Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types

Jason Ernst$^{1,2,3,4}$ and Manolis Kellis$^{5,6}$

$^1$Department of Biological Chemistry, David Geffen School of Medicine, University of California, Los Angeles
$^2$Eli & Edythe Broad Center of Regenerative Medicine and Stem Cell Research at UCLA
$^3$Jonsson Comprehensive Cancer Center, University of California, Los Angeles
$^4$Molecular Biology Institute, University of California, Los Angeles
$^5$MIT Computer Science and Artificial Intelligence Laboratory
$^6$Broad Institute of MIT and Harvard

Corresponding Author
Manolis Kellis (manoli@mit.edu)
32 Vassar Street, 32D-524
Cambridge, MA 02139
Phone: 617-253-2419

Running Title: Chromatin state, regulator binding, and motifs
Keywords: Chromatin, Epigenomics, Motifs, Transcription Factors
Abstract
The regions bound by sequence-specific transcription factors can be highly variable across different cell types, despite the static nature of the underlying genome sequence. This has been partly attributed to changes in chromatin accessibility, but a systematic picture has been hindered by the lack of large-scale datasets. Here, we use 456 binding experiments for 119 regulators and 84 chromatin maps generated by ENCODE in six human cell types and relate those to a global map of regulatory motif instances for these factors. We find specific and robust chromatin state preferences for each regulator beyond the previously-reported open-chromatin association, suggesting a much richer chromatin landscape beyond simple accessibility. The preferentially-bound chromatin states of regulators were enriched for sequence motifs of regulators relative to all states, suggesting that these preferences are at least partly encoded by the genomic sequence. Relative to all regions bound by a regulator however, regulatory motifs were surprisingly depleted in the regulator’s preferentially-bound states, suggesting additional non-sequence-specific binding beyond the level predicted by the regulatory motifs. Such permissive binding was largely restricted to open-chromatin regions showing histone modification marks characteristic of active enhancer and promoter regions, whereas open-chromatin regions lacking such marks did not show permissive binding. Lastly, the vast majority of co-binding of regulator pairs is predicted by the chromatin state preferences of individual regulators. Overall, our results suggest a joint role of sequence motifs and specific chromatin states beyond mere accessibility in mediating regulator binding dynamics across different cell types.

Introduction

While the genome sequence of each human cell is invariant across nearly all cell types of the human body, the morphology and function of each cell is dramatically different, owing to their differential regulation and gene expression patterns. At the molecular level, the binding landscape of a given regulator can be extremely dynamic, even though its sequence specificity remains unchanged(Harbison, et al, 2004,Zhong, et al, 2010,Mullen, et al, 2011,Trompouki, et al, 2011). This is attributed at least in part to the dynamic chromatin landscape of each cell, via active and repressed regions that can then be epigenetically maintained(John, et al, 2011,Li, et al, 2011,Lickwar, et al, 2012,Lam, et al, 2008,Segal, et al, 2009,Essien, et al, 2009). The chromatin landscape is itself thought to be driven at least in part by the regulators active in each cell type(Siersbaek, et al, 2011,Steger, et al, 2010,Lefterova, et al, 2008,Lupien, et al, 2008). For example, transient over-expression of a small number of transcription factors has been shown sufficient for stable epigenetic reprogramming, which is now commonplace in the generation of induced Pluripotent Stem (iPS) cells(Takahashi, et al, 2006,Meissner, 2010). However, a systematic study of the interplay between regulator binding including both general and sequence specific regulators, chromatin accessibility, and chromatin states defined by histone modification marks has been unfeasible due to the lack of systematic genome-wide regulator binding experiments in multiple cell types with matched chromatin datasets.
This situation changed with the scale-up of the ENCODE project (ENCODE Project Consortium, et al, 2012). First, the genome-wide binding locations of more than 100 regulators have been mapped in one or multiple cell types (Supplementary Tables 1 and 2), identifying thousands of constitutive and variable target locations for each experiment. Second, the chromatin accessibility landscape of matched cell types has been mapped using DNase hypersensitivity and formaldehyde-based FAIRE (Song, et al, 2011, Hesselberth, et al, 2009). Third, at least eight histone modification marks have been mapped in the same cell types that can be used to pinpoint distinct chromatin functions such as enhancer and promoter regions. These data have individually highlighted the remarkable fact that in a given cell type, only a very small percentage of the 3 billion bases of the genome have robustly detectable regulator binding, accessible chromatin, or histone marks denoting active regulatory elements. Strong relationships between each pair of data types have been previously reported and regulatory motifs have been shown to be over-represented (enriched) within both active chromatin marks and regions of regulator binding (John, et al, 2011, Lupien, et al, 2008, Robertson, et al, 2008, Xi, et al, 2007, Boyle, et al, 2008, Ernst, et al, 2010, Ernst, et al, 2011, Wu, et al, 2011, modENCODE Consortium, et al, 2010, Heintzman, et al, 2007, Heintzman, et al, 2009, Hon, et al, 2008). However, the dynamic changes in regulator binding and active regulatory elements across cell types are far from understood at the systems level.

In this paper, we integrate this vast collection of histone modification, chromatin accessibility, and regulator binding including both transcription factor and general binding, as well as regulatory motif information to systematically study dynamic regulatory binding across multiple cell types. We leverage 84 genome-wide datasets of chromatin related mark patterns and 456 regulator binding experiments for 119 different regulators generated by the ENCODE project in six human cell types, and 2.76 million total motif instances from 51 different positional weight matrices (ENCODE Project Consortium, et al, 2012, Kheradpour et al, in preparation). We use a chromatin state model that summarizes biologically-meaningful combinations of chromatin marks (Hoffman, et al, 2013), shown to more directly correlate with diverse functional elements (Ernst, et al, 2010, Jaschek, et al, 2009, Thurman, et al, 2007). The model used here integrates both histone modification and chromatin accessibility into a joint 25-state model, which enables us to distinguish diverse classes of open chromatin. We use this rich chromatin annotation to study the interplay between chromatin state, regulator binding, and regulatory motifs across multiple cell types, revealing numerous new insights:

