ABSTRACT

Title of dissertation: *CROWDSOURCING*: A NOVEL GROUP-LEVEL MECHANISM STRUCTURES CHROMATIN AND FOSTERS GENE-COMPLEX ACTIVATION

Justin Lewis Malin, Doctor of Philosophy, 2015

Dissertation directed by: Professor Sridhar Hannenhalli
Department of Cell Biology and Molecular Genetics

Transcriptional regulation of a co-expressed gene network often relies on adoption of a three-dimensional conformation, dubbed a 'chromatin hub' or 'regulatory archipelago', which radically reduces spatial distances between genomically remote enhancers and gene targets, as well as among enhancers. While the advantage of spatial proximity for fostering pairwise interactions is self-evident, there has been limited exploration within archipelagos of higher-order interactions. Here we probe the evidence for a novel and group-level mechanism which, we hypothesize, is emergent when numerous coordinately-acting regulatory enhancers, mediated by chromatin, converge in space. Based on functional human genomic data and biophysical modeling, and using a set of 40 enhancer archipelagos we identified through shared activity across 37 tissues, we show that three-dimensional juxtaposition of dozens of genomically dispersed binding sites for a given transcription factor (TF) can briefly 'trap' diffusing TF proteins, eliciting a spike in local TF concentration and a two-fold boost in its DNA occupancy at member enhancers. We find substantial evidence for the role of this 'crowdsourcing' effect in tissue-specific gene-

complex activation, and in the process, offer the first evidence for a predictable group-level modulator of TF occupancy that operates independently of genomic distance. In turn, crowd-sourcing proves a surprising answer to the paradoxical source of binding specificity for degenerate TFs, in general, and various master regulator TFs, in particular. Additionally, we show that crowdsourcing likely contributes to super-enhancer functionality and speculate on crowdsourcing's role in coordinating collectives of super-enhancers in cell lineage determination. Finally, we ask whether the biophysical impact of crowdsourcing also flows in the opposite direction. Here we find, likely mediated by elevated TF concentrations, that coordinately acting enhancers adopt a more compact conformation, stereotypical of activated gene complexes. Together, we find compelling evidence for a novel and pervasive regulatory mechanism that is emergent at the level of co-expressed gene module and which, both, mediates and is mediated by higher-order chromatin structure.

Crowdsourcing: a novel, group-level mechanism structures chromatin and fosters gene-complex activation

Justin Lewis Malin

Dissertation submitted to the Faculty of the Graduate School of the

University of Maryland, College Park, in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

2015

Advisory committee:

Professor Sridhar Hannenhalli, Chair

Professor Kan Cao

Professor Hector Corrada Bravo

Professor Steve Mount

Dr. Ivan Ovcharenko

Professor Michelle Girvan, Dean's Representative

Preface

When Watson and Crick's Central Dogma posited that life was inscribed in a 4-letter alphabet, it sparked a revolution that has since privileged an informatics and linear perspective in the study of the genome. The advent of high-throughput sequencing has done little to change this.

In the last decade, however, there has been a rapidly growing appetite in the community to look beyond sequence. While it has always been clear to some that a full understanding of transcriptional regulation required both structure-based and sequence-based insights, there have been technical challenges in integrating data from the two domains. With the introduction and rapid adoption of chromatin conformation capture (3C) techniques, however, it has become possible to not only probe chromatin's three-dimensional structure in unprecedented detail and scale, but to do so through the lens of sequence. Based on this technology several novel mechanisms have been elucidated though which the genome's topology during interphase is directly implicated in transcriptional regulation. It is in this young tradition that the current work is situated.

Dedication

To my mother and my father –

with love, admiration,

and immense gratitude

Acknowledgments

I am extremely grateful to my advisor Sridhar Hannenhalli for the opportunity to join his lab and its highly stimulating atmosphere, and for his generous financial support. I have grown into a researcher thanks in no small part to his dedication and his incisive and abundant feedback, and for the superb example he has provided of how to do research. I am also grateful to my committee members, all of whom have been gracious with their time, making helpful suggestions along the way. Thank you to my collaborators Daphne Ezer, Xiaoyan Ma, Hiren Karathia, and Seung Gu Park for their vital contributions. My work is undoubtedly better because of them. I am very appreciative of all my colleagues in the Hannenhalli lab, past and present, who have helped make the last four years special, intellectually stimulating, and even fun. They have each been helpful and supportive, but especially Avinash Das and Kun Wang, with whom I started in the Hannenhalli lab, and Seung Gu Park. Thank you also to Carl Kingsford and to my former lab mates in the Kingsford lab for introducing me to the fascination and power of computational biology.

In addition, I'd like to thank Howard Savage for his enduring friendship and support during this chapter of my life, and the chapter before that, along with his whole family, who have always made me feel like part of theirs. I am thankful to all my friends, most now scattered, though always with me. Thank you to my family for their unwavering support and many kindnesses over the years, with a special thank you to Randi and John Buergenthal for the countless gracious invitations, making holidays that much more special.

Finally, to my father and mother, I offer my most heartfelt gratitude for making this late-life plunge back into school so much smoother than it might have been, asking

for nothing while giving so much of themselves. I feel very lucky to have such wonderful parents, and am grateful for the many lessons I have learned from them along the way.

Table of contents

List of Figures

List of Tables

Crowdsourcing: the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group rather than from traditional suppliers.

--*redacted from Merriam Webster*

Chapter 1: **Introduction**

Non-coding regions

Recent work has exposed the tension between the pinpoint spatiotemporal control of cell fate determination and the far "leakier" elements of transcriptional regulation underlying it (Spitz and Furlong, 2012). But this noise should not obscure the deep causal link between development and transcriptional regulation. In mouse, by one estimate, variability in overall transcript abundance among individuals accounts for 60% of variation in protein abundance (Maier et al., 2009), the building blocks of cells, while more than 40% of variability in protein abundance among species can be explained by variability in orthologous transcript levels (Vogel and Marcotte, 2012). Eukaryotes have evolved complex mechanisms to modulate transcriptional output in response to integrated developmental and environmental cues (Adelman and Lis, 2012). Elucidating these regulatory mechanisms has been the focus of considerable attention since the first such mechanism was characterized more than 50 years ago, for the lactase-coding lac operon in *E. coli* (Jacob and Monod, 1961). Due largely to the relatively low proportion of candidate regulatory elements in non-coding regions, combined with a dearth of sequence data, this work, until recent decades, focused overwhelmingly on the promoter region immediately upstream of the transcription start site, where the enzyme polymerase II is initially recruited as a required step before elongation. Reinforcing this promoter-centric

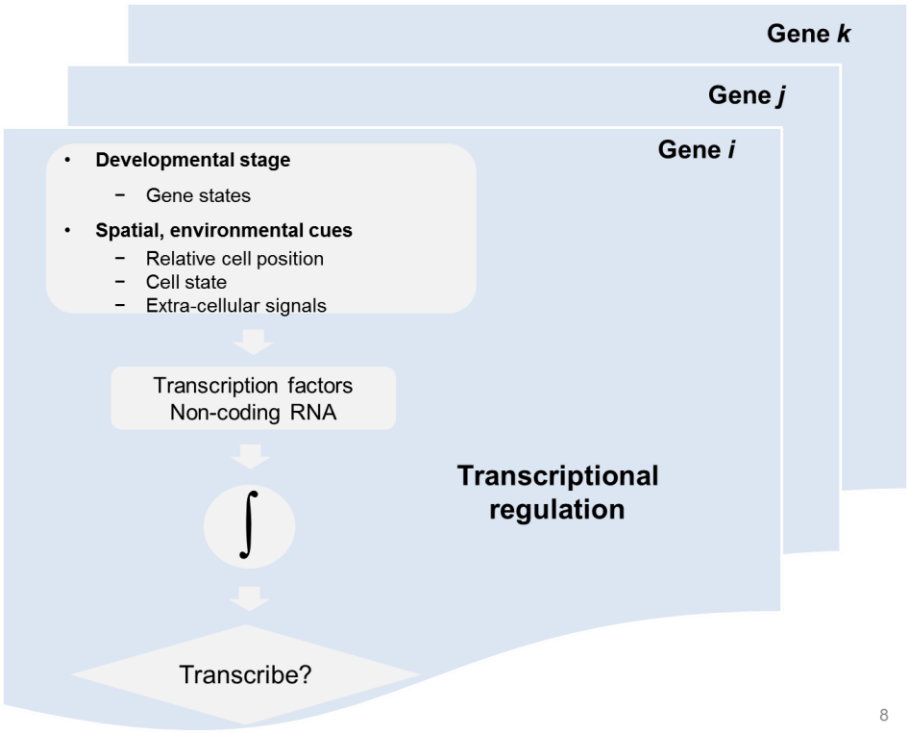
paradigm, arguably, was a prevailing informatic, one-dimensional view of the genome, ushered in with the discovery of DNA's secondary structure in 1953.

A key corrective to this promoter-centric viewpoint, thanks to greater availability of data, has been multi-species sequence alignment. By comparing observed conservation levels to levels predicted by a neutral model of mutation, non-coding regions covering 5 to 10 percent of the genome have been identified, many quite distant from any known gene, that are subject to high levels of purifying selection (Dermitzakis et al., 2005; Lindblad-Toh et al., 2011). Further strengthening the case for their *in vivo* functionality, a significant percentage of the conserved non-coding regions tested drive reporter gene expression (Li et al., 2010; Nobrega et al., 2003), while many others harbor disease-associated single nucleotide polymorphisms identified by genome-wide association studies (Cooper and Shendure, 2011). To be sure, the correlation between sequence conservation and functional conservation appears to be modest, with each, alternately, overly and insufficiently conservative in predicting the other, depending on tissue and other factors (Blow et al., 2010; Nelson and Wardle, 2013). The recent ENCODE project, however, challenged the existence of correlation altogether when, based on evidence of TF binding and transcription, they asserted that as much as 90% of the human genome was 'functional' (Bernstein et al., 2012), despite minimal conservation. The ensuing outcry ignited a conversation on how best to define 'functional' in a genomic context (Doolittle, 2013). Regardless of where the science settles, far richer data together with heightened awareness of distal regions' potential in a three-dimensional framework to regulate transcription has led to unprecedented attention on non-coding regions.

