High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells

Alan P Boyle, Lingyun Song, Bum-Kyu Lee, et al.

Genome Res. published online November 24, 2010
Access the most recent version at doi:10.1101/gr.112656.110

High-resolution genome-wide *in vivo* footprinting of diverse transcription factors in human cells

Alan P. Boyle[1], Lingyun Song[1,2], Bum-Kyu Lee[3], Darin London[1], Damian Keefe[4], Ewan Birney[4], Vishwanath R. Iyer[3], Gregory E. Crawford[1,2,*], and Terrence S. Furey[1,*]

[1]Institute for Genome Sciences & Policy, Duke University, Durham, North Carolina 27708, USA
[2]Department of Pediatrics, Division of Medical Genetics, Duke University, Durham, North Carolina 27708, USA
[3]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, Section of Molecular Genetics and Microbiology, University of Texas at Austin, Austin, Texas 78712, USA
[4]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1RQ, UK

[*]**C**orrespondence: terry.furey@duke.edu (T.S.F), greg.crawford@duke.edu (G.E.C)

Running title: Genome-wide *in vivo* footprinting

Keywords: genomics, transcription, gene regulation, footprinting

Manuscript corresponding Author:
Terrence Furey, PhD
Assistant Professor
Duke Institute for Genome Sciences & Policy
Duke University
CIEMAS
101 Science Drive
Box 3382
Durham, NC 27708
T 919-668-4477
F 919-668-0795

Current Address:
Assistant Professor
Department of Genetics and Department of Biology
Carolina Center for Genome Sciences
University of North Carolina at Chapel Hill
5022 Genetic Medicine Building
CB#7264
Chapel Hill, NC 27599-7264
T 919-966-7033

Abstract

Regulation of gene transcription in diverse cell types is largely determined by varied sets of cis-elements where transcription factors bind. Here we demonstrate that data from a single high-throughput DNaseI hypersensitivity assay can delineate hundreds of thousands of base-pair resolution *in vivo* footprints in human cells that precisely mark individual transcription factor-DNA interactions. These annotations provide a unique resource for the investigation of cis-regulatory elements. We find that footprints for specific transcription factors correlate with ChIP-seq enrichment and can accurately identify functional vs. non-functional transcription factor motifs. We also find that footprints reveal a unique evolutionary conservation pattern that differentiates functional footprinted bases from surrounding DNA. Finally, detailed analysis of CTCF footprints suggests multiple modes of binding and a novel DNA binding motif upstream of the primary binding site.

Supplementary Materials Included

Introduction

Demarcation of transcription factor binding is key to the understanding of gene expression and whole regulatory networks within a cell. This is a particularly daunting task since it is estimated that there are approximately 1,500 transcription factors in the human genome (Vaquerizas et al. 2009). A number of methods have been developed to identify the location of transcription factor binding, such as ChIP-seq, position weight matrices (PWMs), electrophoretic mobility shift assays (EMSA), and footprinting using DNaseI or dimethylsulfate.

While these methods are extremely powerful and complementary, each method has limitations. For example, ChIP-seq requires a large number of cells, a high-quality antibody (or epitope tagged version), and is unable to resolve DNA-protein interactions at base-pair resolution. PWMs model DNA binding site sequence preferences, commonly referred to as "motifs", for different transcription factors. Since most transcription factor motifs are 4-8 bases in length, these annotations often include large numbers of predicted sites with low specificity. In addition, PWMs are only available for a fraction of factors. Electrophoretic shift assays (EMSA) test whether any fragment of DNA can bind to nuclear extracts or purified single proteins. However, this *in vitro* assay may not be accurate if multiple factors or DNA segments are required for binding. Traditional footprinting assays accurately identify the precise binding sites of any factor. However, this low-throughput method is highly technical and can only analyze a single small region (< 1kb) at a time. Together, this indicates that additional methods are clearly needed to better understand global gene regulation.

Mapping DNaseI hypersensitive (HS) sites across the genome using a deep sequencing approach (DNase-seq) identifies a broad variety of active cis-regulatory elements (Gross and Garrard 1988; Wu 1980; Boyle, Davis, et al. 2008). DNase-seq identifies sites of DNase digestion at single base resolution, even though this data is typically smoothed to identify larger DNaseI HS sites (~200 bp). Previously it was shown that similarly derived DNaseI digestion data could identify individual binding sites in *Saccharomyces cerevisiae* based on the protection of short stretches of nucleotides with the larger HS sites (Hesselberth et al. 2009). Likewise, in human we observed the raw distribution of sequence tag locations within each HS site is not uniform reflecting *in vivo* protection of DNA by individually bound proteins, similar to traditional DNaseI footprinting assays.

Here, we describe DNaseI footprints identified from DNase-seq data generated from 7 similar (lymphoblastoid cell lines (McDaniell et al. 2010)) and 5 diverse (K562, HelaS3, HUVEC, NHEK, and embryonic stem cell) human cell types (available at http://www.genome.duke.edu/labs/furey/datasets/). We show that DNaseI footprints are reproducible, robust, and accurate at identifying and annotating hundreds of thousands of putative protein binding sites genome-wide. Footprinting data alone cannot annotate every site for every known and unknown factor, but it is an important complement to ChIP-seq and conservation data that provides valuable protein/DNA interaction information. Together, these enable an even more comprehensive accounting and characterization of active cis-elements with base-pair resolution.

Results

Transcription factor binding sites are depleted for DNaseI cleavage sites

DNase-seq data was generated and uniformly processed from multiple independent replicates as part of the human ENCODE project (ENCODE Project Consortium 2004) (Supplementary Table 1A). To assess the ability of DNase-seq to identify footprints, we first investigated general digestion patterns around published motifs for transcription factors with known PWMs (Kim et al. 2007; Bryne et al.

2008; Matys et al. 2006; Newburger and Bulyk 2009). We determined all positions in the genome matching these motifs, referred to from here on as motif predicted binding sites. We then calculated the number of DNaseI cuts at each base pair within and surrounding all motif predicted binding sites for a particular factor across the genome. At this cumulative level, we clearly detected a footprint characterized by a lack of DNaseI digestion within these sites for many individual transcription factors (Fig. 1A for CTCF, Supplementary Fig. 1A for multiple other factors). These footprints were detected at the cumulative level even though presumably many of the motif predicted binding sites were not bound by transcription factors. This aggregate detection of footprints was only seen for factors with high information content motifs whose predictions are significantly enriched for functional sites in the genome (Supplementary Fig. 1A). Approximately 30% of motif predicted CTCF binding sites corresponded with a footprint signal. Factors with low information content motifs (shorter and/or less complex) generate many false positives that mask the cumulative footprinting signal. For example, less than an estimated 0.1% of over 40 million short (4 bases), information-poor GATA1 motif predicted binding sites showed evidence of footprinting (data not shown).

To determine whether functional binding sites could be determined based solely on DNase-seq data, we used k-means clustering to divide up motif predicted CTCF binding sites that overlap or do not overlap a footprint (see Methods). We compared these sets to CTCF ChIP-seq data collected from the same cell growth. Motif predicted CTCF binding sites with a footprint signal were highly enriched for ChIP-seq signal (Fig. 1A), whereas those without footprint evidence displayed almost no ChIP-seq signal (Fig. 1B).