- We find several distinct classes of chromatin state preferences for different regulators, which are stable across experimental conditions and across different cell types.
- We find different chromatin state preferences for locations of cell type unique, shared, and excluded binding, respectively enriched in enhancer, promoter, and repressed states.
- We find that the chromatin states preferentially bound by regulators show strong enrichment for their cognate regulatory sequence motifs.
• Surprisingly however, relative to all regions bound by a given regulator, binding events within preferred chromatin states are more often depleted for regulatory motif instances, suggesting additional non-specific binding that does not rely on specific regulatory motif instances.
• Moreover, we find non-specific binding primarily in those accessible chromatin states that also contain active histone modification marks, suggesting that permissive binding is associated with specific modifications rather than open chromatin alone.
• Lastly, we find that previously-reported pairwise enrichment in the binding locations of regulator pairs can potentially be explained to a large extent by similar chromatin state preferences.

Overall, our results suggest a previously-unappreciated diversity in the chromatin state preferences of different transcription factors that likely underlies non-permissive binding, potentially mediates interactions between multiple regulators, and facilitates the cell type-specific and cell type-restricted regulator binding.

Results

Chromatin landscape reveals diverse classes of accessible chromatin regions.

To study the dynamic nature of regulator binding and chromatin states, we focused on six human cell types, consisting of lymphoblastoid (Gm12878), cervix adenocarcinoma (HeLa-S3), liver carcinoma (HepG2), umbilical vein endothelial cells (Huvec), myelogenous leukemia (K562), and embryonic stem cells (H1-hESC). These were prioritized as Tier 1 and Tier 2 cells in the ENCODE project, and thus benefit from extensive experimentation across all ENCODE groups, enabling integration across histone modifications, chromatin accessibility, transcription factor binding, and gene expression datasets, even though these were generated by different ENCODE production groups.

In each of these cell lines, we characterize the chromatin landscape by integrating 14 genome-wide chromatin tracks. These include eight histone modification marks: mono-, di-, and tri-methylation of histone 3 lysine 4 (H3K4me1, H3K4me2 and H3K4me3), typically associated with enhancer and/or promoter regions; acetylation of histone 3 lysine 9 and 27 (H3K9ac and H3K27ac) that mark active regulatory elements; the repressive mark H3K27me3; H3K36me3 and H4K20me1 associated with gene bodies (Ernst, et al, 2011, Barski, et al, 2007, Wang, et al, 2008). These also include three tracks of chromatin accessibility, typically associated with increased regulator binding: single-cut DNase (Song, et al, 2011), double-cut DNase (Hesselberth, et al, 2009), and sonication (Song, et al, 2011) assays. Lastly, they include binding of two general regulators: CTCF, associated with insulator and other functions; and RNA polymerase POL2; and an input control.

We used chromatin states (Ernst, et al, 2010) from ChromHMM (Ernst, et al, 2012) to summarize biologically-meaningful combinations of those marks into 25 chromatin states (Hoffman, et al, 2013), which were consistently defined across cell types. Briefly, the states consist of: Transcription start site
These states are defined by the frequency of each mark in each chromatin state (emission probabilities), and the frequency with which states are found adjacent to each other (transition probabilities) (Supplementary Figure 1). The state definitions are constant across cell types, as the model was learned jointly by a virtual concatenation of the six cell types, but the state assignments are cell type-specific, as they depend on the specific combination of chromatin marks observed in a given cell type. Different states cover very different fractions of the genome: individual promoter, enhancer, open chromatin, and CTCF states usually cover less than 1% of the genome, the set of transcribed states (14-19) on average cover roughly 10% of the genome, and the low and quiescent states 23 and 24 together cover approximately 70% of the genome (Supplementary Figure 2).

While transcription factor binding is known to be generally associated with regions of open chromatin (Song, et al, 2011), our chromatin state annotations suggest a much more complex picture with several types of open chromatin. In fact, at least nine chromatin states showed DNase hypersensitivity emission parameter frequencies ~50% or greater (Supplementary Figure 1), including: promoter states (1_Tss, 4_PromP), enhancer states (5_Enh, 8_EnhW), DNase-only regions lacking other histone marks (9_DNaseU, 10_DNaseD), CTCF binding regions (12_CtcfO), the specific repression state frequently enriched in promoter regions (20_ReprD), and the artifact state (25_Art). This diversity of open chromatin states suggests that a more complex relationship may exist between transcription factor binding and chromatin, beyond simply a general preference for accessible chromatin regions, which we explore next by studying the preferences of each transcription factor for each chromatin state.

Regulators show multiple distinct chromatin state enrichment patterns

As expected, we find that regulators are typically most enriched for states with open chromatin. In fact, after collapsing multiple experiments for the same regulator and excluding CTCF and POL2RA, we find that 104 of 115 regulators (90%) with cell type matched chromatin data available showed a maximum enrichment in either the active promoter state 1_Tss (59 regulators) or the strong enhancer state 5_Enh (45 regulators). However, the extent of each regulator’s preference for the two states varied substantially for different regulators, as did their enrichment for additional open chromatin states.
We recognized common patterns of chromatin state preference (Figure 1, Supplementary Figure 3), using a k-means clustering algorithm and selecting 12 clusters (Supplementary Table 3, Supplementary Figure 4, see Methods). Regulators profiled in multiple cell types and conditions showed generally consistent enrichment patterns, with two thirds of the individual regulator experiments showing highest similarity to the cluster center where the corresponding regulator was assigned (Supplementary Figure 5).

Four of these clusters (C1-C4) had their strongest relative preference for promoter states:

- Regulators in cluster C1 showed almost exclusive binding preference in the 1_Tss promoter state. This cluster included all six factors annotated as ‘general Pol II associated factor, not site specific’ (p<0.001) (Wang, et al, 2012), consistent with non-specific binding that may be partly mediated by the chromatin landscape at transcription start sites.