Trans-acting factors and binding

The recruitment in mammals of RNA polymerase to the promoter depends on a minimal set of six general DNA-binding proteins, or transcription factors – TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH (Orphanides et al., 1996). The remainder of the 1500 known TFs (Boyle et al., 2011) have a more conditional role in recruiting RNA polII: they encode, through their nuclear abundance, the environmental and developmental state of the cell and its milieu (Levine, 2010). These signals are then translated by promoters and distal cis-regulatory regions into output that modulates polII recruitment or elongation (Figure 1-1). For example, estrogen receptor ER-alpha, a model steroid receptor TF, is expressed in a number of tissues where it recognizes estrogen. Upon binding in the

Figure 1-1. Flow of environmental and development state information through transcription factors in mediating gene transcription.



cytoplasm, ER-alpha relocates to the nucleus where, together with co-activators, it binds DNA as a dimer, fostering up- or down-regulation of dozens of genes (Moggs and Orphanides, 2001).

Transcription factors are extraordinarily evolutionarily conserved, particularly their binding domains, with relatively few TFs novel in human alone (Neph et al., 2012a; Stergachis et al., 2014; Stewart et al., 2012). TFs are organized into families on the basis of shared DNA-binding domain morphology, which can often be traced to ancient duplication events of their encoding genes. The family of homeotic (HOX) TFs, which together guide selection among alternative pathways in elaborating the body plan, is a classic example whose 60-nucleotide homeodomain exhibits particularly deep homology. This phylogenetic conservation was substantiated in a well-known experiment wherein fly HOX-gene null mutants were rescued by insertion of the orthologous coding region extracted from chicken (Lutz et al., 1996).

Not surprisingly, TFs in the same family recognize similar binding sites. Decades of exploring transcription factor binding domain preferences have revealed that a domain recognizes DNA through a combination of a sequence's 'base readout' and its 'shape readout.' Base readout dictates the formation of hydrogen bonds and hydrophobic attractions between TF amino acid side chains and the edges of a given base pair. Shape readout, conversely, arises from interaction among base pairs, and the higher-order 3-D structure (e.g. location in major vs. minor groove) (Slattery et al., 2014).

There have been dozens of computational models published for predicting TF binding preference, many of which eschew mechanistic complexity in favor of empiricism (Hannenhalli, 2008). The most prevalent model, the position weight matrix, is also likely the simplest (Stormo and Zhao, 2010): after clustering sequences linked to observed binding events (from microarray, ChIP-Seq or, more recently, ChIP-exo), sequences in a cluster are aligned and a probability derived at each position for each nucleotide. Despite the oversimplifications of this model – for example, it does not explicitly account for the likely interdependence between neighboring positions – the position weight matrices (PWMs), or 'motifs', predict with reasonable accuracy in vitro binding of isolated TFs to naked DNA (Stormo and Zhao, 2010) .

Due to a complex binding landscape, however, which includes factors such as dependency on cooperative binding, predictive accuracy of PWMs *in vivo* is much lower (Hannenhalli, 2008; Yáñez-Cuna et al., 2012). Nonetheless, as their sensitivity *in vivo* is much higher than their specificity (D'haeseleer, 2006), motifs can be used with a threshold for match quality as an initial screen for putative binding sites (Levy and Hannenhalli, 2002). As such, TFs are modeled as interacting with only a discrete set of sites. Model predictions can then be compared with *in vivo* TF binding that has been observed directly. In a high-throughput environment, this often means evidence from a ChIP-Seq assay, in which DNA-protein complexes are cross-linked, sheared, then retrieved by chromatin immunoprecipitation with an appropriate antibody, and the resulting library of bound DNA sequenced. 'Peaks' of overlapping sequence reads that rise significantly above background level signal a DNA-bound protein (Furey, 2012).

Histone modifications & chromatin accessibility

TFs compete against not only other TFs for DNA recognition sites, but against histone proteins (Levo and Segal, 2014). Ubiquitous histone octamers provide the genome its lowest-order of organization – 157 bp of DNA is spooled around each complex at semi-regular intervals – while also providing a way for TF accessibility to the DNA to be regulated through their dynamic displacement (Lelli et al., 2012). Histones – primarily lysine residues in the H3 member of the histone octamer – are often adorned with a number of biochemical modifications such as methylation, acetylation, and ubiquitination, which further impact the local chromatin's accessibility to TF binding. There has been great interest in interpreting this 'histone code', and to date, there have been partial successes, for example marks or combinations of marks have been associated with inactive heterochromatin, active enhancers, and an enhancer state 'poised' between active and inactive' (Jaenisch and Bird, 2003). Importantly, though, histone modifications are not root causal agents but are deposited and removed by chromatin modifying enzymes that have been recruited by bound TFs, RNA polymerase, or other proteins. Hence their presence, and interpretability, is inevitably noisy (Henikoff and Shilatifard, 2011; Wang et al., 2011).

Fortunately, chromatin's overall accessibility can be measured directly. The most widely used technique for this is the DNase-hypersensitivity assay, in which DNA is subjected to DNase I enzyme and the cleaved fragment ends aligned to the genome. Based on enrichment for cleaved ends, Dnase hypersensitive (DHS) regions are identified. Ranging in length from a few hundred to a few thousand base pairs, such DHS regions are

stereotypically de-condensed and transcriptionally active euchromatin, often featuring TF binding.

By substantially increasing the DHS assay's depth of coverage, it becomes tractable to resolve 'footprints' of individual TF binding events. By carefully matching these single-base resolution footprints to independently identified motif instances for a set of TFs, individual TF binding events can be estimated (Neph et al., 2012a). While ChIP-based methods remain the gold standard for identifying bound TFs (Adli and Bernstein, 2011), digital footprinting is an excellent complement, as it can simultaneously estimate binding by hundreds of distinct TFs with a known motif, all in a single experiment per tissue.

Enhancers (distal cis-regulatory modules)

During development and beyond, gene networks interpret cellular and developmental state through the combinatorial interactions between TFs and DNA. Less than 10 percent of bound TFs, it turns out, are found in promoter regions (Neph et al., 2012a). The vast majority, and a large share of the imputed regulation, instead, falls to cis-regulatory modules (CRMs) that are distal to their target gene, each harboring dozens to several hundred putative binding sites stretching over 100 to several thousand base pairs (Yáñez-Cuna et al., 2013). Distal CRMs can be classified by function –insulators, tethering elements, enhancers, and silencers (Spitz and Furlong, 2012). Following common practice, I will refer collectively to distal CRMs that act as enhancers or silencers of target gene expression as 'enhancers.'

Enhancers were discovered in 1981 and earned their name from the observation that a cloned beta-globin gene expressed 200-fold more transcripts when it was accompanied

by a 72bp sequence that, in its native virus, ‘enhanced’ expression (Banerji et al., 1981). Most animal genes are thought to interact with at least one such enhancer region. This dependence is, up to a point, distance and orientation independent with respect to the promoter (Bulger and Groudine, 2011). When removed, transcription stereotypically drops to a basal level far below wild type levels, but does not cease (Bulger and Groudine, 2011).

Identifying enhancers

Interestingly, enhancers bear much in common with promoters (Andersson et al., 2015). It has been argued that, as regulatory complexity increased in metazoa, enhancers were an evolutionary response to the increasing inadequacy of the limited ‘real estate’ available next to the transcription start site. This is consistent with the observations that genes that are highly responsive to variation in spatiotemporal cues interact with enhancers far more often than do housekeeping genes, with their relatively constitutive expression (Pan et al., 2010a).

Because enhancers are not constrained to flank their target gene, identifying them has been more challenging than identifying promoters. Complicating the task further, in stark contrast to the high conservation exhibited by TFs and their binding motifs, enhancer sequences have experienced massive turnover in mammals over the last 100 million years (Stergachis et al., 2014). This is deceptive, though. Even when an enhancer is functionally conserved across phyla – active in the same embryonic domain and responsive to the same TFs – the motif arrangement, or grammar, may not be conserved (Junion et al., 2012). This suggests a model of enhancer activity in which bound TFs interact with the transcription apparatus somewhat independently of one another (i.e., the

'billboard' model) (Yáñez-Cuna et al., 2013). In a related scenario (the 'TF collective' model), there is tight interdependence among a cohort of TFs, however protein-protein interactions among them bestow flexibility as to which cohort members bind the enhancer directly and which TFs merely bind a cohort member (Junion et al., 2012).

Currently, the most-used high-throughput approaches to detect enhancers involve 1. searching for accessible chromatin (Sheffield et al., 2013); 2. ChIP-Seq identification of indirectly enhancer-bound co-activator P300 (Visel et al., 2009); and 3. unsupervised machine learning based on a set of chromatin features that include 1. and 2., followed by identification of the non-coding cluster(s) with high neighbor gene transcription (Ernst and Kellis, 2012). The current gold standard for verifying an enhancer calls for transfecting into an embryo a putative enhancer sequence fused to a reporter gene with a minimal promoter (Nelson and Wardle, 2013; Pennacchio et al., 2006). Of course, no method is without weaknesses. Where the computational evidence is noisy and somewhat indirect, transgenic approaches fail to account for an enhancer's *in vivo* context. But in combination, where computational identification is followed by transgenic validation, the individual weaknesses are significantly mitigated (Nelson and Wardle, 2013). Even in the face of positive enhancer identification, however, part of an enhancer may remain obscured. For example, a minimal enhancer can be defined for the *eve* gene in fly capable of recapitulating expression patterns seen for the endogenous gene. But without an additional 1-2Kb of flanking region, the enhancer will not recapitulate stereotypical robustness to typical temperature fluctuations (Ludwig et al., 2011). This points to a long-distance regulatory landscape that may be even more widespread than often assumed,

with numerous enhancer elements contributing either modularly or continuously to refining gene expression (Spitz and Furlong, 2012).

Enhancer–promoter interaction

Via selective TF recognition, enhancers effectively integrate, and integrate over, the myriad state values of the cell, and interpret them for their target gene(s) (Slattery et al., 2014). This enhancers do by increasing the presence of select TFs in the vicinity of the targeted promoter (Pombo and Dillon, 2015). While a detailed mechanistic understanding of enhancer function and enhancer-promoter interaction are still lacking, clues have emerged, such as the aforementioned independence of distance and orientation, and the oft-seen flexibility in motif grammar. Primarily, two non-mutually exclusive models are invoked to account for how enhancers and gene promoters bridge their intervening distance to interact:

1. promoter tracking, in which TFs are first recruited to an enhancer and then slide along the chromatin until reaching the proximal promoter (Hatzis and Talianidis, 2002), consistent with the facilitated diffusion model of TF dynamics (Wunderlich and Mirny, 2008).