ChIP-seq signal was significantly stronger in regions that overlapped footprints compared to ChIP-seq peaks without footprints ($p<2.2e-16$, Supplementary Fig. 1B). Similarly, motif-predicted sites that overlapped footprints had higher PWM scores than those without footprints ($p<2.2e-16$, Supplementary Fig. 1C). It is important to note that the strength of the PWM score only partially predicts *in vivo* binding by ChIP-seq (Supplementary Fig 1D) compared to using footprint data as a guide (Fig. 1A-B). General correspondence of footprints, ChIP-seq signal, and PWM strength indicate that all three data describe biologically relevant and important characteristics of transcription factor binding, which is likely related to increased protein binding affinity and/or increased occupancy throughout the cell population.

Identification of individual footprints

While cumulative plots provide summary validation for DNaseI footprinting, we also developed a 5-state hidden Markov model (HMM; see Methods) in order to identify individual footprints throughout the genome (Supplementary Fig. 1E). This HMM identified small regions within DNaseI HS sites where there was reduced DNaseI digestion (footprints) as compared to adjacent bases (see Methods). Footprints were identified in individual cell types and well in pooled lymphoblastoid data. In general, we found that signals in all lymphoblastoid cell lines were extremely similar (Supplementary Figure 2). We were conservative in our delineation of footprints to reduce the number of false positives. The number of footprints per cell type ranged from 100,000-325,000. Variation in the number of footprints identified appears to be primarily due to differences in the number and average size of DNaseI HS sites annotated in each cell line (Supplementary Table 1).

We identified the putative factors bound to each footprint using STAMP (Mahony and Benos 2007) in conjunction with motifs that are publicly available in the JASPAR (Bryne et al. 2008), TRANSFAC (Matys et al. 2006) and UniPROBE (Newburger and Bulyk 2009) PWM databases. In total, there were 476 PWMs representing 398 distinct factors, some of which represent the binding of multi-protein

complexes. We required motif matches with p-values $<10^{-6}$ and allowed footprints to be labeled with multiple factors if each separate motif matched with scores below this threshold. Using this strict criterion, only 21-26% of predicted footprints were annotated by a currently known motif (Supplementary Table 1). Our strict motif criteria and more generally the incomplete knowledge of sequence preferences for all DNA binding proteins likely contributed to this low rate of annotation. Using more lenient motif match criteria would increase the number of annotated sites, but would also increase the rate of incorrect annotations. We note that these annotations represent candidate binding factors. Those factors with large, information-rich motifs, like CTCF and REST, are more likely to be correctly labeled. In many cases, multiple distinct factors have very similar binding site motifs (Mahony et al. 2007). Thus, annotations of factors with smaller, information-poor motifs or whose motifs are similar to those of other factors may be less accurate.

To demonstrate the accuracy of our model, we show that predicted footprints from DNase-seq data pooled from seven distinct lymphoblastoid cell lines perfectly match previously identified NRF1, SP1, AP-2, and Myc footprints near the *FMR1* (fragile X mental retardation 1) promoter (Figure 1C) (Drouin et al. 1997). We also found that CTCF footprints corresponded extremely well to individual CTCF binding sites detected both by ChIP-seq as well as by motif prediction (Fig. 1D). To more globally determine the accuracy of our model, we used ChIP-seq data for CTCF, REST, GABP, and SRF, and determined the positive predictive value (PPV) of motifs that were 1) present across the entire genome, 2) found within a DNaseI HS site, or 3) found within a footprint (Figure 2 and Supplementary Table 2). The motifs with a corresponding ChIP-seq peak were considered functional (true positives), while the motifs with no ChIP evidence were considered not functional (true negatives). Predicted CTCF and REST footprints had a PPV of >98%, while predicted GABP and SRF footprints had a PPV of >50%. The reduced PPV for GABP and SRF footprints may be due to DNaseI and ChIP data originating from non-matched cell types (Valouev et al. 2008), or may be due to these factors having binding motifs with lower information content. However, we note that for GABP and SRF, the footprint PPV significantly outperforms the PPV using a purely sequence-based motif approach by 20-50 fold (Figure 2). Using stricter PWM criteria to identify positives and negatives does not significantly affect the PPV for CTCF and REST footprints, but it does increase the PPV for GABP and SRF footprints to approximately 80% (Supplementary Figure 3). Footprints also are much more accurate at identifying ChIP-seq peaks compared with simply using motifs that are present in a DNaseI HS site (Figure 2 and Supplementary Figure 3). Sensitivity and specificity measurements show similar results (Supplementary Table 2). These observations indicate that DNase-seq footprinting accurately identifies active transcription factor binding sites.

The reproducibility of our DNaseI footprinting method is evident by footprint annotations across two lymphoblastoid lines being much more correlated than between cell lines of different lineages (Supplementary Fig. 4A). In fact, these two lines showed higher correlation based on footprint annotations than based on gene expression levels (Supplementary Fig. 4B). Additionally, DNaseI footprint annotations and CTCF ChIP-seq data were also generated for K562, HelaS3, NHEK, HUVEC, and embryonic stem cells. Even though there were less DNaseI sequences generated than for the combined lymphoblastoid data (Supplementary Table 1), we found that the accuracy of predicted CTCF footprints remained high in all cell lines as evidenced by strong positive predictive value rates (94%-99%) for CTCF ChIP-seq signals (Supplementary Table 3).

As mentioned previously, the number of footprints identified in each cell type is dependent on the DNaseI HS site annotation. Higher numbers of DNase-sequences improves the accuracy of the footprint annotation within these HS sites. For example, the PPV for CTCF is higher and more

footprints are annotated with factors by STAMP in the combined lymphoblastoid data compared to individual lymphoblastoid cell lines (Supplementary Tables 3 and 1, respectively).

Preferences in footprint locations relative to genes and each other

Many transcription factors have been more closely associated with binding proximal promoter regions while others have shown a preference for distal regions. Using the distribution of footprints from 89 factors that are annotated in at least 100 distinct locations in the combined lymphoblastoid data, we determined the expected number and enrichment/depletion of footprinted sites for a single factor in promoter or non-promoter categories (Supplementary Table 4). Not surprisingly, several well-known factors including SP1, AP-2, USF, and GABPA were enriched in promoter regions. These have been previously associated with basal promoter activity in a large variety of genes with diverse functions. In contrast, other factors are depleted in promoters and enriched in non-promoter regions including Fos, STAT1, IRF1, IRF2, HSF, and CTCF. Many of these latter factors are involved in more specialized cellular functions, in this case in lymphoblastoid cells, and are often activated in response to an external stimuli.

Since many factors bind within complexes or work co-operatively, we asked if any two factors were preferentially bound within the same DNaseI HS site more often than chance. When analyzing each pair of factors, individual footprints labeled with both factors were discarded to correct for factors with very similar motifs. We found 35 combinations of factors that significantly co-localize in lymphoblastoid cells (Supplementary Table 5). Co-localization graphs depicting these relationships show that the factors divide neatly into two clusters. The first, larger cluster consists of twelve factors including SP-1, AP-2, Mycn, and GABP (Supplementary Fig. 5A). Approximately 87% of co-occurrences for these factors map within promoter regions in lymphoblastoid cells. In addition, on average in each of the other five diverse cell types, over 97% of the DNaseI HS sites containing these co-occurrences were also annotated. In contrast, the second, smaller cluster consists of 6 factors known to regulate genes in response to external stimuli (Supplementary Fig. 5B). Interestingly, 87% of the second cluster instances are found in non-promoter regulatory regions. Only 33% of the DNaseI HS sites containing these co-localized footprints were similarly identified on average in non-lymphoblastoid cell types. Therefore, very distinct combinations of factors appear to co-localize in ubiquitously open promoters in contrast to cell-type specific non-promoter regions.