- Regulators in clusters C2 and C3 also had a preference for the 1_Tss state, but relative to C1 had a stronger preference for the more repressive 4_PromP state (both C2 and C3), and for 25_Art (C3 only). Cluster C2 contained both NFYA and NFYB, which we had previously predicted to have repressive activity in enhancer states (Ernst, et al, 2011), suggesting they may also show repressive roles in poised promoters.

- Regulators in C4 showed the strongest preference for 1_Tss, but also a weaker preference for 5_Enh. This cluster contained both helix-loop-helix heterodimers MAX-MYC and USF1-USF2. The presence of TFIID-interacting USF1 and USF2 in this cluster suggests a possibly under-appreciated role in distal as well as proximal regulatory regions (Rada-Iglesias, et al, 2008).

Three clusters (C5-C7) had strong relative enrichment for enhancer states, and varying levels of promoter enrichment:

- Regulators in C5 had a balanced preference for both state 1_Tss and 5_Enh, and likely play dual roles in enhancers and promoters. C5 includes all four regulators that are part of the SWI/SNF chromatin remodeling complex (p<0.05), consistent with previous reports suggesting both enhancer and promoter roles (Euskirchen, et al, 2011).

- Cluster C6 had regulators with a much stronger preference for enhancers than promoters or other states. Of the 17 regulators in this cluster, 13 are known to be involved in the regulation of developmental processes (p<0.01). The cluster also enriched for genes involved in intracellular transport (p<0.01). Among the C6 genes was EP300 a common co-regulator found in large numbers of enhancers (Visel, et al, 2009).

- C7 regulators showed the strongest preference for enhancer state 5_Enh, but also a weaker preference for 8_EnhW, 9_DNaseU and 1_Tss. This cluster contained 8 of 12 leucine zipper domain
regulators (p<0.001), and was also significantly enriched for positive regulation of cell differentiation genes (p<0.01).

The remaining clusters (C8-C12) contained a number of regulators that did not fit the usual enhancer or promoter pattern:

- Cluster C8 regulators had the strongest enrichment for 12_CtcfO and contained RAD21 and SMC3, both members of the cohesin complex, which is known to interact with CTCF(Wendt, et al, 2009). It also contained the CTCF paralog CTCFL, and ZNF143, which has been previously implicated with CTCF(Gerstein, et al, 2012).

- Cluster C9 regulators were most enriched in the 9_DNaseU open-chromatin state that lacks any activating histone marks, and contained two regulators, TRIM28 and SETDB1, suggesting a potential repressive role upon binding. Indeed, both are known to be associated with chromatin gene-silencing (Schultz, et al, 2002).

- Cluster C10 regulators was associated with multiple states with DNase but lacking active histone marks and consisted of a single regulator, REST (NRSF), which has a known role in gene silencing.

- Cluster C11 also consisted of a single regulator SUZ12, a polycomb protein, which had a strong enrichment for the 4_PromP and 20_ReprD which is expected given that these states are associated with high levels of H3K27me3.

- Lastly, C12 regulators had relatively high enrichment for the ‘artifact’ state 25_Art, and indeed three of the four regulators (BRF2, HDAC8, and ZZZ3) were independently flagged by the ENCODE consortium as being of medium quality(Landt, et al, 2012, Kundaje et al, in preparation). The remaining regulator ZNF274 also had a notable enrichment for the 3’ gene body state 17_Gen3’ consistent with its involvement in recruiting a methyltransferase to the 3’ end of ZNF genes(Frietze, et al, 2010).

Dynamic enrichment patterns across cell types are both regulator and cell type driven.

We next expanded our chromatin state enrichment analysis to incorporate multiple cell types. For each regulator, we directly compared the chromatin state enrichment across different cell types (Figure 2) extending our analysis focusing on matched regulator-chromatin state experiments to study vectors of enrichments across six cell types (see Methods). We ordered the resulting matrix to minimize the total correlation-based distance between neighboring rows using an instance of the traveling salesman problem (see Methods) which showed greater coherence in both cell type and regulator streaks relative to optimal leaf ordering of a hierarchical clustering solution(Bar-Joseph, et al, 2001) (Supplementary Figure 6). We also generated an unbiased ordering at the individual experiment level to study whether common cell types or regulators were preferentially consecutively ordered, and to
highlight unexpected similarities in relative or absolute enrichment patterns (Supplementary Figures 7-10).

Most cell types exhibited substantial coherence in grouping together different regulators (Figure 2, Table 1). For example, 85% of the regulators in HepG2 fell into one of two groups preferentially enriched in states 1_Tss or 5_Enh. However, we also observed cases in which multiple experiments of the same regulator profiled in different cell types were grouped together (Table 2). Notable among these was eight of the nine REST experiments and all four ZNF274 experiments in different cell types, both of which play repressive roles.

We also observed constitutive regulator enrichments across cell types, particularly in states 1_Tss and 12_CtcfO, known to be less dynamic than enhancer states (Heintzman, et al, 2009). We confirmed at the individual experiment level that this was associated with invariant chromatin states, rather than simply independent enrichments at similar levels in each cell type (Supplementary Figure 7). Ordering experiments by cell type shows that each regulator is most enriched in enhancers active in the cell type in which it was profiled (Supplementary Figure 11), and ordering by regulator highlights dynamic changes in enrichment across cell types (Supplementary Figure 12) primarily for enhancer states, and to a lower degree for promoter states.

Motif depletion suggests abundant non-specific binding in permissive chromatin states.

To understand the potential role of DNA sequence in guiding observed transcription factor-chromatin associations, we studied the absolute motif enrichment in each chromatin state for regulators with a known regulatory motif, compared to the rest of the genome using control motifs (see Methods, Supplementary Figure 13). The states for which the greatest number of regulators show significant absolute motif enrichment were 1_Tss, 4_PromP, 5_Enh, 8_EnhW, 9_DNaseU, and 10_DNaseD (Figure 3a), corresponding well to the states showing the greatest transcription factor binding enrichment (with the exception of 10_DNaseD, Supplementary Figure 14). This property also held when considering motifs instances in aggregate across all regulators (Supplementary Figure 15). Moreover, the chromatin state enrichments for regulatory motifs mirrored their chromatin state binding preferences for the corresponding transcription factor clusters (Supplementary Figure 16). Together, these results suggest that the chromatin state preferences of sequence-specific regulators are at least in part encoded by genome sequence.