2. chromatin looping, in which the intervening chromatin is looped out and the enhancer comes into physical proximity of the promoter (Ptashne, 1986)

The presence of looping, in particular, has been substantiated by numerous reports (Ptashne, 1986) (Vakoc et al., 2005) (Deng et al., 2012) (Tolhuis et al., 2002), with significant insight into mechanistic details (Song et al., 2010) (Kagey et al., 2010a). Ultimately, which of the two mechanisms underlies communication between an enhancer

and promoter may be a function of such factors as the genomic distance between the pair. Of the two mechanisms, chromatin looping alone is associated with higher-order chromatin topology – critical for this work – hence, I will not discuss tracking further.

Chromatin looping between an enhancer and promoter appears, ultimately, to be an instance of a more generalized phenomenon. Loops also form between an enhancer and an insulator as a way to pre-empt enhancer regulatory activity (Wallace and Felsenfeld, 2007). Moreover, insulator-mediated looping, has been shown to contribute to the higher order structuring of the genome, imposing modularity on genomic interaction through the creation of 'topologically associated domains' averaging 1-3 Mb each (Dixon et al., 2012) (Junier et al., 2010) (Filippova et al., 2013).

The single known insulator protein in mammals, CTCF, has been intimately linked to loop formation through its recruitment of cohesin which, together with Mediator complex are largely responsible for forming and stabilizing chromatin loops in metazoans (Kagey et al., 2010a; Seitan et al., 2013). This includes loops that are highly tissue-specific as well as loops that appear to be retained across multiple cell types (DeMare et al., 2013). Recent work has also highlighted the recruitment of cohesin in the absence of CTCF (Schmidt et al., 2010).

Determining higher order chromatin structure using Hi-C

These and other recent findings highlighting 3-dimensional chromatin structure have been largely enabled by the introduction of chromatin conformation capture techniques by (Dekker et al., 2002), with a high-throughput version(s) coming out a few years later (Dostie et al., 2006). These techniques use formaldehyde-mediated cross-linking to

identify contact between genomic loci. After restriction enzyme digestion and extraction of the cross-linked chromatin fragments, fragments are ligated, and ligation products that align to non-contiguous regions – marking a putative interaction – tallied. By measuring the population-normalized frequencies of such interaction pairs, a global view emerges of genome-wide interaction. The disadvantage of this technology is its low sensitivity to interactions that are short-lived or highly specialized, as interactions must be detectable when averaged over tens of thousands of cells (Mercer and Mattick, 2013). Nonetheless, this technique has been instrumental in confirming and identifying *de novo* stable contacts between pairs of loci, ranging from contacts between the model beta-globin locus control region and gene promoter 60kb downstream (Vakoc et al., 2005), to the dense skein of interactions characteristic among super-enhancer sub-components (Heinz et al., 2015).

Higher order coordination and regulatory archipelagos

As confirmed by Hi-C, enhancers often interact with more than one gene, while genes typically receive input from multiple enhancers (Sanyal et al., 2012). This is a symptom of a deeper truth: across all clades of life, genes tend to be expressed as elements of larger interdependent networks. Early efforts to elucidate networks of coordinately-active genes applied cDNA micro-array data to identify genes with expression levels correlated across a series of conditions or cell types (Bar-Joseph et al., 2003). Genes within a co-expressed module, in turn, exhibited high functional coherence and their promoters recognized common ‘master-regulator’ TFs (Bar-Joseph et al., 2003). In a similar vein, co-active cardiac enhancers have been found to recognize similar TFs (Narlikar et al., 2010). This is broadly consistent with a coordinating role for an enhancer network that underlies a

gene network (Taher et al., 2013). But establishing a link between enhancers and target genes is non-trivial. Moreover, to date, no network-level analysis has been reported for enhancers, i.e. no regulatory analog to gene module analysis. We propose such an analysis, starting by identifying pairs and then clusters of enhancers with correlated activity across multiple tissues, where activity will be estimated by chromatin accessibility (DHS).

Regulatory archipelagos

There have been various recent reports, based on chromosomal conformation capture (4C, 5C, HiC, ChIA-PET) of chromatin looping combining at a higher-level of organization (Markenscoff-Papadimitriou et al., 2014; Montavon et al., 2011; Vernimmen, 2014). As a chromatin loop suggests a pairwise functional interaction, a network, or 'archipelago', of pairwise loops is suggestive of higher-order interactions. At the mouse *HoxD* gene cluster in digit cells, many-to-many interactions were observed between the *Hoxd* genes and enhancers in the flanking gene desert, in addition to interactions among enhancers (Montavon et al., 2011). In five human cell lines, in addition to enhancer-promoter and enhancer-enhancer interactions, abundant promoter-promoter interactions were detected (Li et al., 2012) (Zhang et al., 2013). Although, it is not possible to distinguish between transient, dynamic interactions and simultaneous, stable interactions, this does not substantively alter the functional interpretation of archipelagos. Together, the observed cis-cooperativity has been shown to be a source of regulatory buffering against environmentally-mediated fluctuations in TF abundance (Perry et al. 2010). The combinatorial actions of enhancers with shared but non-identical TF BS are also thought to further refine target gene expression more than a single

enhancer (Montavon and Duboule, 2012). Based largely on ‘super-enhancers’ – regulatory regions dense with constituent enhancers, and implicated in cell fate determination – an additional model posits that as the number of interacting enhancers increases, so too does aggregate enhancer output and total target gene expression. The mechanism underlying this last relationship, however, has not been identified (Andersson et al., 2015).

While spatial proximity is the norm for active and actively transcribing archipelagos, the same enhancers show substantially decreased proximity in embryonic domains and mature tissues where they are not active (Montavon et al., 2011; Schwarzer and Spitz, 2014; Spitz and Furlong, 2012). The spatial proximity thus appears to be conditional, however it is not clear what drives the enhancers to co-localize: while there have been intriguing insights into the roles of cohesin and Mediator in chromatin loop formation, these alone do not account for the coordinated loop formation that defines archipelagos.

Regulatory complexity and the challenge of specificity

The recent identification of regulatory archipelagos through chromatin capture techniques, in fact, mirrors two decades of experimental findings based on imaging that show the bulk of transcriptional activity occurs in discrete nuclear foci. Termed ‘transcription factories’, these subnuclear compartments concentrate polymerases and other transcriptional resources, and feature unusually high levels of RNA transcription (Chakalova and Fraser, 2010; Cook, 1995). Indeed, the presence of targeted chromatin looping and, at a higher organizational level, three-dimensional archipelagos and transcription factories appears to be a signal difference between eukaryotes and prokaryotes. Complete reliance on one-dimensional regulatory mechanisms such as the

bacterial operon is simply not compatible with the much higher combinatorial complexity characteristic of eukaryotic gene regulation (Daniel et al., 2014).

TF specificity

Regardless of organization, with greater complexity comes a larger genome (albeit the converse is not true (Pagel and Johnstone, 1992)) and with a larger genome, the increased challenge of specificity – ensuring the precise spatiotemporal targeting of regulatory actions. For a transcription factor, specificity encapsulates the unlikelihood an instance of the motif is found in a genome by chance. This is usually calculated as relative entropy, an information theoretic quantity that measures divergence between a motif's base frequencies and those in the genomic background (D'haeseleer, 2006). Notably, there is a strong positive relationship between a motif's specificity and its affinity for its best-matched sequence. Longer motifs containing more of the rarer, and double-hydrogen bond forming, guanine and cytosine tend to be more specific and bind more strongly, while shorter motifs containing more adenine and thymine tend to be more degenerate and bind more weakly.

Interestingly, BS for TFs with degenerate motifs and, hence, which are weakly binding, numerically dominate the regulatory landscape (He et al., 2012). A number of hypotheses have been advanced, including evolutionary expedience (He et al., 2012); a consequence of mutation-selection balance (Stewart and Plotkin, 2013); and greater compatibility of weak binding with transient or context-specific events (Spitz and Furlong, 2012). Notwithstanding the driving force, it presents a clear paradox which begs explanation – how do TFs with degenerate motifs distinguish between their bona fide sites and the many (10^3 to 10^5) other promiscuous, but energetically equivalent sites (Levine, 2010; Z

Wunderlich, 2009)? This challenge is limited mainly to metazoans. Binding motifs in bacteria and yeast have higher mean information content (higher relative entropy) and, more centrally, they are often sufficiently informative to specify a unique binding site given the vastly smaller genome (Stewart and Plotkin, 2013; Z Wunderlich, 2009).

It turns out that only a small fraction of metazoan recognition sites, particularly for degenerate motifs, functionally bind their cognate TF (Dror et al., 2015; Levine, 2010), despite relatively high nuclear abundances of TF proteins (Biggin, 2011). This suggests that there are other factors mediating *in vivo* recognition of a TF's functional target. Recent work has, indeed, highlighted several such factors:

1. *GC-content:* High GC content in the flanking sequence surrounding a putative site, or more generally, base composition in the flanking sequence that mirrors the base composition of the putative site greatly increase likelihood of occupancy (Dror et al., 2015; White et al., 2013).

2. Cooperative binding: Protein-protein interactions with a neighboring TF serve to stabilize binding and, hence, increase occupancy (Kazemian et al., 2013) (Slattery et al., 2014).