We repeated these analyses on each of the five other cells lines. Interestingly, we found a core set of nine combinations involving seven factors (SP1, AP-2, Mycn, Pax4, USF, Arnt, RREB1) that significantly co-localized in essentially all cell types (Supplementary Table 6, Supplementary Fig. 5A – bold lines). Another 22 pairs of factors were commonly found in 2 or more cell lines, and several co-occurrences were limited to a single cell-type (Supplementary Table 6). As mentioned previously, greater sequencing depth appears to influence the ability to and accuracy of footprint annotations. Besides the combined lymphoblastoid data, only the H1 embryonic stem cells were sequenced to a depth of at least 100 million sequences. Thus, the combined lymphoblastoid and H1 lines had far greater numbers of significantly co-occurring factors. In the H1 data, we also see co-localized factors for which footprints for both are primarily found (>89%) in non-promoter distal regions (Supplementary Table 6, Supplementary Fig. 5C). Interestingly, these primarily consist of NFκB or its subunits (Rel, c-Rel) in combination with another factor. Presumably, deeper sequencing in other cell lines would likewise reveal more cell-type specific combinations in this and other cell types.

Cell-type specific footprinting patterns

By mapping footprints across various cell lines, we used cumulative plots similar to Figure 1 to detect factors similarly utilized in all cell types as well as those that differed in a cell-type specific manner. Factors used in all cell types displayed consistent footprinting signals in all cell types, while cell-type specific footprints showed a diminished or distinct lack of a footprint signal in one or more cell types. For example, REST showed a footprint pattern in all cell types (Fig. 3A), while TLX1-NFIC and IRF2 displayed cell type-specific patterns (Fig. 3B-C). These cell-type specific footprinting patterns are highly correlated with gene expression differences (data not shown), and are supported by previous studies. For example, REST is known to repress neuronal genes in all non-neuronal cell types, and IRF2 is an interferon regulatory transcription factor known to be involved in the development of immune-related cells, including B cells (Tamura et al. 2008). The homeoprotein TLX1 is known to interact with the CCAAT binding transcription factor NFIC (N Zhang et al. 1999). We clearly see footprints for the TLX1/NFIC complex in K562, HeLaS3, HUVEC and NHEK cells, but not lymphoblastoid or embryonic stem cells. We do not detect a difference in the mean expression of *TLX1* across these cell lines (μ_1=5.47 $\log_2$ expression for cells with footprints, μ_2=5.41 for cells without footprints), but we do see a nearly three fold increase in the expression of *NFIC* in those cell types with footprints (μ_1=9.23, μ_2=7.73). This suggests that the differential binding of the TLX1/NFIC complex in these cell types identified by the footprinting data is likely mediated by *NFIC* expression.

Relationship to gene expression

We compared DNase-seq footprints to gene expression patterns using RNA isolated from the same growths of the six diverse cell lines. Transcription factors with low expression signals in a cell line had fewer predicted footprints while highly expressed factors were over-represented in the number of predicted footprints (Supplementary Fig. 4C). This trend was especially striking in the top quartile of highly expressed factors where significantly higher numbers of footprints were predicted than for factors in the second quantile ($p < 2.381e-15$) and even more so when compared with factors in the bottom quantile ($p < 8.381e-16$) (Supplementary Fig. 4D). Interestingly, *CTCF* was one of the most highly expressed genes and displayed the most footprints in all cell types (Supplementary Fig. 4C). However, we find that high levels of transcription factor expression do not necessarily imply a corresponding high number of annotated footprints.

Evolutionary conservation of footprints

Previous studies have shown that many functional binding sites are more conserved than background at the sequence level (Hesselberth et al. 2009; Yueyi Liu et al. 2004). Many computational predictors of novel regulatory elements often attempt to incorporate this information (Boffelli et al. 2003). To date, this has proved difficult, likely due to the short, degenerate nature of binding motifs for many factors and the inability to precisely locate their positions.

To assess patterns of evolutionary conservation within our predicted footprints, we analyzed the conservation profiles using PhastCons (Siepel et al. 2005) for sites corresponding to each specific factor. We found a significant increase in conservation directly within the footprint, which is above the average level of conservation across an entire DNaseI hypersensitive site (Fig. 4A). For most factors, we detected a marked drop in conservation approximately 10bp immediately flanking the footprint. Beyond this drop, conservation increased again before gradually decreasing to background levels creating a “shoulder” in this signal. This unusual conservation pattern was not observed in footprints identified by DNase-seq in *Saccharomyces cerevisiae* (Hesselberth et al. 2009), and was detected for most individual factors (Fig. 4B, Supplementary Fig. 6C, and data not shown). This “shoulder” is notably absent from CTCF, which only displays a single peak at the footprint (Fig. 4C). This is not an artifact of the large number of annotated CTCF sites as the shoulder is still absent when

considering only CTCF footprints at promoters or a small fraction of random CTCF sites (Supplementary Fig. 6D). The sharp conservation pattern in footprints can also be seen for individual footprinted regions (Fig. 1C-D). The drop in conservation between the footprint and "shoulder" can also be seen by analyzing conservation patterns in the DNA that corresponded to each state in our footprint HMM model (Supplementary Fig. 1E). DNA labeled by the "footprint" state (FP) is highly conserved, whereas DNA adjacent to footprints but still within regulatory regions (UP and DOWN) shows much lower conservation (Supplementary Fig. 6A). In general, higher evolutionary conservation in DNaseI HS regions (Supplementary Fig. 6B) is likely due to the presence of multiple functional sites in the larger cis-regulatory modules or promoter regions. The drop in conservation immediately adjacent to most footprints suggests that steric hindrance prevents the binding of nearby factors and may have resulted in relaxed selection pressure for those bases.

CTCF footprints display unique binding characteristics

CTCF is a unique transcription factor that has been shown to display diverse regulatory roles including insulator, enhancer, and repressor activity (Phillips and Corces 2009). CTCF contains eleven zinc fingers and several studies have discovered a highly conserved GC-rich 20 base pair motif associated with many, but not all, detected CTCF binding sites (Kim et al. 2007). Mutational studies have demonstrated that in at least some instances, additional zinc fingers contact sequence upstream of this GC-rich core that may be necessary for binding (Filippova et al. 1996) or appear to stabilize binding (Quitschke et al. 2000) of CTCF. A significant fraction of CTCF binding sites do not contain this primary motif or any other motif, indicating that CTCF binds indirectly to some regions of the genome.

It has been demonstrated that footprint digestion patterns can reflect the actual structural interaction of a factor with the DNA (Hesselberth et al. 2009). Orienting and aligning CTCF footprints based on the direction of the 20 base pair motif, we found that CTCF has a very unique footprinting profile (Fig. 5A). CTCF footprints were extremely depleted for DNaseI digestion in the 20 base pair region that corresponds to the previously characterized GC-rich binding motif. Unlike other factors where the frequency of DNaseI cuts rises sharply at the boundaries of the motif, the signal upstream region of the CTCF motif increases more gradually, indicating that CTCF protection extends beyond the 20bp motif. A similar phenomenon was detected to a lesser extent in the 3' downstream direction. The total length of these combined protected regions agrees with a single footprinted CTCF site that displayed 50-60 bases of general protection (Phillips and Corces 2009).