Surprisingly however, the two most often maximally preferentially bound states, 1_Tss and 5_Enh, showed a depletion of regulatory motif instances relative to all bound regions (Figure 3b, Supplementary Figure 15). This relative depletion implies that regulator binding in these regions exceeds the level predicted by regulatory motifs, suggesting that in addition to motif-driven binding, these states are conducive to non-specific binding. A similar depletion of strong CTCF motifs was noted in the presence of active modifications (Essien, et al, 2009), and of many regulator motifs in hotspots of
regulator binding in fly (modENCODE Consortium, et al, 2010) and human (Yip, et al, 2012). Indeed, states 1_Tss, 5_Enh, and 25_Art contained on average about half of hotspot regions (Supplementary Figure 17). However, even amongst hotspot regions, those overlapping states 1_Tss, 5_Enh, and 25_Art were less likely to contain motifs (Supplementary Figure 18), suggesting that chromatin states contain additional properties facilitating transcription factor binding than explained by high-occupancy binding alone. Enhancer state 5_Enh that showed the most dynamic binding across cell types also showed one of the strongest motif depletion signals (Figure 3b, Supplementary Figure 15), consistent with permissive binding facilitating condition specificity.

In contrast, relative motif depletion was not found in states lacking active marks. Instead, repressive states (20_ReprD, 21_Repr, 22_ReprW) and low-activity states (23_Low, 24_Quies) showed motif enrichments relative to all bound regions. This suggests that for binding to occur in these regions, a regulatory motif is more often required, and thus that regulator binding within these regions is more likely motif-dependent. In other words, binding appears to be more frequently sequence-mediated in regions that are not permissive or chromatin mediated.

Importantly, the two DNase-associated open-chromatin states that lack active chromatin marks showed a signature of non-permissive, motif-dependent binding characteristic of repressive states. Recall that States 9_DNaseU and 10_DNaseD showed signatures of open chromatin, but had limited H3K4 methylation marks characteristic of enhancer and promoter regions, and also lacked the H3K9 and H3K27 acetylation marks associated with activation of both enhancer and promoter regions. These states showed strong relative motif enrichments compared to all bound regions across all factors, and showed more factors with relative enrichment than with relative depletion, suggesting that they are strongly motif-dependent.

In summary, promoters, enhancers, and open chromatin states lacking active histone modifications all show strong absolute enrichments for regulatory motifs (e.g. CEBP motif for CEBP bound sites in HepG2, Figure 3c), suggesting that motifs at least partially determine regulator binding preferences in different chromatin states. However, motifs are depleted in promoter and enhancer peaks, relative to all peaks, suggesting permissive binding in active states restricted to open chromatin regions that also show active histone modification marks (Figure 3d for CEBP).

For regulators profiled in multiple cell types, we also analyzed the chromatin state enrichments in a given cell type for commonly-bound and differentially-bound sites. We defined: (1) ‘shared’ sites, bound in the cell type and at least one other of the six primary cell types; (2) ‘unique’ sites, bound only in the considered cell type and none of the other primary cell types considered; and (3) ‘excluded’ sites, that were bound in at least one other cell type but not the one considered (Figure 4, Supplementary Figure 19-20). For example, the transcription factor CEBPB shows very different chromatin state enrichments for these different classes of bound sites (Figure 4a): for 1_Tss, enrichment is strongest for shared sites, consistent with constitutive binding in promoter regions; for 5_Enh, enrichment is
strongest for unique sites, consistent with dynamic binding in enhancer regions; for 10_DNaseD, enrichment is strongest for excluded sites, consistent with specific repression; for 24_Quies, enrichment is weak throughout, consistent with lack of binding. Considering all regulators and all states, uniquely bound locations are preferentially found in States 5_Enh and 8_EnhW, ‘shared’ locations that are bound in multiple cell types are preferentially found in state 1_Tss, and ‘excluded’ locations that are bound only in other cell types are more enriched than the other two classes in states 9_DNaseU, 10_DNaseD, and 20_ReprD, suggesting they are specifically repressed (Figure 4b). We also found that regions of shared binding were more likely to contain regulatory motifs (Figure 4c), consistent with sequence driven binding, while dynamically-bound regions showed lower enrichments, as expected for dynamic binding since motifs are by definition invariant across cell types.

Common chromatin state preferences predictive of regulator co-binding.

We next studied the role of chromatin in the pairwise binding of regulators. As reported in previous studies (Gerstein, et al, 2012, Moorman, et al, 2006, Li, et al, 2008), we found extensive co-binding enrichments ranging from 50-fold and 500-fold on average for most pairs of regulators (Figure 5a), which held for both pairs of regulators and pairs of individual experiments (Supplementary Figures 21-22). The computationally-ordered pairwise enrichment matrix showed strong internal structure, revealing several subgroups of regulators with even higher pairwise overlap to each other, as reported in previous studies (Negre, et al, 2011), but whose origin has remained unclear.

The chromatin state preferences of regulators in each co-occupancy group revealed a potential chromatin state basis of co-occupancy groups, as regulators in the same subgroup showed preferences for the same set of chromatin states. For example, amongst the regulators mapped in HepG2 (Figure 5a), a large co-binding group was dominated by regulators almost exclusively in cluster C1 of Figure 1 that bind primarily promoter states while another large group consisted of regulators in clusters C5, C6, and C7 that primarily bind enhancer states.

