–989.

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Eyre-Walker, A., and Bulmer, M. (1993). Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic acids research *21*, 4599–4603.

Felsenfeld, G. (1996). Chromatin unfolds. Cell *86*, 13–19.

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

Filion, G. J., Van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., Brugman, W., De Castro, I. J., Kerkhoven, R. M., Bussemaker, H. J., et al. (2010). Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. Cell *143*, 212–224.

Filippova, G. N., Fagerlie, S., Klenova, E. M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P. E., Collins, S. J., and Lobanenkov, V. V (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. Molecular and Cellular Biology *16*, 2802–2813.

Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S., and MacCoss, M. J. (2006). Analysis of Peptide MS/MS Spectra from Large-Scale Proteomics Experiments Using Spectrum Libraries. Analytical Chemistry *78*, 5678–5684.

Fried, M., and Crothers, D. M. (1981). Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. Nucleic acids research *9*, 6505–6525.

Friedman, J. R., Fredericks, W. J., Jensen, D. E., Speicher, D. W., Huang, X. P., Neilson, E. G., and Rauscher, F. J. (1996). KAP-1, a novel corepressor for the highly conserved KRAB repression domain. Genes & Development *10*, 2067–2078.

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D. A., et al. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature *493*, 216–220.

Fusaro, V. A., Mani, D. R., Mesirov, J. P., and Carr, S. A. (2009). Prediction of high-responding peptides for targeted protein assays by mass spectrometry. Nature Biotechnology *27*, 190–198.

Galas, D. J., and Schmitz, A. (1978). DNAase footprinting a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Research *5*, 3157–3170.

Garner, M. M., and Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic acids research *9*, 3047–3060.

Gascoyne, D. M., Long, E., Veiga-Fernandes, H., De Boer, J., Williams, O., Seddon, B., Coles, M., Kioussis, D., and Brady, H. J. M. (2009). The basic leucine zipper transcription factor E4BP4 is essential for natural killer cell development. Nature immunology *10*, 1118–1124.

Gebhardt, J. C. M., Suter, D. M., Roy, R., Zhao, Z. W., Chapman, A. R., Basu, S., Maniatis, T., and Xie, X. S. (2013). Single-molecule imaging of transcription factor binding to DNA in live mammalian cells. Nature Methods *10*, 421–426.

Gerstein, M. B., Lu, Z. J., Van Nostrand, E. L., Cheng, C., Arshinoff, B. I., Liu, T., Yip, K. Y., Robilotto, R., Rechtsteiner, A., Ikegami, K., et al. (2010). Integrative Analysis of the Caenorhabditis elegans Genome by the modENCODE Project. Science *330*, 1775–1787.

Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K., and Weissman, J. S. (2003). Global analysis of protein expression in yeast. Nature *425*, 737–741.

Gilbert, W., and Müller-Hill, B. (1967). The lac operator is DNA. Proceedings of the National Academy of Sciences of the United States of America *58*, 2415–2421.

Gleason, D. F., and Mellinger, G. T. (1974). Prediction of prognosis for prostatic adenocarcinoma by combined histological grading and clinical staging. The Journal of urology *111*, 58–64.

Goldstein, A. S., Stoyanova, T., and Witte, O. N. (2010). Primitive origins of prostate cancer: in vivo evidence for prostate-regenerating cells and prostate cancer-initiating cells. Molecular oncology *4*, 385–396.

Gordân, R., Shen, N., Dror, I., Zhou, T., Horton, J., Rohs, R., and Bulyk, M. L. (2013). Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape. Cell Reports.

Goshima, N., Kawamura, Y., Fukumoto, A., Miura, A., Honma, R., Satoh, R., Wakamatsu, A., Yamamoto, J., Kimura, K., Nishikawa, T., et al. (2008). Human protein factory for converting the transcriptome into an in vitro–expressed proteome. Nature Methods *5*, 1011–1017.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. Journal of molecular biology *313*, 903–919.

Gouy, M., and Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. Nucleic acids research *10*, 7055–7074.

Graf, M., Bojak, A., Deml, L., Bieler, K., Wolf, H., and Wagner, R. (2000). Concerted action of multiple cis-acting sequences is required for Rev dependence of late human immunodeficiency virus type 1 gene expression. Journal of virology *74*, 10822–10826.

Graf, T., and Enver, T. (2009). Forcing cells to change lineages . Nature *462*, 587–594.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., and Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic acids research *9*, r43–74.

Gray, N. K., and Hentze, M. W. (1994). Regulation of protein synthesis by mRNA structure. Molecular biology reports *19*, 195–200.

Griffiths, R. C., and Tavare, S. (1998). The age of a mutation in a general coalescent tree. Commun. Stat. Stoch. Models *14*, 273–295.

Grignani, F., Ferrucci, P. F., Testa, U., Talamo, G., Fagioli, M., Alcalay, M., Mencarelli, A., Peschle, C., and Nicoletti, I. (1993). The acute promyelocytic leukemia-specific PML-RAR alpha fusion protein inhibits differentiation and promotes survival of myeloid precursor cells. Cell *74*, 423–431.

Grignani, F., De Matteis, S., Nervi, C., Tomassoni, L., Gelmetti, V., Cioce, M., Fanelli, M., Ruthardt, M., Ferrara, F. F., Zamir, I., et al. (1998). Fusion proteins of the retinoic acid receptor-alpha recruit histone deacetylase in promyelocytic leukaemia. Nature *391*, 815–818.

Gross, D. S., and Garrard, W. T. (1988). Nuclease hypersensitive sites in chromatin. Annual Review of Biochemistry *57*, 159–197.

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

Groudine, M., and Weintraub, H. (1982). Propagation of globin DNAase I-hypersensitive sites in absence of factors required for induction: a possible mechanism for determination. Cell *30*, 131–139.

Gu, W., Zhou, T., and Wilke, C. O. (2010). A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS computational biology *6*, e1000664.

Gupta, S., Stamatoyannopoulos, J., Bailey, T., and Noble, W. (2007). Quantifying similarity between motifs. Genome Biology *8*, R24+.