Interestingly, we detected a 10 base pair spike in DNaseI hypersensitivity immediately upstream of the primary motif that is only present on the positive strand (Fig. 5B). This spike was previous reported by traditional DNaseI footprinting on a single CTCF binding site in the promoter of the Amyloid precursor protein gene, but was not shown to be strand-specific (Quitschke et al. 2000). We found that this spike was not present in all footprints. Of the 4,122 CTCF footprints where we additionally required the footprint to overlap 95% of the 20 base pair motif, approximately 80% showed evidence of a spike in hypersensitivity, while the remaining 20% did not (Fig. 5A). Approximately 60-80% of CTCF footprints that contained the spike in lymphoblastoid cells also contained the spike in the other 5 diverse cell types. DNA strand-specific footprint analysis displays very distinct patterns of digestion within and across these two sets of footprints (Fig. 5B-C) and suggests alternative ways that CTCF associates with DNA.

We analyzed different portions of the larger CTCF footprint for secondary motifs and identified SCTGCAST, a motif similar to that recently described for the zinc finger protein Zbtb3 (Badis et al. 2009), approximately 10bp (or one DNA helical turn) upstream from the 5' end of the CTCF motif

(Fig. 5A). This motif appears in 10-20% of sites with the spike in hypersensitivity but is not present in the set of CTCF footprints lacking the upstream spike. Excluding CTCF motifs that have an adjacent Zbtb3 motif does not eliminate the upstream footprint, indicating that Zbtb3 is not the only putative factor that binds upstream of CTCF, and it is not likely contributing to the spike digestion pattern upstream of the main CTCF motif.

Discussion

We have shown that genome-wide *in vivo* DNaseI footprinting can precisely identify a large number of specific cis-regulatory protein binding events in human *in a single experiment*. DNase-seq works well with as little as 1 million cells, which will be useful in studying rare primary cell types that are limited in number. We have generated publicly available footprint annotations for seven similar and six diverse cell lines for which DNase-seq data was generated as part of the ENCODE project. Even though comprehensiveness of footprint annotations is likely dependent on the number of sequences from DNase-seq and the level of background noise in the experiment, we show evidence that even a relatively small number of sequences provide a good initial annotation. This indicates that high-throughput DNase-footprint identification is possible for genomes much larger than yeast. As sequencing continues to get less expensive, we expect to be able to obtain even more accurate footprint maps.

Since the binding affinity of a factor to DNA can affect the relative amount of protection from DNaseI digestion, this may affect our ability to annotate all transcription factor binding sites. This is supported by our observation that ChIP-seq sites that do not overlap DNaseI footprints have weaker ChIP signals. Relative intensity of footprints may therefore provide another source of data to measure binding affinity and occupancy for various factors across the genome. We have also shown that DNase-seq footprints can distinguish more precisely how factors like CTCF may interact with DNA in different ways and may bind in conjunction with other novel factors. These types of information are distinctly different than that typically extracted from ChIP-seq data.

DNaseI footprint annotation relies on known PWMs to predict what individual factors are bound within each footprint. Since DNA binding preferences are available for only a fraction of known factors in the public JASPAR (Bryne et al. 2008), TRANSFAC (Matys et al. 2006) and UniPROBE (Newburger and Bulyk 2009) PWM databases, we believe this contributes to our ability to only annotate 25% of footprints. We anticipate that continued PWM discovery using both *in vivo* and *in vitro* assays (ChIP, SELEX, dsDNA arrays, etc.) will increase our knowledge of binding preferences for factors enabling a more accurate and complete annotation of DNaseI footprints. We believe that identifying *de novo* computational motifs in unlabeled footprints will also be an important part of this endeavor.

While DNase-seq footprinting offers a powerful method to identify transcription factor binding sites, it has a number of limitations. For example, it will not likely be able to precisely identify different transcription factors that bind to the same motif. It will also not identify proteins that indirectly bind DNA via interactions with other DNA binding proteins. This indirect binding likely explains why many binding sites identified by ChIP experiments often lack a recognizable motif. Therefore, to more fully identify and understand how complexes bind DNA, future studies will need to integrate multiple complementary genome-wide datasets, including DNaseI footprints, ChIP-seq for many factors, PWMs, and other datasets such as those that have recently identified combinatorial binding interactions between large numbers of transcription factors (Ravasi et al. 2010).

DNase-seq footprinting represent a significant advance towards the better understanding of the location, identity, and affinity for hundreds of thousands of cis-regulatory elements genome-wide. For most footprints identified here, it is unknown what trans factor binds to each site. While these data provide a scaffold to more fully annotating the genome, other complementary methods will be required for complete annotation. However, regardless of their identity, we now have evidence of specifically where factors are interacting with the DNA, which is an important step in understanding the components that regulate global gene expression.

Methods

Cell Line Growth

The source of cells, catalog numbers, and extensive cell growth protocols for all cell types described here can be found on the UCSC Genome browser ENCODE website (http://genome.ucsc.edu/ENCODE/cellTypes.html).

RNA expression

Total RNA was isolated from these cells using trizol extraction followed by cleanup on RNEasy column (Qiagen) that included a DNaseI step. The RNA was checked for quality using a nanodrop and an Agilent Bioanalyzer . RNA (1ug) was then processed according to the standard Affymetrix Whole transcript Sense Target labeling protocol that included a riboreduction step. The fragmented biotin-labeled cDNA was hybridized over 16 hours to Affymetrix Exon 1.0 ST arrays and scanned on an Affymetrix Scanner 3000 7G using AGCC software. The resulting cell files were analyzed for quality using Affymetrix Expression Console software. All expression data was submitted to the Gene Expression Omnibus (GSE15805).

DNaseI Assay

DNase-seq was performed as previously described (Song and Crawford 2010). Briefly, cells were lysed with NP40 and digested with optimal amounts of DNaseI enzyme. DNaseI ends were made blunt and ligated to biotinylated linkers containing an MmeI restriction site. After digesting with MmeI, DNaseI digested ends were enriched on a streptavidin magnetic beads (Invitrogen) and ligated to a second set of linkers. DNA was lightly amplified and sequenced using the Illumina GAII. All DNase-seq data has been made publicly available on the UCSC Genome Browser (Regulation Group, Open Chromatin track).

CTCF ChIP-seq Assay

10^8 cells were fixed for 10 min at room temperature by adding formaldehyde (1% final concentration). Formaldehyde was deactivated with 2.5 M glycine (125 mM final concentration). The cross-linked cells were lysed and sonicated 3 times for 10 min with a Bioruptor (Diagenode), which generated an average size of 500 bp DNA fragments. Chromatin immunoprecipitation was performed with the sonicated cell lysate to purify CTCF-DNA complexes, using an anti-CTCF antibody (Millipore 07-729). Crosslinks in the immunoprecipitated DNA protein complexes were reversed by incubation at 65^{o}C overnight. These samples were then treated with RNase A (Ambion) and Proteinase K (Invitrogen), followed by a phenol-chloroform extraction and ethanol precipitation. 30 ng of immunoprecipitated DNA was used to construct the library for Illumina sequencing.

Raw sequence processing

Raw sequence data from technical replicates for each cell type were combined and aligned to hg18

with MAQ (Li et al. 2008). We retained all sequences that aligned to at most 4 genomic locations and that contained at most 2 mismatches. Those sequences that aligned to multiple genomic locations were randomly assigned one of these locations by MAQ.