Remarkably, after controlling for chromatin state preferences, genome-wide regulator co-associations were dramatically reduced (Figure 5b). Co-binding enrichments were reduced ~30-fold on average, and the strong structure of co-binding associations was lost (Figure 5, Supplementary Figures 21-22, see Methods). Previous work has noted extensive co-binding within DNase hypersensitive regions (Li, et al, 2011, Kaplan, et al, 2011). However we find that conditioning on DNase accessibility still leaves substantial block structures in the co-association that are eliminated when including chromatin state information (Supplementary Figure 23). Only small groups of co-associations remained that could not easily be explained by common chromatin state preferences for the corresponding regulators. For example, these factors included MAFF and MAFK proteins in HepG2 which can be involved in repression (Blank, 2008) and were found together in heterochromatic regions; SETDB1, KAP1, and ZNF274 in K562, three factors known to interact (Frietze, et al, 2010), and POL3 associated factors in K562 and HeLa-S3 (Figure 5, Supplementary Figures 21-22). Even when co-binding groups
mirrored chromatin state preferences, the direction of causality remains unclear. It is possible that co-binding regulators help define chromatin states, or that chromatin states facilitate regulator interactions.

**Discussion**

We undertook a global analysis of the relationship between regulator binding across multiple cell types, the dynamic chromatin landscape, and regulatory motifs. We integrated histone modifications, DNase hypersensitivity, FAIRE, CTCF, and POL2 to define chromatin states, revealing several classes of open chromatin, some of which were surprisingly devoid of histone modification marks. All chromatin states with open chromatin are enriched in regulator binding, but different regulators show distinct preferences for specific subsets of chromatin states: some bind primarily promoter regions such as NRF1 and TAF1; others bind in enhancers regions, such as JUN and FOXA1; others bind equally in both, as members of the SWI/SNF chromatin remodeling complex; and a smaller number of regulators bind insulators such as RAD21, SMC3, and ZNF143, or other regions such as SETDB1, TRIM28, SUZ12, and REST.

Chromatin state preferences are highly stable for a given regulator across different cell types and conditions, and are reflected in the underlying regulatory motif enrichments for the corresponding regulators. However, specific enhancer and promoter states that showed enrichment for transcription factor binding were also least likely to contain regulatory motifs relative to all bound regions. This suggests that regulatory sequence motifs may help define promoter and enhancer states, but that once these chromatin states are established, they in turn provide a permissive environment for additional binding that does not require regulatory motifs. Cooperative and non-sequence specific binding within such a generally permissive environment can also explain occurrences of regulator binding peaks that lack a sequence motif. Consistent with this latter possibility we find that bound locations lacking a motif interact with locations bound by the same transcription factor and containing a motif, based on 5C interaction data (Supplementary Table 4, see Methods). Importantly, open chromatin states that lack histone modifications did not show evidence of non-specific regulator binding, suggesting that active histone marks, rather than open chromatin, may enable permissive regulator binding.

While co-binding patterns between pairs of regulators have garnered much attention in previous studies, we find that most of these patterns can potentially be explained by similar chromatin state preferences for individual regulators. Once chromatin state preferences of each regulator are accounted for, co-binding enrichments are reduced by more than an order of magnitude, revealing only few instances of co-binding regulators that are difficult to explain by chromatin preferences alone. These results do not imply that their regulator co-occurrence patterns are not meaningful. In fact, chromatin may simply provide a mechanism for guiding functionally-related regulators to the same locations of the genome, and thus enabling their joint activity even in the absence of any direct protein-protein interactions between them. Conversely, interactions between regulators may underlie common
chromatin state preferences by mutual recruitment. In either case, it is important to recognize that regulator co-binding should be studied in the context of chromatin that may explain, or facilitate, regulator interactions.

Going forward, a systematic understanding of the joint role of DNA sequence information and epigenetic modifications will be paramount in understanding the molecular basis of human disease. On one hand, single-nucleotide polymorphisms (SNPs) have been shown to result in highly pronounced changes in regulator binding, chromatin accessibility, and gene expression across individuals (Montgomery, et al, 2010, Degner, et al, 2012). On the other hand, epigenetic changes in the lifetime of an individual, and sometimes spanning decades and generations, can lead to reproducible effects on metabolism and health even without underlying genomic alterations. Understanding how genome sequence and chromatin act jointly to specify the dynamic landscape of active and repressive regulatory elements across individuals and cell types will be needed to decipher the regulatory, molecular, and organismal phenotypes that underlie human disease in the context of genetic and epigenomic variation.

Methods

Inferring Chromatin States

The chromatin state model, is the 25-state ChromHMM (Ernst, et al, 2012) model described in (Hoffman, et al, 2013). Details on the data processing and model learning can be found in that paper.

Computing Regulator Binding Enrichments

For binding peak calls we used the standardized peak calls produced by the ENCODE consortium (Gerstein, et al, 2012, Kundaje et al., in preparation) using the SPP peak caller (Kharchenko, et al, 2008). To compute the enrichment for a peak call-data set in a specific chromatin state and cell type, \( s \), we computed the enrichment for transcription factor binding as \( \frac{a_s}{b} / \frac{c_s}{d} \) where \( a_s \) is the total number of bases in a peak call in \( s \), \( b \) the total number of bases in a peak call, \( c_s \) the total number of bases in \( s \), and \( d \) the total number of bases for which the segmentation was defined.

Clustering of Cell-Type Matched Enrichments

We clustered a single vector of enrichment values for each regulator with matched cell type chromatin data excluding CTCF and POLR2A. If multiple experiments were available for the regulator we averaged them using the median. If there was an even number of enrichment values we used the geometric mean to average the two middle values. To cluster the experiments based on cell type matched enrichments of chromatin state and binding we used the kmeans function in Matlab. We used the correlation distance function, 10 random restarts, and singleton as the empty action. We tried between 2 and 20 clusters and focused on 12 clusters since the patterns were largely homogenous and peaked for enrichment in protein-protein interactions (Supplementary Figure 4).
**Ordering Rows of the Dynamic Enrichment Heatmap**

We ordered the rows of the dynamic enrichment heatmap to minimize the distance between rows. The distance between two rows was defined using the distance metric $\sqrt{1-\text{corr}(x,y)}$, where $x$ and $y$ are the vectors of the 150 fold enrichment values for two experiments. Finding such an ordering can be made trivially equivalent to the computational traveling salesman optimization problem (Biedl, et al, 2001, Applegate, et al, 2006) of finding a minimum cycle that visits every city once by adding a dummy city with zero distance to every other city (Supplementary Figure 24). We applied a specially designed optimal traveling salesman problem solver, Concorde (http://www.tsp.gatech.edu/concorde.htm) (Applegate, et al, 2006), which despite the general problem being NP-hard was able to find an optimal solution to our instance of the problem in less than a minute using a single CPU. When forming the rows at the matrix at the regulator level we averaged enrichments for experiments on the same regulator in the cell type using the median and treating the various GM cell types as the same.