Hackett, J. A., Sengupta, R., Zylicz, J. J., Murakami, K., Lee, C., Down, T. A., and Surani, M. A. (2013). Germline DNA Demethylation Dynamics and Imprint Erasure Through 5-Hydroxymethylcytosine. Science *339*, 448–452.

Halupa, A., Bailey, M. L., Huang, K., Iscove, N. N., Levy, D. E., and Barber, D. L. (2005). 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.

Hammar, P., Leroy, P., Mahmutovic, A., Marklund, E. G., Berg, O. G., and Elf, J. (2012). The lac repressor displays facilitated diffusion in living cells. Science (New York, N.Y.) *336*, 1595–1598.

Harvey, W. (1651). Exercitationes de generatione animalium (London: Typis Du-Gardianis; Impensis O. Pulleyn).

Hathaway, N. A., Bell, O., Hodges, C., Miller, E. L., Neel, D. S., and Crabtree, G. R. (2012). Dynamics and memory of heterochromatin in living cells. Cell *149*, 1447–1460.

Hawkins, R. D., Hon, G. C., Lee, L. K., Ngo, Q., Lister, R., Pelizzola, M., Edsall, L. E., Kuan, S., Luu, Y., Klugman, S., et al. (2010). Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. Cell stem cell *6*, 479–491.

Hebbes, T. R., Thorne, A. W., and Crane-Robinson, C. (1988). A direct link between core histone acetylation and transcriptionally active chromatin. The EMBO Journal *7*, 1395–1402.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., Ye, Z., Lee, L. K., Stuart, R. K., Ching, C. W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Hellman, A., and Chess, A. (2007). Gene body-specific methylation on the active X chromosome. Science (New York, N.Y.) *315*, 1141–1143.

Henikoff, S., Henikoff, J. G., Sakai, A., Loeb, G. B., and Ahmad, K. (2009). Genome-wide profiling of salt fractions maps physical properties of chromatin. Genome Research *19*, 460–469.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., Thurman, R. E., Neph, S., Kuehn, M. S., Noble, W. S., et al. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nature Methods *6*, 283–289.

Hirszfeld, L., Halber, W., and Rosenblat, J. (1932). Untersuchungen uber Verwandtschaftsreaktionen zwischen embryonal und Krebsgewebe. II. Menschenembryo und Menschenkrebs. ZeitschriftfUr Immunitatsforschung und Experimentelle Therapie *75*, 209–216.

Hurtado, A., Holmes, K. A., Ross-Innes, C. S., Schmidt, D., and Carroll, J. S. (2011). FOXA1 is a key determinant of estrogen receptor function and endocrine response. Nature genetics *43*, 27–33.

Hyde, J. E., and Walker, I. O. (1975). Covalent cross-linking of histones in chromatin. FEBS letters *50*, 150–154.

Hyder, S. M., Nawaz, Z., Chiappetta, C., Yokoyama, K., and Stancel, G. M. (1995). The protooncogene c-jun contains an unusual estrogen-inducible enhancer within the coding sequence. The Journal of biological chemistry *270*, 8506–8513.

Ikemura, T. (1981a). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. Journal of molecular biology *146*, 1–21.

Ikemura, T. (1981b). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. Journal of molecular biology *151*, 389–409.

Iliopoulos, D., Hirsch, H. A., and Struhl, K. (2009). An epigenetic switch involving NF-kappaB, Lin28, Let-7 MicroRNA, and IL6 links inflammation to cell transformation. Cell *139*, 693–706.

Imada, K., Bloom, E. T., Nakajima, H., Horvath-Arcidiacono, J. A., Udy, G. B., Davey, H. W., and Leonard, W. J. (1998). Stat5b is essential for natural killer cell-mediated proliferation and cytolytic activity. The Journal of experimental medicine *188*, 2067–2074.

Irie, N., and Kuratani, S. (2011). Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. Nature Communications *2*, 248.

Itzkovitz, S., and Alon, U. (2007). The genetic code is nearly optimal for allowing additional information within protein-coding sequences. Genome research *17*, 405–412.

Iyengar, S., Ivanov, A. V, Jin, V. X., Rauscher, F. J., and Farnham, P. J. (2011). Functional analysis of KAP1 genomic recruitment. Molecular and Cellular Biology *31*, 1833–1847.

Izzo, A., Kamieniarz, K., and Schneider, R. (2008). The histone H1 family: specific members, specific functions? Biological Chemistry *389*, 333–343.

Jackson, D. A., Hassan, A. B., Errington, R. J., and Cook, P. R. (1993). Visualization of focal sites of transcription within human nuclei. The EMBO journal *12*, 1059–1065.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. Journal of Molecular Biology *3*, 318–356.

John, S., Sabo, P. J., Thurman, R. E., Sung, M.-H. H., Biddie, S. C., Johnson, T. A., Hager, G. L., and Stamatoyannopoulos, J. A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nature Genetics *43*, 264–268.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497–1502.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Research *20*, 861–873.

Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K. R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-Binding Specificities of Human Transcription Factors. Cell *152*, 327–339.

Joshi, R., Passner, J. M., Rohs, R., Jain, R., Sosinsky, A., Crickmore, M. A., Jacob, V., Aggarwal, A. K., Honig, B., and Mann, R. S. (2007). Functional specificity of a Hox protein mediated by the recognition of minor groove structure. Cell *131*, 530–543.

Kadauke, S., Udugama, M. I., Pawlicki, J. M., Achtman, J. C., Jain, D. P., Cheng, Y., Hardison, R. C., and Blobel, G. A. (2012). Tissue-specific mitotic bookmarking by hematopoietic transcription factor GATA1. Cell *150*, 725–737.

Kadonaga, J. T., Carner, K. R., Masiarz, F. R., and Tjian, R. (1987). Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. Cell *51*, 1079–1090.

Kalinka, A. T., Varga, K. M., Gerrard, D. T., Preibisch, S., Corcoran, D. L., Jarrells, J., Ohler, U., Bergman, C. M., and Tomancak, P. (2010). Gene expression divergence recapitulates the developmental hourglass model. Nature *468*, 811–814.

Kall, L., Canterbury, J. D., Weston, J., Noble, W. S., and MacCoss, M. J. (2007). Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nature Methods *4*, 923–925.