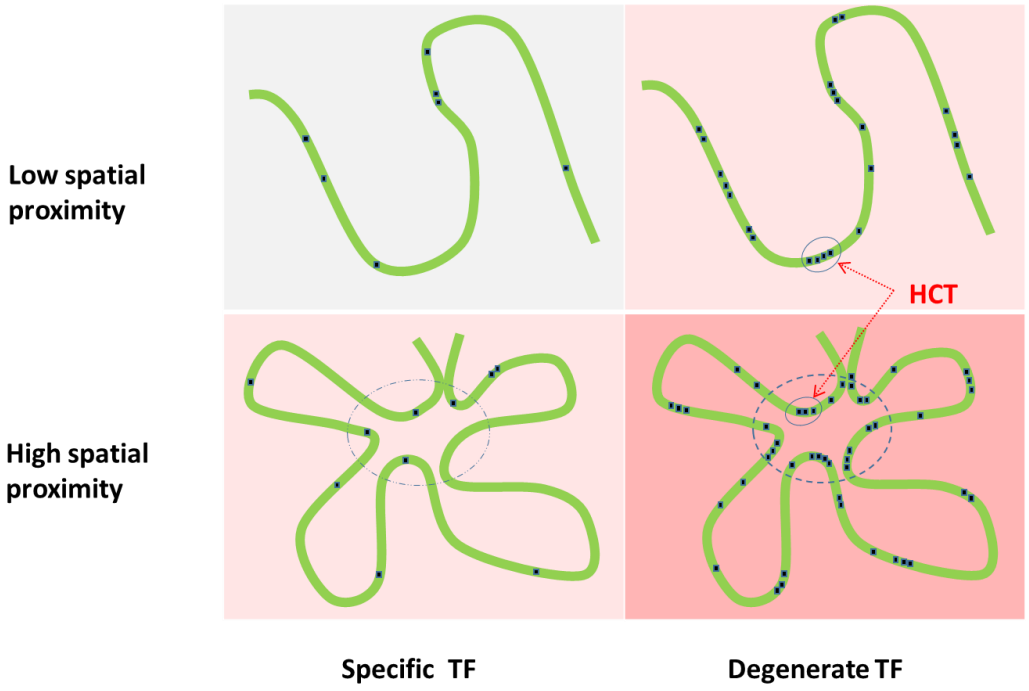
3. Homotypic clusters of BS: Genomic clusters of binding sites for the same TF, based on a facilitated diffusion model of TF dynamics, effectively trap a transcription factor into diffusing back and forth along the 1-D chromatin, thereby increasing both occupancy and the local TF concentration (Ezer et al. 2014; Brackley et al. 2012; Dror et al. 2015).

Cooperative binding and homotypic clusters of TF BS are particularly common among degenerate TFs. Notably, these additional features all reside in a binding site's genomic

flanking region. No study to date, however, has examined the effect on a binding site's occupancy of its *spatial* context.

Specifically, a homotypic cluster's impact on binding is governed by its binding site abundance in a limited genomic region (Brackley et al., 2012). The impact of binding site abundance in a limited nuclear space, such as a regulatory archipelago, is not known. Homotypic clusters are predominantly degenerate motifs (Dror et al., 2015), hence, the 'crowding' of BS expected from increased spatial proximity in an archipelago should accrue predominantly for degenerate BS (Figure 1-2).

Figure 1-2. Cartoon comparing predicted BS 'crowding' in a regulatory archipelago for specific TFs and degenerate TFs. HCT: homotypic cluster of TF BS



Organization of Thesis

In Chapter 2, using data from the ENCODE project, we test a novel algorithm that resolves pairs of coordinately active enhancers based on their activity profiles across

several dozen representative cell types. From correlated pairs, we identify correlated clusters of enhancers, which exhibit multiple hallmarks of coordinate regulation, including spatial proximity.

In Chapter 3, we test whether spatially proximate but genomically distal homotypic binding sites impact occupancy observed at a given TF BS. We find that, indeed occupancy scales with the abundance of spatially proximate homotypic BS or, similarly, the degeneracy of the TF's motif. Through biophysical modelling we show that spatial proximity induces a stronger, generalized 3-D version of the mechanism known to boost TF occupancy and concentration in 1-D homotypic clusters, consistent with our observations. Moreover, in contrast to the genomically hard-wired 1D version, spatial homotypic clusters are conditioned on the chromatin's conformation. Accordingly, we find that the archipelago-centered occupancy boost is much more cell type-specific.

In Chapter 4, we scale up from binding sites to enhancers and whole archipelagos in order to test for downstream functional impact of the occupancy boost observed in Chapter 2. We find evidence of strongly divergent behavior between enhancers enriched, and alternatively, depleted for degenerate motifs; enriched enhancers have much higher chromatin accessibility, putative target gene expression, and are subject to much higher purifying selection. Together with the unusually high responsiveness of archipelago-wide activity to degenerate TF availability, we infer that the occupancy boost characteristic of spatial homotypic clusters fosters archipelago upregulation.

Active archipelagos have been shown to be more spatially compact compared to their ground state conformation. In chapter 5, we ask whether this compaction can be

explained by a feedback loop involving the demonstrated occupancy boost. Specifically, we test whether the increase in local degenerate TF occupancy and TF concentration demonstrated in Chapter 3, in turn, induce increased chromatin looping through either of two mechanisms. We present preliminary evidence that it does, resulting in a more compact archipelago in its active state. Finally, in chapter 6 we conclude with overall perspective and potential future directions.

Chapter 2: **Enhancer networks revealed by correlated DNAse hypersensitivity states of enhancers**

Abstract

Mammalian gene expression is often regulated by distal enhancers. However, little is known about higher order functional organization of enhancers. Using ~100K P300-bound regions as candidate enhancers, we investigated their correlated activity across 72 cell types based on DNAse hypersensitivity (DHS). We found widespread correlated activity between enhancers, which decreases with increasing inter-enhancer genomic distance. We found that correlated enhancers tend to share common transcription factor (TF) binding motifs, and several chromatin modification enzymes preferentially interact with these TFs. Presence of shared motifs in enhancer pairs can predict correlated activity with 73% accuracy. Also, genes near correlated enhancers exhibit correlated expression and share common function. Correlated enhancers tend to be spatially proximal. Interestingly, weak enhancers tend to correlate with significantly greater numbers of other enhancers relative to strong enhancers. Furthermore, strong/weak enhancers preferentially correlate with strong/weak enhancers respectively. We constructed enhancer networks based on shared motif and correlated activity and show significant functional enrichment in their putative target gene clusters. Overall, our analyses shows extensive correlated activity among enhancers and reveals clusters of enhancers whose activities are coordinately regulated by multiple potential mechanisms involving shared TF binding, chromatin modifying enzymes and 3D chromatin structure, that ultimately co-regulate functionally linked genes.

Introduction

Eukaryotic transcription is intricately regulated at multiple levels, including epigenomic modifications, chromatin reorganization, and sequence specific binding of TF to either proximal promoter regions or to distal enhancer/repressor regions of a gene (Maston et al., 2006; White, 2011). Distal enhancers can regulate their target genes from long distances, the most extreme case being the Shh gene's enhancer at ~1Mb away, and are especially important in regulating critical developmental genes (Lettice, 2003; Naranjo et al., 2010). Recent advances in sequencing technologies have revealed that cell-specific enhancers are often marked by P300 binding (a histone acetyltransferase and transcription coactivator) (May et al., 2011; Visel et al., 2009), as well as other epigenomic marks such as DNAse hypersensitivity (DHS), H3K4me1, H3K27ac, etc. (Heintzman et al., 2009a; Zentner et al., 2011). Various combinations of these marks have been used to generate genome-wide catalogs of potential cell type specific distal enhancers (Heintzman et al., 2009b). However, the target genes of the distal enhancers remain unknown for the most part. Moreover, the mechanisms by which distal enhancers regulate the expression of their target genes are not completely understood.

Functionally linked genes, e.g., components of a biological pathway or a protein complex, tend to be co-expressed and are presumed to be co-regulated (Berman et al., 2004; Liu et al., 2009; Stuart, 2003; Wasserman and Fickett, 1998) Gene networks based on co-expression patterns of gene pairs across multiple conditions and/or cell types reveal intricate organization of genes into pathways and functional groups (Dewey et al., 2011). Similar to functionally related genes, functionally related enhancers, i.e., those regulating functionally related genes, share TF binding sites and are likely to have spatio-temporal

coordinated activity (Narlikar et al., 2010). A network-level analysis of coordinated activities of distal enhancers has not been reported and such an analysis is likely to reveal higher order organization of a global transcriptional regulatory network mediated by distal enhancers. Analogous to using expression level to quantify transcriptional activity of a gene, DHS of an enhancer region has been proposed as a proxy for its condition-specific regulatory (Heintzman et al., 2009b; Li et al., 2011; Pique-Regi et al., 2011) Under the ENCODE project, whole genome DHS profiles have been generated for dozens of human cell type (Bernstein et al., 2012). Analogous to using cross-condition expression correlation to infer gene networks, cross-condition DHS correlation can be used to infer enhancer networks. Indeed, a recent report has shown the effectiveness of using cross-condition DHS correlation between distal enhancers and gene promoters to identify distal enhancers of genes (Sanyal et al., 2012).

Tissue-specific enhancers are often marked by P300 binding. Most of the tested P300 bound regions in mouse embryonic forebrain, midbrain and limb tissue were shown to function as enhancers in transgenic mice (Visel et al., 2009). Thus, a genome-wide profile of P300 bound regions provides a reasonable approximation for candidate enhancer regions. Starting with ~100,000 P300 bound regions in one or more out of 4 cell types as candidate enhancers, here we perform a detailed network-level analysis of enhancers based on their DHS correlation across 72 cell types. We identified a large set of enhancer pairs whose DHS level was significantly correlated across cell types, even after controlling for autocorrelation of DHS along the chromosome. We found that **(i)** correlated enhancers tend to share common TF binding motifs. **(ii)** Several chromatin modification enzymes preferentially interact with TFs whose binding sites co-occur in

pairs of correlated enhancers. **(iii)** Presence of shared motifs can discriminate between correlated and uncorrelated enhancer pairs with 73% accuracy. **(iv)** Using the gene closest to an enhancer as its putative target, we found that the targets of correlated enhancers have correlated expression and are involved in common biological processes. **(v)** Based on Hi-C data on chromatin spatial interaction in two different cell types, we found that correlated enhancers are spatially proximal significantly more often than expected. **(vi)** Strong enhancers, those with higher expression levels of the nearest gene, tend to be correlated with fewer enhancers than weak enhancers but preferentially correlate with other strong enhancers, while weak enhancers are correlated with a greater number of enhancers and preferentially correlate with other weak enhancers. **(vii)** We constructed enhancer networks based on correlated activity and shared TF motifs, and found significant enrichment of specific biological processes among the putative gene targets of the enhancer modules.