**Motif Analysis**

The motif instance were obtained from http://compbio.mit.edu/encode-motifs (ENCODE Project Consortium, et al, 2012, Kheradpour et al, in preparation). The motif instance enrichments were computed relative to a set of selected permuted control motif instances based on the approach described in (Kheradpour, et al, 2007) without using motif conservation. If a regulator had multiple known motifs associated with it, the motif which had the greatest chromatin state fold enrichment in any cell type was selected and used consistently in the analysis. The p-value significance of motif instance overlap with a chromatin state was computed using a binomial distribution. The null probability of a motif falling into a specific chromatin state was the ratio of the number of control motif instances overlapping the state to the total number of control motif instances, denoted $f_s$. For determining the significance after first conditioning on a peak call, there generally were too few control motif instances overlapping peak calls to obtain robust null estimates based on the frequency of control motifs within peaks. Instead to determine the frequency, the assumption was made that conditioned on a chromatin state the random chance expectation of a motif within a peak or outside a peak was uniform after controlling for the number of bases considered in the motif scanning. The null probability of motifs in peaks for a state $s$ was computed by first computing the ratio of the number of bases considered in motif scanning overlapping a peak call within state $s$ to the total number of such bases overlapping a peak call, denoted by $p_s$. This was then adjusted by $f_s$ and the total fraction of bases included in the motif scanning falling in the state, denoted by $e_s$, using the formula:

$$\frac{p_s f_s}{\sum_i p_i f_i / e_i}$$
The summation is over all states. For computing both the motif enrichments in chromatin states and conditioned on peaks, a p-value significance of 0.001 was used for testing separately enrichment and depletion. The cell type of the chromatin state corresponds to the cell type in which the regulator was profiled. If multiple experiments on the same regulator were conducted, then each experiment was counted inversely proportional to the number of experiments conducted on the transcription factor.

The aggregate motif fold enrichment for a chromatin state was computed as $\frac{a_s}{b} / \frac{c_s}{d}$ with $a_s$ the total number of motif bases in state $s$, $b$ the total number of motif bases, $c_s$ the total number of control motif bases in state $s$, and $d$ the total number of control bases. The enrichment conditioned on a peak call was computed the same way except restricting $a_s$, $b$, $c$, and $d$ to only bases that fell within a peak call. For computing the cluster motif enrichments each motif was counted once for each cell type in which there was a corresponding regulator experiment assigned to the cluster. The geometric mean of the enrichments after adding a pseudocount of one was used. For the motif usage in different classes of bound regions the median fold enrichment relative to control motifs across regulators was reported excluding CTCF from the analysis.

We compared the motif usage within peaks within states 1_Tss, 5_Enh, and 25_Art to the other 22 states restricted to HOT regions defined by (Yip, et al, 2012) in Gm12878, H1-hESC, HeLa-S3, HepG2, and K562. For each experiment in one of these five cell types corresponding to a regulator with a motif defined we computed the proportion of bases included in motif scanning that fell within peaks within HOT regions in states 1_Tss, 5_Enh, and 25_Art to all bases included in the motif scanning that fell within peaks within HOT regions. We then determined based on a binomial test using this proportion at a p<0.01 if there was a significant number of motifs in states 1_Tss, 5_Enh, and 25_Art or the other 22 states out of the total number of motifs that fell within peaks within HOT regions. If multiple experiments were conducted for the same regulator each was counted inversely to the number of experiments conducted on it.

**Pairwise Enrichment Calculations**

We computed the raw pairwise enrichments based on peak overlap at the base level. Let $a$ and $b$ be the number of bases in peaks for transcription factors A and B respectively. Let $c$ be the number of bases in their intersection. Let $d$ be the size of the genome and $\varepsilon$ be a pseudocount of 100 bases for smoothing. In this case the pairwise enrichment is $\frac{c + \varepsilon}{(a \times b) / d + \varepsilon}$. To compute the conditional pairwise enrichments let $a_s$ and $b_s$ denote the number of bases in peaks for transcription factor A and B respectively in state $s$, and $d_s$ the total number of bases in state $s$, then the conditional enrichment is $\frac{c + \varepsilon}{(\sum_s (a_s \times b_s) / d_s) + \varepsilon}$. When computing pairwise enrichments between pairs of regulators we averaged using the median all pairs of experiments involving both of the regulators. To compute the conditional enrichments based on DNase data we used peak calls from one replicate of the University of Washington DNase data (Thurman, et al, 2012). The same formulas above applied, where
when conditioning on DNase it was effectively equivalent to a two state model of being in a DNase peak or not. When also including state information with DNase the same formulas also applied, but the number of states effectively doubled by partitioning each state into the portion overlapping a DNase peak or not. The ordering of the heatmaps was determined using the traveling salesman formulation as was also used for ordering the dynamic enrichment matrix.

**Protein-protein interaction data**

The protein-protein interactions used were the Biogrid version 3.1.87 physical interaction data sets (Stark, et al, 2011). Fold enrichments were computed two ways. Let $n$ be the total number of known protein-protein interactions that fall into a cluster, $N$ the total number of known protein-protein interactions involving regulators considered, $K$ the number of clusters, $r$ the number of regulator pairs that are in the same cluster, and $R$ the total number of pairs of regulators. Under the uniform expectation enrichment was $K(n/N)$, and when conditioned on the cluster size it is $(n/N)/(r/R)$.