Kaplan, T., and Biggin, M. D. (2012). Quantitative models of the mechanisms that control genome-wide patterns of animal transcription factor binding. Methods in cell biology *110*, 263–283.

Karin, M., Haslinger, A., Holtgreve, H., Richards, R. I., Krauter, P., Westphal, H. M., and Beato, M. (1984). Characterization of DNA sequences through which cadmium and glucocorticoid hormones induce human metallothionein-IIA gene. Nature *308*, 513–519.

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

Kattman, S. J., Witty, A. D., Gagliardi, M., Dubois, N. C., Niapour, M., Hotta, A., Ellis, J., and Keller, G. (2011). Stage-specific optimization of activin/nodal and BMP signaling promotes cardiac differentiation of mouse and human pluripotent stem cell lines. Cell Stem Cell *8*, 228–240.

Keshishian, H., Addona, T., Burgess, M., Kuhn, E., and Carr, S. A. (2007). Quantitative, Multiplexed Assays for Low Abundance Proteins in Plasma by Targeted Mass Spectrometry and Stable Isotope Dilution. Molecular & Cellular Proteomics *6*, 2212–2229.

Khan, A. H., Lin, A., and Smith, D. J. (2012). Discovery and characterization of human exonic transcriptional regulatory elements. PloS one *7*, e46098.

Kim, J., Chu, J., Shen, X., Wang, J., and Orkin, S. H. (2008). An extended transcriptional network for pluripotency of embryonic stem cells. Cell *132*, 1049–1061.

Kim, K., Doi, A., Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M. J., Ji, H., Ehrlich, L. I. R., et al. (2010). Epigenetic memory in induced pluripotent stem cells. Nature *467*, 285–290.

Kim, T. H., Abdullaev, Z. K., Smith, A. D., Ching, K. A., Loukinov, D. I., Green, R. D., Zhang, M. Q., Lobanenkov, V. V, and Ren, B. (2007). Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. Cell *128*, 1231–1245.

Klenova, E. M., Nicolas, R. H., Paterson, H. F., Carne, A. F., Heath, C. M., Goodwin, G. H., Neiman, P. E., and Lobanenkov, V. V (1993). CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms. Molecular and Cellular Biology *13*, 7612–7624.

Kneale, G. G., and Wijnaendts van Resandt, R. W. (1985). Time-resolved fluorescence of bacteriophage Pf1 DNA-binding protein. Determination of oligonucleotide and polynucleotide binding parameters. European journal of biochemistry / FEBS *149*, 85–93.

Krivtsov, A. V, Twomey, D., Feng, Z., Stubbs, M. C., Wang, Y., Faber, J., Levine, J. E., Wang, J., Hahn, W. C., Gilliland, D. G., et al. (2006). Transformation from committed progenitor to leukaemia stem cell initiated by MLL-AF9. Nature *442*, 818–822.

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology *305*, 567–580.

Krzywinski, M., Birol, I., Jones, S. J., and Marra, M. A. (2011). Hive plots--rational approach to visualizing networks. Briefings in bioinformatics.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. Genome Research *19*, 1639–1645.

Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols *4*, 1073–1081.

Kuo, M. T., Mandel, J. L., and Chambon, P. (1979). DNA methylation: correlation with DNase I sensitivity of chicken ovalbumin and conalbumin chromatin. Nucleic Acids Research *7*, 2105–2113.

Lamond, A. I., and Earnshaw, W. C. (1998). Structure and function in the nucleus. Science *280*, 547–553.

Lang, G., Gombert, W. M., and Gould, H. J. (2005). A transcriptional regulatory element in the coding sequence of the human Bcl-2 gene. Immunology *114*, 25–36.

Lange, V., Picotti, P., Domon, B., and Aebersold, R. (2008). Selected reaction monitoring for quantitative proteomics: a tutorial. Molecular Systems Biology *4*.

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

Lazarovici, A., Zhou, T., Shafer, A., Dantas Machado, A. C., Riley, T. R., Sandstrom, R., Sabo, P. J., Lu, Y., Rohs, R., Stamatoyannopoulos, J. A., et al. (2013). Probing DNA shape and methylation state on a genomic scale with DNase I. Proceedings of the National Academy of Sciences.

Leder, C., Kleinschmidt, J. A., Wiethe, C., and Müller, M. (2001). Enhancement of capsid gene expression: preparing the human papillomavirus type 16 major structural gene L1 for DNA vaccination purposes. Journal of virology *75*, 9201–9209.

Lee, M. E., Temizer, D. H., Clifford, J. A., and Quertermous, T. (1991). Cloning of the GATA-binding protein that regulates endothelin-1 gene expression in endothelial cells. The Journal of Biological Chemistry *266*, 16188–16192.

Lever, M. A., Th'ng, J. P. H., Sun, X., and Hendzel, M. J. (2000). Rapid exchange of histone H1.1 on chromatin in living human cells. Nature *408*, 873–876.

Levin, M., Hashimshony, T., Wagner, F., and Yanai, I. (2012). Developmental milestones punctuate gene expression in the Caenorhabditis embryo. Developmental cell *22*, 1101–1108.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) *25*, 2078–2079.

Li, X., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendriks, C. L., et al. (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS biology *6*, e27.

Liao, W. P., Uetzmann, L., Burtscher, I., and Lickert, H. (2009). Generation of a mouse line expressing Sox17-driven Cre recombinase with specific activity in arteries. Genesis *47*, 476–483.

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., et al. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science *326*, 289–293.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Lister, R., Pelizzola, M., Kida, Y. S., Hawkins, R. D., Nery, J. R., Hon, G., Antosiewicz-Bourget, J., O'Malley, R., Castanon, R., Klugman, S., et al. (2011). Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. Nature *471*, 68–73.

Lyons, I., Parsons, L. M., Hartley, L., Li, R., Andrews, J. E., Robb, L., and Harvey, R. P. (1995). Myogenic and morphogenetic defects in the heart tubes of murine embryos lacking the homeo box gene Nkx2-5. Genes & development *9*, 1654–1666.