Aligned sequences were filtered using three different methods. First, all sequences that fall into regions where the human genome assembly under-represents the true amount of a particular sequence, namely satellites in pericentromeric and subtelomeric regions, were removed. These locations can be found on the UCSC browser (table wgEncodeDukeRegionsExcluded). Next, when more than five sequences aligned to a single location, this count was reduced to five as based on fitting the sequence data to a Poisson distribution, it is highly unlikely that more than five of the same sequence would be present and likely represents experimental artifact. Finally, we removed all sequences from regions where 70% of sequences in a 31bp region fall in a 5bp window. We believe these represent experimental artifacts likely due to incorrect PCR amplification.

DNaseI hypersensitive regions identified by DNase-seq
Regions of significant enrichment of DNaseI tags were identified using the F-seq peak caller (Boyle, Guinney, et al. 2008). F-seq identifies regions of high density of sequence reads as compared to a random background distribution of reads. We adjusted the background of F-seq based on input sequence from each cell line and an alignability background (UCSC Browser, table wgEncodeDukeUniqueness20bp). The distribution of base pair F-seq scores was fit to a gamma distribution, and the score corresponding to a p-value of 0.05 was used to discretely define DNaseI hypersensitive sites.

CTCF, SRF, REST, and GABP binding sites identified by ChIP-seq
CTCF binding sites defined by significant enrichment of sequences from ChIP-seq were similarly identified as for DNase-seq using F-seq. A p-value cut-off was also used to define discrete binding regions and the maximium F-seq score was assigned to each peak. Factor binding sites for SRF, REST, and GABP defined in their original publication within Jurkat human T lymphoblasts were used for these factors (Valouev et al. 2008).

Mapping motif binding sites
PWMs from the JASPAR (Bryne et al. 2008), free TRANSFAC (Matys et al. 2006), and UniPROBE (Newburger and Bulyk 2009) databases as well as the Ren lab CTCF motif (Essien et al. 2009) were used to initially annotate potential factor binding sites. There is redundancy between these sets, but because of slight motif differences, all PWMs were used.

Each mapping produced a bit-score that was used to filter the motif matches. We set the cutoff for matches as the lowest of either 70% of maximum possible bit score or 90% of functional depth, where the functional depth is defined as the difference between the maximum possible and minimum possible score for a particular PWM. Finally, if the PWM had a 0 value for any base (AGCT) in a particular position this was respected meaning no mappings for this motif may contain this 0-scored base in this position.

Clustering CTCF motifs based on footprinting data
For each base +/- 200 bases surrounding the 20bp CTCF motif, we determined the number of DNaseI cut sites. Using this 420-value vector, CTCF motifs were split into two clusters using k-means clustering implemented in R (kmeans package, k=2).

Hidden Markov Model (HMM) to identify footprints

Prior to input to the HMM, aligned raw sequence data were pre-processed by normalizing and smoothing the input sequences as follows. First, the number of sequence tags at each base was normalized. This normalization consisted of dividing counts at each base by the mean of all non-zero sequence counts in a 1-kilobase region surrounding that base. Next, these values were fit to a 2nd-order polynomial based on the 8 surrounding bases using the Savitzk-Golay filter. In this way, these values corresponded to the slope of a curve representing the relative change in the density of DNaseI cuts.

The HMM consisted of 5 states where emissions from each state corresponded to features of the values described above (Supplementary Fig. 1E). Each footprint was expected to be in a region of low DNaseI digestion with a reduced density of DNaseI cuts surrounded by increased densities of cuts to either side. This pattern was captured by a path through the model states starting with the background DNaseI hypersensitive state ("HS"), transitioning to the state with increasing DNaseI digestion ("UP"), followed by the state with decreasing DNaseI digestion ("DOWN"), then the footprint state ("FP") and again through the UP and DOWN states. The transition and emission probabilities were initially trained for the combined lymphoblastoid cell lines with experimental data from the previously studied promoter of the FMR1 gene. The footprinting of the FMR1 region was small and specific to lymphoblastoid cells, therefore we chose to use a set of predicted sites from the combined lymphoblastoid data to train each additional line. Footprints from the 1000 highest scoring DNaseI HS sites from chromosome 6 were chosen with the assumption that these strong sites will primarily consist of ubiquitous regulatory features based on our analyses of DNaseI HS sites across cell lines (data not shown). Emission probabilities for all other cell lines were trained using these sites using maximum likelihood training. Emission and transition probabilities for all models are listed in Supplementary Table 7. HMM software used to implement these models can be found at http://www.kanungo.com/software/software.html (Kanungo 1999).

Each identified DNaseI hypersensitive region were annotated with footprints using posterior decoding based on the trained HMM model. For display purposes, 3bp was added to either side of the called footprint. STAMP was then used to label each footprint based on known PWMs. Each motif label required a STAMP calculated p-value match of < 1e-06 to be labeled with a particular factor. All factors meeting this threshold were included in the factor label and sorted by significance. Note, the motif was required to be fully contained within the footprinted region in order to be annotated by STAMP.

Conservation Values and Plots

The first base of all footprints was set to the 0 position. Footprint conservation scores were then down-sampled or up-sampled to create 20bp of discrete values for each footprint (essentially shrinking or spreading each footprint conservation scores to the same genomic size). The upstream values were then calculated starting at the beginning of the footprint and the downstream values were calculated starting at the end of the footprint. Vertebrate plots were created using the UCSC table phastCons44way, mammal plots were created using the UCSC table phastCons44wayPlacental, and primate plots were creating using the UCSC table phastCons44wayPrimates.

Acknowledgements

We would like to thank Lisa Bukovnik, Tonya Severson, and Fangfei Ye at the Duke sequencing core facility. We thank Sridar Chittur and Scott Tenenbaum at the University of Albany-SUNY for help

with RNA expression studies. This research is supported by a NHGRI grant U54HG004563 to TSF, VRI, EB, and GEC.

Figure Legends

Figure 1. DNase-Seq identifies protein-DNA footprints. All potential CTCF binding sites were identified genome-wide using motif matching and compiled such that their 5' end was set at position zero. Cumulative DNase-seq and CTCF ChIP-seq signals within 500 bp of each site in both directions was determined. A) CTCF motifs that have a DNaseI footprint (red) also display high CTCF ChIP-Seq signal (green). B) CTCF motifs that have no footprint have greatly reduced CTCF ChIP-Seq signal. C) Footprinting using DNase-Seq accurately identified footprints within the FMR1 promoter region previously mapped using traditional *in vitro* DMS footprinting. Dips in raw DNase-Seq signal and annotated footprints correspond perfectly with previously identified footprints (gray boxes) (Drouin et al. 1997). The PhastCons annotation shows increased levels of evolutionary conservation within called footprints. D) A representative individual region displaying a DNaseI footprint matching a known CTCF binding motif (gray box) with a strong corresponding CTCF ChIP-Seq signal. See also Supplementary Figure 1A.

Figure 2. Accuracy of footprinting model. Positive predictive value (PPV) was calculated for predictions of four factors: CTCF, REST, GABP, and SRF. True positives were determined by ChIP-seq peaks with a matching motif while true negatives were determined by motifs without corresponding ChIP-seq peaks. PPV is shown for predictions using only PWMs (all PWMs are considered an actual binding site), PWMs that map within DHS sites (all PWMs within a DNaseI hypersensitive site are considered actual binding sites while those PWMs outside of DNaseI hypersensitive sites are considered negatives), and PWMs that map within footprints (all PWMs within a footprint are considered actual binding sites while those PWMs outside of footprints are considered negatives). The total number of PWMs mapped to the genome for each factor is listed in parentheses.