Overall, our analysis suggests that functionally linked genes may be co-regulated by distal enhancers whose activities are regulated by common sets of TFs and mediated by both 3D chromatin structure as well as chromatin modification enzymes. Our work represents the first investigation of enhancer networks based on correlated activity across multiple cell types.

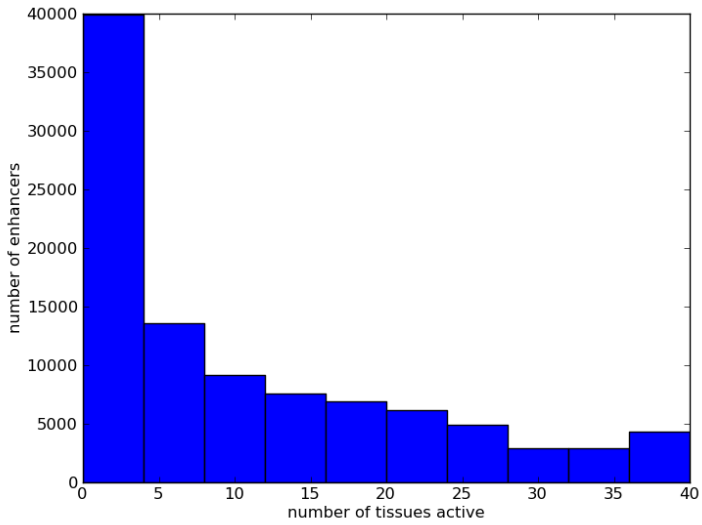
Results

Data overview

P300 binding has been shown to be a reliable marker of tissue specific enhancers (Visel et al., 2009). As a starting set of candidate enhancers we obtained 98,353 P300 peaks in 4 different cell types (see M&M). We extracted genome-wide DHS broad peak data for 72

tissue types in the ENCODE database (Bernstein et al., 2012) and clustered the 72 tissues into 37 representatives (Appendix Table 1) based on genome-wide correlation (see M&M). Enhancers vary broadly (0-37 tissues) in the number of tissues in which they overlap a DHS peak (Figure 2-1). For each enhancer, we constructed a DHS profile as a binary vector of length 37 corresponding to 37 cell types, by setting the DHS value to 1 if the enhancer region overlapped a DHS peak in the particular tissue; otherwise it was set to 0. This procedure yielded a 98,353 x 37 enhancer 'activity' matrix, with rows corresponding to enhancers, columns to tissue (or cell) types.

Figure 2-1. Activity per enhancer. Histogram shows the number of tissues (x-axis) in which a given enhancer is active (out of 37 tissues possible).



Identifying enhancers with correlated activity

We quantified correlated activity for a pair of enhancers using the information theoretic measure *Mutual Information* (*MI*) using DHS in 37 tissues (see M&M). However, *MI* can be biased towards enhancer pairs that are near each other on the genome, if DHS regions are long or tend to cluster on the genome. We tested this by selecting intra-chromosomal

pairs using 100,000 random genomic segments and computing their *MI*. Figure 2-2 shows that the fraction of segment-pairs with *MI* > 0.4 decays monotonically with increasing inter-segment distance, suggesting autocorrelation of DHS along the genome; the same trend holds for other *MI* thresholds. The same trend also holds for the 35 million enhancer pairs tested, but crucially, the fraction of enhancer pairs with high *MI* is greater than that of random genomic segments (represented by yellow and gray bars, respectively, in Figure 2-2). We controlled for the observed cell type-specific DHS autocorrelation to detect significantly correlated enhancer pairs (see M&M and Figure 2-3). We consider six distance-bins ranging from 20 Kb to '>12.5 Mb' (Figure 2-4) and within each distance-bin, we identify significantly correlated enhancer pairs by estimating a nominal False Discovery Rate (FDR) (Reiner et al., 2003) by comparing *MI* scores for actual and control pairs (see M&M).

Figure 2-2. Generating the synthetic enhancer data to account for autocorrelation.

(A) Starting with a large set of random genomic regions and their DHS profiles across 37 cell types, we estimate, for each cell type separately, the conditional probability of observing DHS at a location *Y'* given the DHS status at another location *X* at distance *d* from *X*. **(B)** Given a pair of enhancer DHS profiles (*X*,*Y*), we generate a synthetic pair of DHS profiles as (*X,Y'*) where *Y'* is randomly generated from *X* and the conditional probabilities estimated in (A). See text for further details. Blue: DHS=1 (open chromatin); white: DHS = 0 (closed chromatin)

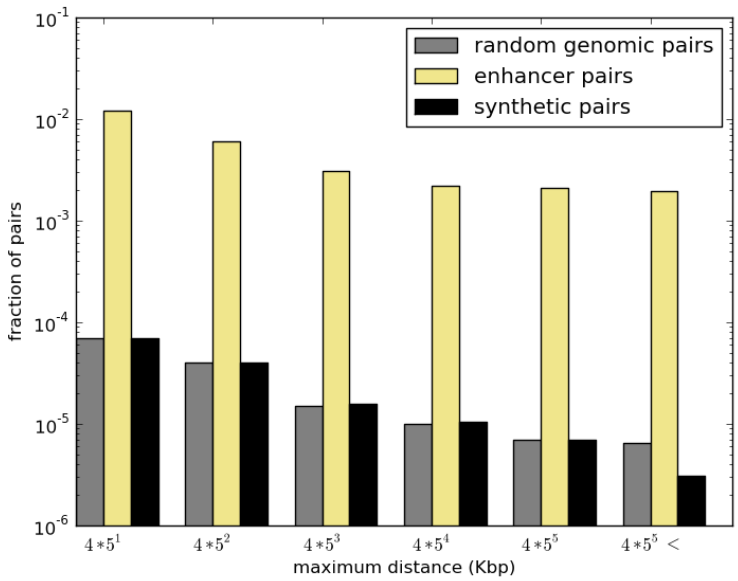


Figure 2-3. Mutual information of chromatin states is higher among enhancer pairs than background pairs, and it decreases monotonically with increasing distance. Plot shows the relationship between inter-enhancer genomic distance and the number of actual and synthetic enhancer pairs with *MI* above 0.4 across 37 representative cell types. Enhancer pairs (yellow) were selected from 98,000 enhancers identified based on P300 ChIP-Seq peaks by exhaustively pairing all enhancers sharing the same chromosome and <12.5 Mb apart. Five million additional pairs were sampled for distances >12.5 Mb, as well as 1 million inter-chromosomal pairs. As a negative control the DHS vector of a randomly chosen member of each enhancer pair was used as a seed to generate a paired synthetic DHS vector by conditioning on observed cell type-specific DHS autocorrelation along the genome. This resulted in 1 synthetic enhancer pair (black) for each enhancer pair; pairs of random genomic segments (gray) were generated in the same fashion as enhancer pairs by drawing from 100,000 random genomic segments of mean length 500 bp. *MI* of 0.4 roughly corresponds to FDR 0.01 (see text).

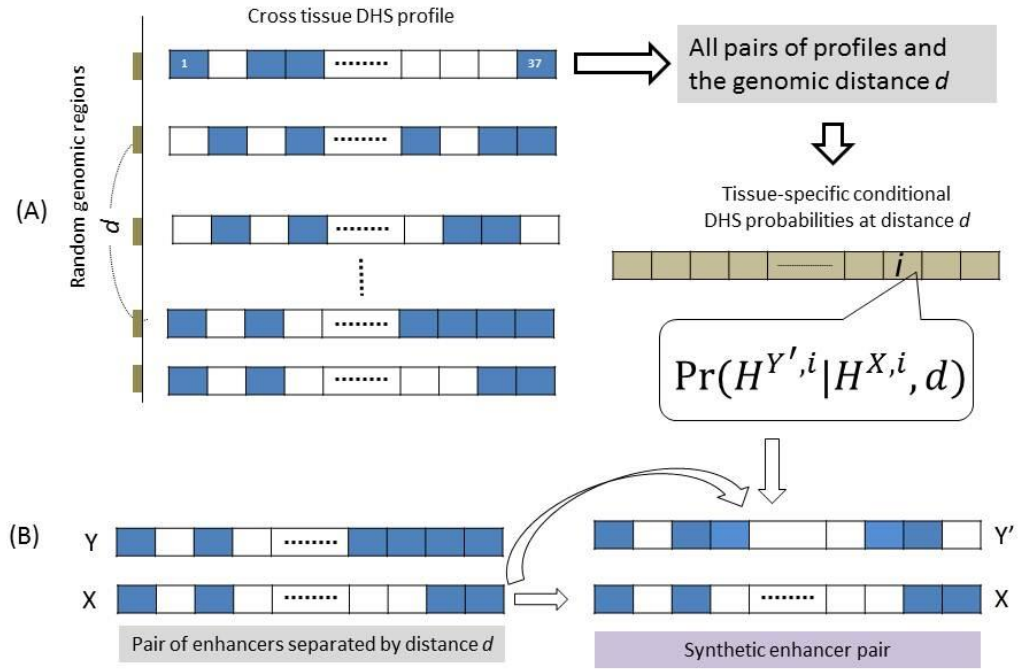
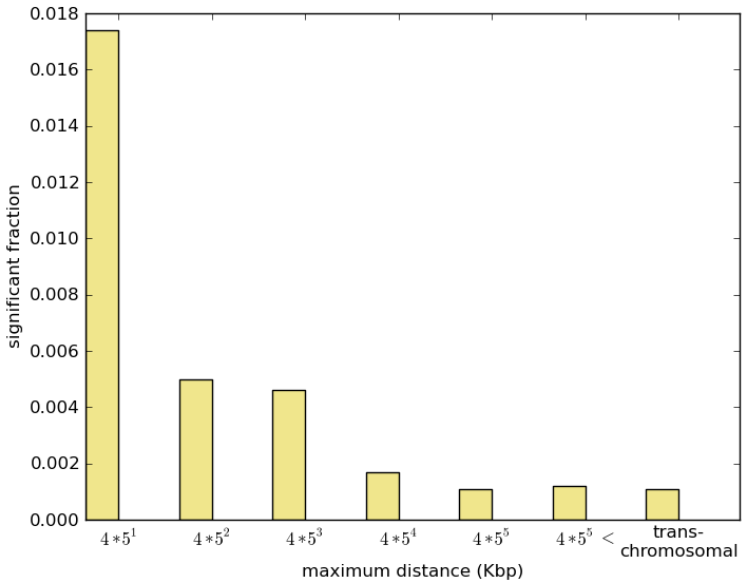