Maclean, B., Tomazela, D. M., Abbatiello, S. E., Zhang, S., Whiteaker, J. R., Paulovich, A. G., Carr, S. A., and Maccoss, M. J. (2010). Effect of collision energy optimization on the measurement of peptides by selected reaction monitoring (SRM) mass spectrometry. Analytical Chemistry *82*, 10116–10124.

MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M., Finney, G. L., Frewen, B., Kern, R., Tabb, D. L., Liebler, D. C., and MacCoss, M. J. (2010). Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. Bioinformatics *26*, 966–968.

MacNeill, C., French, R., Evans, T., Wessels, A., and Burch, J. B. (2000). Modular regulation of cGATA-5 gene expression in the developing heart and gut. Developmental Biology *217*, 62–76.

Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database : the Journal of Biological Databases and Curation *2011*, bar009.

Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., et al. (2006). Computational prediction of proteotypic peptides for quantitative proteomics. Nature Biotechnology *25*, 125–131.

Malovannaya, A., Lanz, R. B., Jung, S. Y., Bulynko, Y., Le, N. T., Chan, D. W., Ding, C., Shi, Y., Yucer, N., Krenciute, G., et al. (2011). Analysis of the Human Endogenous Coregulator Complexome. Cell *145*, 787–799.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics (Oxford, England) *26*, 2867–2873.

Manning, G. S. (1969). Limiting Laws and Counterion Condensation in Polyelectrolyte Solutions I. Colligative Properties. The Journal of Chemical Physics *51*, 924.

Manning, G. S. (1978). The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. Quarterly Reviews of Biophysics *11*, 179–246.

Marx, V. (2012). Targeted proteomics. Nature Methods *10*, 19–22.

Matys, V., Kel-Margoulis, O. V, Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Research *34*, D108–D110.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., et al. (2012a). Systematic localization of common disease-associated variation in regulatory DNA. Science (New York, N.Y.) *337*, 1190–1195.

Maurano, M. T., Wang, H., Kutyavin, T., and Stamatoyannopoulos, J. A. (2012b). Widespread site-dependent buffering of human regulatory polymorphism. PLoS genetics *8*, e1002599.

May, D., Blow, M. J., Kaplan, T., McCulley, D. J., Jensen, B. C., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., et al. (2011). Large-scale discovery of enhancers from human heart tissue. Nature Genetics *44*, 89–93.

Mazia, D., and Jaeger, L. (1939). Nuclease Action, Protease Action and Histochemical Tests on Salivary Chromosomes of Drosophila. Proceedings of the National Academy of Sciences of the United States of America *25*, 456–461.

McDaniell, R., Lee, B.-K., Song, L., Liu, Z., Boyle, A. P., Erdos, M. R., Scott, L. J., Morken, M. A., Kucera, K. S., Battenhouse, A., et al. (2010). Heritable individual-specific and allele-specific chromatin signatures in humans. Science *328*, 235–239.

McKnight, S., and Tjian, R. (1986). Transcriptional selectivity of viral genes in mammalian cells. Cell *46*, 795–805.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics (Oxford, England) *26*, 2069–2070.

McNally, J. G., Müller, W. G., Walker, D., Wolford, R., and Hager, G. L. (2000). The glucocorticoid receptor: rapid exchange with regulatory sites in living cells. Science *287*, 1262–1265.

Zur Megede, J., Chen, M. C., Doe, B., Schaefer, M., Greer, C. E., Selby, M., Otten, G. R., and Barnett, S. W. (2000). Increased expression and immunogenicity of sequence-modified human immunodeficiency virus type 1 gag gene. Journal of virology *74*, 2628–2635.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004). Superfamilies of evolved and designed networks. Science *303*, 1538–1542.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science *298*, 824–827.

Mirny, L. A. (2010). Nucleosome-mediated cooperativity between transcription factors. Proceedings of the National Academy of Sciences of the United States of America *107*, 22534–22539.

Misteli, T. (2001). Protein dynamics: implications for nuclear architecture and gene expression. Science *291*, 843–847.

Misteli, T. (2013). The Cell Biology of Genomes: Bringing the Double Helix to Life. Cell *152*, 1209–1212.

Misteli, T., Gunjan, A., Hock, R., Bustin, M., and Brown, D. T. (2000). Dynamic binding of histone H1 to chromatin in living cells. Nature *408*, 877–881.

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

modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P. V, Kheradpour, P., Negre, N., Eaton, M. L., Landolin, J. M., Bristow, C. A., Ma, L., et al. (2010). Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science *330*, 1787–1797.

Molkentin, J. D., Lin, Q., Duncan, S. A., and Olson, E. N. (1997). Requirement of the transcription factor GATA4 for heart tube formation and ventral morphogenesis. Genes & Development *11*, 1061–1072.

Monahan, K., Rudnick, N. D., Kehayova, P. D., Pauli, F., Newberry, K. M., Myers, R. M., and Maniatis, T. (2012). Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of protocadherin-α gene expression. Proceedings of the National Academy of Sciences of the United States of America *109*, 9125–9130.

Monty, K. J., Litt, M., Kay, E. R., and Dounce, A. L. (1956). Isolation and properties of liver cell nucleoli. The Journal of Biophysical and Biochemical Cytology *2*, 127–145.

Morgan, T. H. (1901). Regeneration in the Egg, Embryo, and Adult. The American Naturalist *35*, 949–973.

Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., Young, R. A., and Bulyk, M. L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nature Genetics *36*, 1331–1339.

Naidu, P. S., Ludolph, D. C., To, R. Q., Hinterberger, T. J., and Konieczny, S. F. (1995). Myogenin and MEF2 function synergistically to activate the MRF4 promoter during myogenesis. Molecular and Cellular Biology *15*, 2707–2718.

Nakajima, H., Liu, X. W., Wynshaw-Boris, A., Rosenthal, L. A., Imada, K., Finbloom, D. S., Hennighausen, L., and Leonard, W. J. (1997). An indirect effect of Stat5a in IL-2-induced proliferation: a critical role for Stat5a in IL-2-mediated IL-2 receptor alpha chain induction. Immunity *7*, 691–701.