Figure 3. Identification of cell type specific footprints. Cumulative DNase-seq footprinting signals were determined across seven different cell lines for REST (A), TLX1-NFIC (B), and IRF2 (C). For each factor, the same set of motifs was used for all seven cell types. DNase-seq read counts were calculated in the regions surrounding these motifs similar to Fig. 1. Regions shaded in gray represent cell types that display reduced footprinting signal. HUVEC IRF2 shows moderate footprinting signal (light gray). Note that for REST, all cell lines display consistent signals.

Figure 4. Conservation of sequence in and around DNaseI footprints. A) In general, footprints contain a strong sequence conservation signal with a nearby "shoulder" of conservation around all footprinted regions (black). Between the conservation peak and shoulder is a region with a marked decrease in conservation. This conservation pattern is not detected when the signal is centered on DNaseI hypersensitive sites (red). The average conservation signal across the genome is shown in green. B) The conservation pattern for a single factor, NFYA, displays the characteristic drop in conservation surrounding the footprint. C) This conservation pattern is not detected around CTCF footprints, which shows relatively little conservation outside the highly conserved footprint. See also Supplementary Figure 6.

Figure 5. High-resolution analysis of CTCF binding sites. A) Cumulative footprinting signal at all CTCF motif predicted site that includes sites with and without a large increase in DNaseI digestion upstream of the CTCF motif. The light gray bar indicates the location of the known CTCF motif. The dark gray bar represents the location of a novel binding motif. The novel binding motif was only

detected in CTCF footprints that contain the small upstream region with a spike in DNaseI hypersensitivity (HS). Note that the entire protected region is approximately 50-60 bases. B) Strand-specific DNase-seq signal for the subset of CTCF motif identified sites that contain the upstream DNaseI HS spike. The DNaseI HS spike is only detected on the positive strand. C) Similarly for the CTCF motif identified sites without the upstream DNaseI HS spike. The diagram below each plot in B) and C) illustrates the estimated strand specific protected regions surrounding the CTCF motif predicted sites.

Figure 1.

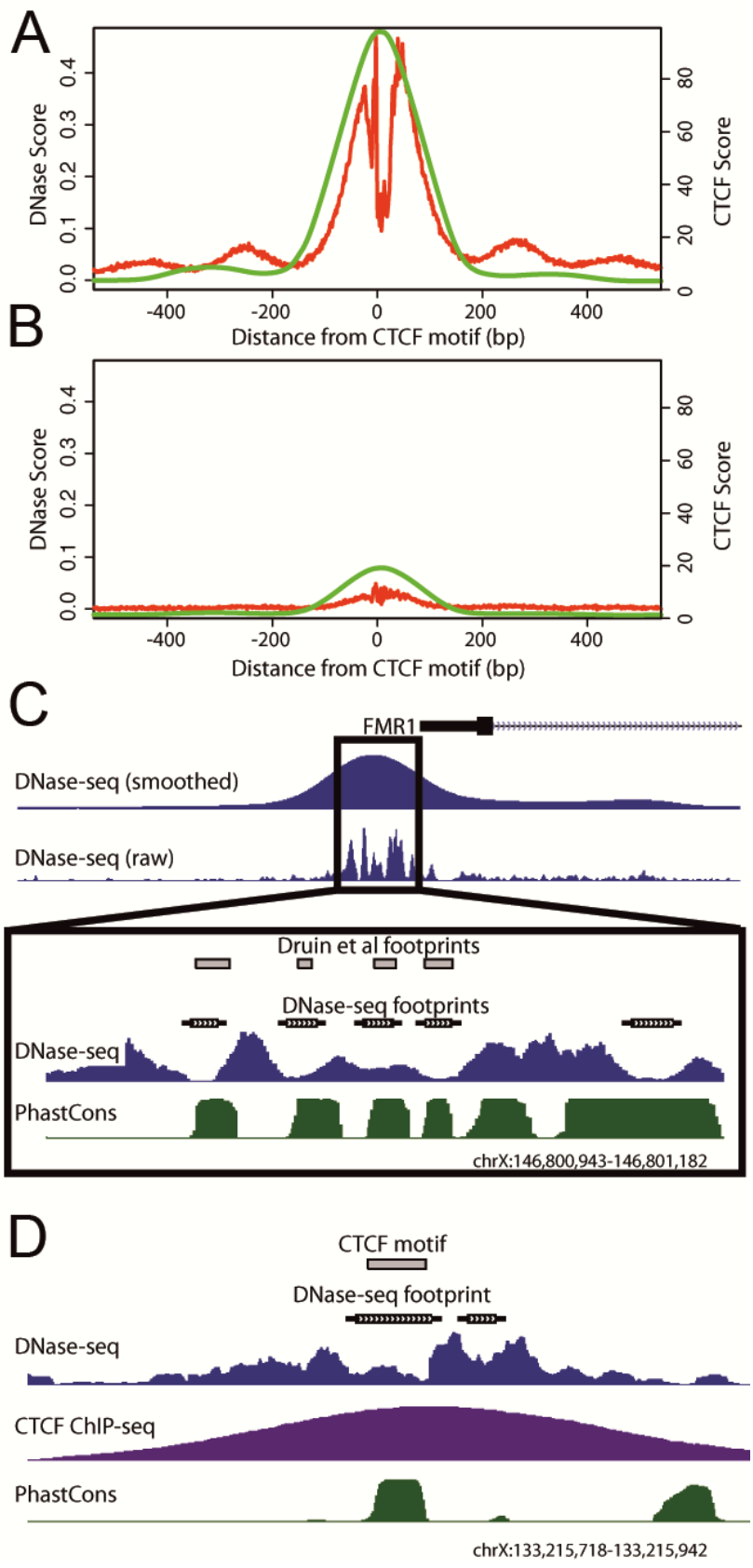


Figure 2.