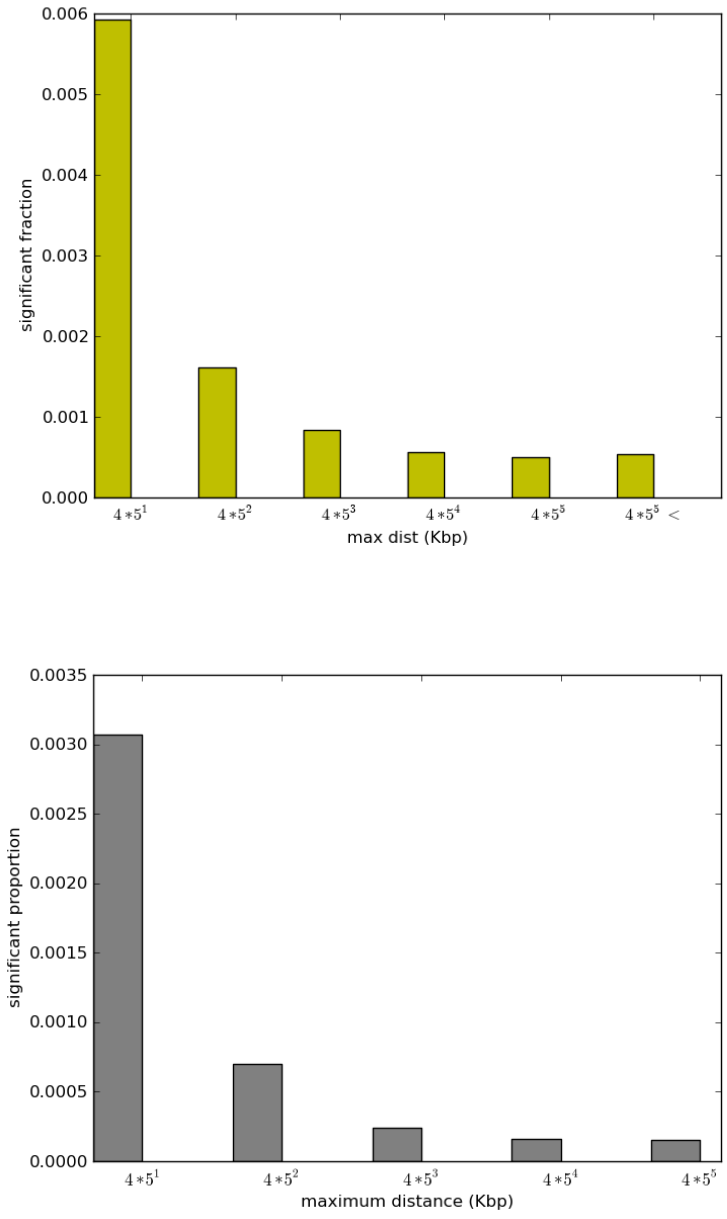
Figure 2-4. Chromatin states of a large number of enhancer pairs are significantly correlated. The plot shows the fraction of pairs with significant mutual information (*MI*) as a function of inter-enhancer distance. Significant enhancer pairs were identified by setting a threshold *MI* for each bin that corresponded to a nominal false discover rate of 0.1% (see text). The plot is based on significant pairs after greedily removing pairs inducing transitive relationships. The percentage of significant enhancer pairs drops with pairwise distance, but stabilizes at ~2 Mb. Moreover, if one of the enhancers in our set overlapped both with a strong and weak chromHMM enhancer, we excluded that enhancer as well as the overlapping chromHMM enhancers from our calculations.



A sizable fraction of enhancer pairs have correlated activity across cell types

We exhaustively assessed ~35 million intra-chromosomal enhancer pairs separated by less than 12.5 Mb; additional sampling at larger distances and across chromosomes suggested that 12.5 Mb ceiling is sufficient to capture general patterns. Despite distance bin-specific FDR control, the fraction of enhancers that are significantly correlated declines with increasing distance (Figure 2-4); after removing transitive relationships (M&M), at FDR of 0.1%, the fraction decreases from 1.7% pairs at 20 Kb to 0.1% for pairs separated by more than 12.5Mb. The corresponding fractions at 5% FDR are 4.8% to 1.3%. A similar trend is also observed when background pairs are pooled across distance bins and a single FDR test is conducted (Figure 2-5 *left*). Similarly, these trends are preserved when we used random trans-chromosomal enhancer pairs as the background to calculate the FDR (Figure 2-5 *right*). Across all bins, at an FDR of 1% we detect a total of 313,757 significant enhancer pairs, covering 32% of enhancers.

Figure 2-5. Fraction of significantly correlated enhancer pairs decreases monotonically with increasing distance between the enhancers when an FDR test is conducted on a common pooled background (*top)*, and on a background of trans-chromosomal pairs (*bottom)*. Bin-wise fractions (y-axis) reflect partitioning of enhancer pairs after significance screen.

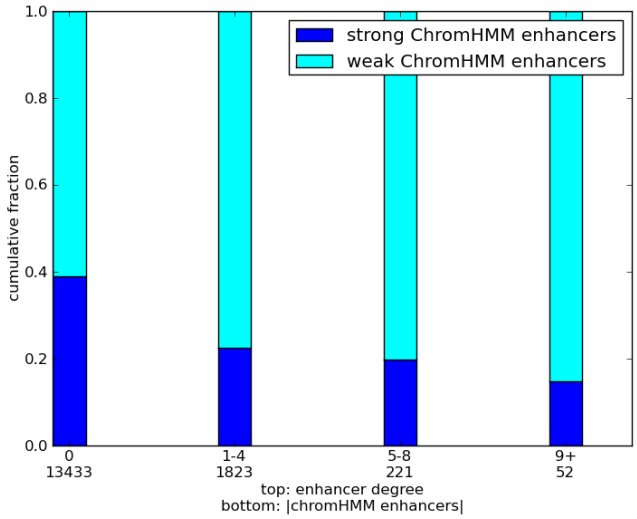


Strong and weak enhancers have different degrees of connectivity and are assortative

Previous studies have shown that low affinity binding sites for individual TFs tend to cluster on the genome (Essien et al., 2009a) and such clustering of binding sites in regulatory regions has been suggested to cooperate to promote overall functionality via multiple mechanisms (Anderson and Freytag, 1991; Coleman and Pugh, 1995; Giniger and Ptashne, 1988; He et al., 2011). Extending this notion to the level of enhancers, we assessed whether weak enhancers have a greater proclivity to cooperate. Ernst *et al* have previously predicted enhancers in the genome based on histone modification patterns using the ChromHMM tool and further classified the enhancers into 'strong' and 'weak' based on cell type-specific expression level of the proximal gene (Ernst and Kellis, 2012). We calculated each enhancer's *degree* as the number of other enhancers it is correlated with and partitioned enhancers into 5 bins based on degrees: 0, 1-4, 5-8, ≥ 9 (other binning schemes do not affect the conclusion). For each bin we calculated the fraction of 'strong' enhancers out of all enhancers overlapping with a ChromHMM enhancer. Figure 2-6 shows that weak enhancers tend to have correlated activity with several other enhancers whereas strong enhancers tend to function in smaller groups. For instance, the percentage of strong enhancers having no correlation partners (44%) is significantly higher than that for the weak enhancers (35%) (Fisher exact test p-value = 1.8e-56). Next we checked whether strong/weak enhancers preferentially interact with other strong/weak enhancers. Even though strong enhancers have fewer interactions, we found that strong enhancers are twice as likely to be correlated with another strong enhancer than expected by chance (Fisher exact test p-value =1.6e-7). Similarly, weak enhancers preferentially interact with other weak enhancers (Fisher exact test p-value =

0.0002). The above results are based on a *MI* FDR threshold of 0.01 but the trend remains significant at FDR = 0.05. Thus, strong and weak enhancers assort with other strong and weak enhancers, respectively.

Figure 2-6. Relative to strong enhancers, weak enhancers are more likely to be coordinately activated with other enhancers. Bar plot shows the relative fractions of all enhancers that are non-ambiguously classified in chromHMM data base as 'weak' or 'strong' enhancers partitioned into 4 groups, based on their degree, i.e., the number of other enhancers with which they are epigenetically highly correlated (FDR 0.0001), which is recorded along top row of x-axis. Numbers on bottom row indicate the total number of non-ambiguously classified chromHMM enhancers in that bin. Note that the determination of whether an enhancer has 0 neighbors was made at a more relaxed FDR 0.05.

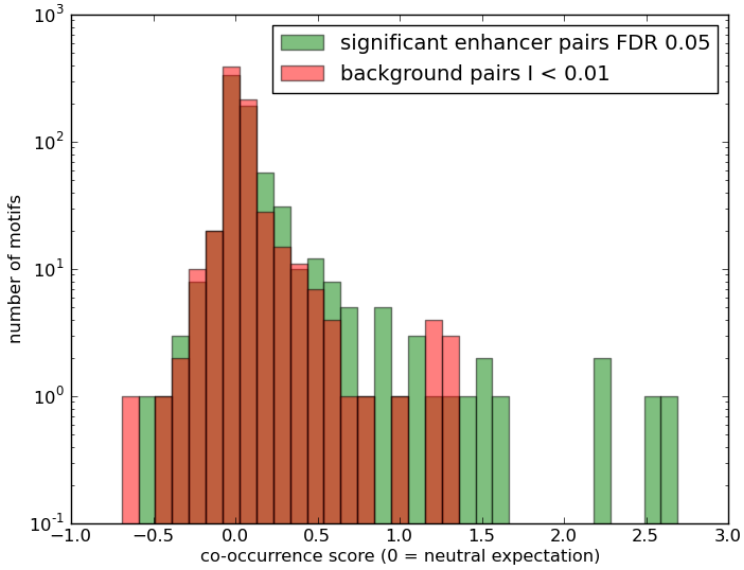


Potential roles of TFs and chromatin modification enzymes in correlated enhancer activity

It is possible that correlated activities of enhancers are mediated by common TFs, as has been shown widely for promoters of co-expressed genes (Liu et al., 2009). We therefore tested whether correlated enhancer pairs harbor common TF binding sites. We created two sets of enhancer pairs: the *foreground* included the significantly correlated enhancer pairs at FDR = 5% (conclusions remain the same at other thresholds) in each distance bin.