Naya, F. J., Black, B. L., Wu, H., Bassel-Duby, R., Richardson, J. A., Hill, J. A., and Olson, E. N. (2002). Mitochondrial deficiency and cardiac sudden death in mice lacking the MEF2A transcription factor. Nature Medicine *8*, 1303–1309.

Nei, M., and Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences of the United States of America *76*, 5269–5273.

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., et al. (2012a). BEDOPS: high-performance genomic feature operations. Bioinformatics *28*, 1919–1920.

Neph, S., Stergachis, A. B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. A. (2012b). Circuitry and Dynamics of Human Transcription Factor Regulatory Networks. Cell *150*, 1274–1286.

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K., et al. (2012c). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83–90.

Newburger, D. E., and Bulyk, M. L. (2009). UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. Nucleic Acids Research *37*, D77–D82.

Nutiu, R., Friedman, R. C., Luo, S., Khrebtukova, I., Silva, D., Li, R., Zhang, L., Schroth, G. P., and Burge, C. B. (2011). Direct measurement of DNA affinity landscapes on a high-throughput sequencing instrument. Nature biotechnology *29*, 659–664.

Nutt, S. L., Heavey, B., Rolink, A. G., and Busslinger, M. (1999). Commitment to the B-lymphoid lineage depends on the transcription factor Pax5. Nature *401*, 556–562.

Ohlsson, R., Renkawitz, R., and Lobanenkov, V. (2001). CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. Trends in genetics : TIG *17*, 520–527.

Orkin, S. H. (1995). Transcription Factors and Hematopoietic Development. J. Biol. Chem. *270*, 4955–4958.

Orrego, M., Ponte, I., Roque, A., Buschati, N., Mora, X., and Suau, P. (2007). Differential affinity of mammalian histone H1 somatic subtypes for DNA and chromatin. BMC biology *5*, 22+.

Paige, S. L., Thomas, S., Stoick-Cooper, C. L., Wang, H., Maves, L., Sandstrom, R., Pabon, L., Reinecke, H., Pratt, G., Keller, G., et al. (2012). A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. Cell *151*, 221–232.

Parmley, J. L., and Hurst, L. D. (2007). Exonic splicing regulatory elements skew synonymous codon usage near intron-exon boundaries in mammals. Molecular biology and evolution *24*, 1600–1603.

Passeron, T., Valencia, J. C., Bertolotto, C., Hoashi, T., Le Pape, E., Takahashi, K., Ballotti, R., and Hearing, V. J. (2007). SOX9 is a key player in ultraviolet B-induced melanocyte differentiation and pigmentation. Proceedings of the National Academy of Sciences of the United States of America *104*, 13984–13989.

Passeron, T., Valencia, J. C., Namiki, T., Vieira, W. D., Passeron, H., Miyamura, Y., and Hearing, V. J. (2009). Upregulation of SOX9 inhibits the growth of human and mouse melanomas and restores their sensitivity to retinoic acid. The Journal of clinical investigation *119*, 954–963.

Patten, B. (1920). The Early Embryology of the Chick (Philadelphia: P. Blakiston's Son & Co. Inc.).

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

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

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*, 499–502.

Persson, H., and Leder, P. (1984). Nuclear localization and DNA binding properties of a protein expressed by human c-myc oncogene. Science (New York, N.Y.) *225*, 718–721.

Petersen, T. N., Brunak, S., Von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature methods *8*, 785–786.

Pevny, L., Simon, M. C., Robertson, E., Klein, W. H., Tsai, S. F., D'Agati, V., Orkin, S. H., and Costantini, F. (1991). Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor GATA-1. Nature *349*, 257–260.

Pfeifer, G. P., and Riggs, A. D. (1991). Chromatin differences between active and inactive X chromosomes revealed by genomic footprinting of permeabilized cells using DNase I and ligation-mediated PCR. Genes & Development *5*, 1102–1113.

Phair, R. D., Scaffidi, P., Elbi, C., Vecerová, J., Dey, A., Ozato, K., Brown, D. T., Hager, G., Bustin, M., and Misteli, T. (2004). Global nature of dynamic protein-chromatin interactions in vivo: three-dimensional genome scanning and dynamic interaction networks of chromatin proteins. Molecular and Cellular Biology *24*, 6393–6402.

Phillips, J. E., and Corces, V. G. (2009). CTCF: master weaver of the genome. Cell *137*, 1194–1211.

Picotti, P., Lam, H., Campbell, D., Deutsch, E. W., Mirzaei, H., Ranish, J., Domon, B., and Aebersold, R. (2008). A database of mass spectrometric assays for the yeast proteome. Nature Methods *5*, 913–914.

Picotti, P., Rinner, O., Stallmach, R., Dautel, F., Farrah, T., Domon, B., Wenschuh, H., and Aebersold, R. (2010). High-throughput generation of selected reaction-monitoring assays for proteins and proteomes. Nature methods *7*, 43–46.

Pirrotta, V., and Ptashne, M. (1969). Isolation of the 434 phage repressor. Nature *222*, 541–544.

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

Prakash, A., Tomazela, D. M., Frewen, B., Maclean, B., Merrihew, G., Peterman, S., and Maccoss, M. J. (2009). Expediting the development of targeted SRM assays: using data from shotgun proteomics to automate method development. Journal of Proteome Research *8*, 2733–2739.

Privalov, P. L., Dragan, A. I., and Crane-Robinson, C. (2011). Interpreting protein/DNA interactions: distinguishing specific from non-specific and electrostatic from non-electrostatic components. Nucleic Acids Research *39*, 2483–2491.

Privalov, P. L., Jelesarov, I., Read, C. M., Dragan, A. I., and Crane-Robinson, C. (1999). The energetics of HMG box interactions with DNA: thermodynamics of the DNA binding of the HMG box from mouse sox-5. Journal of molecular biology *294*, 997–1013.

Pruitt, K. D., Harrow, J., Harte, R. A., Wallin, C., Diekhans, M., Maglott, D. R., Searle, S., Farrell, C. M., Loveland, J. E., Ruef, B. J., et al. (2009a). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome research *19*, 1316–1323.