**Dynamic Binding Analysis**

For a given cell type, nucleotides of ‘unique’, ‘shared’, and ‘excluded’ binding for a regulator were defined if they were profiled in two or more of the Gm12878, K562, HepG2, HeLa-S3, H1-hESC, and Huvec cell types. If multiple experiments were conducted on the same regulator in a cell type they were first combined by taking the union.

**5C-Interaction Analysis**

We used 5C data available on the ENCODE pilot 1% regions (Sanyal, et al, 2012) to investigate if there was a preferential enrichment for sites with a transcription factor peak without a motif to interact with locations with a peak containing a motif. We restricted our analysis to transcription factors which had peaks in one of the four cell types with 5C data available: Gm12878, K562, HeLa-S3, and H1-hESC. We separately considered peaks without motifs on the forward primers (not specific to TSS) interacting with peaks containing motifs on the reverse primers (covering TSS) and vice versa. Only regulators with a minimum of ten possible detected interactions were evaluated. When evaluating the significance of interaction of forward sites without motifs with reverse sites with motifs, we compared the count of observed interactions to the number of interactions when randomizing the interacting reverse primer for a 1000 randomizations. When randomizing the reverse primers they were required to be selected from the same regions as the original interactions, thus to only generate randomized interacting pairs that could have been observed in the real data. A similar analysis was conducted for bound sites without a motif on the reverse primers to interact with forward primers.

**Acknowledgments**
We thank Pouya Kheradpour for assistances with the motifs and Bradley Bernstein for useful discussions related to this work. We thank Anshul Kundaje for data processing; Michael Hoffman, Bill Noble and Ross Hardison for discussions on chromatin states; the ENCODE consortium for sharing datasets. We acknowledge funding for this work provided by NIH grants R01HG004037 and RC1HG005334, NSF Fellowship 0905968 to JE, and the David Geffen School of Medicine at UCLA.
**Figure Legends**

**Figure 1:** Regulator enrichments for each chromatin state in matched cell types. Different regulators show distinct chromatin state preferences. For each regulator with matching chromatin data the average enrichment is shown for each chromatin state (columns). Enrichments have been row-normalized, scaling by the largest enrichment value for each experiment. K-means clustering with 12 clusters produced the clusters labeled C1-C12.

**Figure 2:** Dynamics of regulator enrichment in different chromatin states across cell types. Each row corresponds to one regulator in a given cell type. Each column corresponds to a state-cell type combination. The columns are organized first by state and then by cell type in the order: Gm12878, H1-hESC, HeLa-S3, HepG2, HUVEC, and then K562. The rows have been automatically ordered computationally using a traveling salesman problem instance solver, and reveal both regulator and cell type groups. The fold enrichments have been row-normalized scaled to the maximum enrichment in the row. In the six columns of each group, yellow indicates higher enrichment values and blue lower enrichment values. The next to last column indicates the cell type of the experiment color-coded with all GM cell types colored the same and all other non-tier 1 and 2 cell types colored white. The last column indicates the regulators of the experiments listed consecutively within the same cell type block.

**Figure 3:** Motif enrichment and depletion variation across chromatin states. a. Number of transcription factors with significantly enriched or depleted motif instances in each state at a p-value of 0.001 (see Methods). The maximum value for the y-axis was 79, corresponding to the number of transcription factors considered with regulatory motif instances available. If a transcription factor was profiled multiple times each experiment was counted inversely proportional to the number of times it was profiled. Stars indicate if the number of transcription factors with significantly enriched (depleted) motifs in a state is significant based on a binomial distribution with the number of samples equal to the total number of significant enrichments or depletions in the state and the probability of success equal to the proportion of significant enrichments (depletions) out of all significant enrichments or depletions across all states. Fractional values were first rounded to the nearest integer for the calculation. The p-value cutoff for triple stars was $10^{-6}$ and double stars was 0.01. b. Number of transcription factors with significantly enriched or depleted motif instances in each state conditioning on regions falling within a peak. Stars were computed the same way as in a except a 0.05 p-value cut-off was used for double stars. c. Fold enrichment for CEBPB motifs in four different HepG2 chromatin states. d. Fold enrichments for the motifs within peaks in the same four states relative to the baseline motif enrichment in peaks.

**Figure 4:** Dynamic Binding Enrichments. a. The enrichment in four HepG2 chromatin states for locations of the genome that are bound by CEBPB in HepG2 and another cell type, only in HepG2, and in another cell type but not HepG2. b. The median enrichment over all regulators in Gm12878 for the
three different classes of dynamic binding (for other cell types see Supplementary Figures 19-20) c. The median enrichment of regulatory motifs in bound regions for the different classes of dynamic binding.

Figure 5: Pairwise regulator co-binding enrichments are captured by chromatin state preferences. a. Pairwise regulator co-binding enrichment for all pairs of regulators in HepG2 show strong groups of co-binding. The regulators in each group are typically assigned to the same set of chromatin state cluster preferences from Figure 1. Enrichment levels up to 500 fold are found for the full pairwise enrichment table. Rows of the table have been ordered to maximize correlation of neighboring rows. Black lines correspond to groups of highly enriched pairs of regulators that emerge from this ordering. b. After conditioning on the chromatin state preferences for each regulator (see Methods), the pairwise regulator enrichments are dramatically reduced. c. and d. The same as a and b except for K562. Other cell types can be found in Supplementary Figure 21.
### Tables

#### Table 1: Different Regulators in the Same Cell Type Ordered Consecutively. The first column indicates the position in the order from Figure 2, the second column the cell type, and the third column the factors. The different individual GM lines are not differentiated here.