Background enhancer pairs were randomly chosen from enhancer pairs in each distance bin with *MI* < 0.01. Note that, in this context and in what follows, the term *Background* is used to refer to uncorrelated enhancer pairs as opposed to non-enhancer pairs. Next we identified high-scoring binding sites in each enhancer for each of the 981 vertebrate motifs (see M&M) and quantified the tendency of a motif to co-occur in correlated enhancers based on a *co-occurrence score* (see M&M). We found that the overall co-occurrence score distribution for all motifs was significantly higher in the foreground than the background (Figure 2-7; Wilcoxon test p-value = 6.7e-18). Next, we estimated the significance of co-occurrence for each motif in the foreground by comparing observed and expected co-occurrence frequency using a Chi-squared test. After controlling for multiple

Figure 2-7. Motif co-occurrence is greater among correlated enhancers than background non-correlated enhancer pairs. Histogram shows the log enrichment of motif co-occurrence above random expectation for significantly correlated enhancer pairs (FDR 0.01) (green) compared with the same for background pairs (red). The x-axis shows the log of enrichment values, where 0 denotes random expectation, and more positive scores indicate higher enrichment, while negative scores indicate higher depletion. The y-axis show the number of motifs with the indicated level of log enrichment. Background pairs were selected based on mutual information scores < 0.01. "10-1" on the y-axis is an artifact of the drawing tool and simply represents 0.



testing, at FDR = 0.05, we found 153 motifs with significant co-occurrence (M&M). An identical analysis of background enhancer pairs yielded only 39 motifs. We further filtered the 153 motifs down to the 62 most significant motifs by directly comparing the co-occurrence p-values in the foreground and the background using the nominal FDR approach (25) at 5% FDR. Of the 62, 10 were significant in the background. The remaining 52 motifs (Table 1) were used for further analyses.

TABLE 2-1. Motifs with significantly greater co-occurrence in correlated enhancers than **expected** Col 1: TRANSFAC motif ID, col 2: co-occurrence score (see text), col 3: p-value, col 4: multiple testing corrected q-value, col 5: TF name.

Motif	Cooccurrence Score	p-value	q-value	Gene
M00649	9.80E-02	0	1.70E-04	MAZ
M01742	1.20E+00	0	2.10E-04	Zfp206
M00986	3.90E-02	0	3.00E-04	Churchill
M00915	5.40E-01	0	3.80E-04	AP-2
M01028	2.70E+00	0	4.30E-04	NRSF
M01783	6.30E-01	0	4.70E-04	SP2
M00431	1.30E-01	0	5.10E-04	E2F-1
M00008	3.30E-01	0	5.60E-04	Sp1
M01199	6.90E-01	0	6.00E-04	RNF96
M01219	4.60E-01	0	6.40E-04	SP1:SP3
M00925	5.40E-02	0	7.30E-04	AP-1
M01253	7.50E-01	0	8.10E-04	CNOT3
M00189	6.80E-01	0	9.00E-04	AP-2
M00255	3.70E-01	0	9.40E-04	GC_box
M01482	2.60E+00	0	9.80E-04	Nkx3-2
M00716	8.20E-01	0	1.00E-03	ZF5
M01267	6.40E-02	0	1.10E-03	FRA1
M00199	9.20E-02	0	1.10E-03	AP-1
M00196	6.30E-01	0	1.20E-03	Sp1
M00800	8.00E-01	0	1.20E-03	AP-2
M00807	3.20E-01	0	1.30E-03	Egr
M00931	4.80E-01	0	1.30E-03	Sp1
M00933	3.20E-01	0	1.40E-03	Sp1
M00932	5.90E-01	0	1.40E-03	Sp1
M00615	1.90E+00	0	1.50E-03	c-Myc:Max
M01303	3.10E-01	0	1.50E-03	SP1
M01588	2.90E-01	0	1.50E-03	GKLF_(KLF4)
M00322	4.30E-01	0	1.60E-03	c-Myc:Max
M00976	2.20E-01	0	1.60E-03	AhR,_Arnt,_HIF-1
M00720	7.80E-02	0	1.70E-03	CAC-
M01273	4.50E-01	0	1.70E-03	SP4
M01837	1.70E-01	0	1.80E-03	FKLF
M00174	1.10E-01	1.10E-	1.90E-03	AP-1
M00926	3.80E-02	4.40E-	1.90E-03	AP-1
M00428	4.60E-02	6.70E-	2.00E-03	E2F-1
M01593	9.50E-01	1.20E-	2.10E-03	Zfx
M01104	4.60E-02	2.20E-	2.10E-03	MOVO-B
M01177	3.20E-01	1.50E-	2.10E-03	SREBP2
M01230	2.40E-02	1.60E-	2.20E-03	ZNF333
M01816	1.30E-01	5.60E-	2.20E-03	ZBP89
M00940	5.50E-01	4.10E-	2.30E-03	E2F-1
M01597	2.20E-01	9.70E-	2.30E-03	Zfp281
M01045	3.90E-01	2.70E-	2.40E-03	AP-2alphaA
M01162	3.00E-02	1.20E-	2.40E-03	OG-2
M01292	2.00E-02	1.50E-	2.40E-03	HOXA13
M00378	9.90E-02	1.30E-	2.50E-03	Pax-4
M00982	6.80E-01	2.00E-	2.60E-03	KROX
M00644	3.30E-02	3.70E-	2.60E-03	LBP-1
M01714	3.50E-01	4.70E-	2.70E-03	KLF15
M01275	2.40E-02	9.80E-	2.70E-03	IPF1
M01318	1.40E+00	1.60E-	2.70E-03	Irx-3
M00175	4.70E-02	1.90E-	2.80E-03	AP-4

Motif co-occurrence among correlated enhancer pairs confirmed when cell type-specific TF availability screened for

To make the test of co-occurrence more targeted, instances of co-occurrence in a pair were only counted when there was at least one tissue in which both pair members were active and the cognate TF expressed. Motifs were not considered for which binding TF information was not available or that bind to TFs coded for by two or more genes. Approximately one-half (509) of the 981 motifs qualified. TFs were considered expressed in a given tissue if the normalized tag count density exceeded 0, where 0 was chosen due to the lack of any discontinuity in the distribution of tag count densities. (Based on this criterion, on average < 30% of TFs are expressed in each tissue). Under these conditions, there were a total of 67 motifs that co-occurred significantly more often than expected (FDR 5%, based on p-values from Fisher Exact Test) and present in at least 20 pairs, compared to zero motifs that occurred more often than expected in uncorrelated pairs. 20 of the 52 motifs previously found to co-occur out of 981 motifs were among the set of 67, in spite of the reduced test set of motifs. When thresholds of expression higher than 0 were used similar, if fewer, sets of significant motifs resulted (while still no motifs in random pairs significantly co-occurred). Thus, the co-occurrence of motifs is reinforced when cell-type activity is screened for.

Extending test of correlated motifs to enhancer clusters

We next extended the pair-wise motif co-occurrence analyses to clusters of correlated enhancers. Disjoint clusters with at least 10 enhancers were greedily identified such that mean MI for all pairs within the cluster was at least 0.2 (other thresholds do not change the conclusion). Each TRANSFAC motif was assessed for enrichment in each cluster relative to other clusters based on a Fisher Exact Test, and significance was corrected for

multiple testing. At a FDR threshold of 5%, for the 415 clusters, there were 44 instances of cluster-specific enrichments. In contrast, for a background set of 415 clusters using randomly chosen enhancers (mean pairwise *I* within a cluster << 0.1) sampled to match total motif occupancy, mean GC content, and the cluster size of the foreground, there were only 2 instances of cluster-specific enrichment (Figure 2-8).

Correlated enhancer pairs are potentially co-regulated

Co-regulated enhancers tend to share common motifs (Berman et al., 2004). To investigate whether the enhancer pairs with correlated activity are potentially co-regulated, next we tested whether correlated enhancers share significantly greater *numbers* of motifs than expected. We quantified motif overlap between the two enhancers using Jaccard index, defined as the ratio of the sizes of the intersection and the union of the two motifs sets. Separately for each distance-bin we compared Jaccard index values for the highly correlated enhancer pairs with those for pairs in the background using a Wilcoxon rank-sum test. The foreground and the background enhancer pairs were selected as for Result section 5 above. We found that in every distance bin the foreground pairs have a significantly greater fraction of shared motifs, with p-values ranging from 1.6e-04 to 6.1e-33 (Table 2a). The result remains highly significant when we repeated the analysis at the level of motif clusters instead of individual motifs (see M&M). As expected, the difference between foreground and background is amplified when only 52 significantly co-occurring motifs (see above) were used to calculate Jaccard index (Table 2b). These results suggest that not only are co-occurring motifs present more often than expected in correlated enhancer pairs, but that correlated enhancer pairs also share overall greater numbers of motifs than expected. Taken together, this analysis shows that

epigenetically correlated enhancers share TF binding motifs significantly more frequently than expected, suggesting a role for these TFs in co-regulation of the correlated enhancers.

Figure 2-8. Illustrative example of an enhancer cluster. The Figure shows on the genome browser a representative cluster of enhancers comprising 117 enhancer spread throughout chromosome 2. This cluster includes 12 strong (blue ticks) and 54 weak enhancers (red ticks) as annotated by ChromHMM. DHS (black ticks) in 5 representative cell types are shown for all enhancers. The Figure clearly illustrates the correlated activity of these enhancers across the cell types. In addition, this cluster, which was constructed without regard to motif co-occurrence, in fact broadly shared 2 motifs (magenta ticks).