Pruitt, K. D., Tatusova, T., Klimke, W., and Maglott, D. R. (2009b). NCBI Reference Sequences: current status, policy and new initiatives. Nucleic acids research *37*, D32–6.

Ptashne, M. (2007). On the use of the word "epigenetic". Current biology : CB *17*, R233–6.

Ptashne, M. (1967). Specific binding of the lambda phage repressor to lambda DNA. Nature *214*, 232–234.

Ptashne, M., Jeffrey, A., Johnson, A. D., Maurer, R., Meyer, B. J., Pabo, C. O., Roberts, T. M., and Sauer, R. T. (1980). How the lambda repressor and cro work. Cell *19*, 1–11.

Pujadas, E., and Feinberg, A. P. (2012). Regulated noise in the epigenetic landscape of development and disease. Cell *148*, 1123–1131.

Quint, M., Drost, H.-G., Gabel, A., Ullrich, K. K., Bönn, M., and Grosse, I. (2012). A transcriptomic hourglass in plant embryogenesis. Nature *490*, 98–101.

Raff, R. A. (1996). The Shape of Life: Genes, Development, and the Evolution of Animal Form (University of Chicago Press).

Ramachandran, B., Yu, G., Li, S., Zhu, B., and Gulick, T. (2008a). Myocyte enhancer factor 2A is transcriptionally autoregulated. The Journal of Biological Chemistry *283*, 10318–10329.

Ramachandran, N., Raphael, J. V, Hainsworth, E., Demirkan, G., Fuentes, M. G., Rolfs, A., Hu, Y., and LaBaer, J. (2008b). Next-generation high-density self-assembling functional protein arrays. Nature Methods *5*, 535–538.

Ravasi, T., Suzuki, H., Cannistraci, C. V. V, Katayama, S., Bajic, V. B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., et al. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. Cell *140*, 744–752.

Rebhan, M., Chalifa-Caspi, V., Prilusky, J., and Lancet, D. (1997). GeneCards: integrating information about genes, proteins and diseases. Trends in Genetics : TIG *13*, 163.

Record, M. T., Ha, J. H., and Fisher, M. A. (1991). Analysis of equilibrium and kinetic measurements to determine thermodynamic origins of stability and specificity and mechanism of formation of site-specific complexes between proteins and helical DNA. Methods in Enzymology *208*, 291–343.

Record, M. T., Lohman, T. M., and Haseth, P. de (1976). Ion effects on ligand-nucleic acid interactions. Journal of Molecular Biology *107*, 145–158.

Reece-Hoyes, J. S., Diallo, A., Lajoie, B., Kent, A., Shrestha, S., Kadreppa, S., Pesyna, C., Dekker, J., Myers, C. L., and Walhout, A. J. M. (2011). Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping. Nature Methods *8*, 1059–1064.

Dos Reis, M., Savva, R., and Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. Nucleic acids research *32*, 5036–5044.

Rhee, H. S., and Pugh, B. F. (2011). Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. Cell *147*, 1408–1419.

Rigaud, G., Roux, J., Pictet, R., and Grange, T. (1991). In vivo footprinting of rat TAT gene: dynamic interplay between the glucocorticoid receptor and a liver-specific factor. Cell *67*, 977–986.

Ritter, D. I., Dong, Z., Guo, S., and Chuang, J. H. (2012). Transcriptional enhancers in protein-coding exons of vertebrate developmental genes. PloS one *7*, e35202.

Rocha, E., Davie, J. R., Van Holde, K. E., and Weintraub, H. (1984). Differential salt fractionation of active and inactive genomic domains in chicken erythrocyte. The Journal of Biological Chemistry *259*, 8558–8563.

Rockman, M. V, and Wray, G. A. (2002). Abundant raw material for cis-regulatory evolution in humans. Molecular biology and evolution *19*, 1991–2004.

Rolfs, A., Montor, W. R., Yoon, S. S., Hu, Y., Bhullar, B., Kelley, F., McCarron, S., Jepson, D. A., Shen, B., Taycher, E., et al. (2008). Production and sequence validation of a complete full length ORF collection for the pathogenic bacterium Vibrio cholerae. Proceedings of the National Academy of Sciences of the United States of America *105*, 4364–4369.

Sabo, P. J., Hawrylycz, M., Wallace, J. C., Humbert, R., Yu, M., Shafer, A., Kawamoto, J., Hall, R., Mack, J., Dorschner, M. O., et al. (2004). Discovery of functional noncoding elements by digital analysis of chromatin structure. Proceedings of the National Academy of Sciences of the United States of America *101*, 16837–16842.

Samstein, R. M., Arvey, A., Josefowicz, S. Z., Peng, X., Reynolds, A., Sandstrom, R., Neph, S., Sabo, P., Kim, J. M., Liao, W., et al. (2012). Foxp3 exploits a pre-existent enhancer landscape for regulatory T cell lineage specification. Cell *151*, 153–166.

Sander K. (1982). Oogenesis and embryonic pattern formation; known and missing links. In Ontogeny and Phylogeny, B. Goodwin and R. Whittle, eds. (Cambridge: Cambridge University Press.).

Sanders, M. M. (1978). Fractionation of nucleosomes by salt elution from micrococcal nuclease-digested nuclei. The Journal of Cell Biology *79*, 97–109.

Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, A., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., and Odom, D. T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell *148*, 335–348.

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

Scott, E. W., Simon, M. C., Anastasi, J., and Singh, H. (1994). Requirement of transcription factor PU.1 in the development of multiple hematopoietic lineages. Science *265*, 1573–1577.

Sekimata, M., Pérez-Melgosa, M., Miller, S. A., Weinmann, A. S., Sabo, P. J., Sandstrom, R., Dorschner, M. O., Stamatoyannopoulos, J. A., and Wilson, C. B. (2009). CCCTC-binding factor and the transcription factor T-bet orchestrate T helper 1 cell-specific structure and function at the interferon-gamma locus. Immunity *31*, 551–564.

Serov, V. N., Spirov, A. V, and Samsonova, M. G. (1998). Graphical interface to the genetic network database GeNet. Bioinformatics *14*, 546–547.