<table>
<thead>
<tr>
<th>IDs</th>
<th>Regulators</th>
<th>Fraction</th>
<th>Cell Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>31-34</td>
<td>NRC2</td>
<td>2/4</td>
<td>HeLa-S3, HepG2</td>
</tr>
<tr>
<td>40-41</td>
<td>NR1</td>
<td>2/5</td>
<td>H3K27me3, GM</td>
</tr>
<tr>
<td>41-40</td>
<td>NR1</td>
<td>2/4</td>
<td>H3K27me3, GM</td>
</tr>
<tr>
<td>51-52</td>
<td>NR1</td>
<td>2/5</td>
<td>H3K27me3, HepG2</td>
</tr>
<tr>
<td>125-128</td>
<td>RAD21</td>
<td>4/6</td>
<td>HeLa-S3, HepG2, SK-B1, RA, H-895</td>
</tr>
<tr>
<td>246-249</td>
<td>SIRT2</td>
<td>2/2</td>
<td>B720-0, B720-1</td>
</tr>
<tr>
<td>250-257</td>
<td>B1-20</td>
<td>8/8</td>
<td>B720-0, B720-1, GM, HeLa-S3, SK-B1, RA, H-895</td>
</tr>
<tr>
<td>258-261</td>
<td>USF1</td>
<td>2/6</td>
<td>A549, SK-B1, RA</td>
</tr>
<tr>
<td>266-269</td>
<td>YWH</td>
<td>2/2</td>
<td>B720-0, B720-1, GM</td>
</tr>
<tr>
<td>269-270</td>
<td>ZEB1</td>
<td>2/2</td>
<td>B720-0, B720-1, GM</td>
</tr>
</tbody>
</table>

#### Table 2: Different Cell Types with the Same Regulator Ordered Consecutively. The first column indicates the position in the order from Figure 2, the second the regulator, the third the fraction of cell types for the regulator in this consecutive group of cell types, and the fourth the cell types represented in this consecutive group. GM is used to denote any of the GM cell lines.

<table>
<thead>
<tr>
<th>IDs</th>
<th>Regulator</th>
<th>Fraction</th>
<th>Cell Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-21</td>
<td>JUNB, BCL11A, BATE, 5SFL, BFB1, BFB2, BFB3, STAT1, STAT3, FOS, PDGFD, TCF7, PA63, POL2, BAX, MAX, SFR, SP1, TBP, CHD2, RF5X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22-37</td>
<td>JNK1, BCL2, FOS, TAF, JG2T2B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29-43</td>
<td>HLA-A3, TAF, FOS, E2F2, E2F3, BCL3, TCF12, PAX5, MAX, POU5F1, TAF1, BCL11A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44-50</td>
<td>FOS, JUNB, BCL11A, BATE, 5SFL, BFB1, BFB2, BFB3, STAT1, STAT3, FOS, PDGFD, TCF7, PA63, POL2, BAX, MAX, SFR, SP1, TBP, CHD2, RF5X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>51-64</td>
<td>HIF1, BCL11A, BATE, 5SFL, BFB1, BFB2, BFB3, STAT1, STAT3, FOS, PDGFD, TCF7, PA63, POL2, BAX, MAX, SFR, SP1, TBP, CHD2, RF5X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>65-71</td>
<td>GABPA, E2F4, TAF1, ELK4, NR2C2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>72-78</td>
<td>GABPA, E2F4, TAF1, ELK4, NR2C2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>79-88</td>
<td>NRF1, FOXL1, POLR3A, BRF1, GTF2F1, TAF7, TAF1, SRF, USF2, USF1, HSF1, BHLHE40, PPARC1A, ESRRA, HDAC2, SP1, TCF12, RXRA, HNF4A, HNF4G, EP300, TCF7L2, FOXA2, CEBPB, FOSL2, JUND, JUN, MAF, MAFK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>89-90</td>
<td>NRF1, FOXL1, POLR3A, BRF1, GTF2F1, TAF7, TAF1, SRF, USF2, USF1, HSF1, BHLHE40, PPARC1A, ESRRA, HDAC2, SP1, TCF12, RXRA, HNF4A, HNF4G, EP300, TCF7L2, FOXA2, CEBPB, FOSL2, JUND, JUN, MAF, MAFK</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References


### Regulator enrichment in chromatin state (matched cell types, row normalized)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>Art.</th>
</tr>
</thead>
<tbody>
<tr>
<td>regulators</td>
<td>ATFa</td>
<td>BCLAF1</td>
<td>BRCA1</td>
<td>CHD2</td>
<td>E2F1</td>
<td>E2F4</td>
<td>ELK4</td>
<td>ETS1</td>
<td>GABPA</td>
<td>GTF2B</td>
<td>GTF2F1</td>
<td>IRF3</td>
<td>MXI1</td>
<td>NR2C2</td>
<td>NR3C1</td>
<td>NRF1</td>
<td>POLR3G</td>
<td>RDBP</td>
<td>SIN3A</td>
<td>SIX5</td>
<td>SP2</td>
<td>SREBF1</td>
<td>TAF1</td>
<td>TAF7</td>
<td>TBP</td>
<td>THAP1</td>
</tr>
</tbody>
</table>
CEBPB Binding Enrichments

Median Regulator Enrichment in GM12878

Motif usage in different classes of bound nucleotides
Interplay between chromatin state, regulator binding, and regulatory motifs in six human cell types

Jason Ernst and Manolis Kellis

*Genome Res.* published online April 17, 2013
Access the most recent version at doi:10.1101/gr.144840.112

---

**Supplemental Material**

http://genome.cshlp.org/content/suppl/2013/05/03/gr.144840.112.DC1

**Published online April 17, 2013 in advance of the print journal.**

**Peer-reviewed and accepted for publication but not copyedited or typeset; accepted manuscript is likely to differ from the final, published version.**

**Freely available online through the *Genome Research* Open Access option.**

**This manuscript is Open Access. This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.**

**Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.**

---

Advance online articles have been peer reviewed and accepted for publication but have not yet appeared in the paper journal (edited, typeset versions may be posted when available prior to final publication). Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To subscribe to *Genome Research* go to:

http://genome.cshlp.org/subscriptions

© 2013, Published by Cold Spring Harbor Laboratory Press