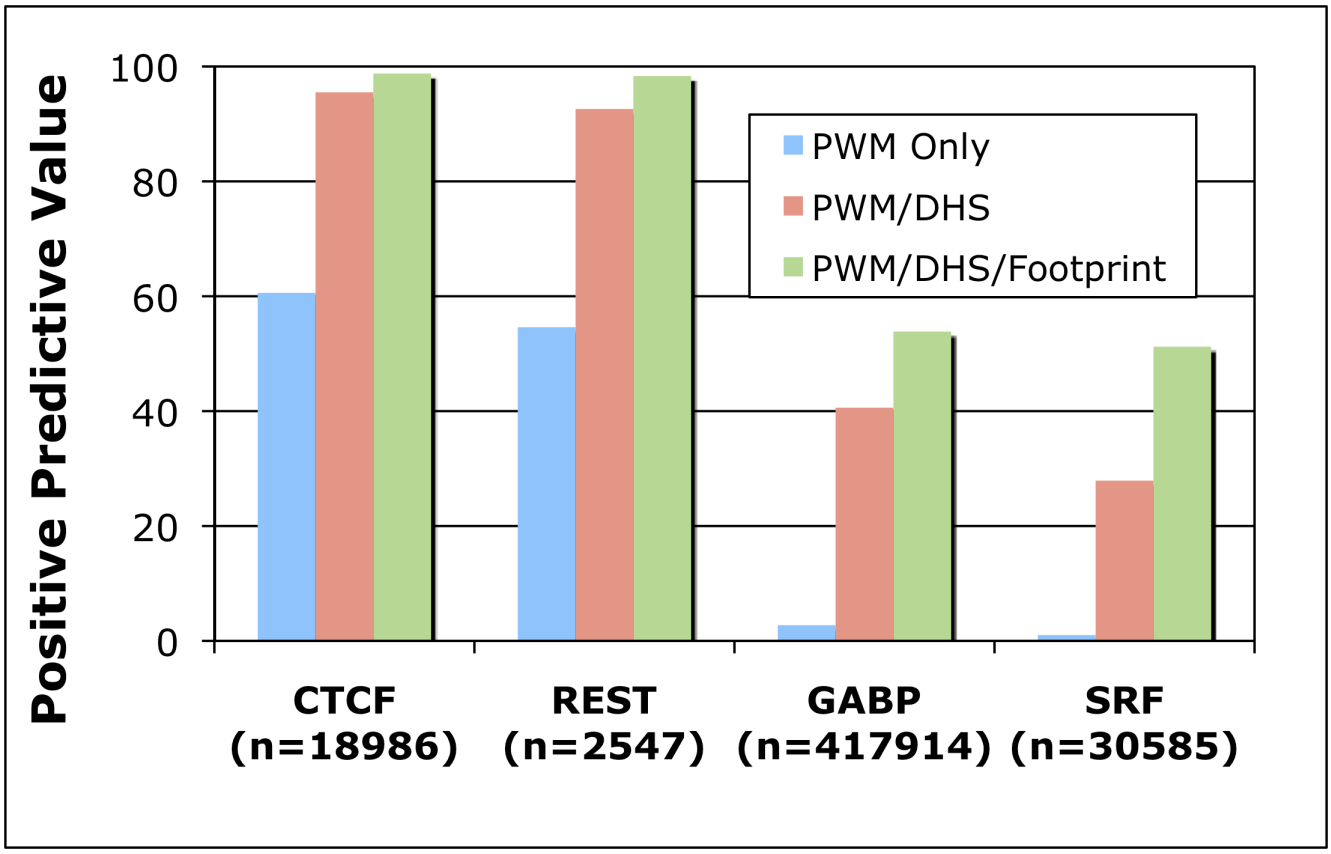


Figure 3.

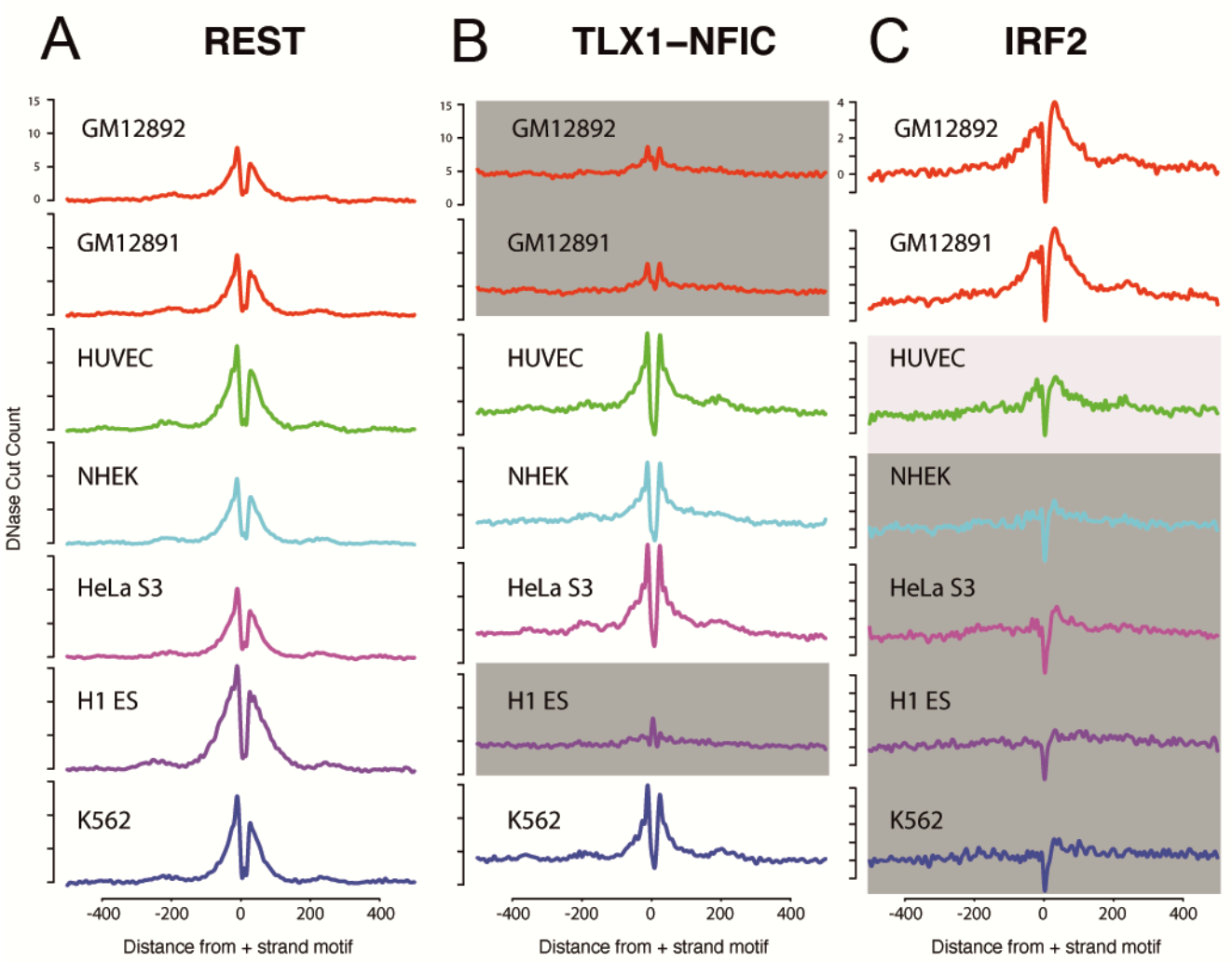


Figure 4.