Severin, F. F., Shanina, N. A., Shevchenko, A., Solovyanova, O. B., Koretsky, V. V, and Nadezhdina, E. S. (1997). A major 170 kDa protein associated with bovine adrenal medulla microtubules: a member of the centrosomin family? FEBS letters *420*, 125–128.

Shabalina, S. A., Spiridonov, N. A., and Kashina, A. (2013). Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic acids research *41*, 2073–2094.

Shahhoseini, M., Favaedi, R., Baharvand, H., Sharma, V., and Stunnenberg, H. G. (2010). Evidence for a dynamic role of the linker histone variant H1x during retinoic acid-induced differentiation of NT2 cells. FEBS letters *584*, 4661–4664.

Sheinerman, F. B., Norel, R., and Honig, B. (2000). Electrostatic aspects of protein-protein interactions. Current Opinion in Structural Biology *10*, 153–159.

Shi, Y., Seto, E., Chang, L. S., and Shenk, T. (1991). Transcriptional repression by YY1, a human GLI-Krüppel-related protein, and relief of repression by adenovirus E1A protein. Cell *67*, 377–388.

Shimazaki, T., Arsenijevic, Y., Ryan, A. K., Rosenfeld, M. G., and Weiss, S. (1999). A role for the POU-III transcription factor Brn-4 in the regulation of striatal neuron precursor differentiation. The EMBO Journal *18*, 444–456.

Siepel, A., Pollard, K., and Haussler, D. (2006). New Methods for Detecting Lineage-Specific Selection. In Lect. Notes Comput. Sci., A. Apostolico, C. Guerra, S. Istrail, P. Pevzner, and M. Waterman, eds. (Berlin/Heidelberg: Springer-Verlag), pp. 190–205.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H. J., et al. (2011). Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins. Cell *147*, 1270–1282.

Slutsky, M., and Mirny, L. A. (2004). Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. Biophysical journal *87*, 4021–4035.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics *27*, 431–432.

Socolovsky, M., Nam, H., Fleming, M. D., Haase, V. H., Brugnara, C., and Lodish, H. F. (2001). Ineffective erythropoiesis in Stat5a(-/-)5b(-/-) mice due to decreased survival of early erythroblasts. Blood *98*, 3261–3273.

Spemann, H. (1918). Über die Determination der ersten Organanlagen des Amphibienembryo I–VI. Archiv für Entwicklungsmechanik der Organismen *43*, 448–555.

Stalder, J., Larsen, A., Engel, J. D., Dolan, M., Groudine, M., and Weintraub, H. (1980). Tissue-specific DNA cleavages in the globin chromatin domain introduced by DNAase I. Cell *20*, 451–460.

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

Stuart, P. E., Nair, R. P., Ellinghaus, E., Ding, J., Tejasvi, T., Gudjonsson, J. E., Li, Y., Weidinger, S., Eberlein, B., Gieger, C., et al. (2010). Genome-wide association analysis identifies three psoriasis susceptibility loci. Nature genetics *42*, 1000–1004.

Sucov, H. M., Lou, J., Gruber, P. J., Kubalak, S. W., Dyson, E., Gumeringer, C. L., Lee, R. Y., Moles, S. A., Chien, K. R., Giguere, V., et al. (1996). The molecular genetics of retinoic acid receptors: cardiovascular and limb development. Biochemical Society Symposium *62*, 143–156.

Swiers, G., Patient, R., and Loose, M. (2006). Genetic regulatory networks programming hematopoietic stem cells and erythroid lineage specification. Developmental Biology *294*, 525–540.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell *131*, 861–872.

Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell *126*, 663–676.

Takata, H., Matsunaga, S., Morimoto, A., Ono-Maniwa, R., Uchiyama, S., and Fukui, K. (2007). H1.X with different properties from other linker histones is required for mitotic progression. FEBS letters *581*, 3783–3788.

Tanaka, M., Schinke, M., Liao, H. S., Yamasaki, N., and Izumo, S. (2001). Nkx2.5 and Nkx2.6, homologs of Drosophila tinman, are required for development of the pharynx. Molecular and cellular biology *21*, 4391–4398.

Tani, M., Shindo-Okada, N., Hashimoto, Y., Shiroishi, T., Takenoshita, S., Nagamachi, Y., and Yokota, J. (1997). Isolation of a novel Sry-related gene that is expressed in high-metastatic K-1735 murine melanoma cells. Genomics *39*, 30–37.

Tatarinov, I. S. (1964). [DETECTION OF EMBRYO-SPECIFIC ALPHA-GLOBULIN IN THE BLOOD SERUM OF A PATIENT WITH PRIMARY LIVER CANCER]. Voprosy meditsinskoĭ khimii *10*, 90–91.

Tate, P. H., and Bird, A. P. (1993). Effects of DNA methylation on DNA-binding proteins and gene expression. Current Opinion in Genetics & Development *3*, 226–231.

Tennessen, J. A., Bigham, A. W., O'Connor, T. D., Fu, W., Kenny, E. E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.

Th'ng, J. P., Sung, R., Ye, M., and Hendzel, M. J. (2005). H1 family histones in the nucleus. Control of binding and localization by the C-terminal domain. The Journal of biological chemistry *280*, 27809–27814.

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

The ENCODE Project Consortium (2011). A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biology *9*, e1001046+.

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

The International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Thoma, F., Koller, T., and Klug, A. (1979). Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. The Journal of Cell Biology *83*, 403–427.

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

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

Tsai, S. F., Martin, D. I., Zon, L. I., D'Andrea, A. D., Wong, G. G., and Orkin, S. H. (1989). Cloning of cDNA for the major DNA-binding protein of the erythroid lineage through expression in mammalian cells. Nature *339*, 446–451.

Turing, A. M. (1937). On Computable Numbers, with an Application to the Entscheidungs problem. Proceedings of the London Mathematical Society *42*, 230–265.

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., and Luscombe, N. M. (2009). A census of human transcription factors: function, expression and evolution. Nature Reviews Genetics *10*, 252–263.

Verbeek, S., Izon, D., Hofhuis, F., Robanus-Maandag, E., Te Riele, H., Van de Wetering, M., Oosterwegel, M., Wilson, A., MacDonald, H. R., and Clevers, H. (1995). An HMG-box-containing T-cell factor required for thymocyte differentiation. Nature *374*, 70–74.

Vernot, B., Stergachis, A. B., Maurano, M. T., Vierstra, J., Neph, S., Thurman, R. E., Stamatoyannopoulos, J. A., and Akey, J. M. (2012). Personal and population genomics of human regulatory variation. Genome Research *22*, 1689–1697.

Voss, T. C., Schiltz, R. L., Sung, M.-H., Yen, P. M., Stamatoyannopoulos, J. A., Biddie, S. C., Johnson, T. A., Miranda, T. B., John, S., and Hager, G. L. (2011). Dynamic exchange at regulatory elements during chromatin remodeling underlies assisted loading mechanism. Cell *146*, 544–554.

Waddington, C. H. (1939). An Introduction to Modern Genetics (New York: The Macmillan Company).

Waddington, C. H. (1935). Cancer and the Theory of Organiser. Nature *135*, 606–608.

Waddington, C. H. (1940). Organisers and Genes (Cambridge University Press).

Waddington, C. H. (1957). The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology (George Allen & Unwin).

Walhout, A. J. M. (2006). Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. Genome Research *16*, 1445–1454.

Wamstad, J. A., Alexander, J. M., Truty, R. M., Shrikumar, A., Li, F., Eilertson, K. E., Ding, H., Wylie, J. N., Pico, A. R., Capra, J. A., et al. (2012). Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. Cell *151*, 206–220.

Wang, H., Maurano, M. T., Qu, H., Varley, K. E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., et al. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. Genome research *22*, 1680–1688.

Warburg, O. (1956). On the origin of cancer cells. Science *123*, 309–314.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association *58*, 236.

Washburn, M. P., Wolters, D., and Yates, J. R. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nature Biotechnology *19*, 242–247.

Weintraub, H. (1985). Assembly and propagation of repressed and depressed chromosomal states. Cell *42*, 705–711.

Weintraub, H. (1984). Histone-H1-dependent chromatin superstructures and the suppression of gene activity. Cell *38*, 17–27.

Weintraub, H., and Groudine, M. (1976). Chromosomal subunits in active genes have an altered conformation. Science *193*, 848–856.

Weisbrod, S., and Weintraub, H. (1979). Isolation of a subclass of nuclear proteins responsible for conferring a DNase I-sensitive structure on globin chromatin. Proceedings of the National Academy of Sciences of the United States of America *76*, 630–634.

Welch, J. J., Watts, J. A., Vakoc, C. R., Yao, Y., Wang, H., Hardison, R. C., Blobel, G. A., Chodosh, L. A., and Weiss, M. J. (2004). Global regulation of erythroid gene expression by transcription factor GATA-1. Blood *104*, 3136–3147.

White, J. G., Southgate, E., Thomson, J. N., and Brenner, S. (1986). The Structure of the Nervous System of the Nematode Caenorhabditis elegans. Philosophical Transactions of the Royal Society B: Biological Sciences *314*, 1–340.

Wingender, E., Dietze, P., Karas, H., and Knüppel, R. (1996). TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites. Nucleic Acids Research *24*, 238–241.

Wu, C. (1980). The 5’ ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature *286*, 854–860.

Wu, C., Bingham, P. M., Livak, K. J., Holmgren, R., and Elgin, S. C. (1979). The chromatin structure of specific genes: I. Evidence for higher order domains of defined DNA sequence. Cell *16*, 797–806.

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338–345.

Xing, Y., Li, C., Li, A., Sridurongrit, S., Tiozzo, C., Bellusci, S., Borok, Z., Kaartinen, V., and Minoo, P. (2010). Signaling via Alk5 controls the ontogeny of lung Clara cells. Development *137*, 825–833.

Yamanaka, S. (2009). Elite and stochastic models for induced pluripotent stem cell generation. Nature *460*, 49–52.

Yang, A., Zhu, Z., Kapranov, P., McKeon, F., Church, G. M., Gingeras, T. R., and Struhl, K. (2006). Relationships between p63 binding, DNA sequence, transcription activity, and biological function in human cells. Molecular cell *24*, 593–602.

Yang, L., Soonpaa, M. H., Adler, E. D., Roepke, T. K., Kattman, S. J., Kennedy, M., Henckaerts, E., Bonham, K., Abbott, G. W., Linden, R. M., et al. (2008). Human cardiovascular progenitor cells develop from a KDR+ embryonic-stem-cell-derived population. Nature *453*, 524–528.

Yang, Z., and Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Molecular biology and evolution *25*, 568–579.

Yu, W., Ginjala, V., Pant, V., Chernukhin, I., Whitehead, J., Docquier, F., Farrar, D., Tavoosidana, G., Mukhopadhyay, R., Kanduri, C., et al. (2004). Poly(ADP-ribosyl)ation regulates CTCF-dependent chromatin insulation. Nature Genetics *36*, 1105–1110.

Yuh, C. H., Ransick, A., Martinez, P., Britten, R. J., and Davidson, E. H. (1994). Complexity and organization of DNA-protein interactions in the 5'-regulatory region of an endoderm-specific marker gene in the sea urchin embryo. Mechanisms of Development *47*, 165–186.

Yun, K., and Wold, B. (1996). Skeletal muscle determination and differentiation: story of a core regulatory network and its context. Current Opinion in Cell Biology *8*, 877–889.

Zandi, S., Mansson, R., Tsapogas, P., Zetterblad, J., Bryder, D., and Sigvardsson, M. (2008). EBF1 is essential for B-lineage priming and establishment of a transcription factor network in common lymphoid progenitors. Journal of immunology *181*, 3364–3372.

Zheng, Y., Josefowicz, S., Chaudhry, A., Peng, X. P., Forbush, K., and Rudensky, A. Y. (2010). Role of conserved non-coding DNA elements in the Foxp3 gene in regulatory T-cell fate. Nature *463*, 808–812.

Zhu, J., Adli, M., Zou, J. Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., De Jager, P. L., et al. (2013). Genome-wide Chromatin State Transitions Associated with Developmental and Environmental Cues. Cell *152*, 642–654.

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nature genetics *39*, 61–69.