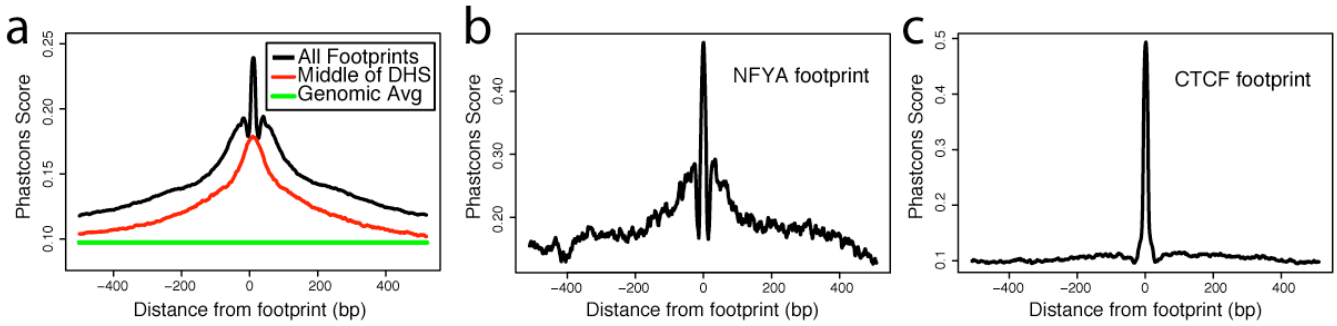
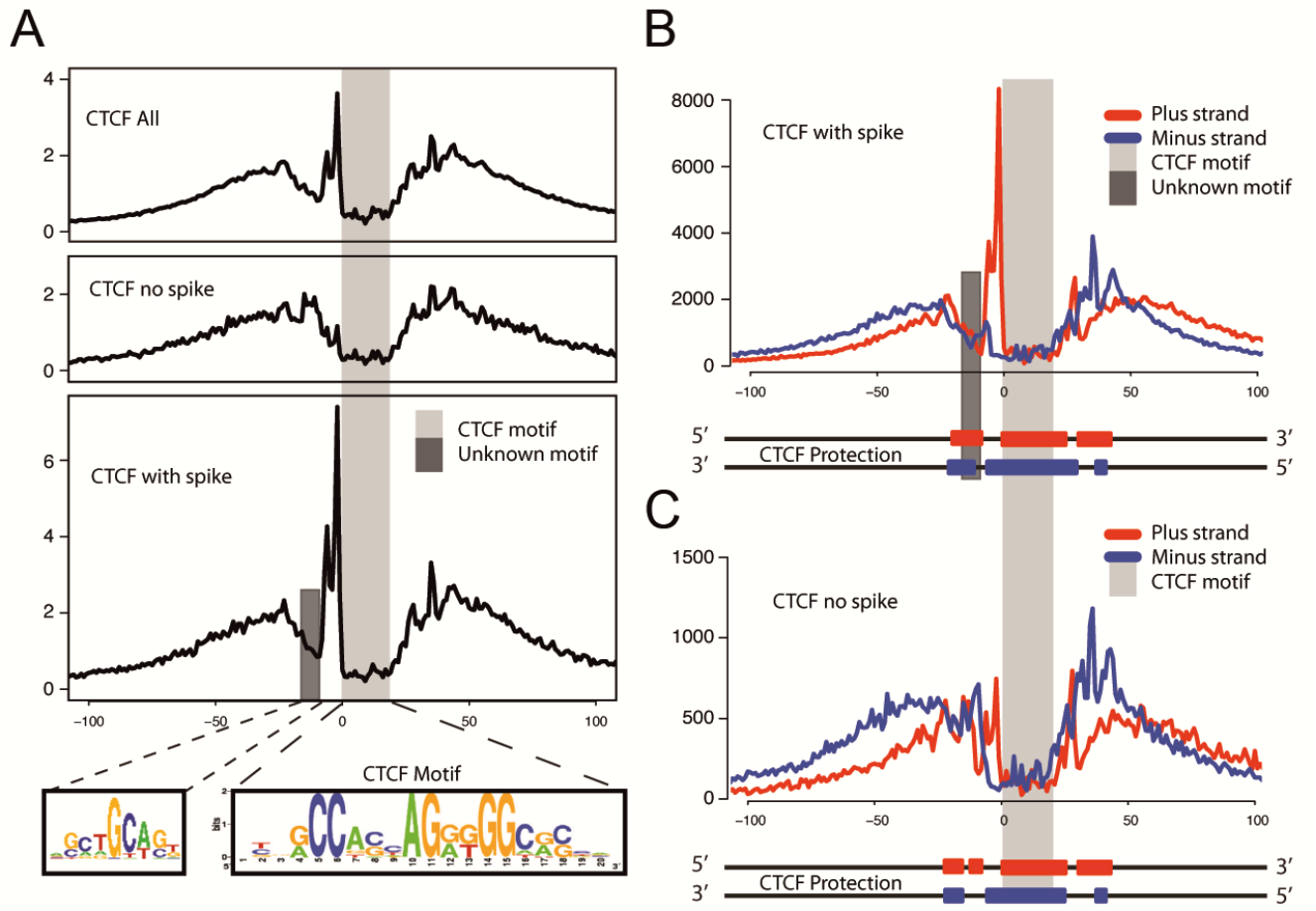


Figure 5.



References

Badis, G. et al. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science* **324**: 1720-1723.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., and Rubin, E.M. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**: 1391-1394.

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.

Boyle, A.P., Guinney, J., Crawford, G.E., and Furey, T.S. 2008. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**: 2537-2538.

Bryne, J.C., Valen, E., Tang, M.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 Res* **36**: D102-106.

Drouin, R., Angers, M., Dallaire, N., Rose, T.M., Khandjian, E.W., and Rousseau, F. 1997. Structural and functional characterization of the human FMR1 promoter reveals similarities with the hnRNP-A2 promoter region. *Hum. Mol. Genet* **6**: 2051-2060.

ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306**: 636-640.

Essien, K., Vigneau, S., Apreleva, S., Singh, L.N., Bartolomei, M.S., and Hannenhalli, S. 2009. CTCF binding site classes exhibit distinct evolutionary, genomic, epigenomic and transcriptomic features. *Genome Biol* **10**: R131.

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. *Mol. Cell. Biol* **16**: 2802-2813.

Gross, D.S., and Garrard, W.T. 1988. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem* **57**: 159-197.

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

Kanungo, T. 1999. UMDHMM: Hideen Markov Model Toolkit. In *Extended Finite State Models of Language*, Cambridge University Press, 1999.

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.

Li, H., Ruan, J., and Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**: 1851-1858.

Liu, Y., Liu, X.S., Wei, L., Altman, R.B., and Batzoglou, S. 2004. Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res* **14**: 451-458.

Mahony, S., Auron, P.E., and Benos, P.V. 2007. DNA Familial Binding Profiles Made Easy: Comparison of Various Motif Alignment and Clustering Strategies. *PLoS Comput Biol* **3**: e61.

Mahony, S., and Benos, P.V. 2007. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* **35**: W253-258.

Matys, V. et al. 2006. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* **34**: D108-110.

McDaniell, R. et al. 2010. Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* **328**: 235-239.

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

Phillips, J.E., and Corces, V.G. 2009. CTCF: Master Weaver of the Genome. *Cell* **137**: 1194-1211.

Quitschke, W.W., Taheny, M.J., Fochtmann, L.J., and Vostrov, A.A. 2000. Differential effect of zinc finger deletions on the binding of CTCF to the promoter of the amyloid precursor protein gene. *Nucleic Acids Res* **28**: 3370-3378.

Ravasi, T. et al. 2010. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell* **140**: 744-752.

Siepel, A. et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research* **15**: 1034-1050.

Song, L., and Crawford, G.E. 2010. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *CSH Protoc* **2010**: pdb.prot5384.

Tamura, T., Yanai, H., Savitsky, D., and Taniguchi, T. 2008. The IRF family transcription factors in immunity and oncogenesis. *Annu. Rev. Immunol* **26**: 535-584.

Valouev, A., Johnson, D.S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R.M., and Sidow, A. 2008. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Meth* **5**: 829-834.

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

Wu, C. 1980. The 5' ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. *Nature* **286**: 854–860.

Zhang, N., Shen, W., Hawley, R.G., and Lu, M. 1999. HOX11 interacts with CTF1 and mediates hematopoietic precursor cell immortalization. *Oncogene* **18**: 2273-2279.