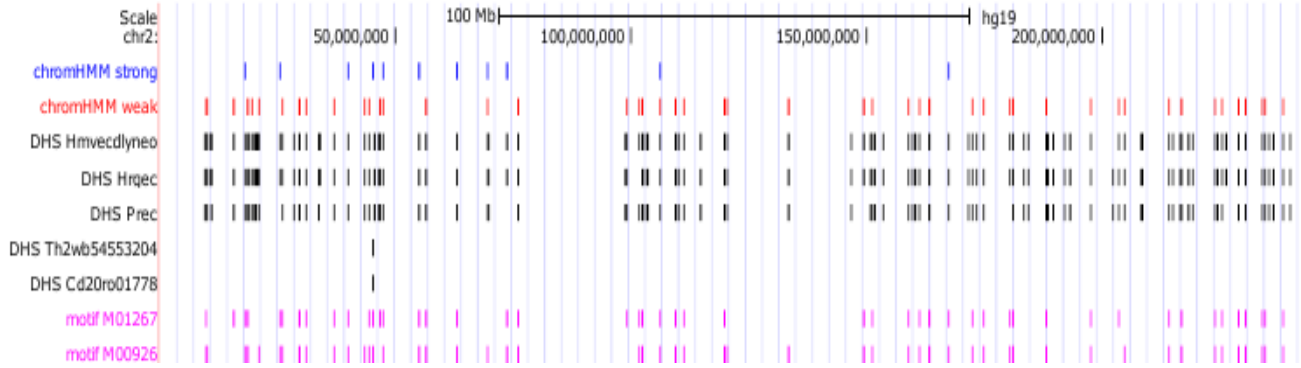


TABLE 2-2. **Motif sharing between coordinated enhancer pairs and the background.** (a) Results of Wilcoxon rank-sum tests comparing the extent of motif overlap in correlated enhancer pairs (FDR 0.0001) to that in background pairs, with one test per distance bin. All 981 vertebrate motifs in the TRANSFAC database were used. (b) same as (a) except that overlap is evaluated only for the significantly co-occurring motifs in correlated enhancers.

(a)

Max dist between Enhancers (kB)	Correlated enhancer pairs (FDR 0.0001)		Background enhancer pairs (I < 0.01)	
	Mean Jaccard (all motifs)	Median Jaccard (all motifs)	Mean Jaccard (all motifs)	Median Jaccard (all motifs)
20	0.32	0.32	0.3	0.3
200	0.32	0.32	0.29	0.28
1000	0.31	0.31	0.29	0.29
20000	0.31	0.31	0.28	0.28
Overall	0.31	0.31	0.29	0.29

(b)

Max dist between Enhancers (kB)	Correlated enhancer pairs (FDR 0.0001)		Background enhancer pairs (I < 0.01)	
	Mean Jaccard (significant motifs)	Median Jaccard (significant motifs)	Mean Jaccard (significant motifs)	Median Jaccard (significant motifs)
20	0.22	0.14	0.12	0
200	0.28	0.2	0.11	0
1000	0.29	0.2	0.11	0
20000	0.3	0.25	0.11	0
Overall	0.28	0.2	0.11	0

Presence of shared motifs is predictive of enhancer DHS correlation

Additionally, we assessed, using machine learning, whether the presence of common motifs can predict correlated activity of a pair of enhancers. For each enhancer pair we

assigned one attribute per motif. The value of the attribute was set to 1 if both enhancers had a motif instance and 0 otherwise. We then trained and tested a support vector machine (SVM) to discriminate between the foreground (FDR 0.01% was used for computational tractability) and the background enhancer pairs, using 10-fold cross validation. When using all 981 motifs as attributes, the SVM achieved an overall average classification accuracy of 73%. Importantly, there was very little reduction in performance (70%) when the model used only the 52 significantly co-occurring motifs (section 5). However, when we used 52 random motifs, the SVM accuracy was reduced to 55%, not much greater than random expectation of 50%. This result suggests that shared occurrence of a specific set of motifs is predictive of correlated enhancer activity.

Interactions between enhancer motifs and chromatin modification enzymes

To further probe the potential involvement of chromatin modification enzymes (CME) in regulating correlated enhancer activities, we assessed CMEs for their preferential interactions with the 52 motifs (Table 1) that significantly co-occur in correlated enhancers. The 52 motifs mapped to 146 unique proteins using TRANSFAC and ENSEMBL databases, while the remaining motifs mapped to 2227 proteins. There are more proteins than motifs due to ambiguous mapping of motifs to isoforms. A list of 828 CMEs was extracted from ENSEMBL database (version 67) based on GO term 'chromatin modification'. Protein-protein interactions were obtained from STRING database using the 'experimental' track. We assessed each of the 828 CMEs for preferential interaction with 146 TFs corresponding to significant motifs relative to the other 2227 TFs, using a Fisher Exact test, followed by multiple testing correction. At FDR = 5% we detected 28 CMEs to preferentially interact with significant TFs (Table 3).

In contrast, there was no CME that preferentially interacted with non-significant TF. This result is especially interesting given that overall, the 146 significant TFs do not interact with CMEs any more than the other 2227 TFs. Overall, this analysis implicates CMEs in correlated enhancer activity.

TABLE 2-3. **Chromatin modifying enzymes (CME) that preferentially interact with significantly co-occurring motifs (Table 2-1). Column 3 denotes the percent of significant motifs interacting with the CME.**

CME	P-value	Interaction Frequency	Description
ENSP00000336750	5.50E-	5.50%	suppressor of Ty 7 (S.
ENSP00000308227	5.90E-	9.60%	high mobility group AT-hook
ENSP00000264709	9.60E-	8.20%	DNA (cytosine-5-)-
ENSP00000362649	1.20E-	16.00%	histone deacetylase 1
ENSP00000231509	1.60E-	12.00%	nuclear receptor subfamily "3,"
ENSP00000349508	2.30E-	6.80%	chromodomain helicase DNA
ENSP00000278823	2.40E-	6.20%	metastasis associated 1
ENSP00000367207	2.90E-	15.00%	v-myc myelocytomatosis viral
ENSP00000343325	2.90E-	5.50%	protein kinase N1
ENSP00000263119	4.20E-	6.20%	calcineurin binding protein 1
ENSP00000362674	5.30E-	5.50%	histone deacetylase 8
ENSP00000334061	5.40E-	6.20%	histone deacetylase 6
ENSP00000386759	7.30E-	6.80%	SET domain containing 2
ENSP00000302967	9.20E-	10.00%	histone deacetylase 3
ENSP00000352516	9.50E-	8.20%	DNA (cytosine-5-)-
ENSP00000284384	1.20E-	6.80%	protein kinase "C," alpha
ENSP00000349049	1.30E-	5.50%	lysine (K)-specific demethylase
ENSP00000225983	1.40E-	8.20%	histone deacetylase 5
ENSP00000381331	1.50E-	9.60%	histone deacetylase 2
ENSP00000371067	2.30E-	8.20%	Janus kinase 2
ENSP00000264606	2.40E-	7.50%	histone deacetylase 4
ENSP00000264010	2.50E-	6.20%	CCCTC-binding factor (zinc
ENSP00000268712	2.50E-	9.60%	nuclear receptor corepressor 1
ENSP00000337088	2.70E-	6.20%	multiple endocrine neoplasia I
ENSP00000356480	2.80E-	5.50%	ring finger protein 2
ENSP00000231487	2.90E-	6.20%	S-phase kinase-associated
ENSP00000263253	3.00E-	15.00%	E1A binding protein p300
ENSP00000267163	3.10E-	9.60%	retinoblastoma 1

Correlated enhancers are spatially proximal

We expect the correlated activity of non-proximal enhancers to be associated with their spatial proximity in the nucleus. We estimated the fraction of correlated enhancer pairs that are spatially proximal based on Hi-C data (GSE18199) (Lieberman-Aiden et al., 2009). We note that the Hi-C data was obtained from human K562 and HIC_gm06690 cell lines, while DHS correlation was obtained across 37 primary cell types. It is known that spatially interacting regions are enriched for DHS (Fang et al., 2009). We controlled for this by ensuring that in each distance bin, the background enhancer pairs were selected such that their average pair-mean DHS across cell types was within 2% of the corresponding average for foreground pairs. We compared foreground and background enhancer pairs in terms of the fraction of pairs that are spatially proximal according to the K562 Hi-C experiment, using a Fisher Exact Test. We found that overall, the foreground enhancer pairs showed a greater coincidence with Hi-C data (p-value = 0.01). Even when we include only the top 10% most confident Hi-C pairs, the p-value = 0.03. When we repeated the above tests using the HIC_gm06690 Hi-C data, the corresponding p-values are 0.02 and 0.009. These results suggest that spatial proximity of the chromosomal regions is associated, albeit weakly, with correlated enhancer activities. The weak association may be due to cell type specificity of spatial proximity (see Discussion).

Genes near correlated enhancers have correlated expression and shared function

We hypothesized that the gene targets of highly correlated enhancers are themselves correlated in their expression. Although the targets of enhancers are largely unknown, as a first approximation, we mapped each enhancer to its nearest gene as a putative target (Thurman et al., 2012). For each gene we obtained from GEO (Barrett et al., 2010) the

normalized RNA-seq transcript counts from 15 of the 72 tissue-types and calculated the Spearman correlation between vectors of transcript counts. For the foreground enhancer pairs at FDR 1% (results are comparable for other FDR thresholds), we found that the median Spearman correlation of expression of the target genes was 0.31, while for the background it was only 0.18 (Wilcoxon rank-sum test p-value = 2.1e-74). It indicates that epigenetically correlated enhancers tend to have co-expressed target genes.

Our analyses thus far suggest that correlated enhancer pairs have (A) a greater motif co-occurrence (section 5), and (B) greater co-expression between their target genes (section 7). Therefore, we assessed directly whether motif co-occurrence in enhancers is predictive of correlated expression in their target genes, regardless of correlated activity of the enhancers. 10,000 enhancer pairs were sampled without regard for their correlation. The Jaccard index for motif sharing between enhancers and gene co-expression for putative target genes was estimated as above. Based on linear regression of expression correlation against the corresponding enhancer pairs' Jaccard indices, we found the two to be highly positively associated with a slope of 0.26 (p-value = 4.4e-26 for null hypothesis that slope = 0), suggesting that shared motifs in enhancers is predictive of their target genes' co-expression.

Next we tested whether targets of correlated enhancers are functionally related. For each enhancer pair, we checked whether target genes, if they are different, share a Gene Ontology (GO) biological process. We only considered specific GO terms including at most 200 genes (this threshold was varied from 200 to 2000). We found that the foreground enhancer pairs consistently share a GO term more frequently than the background; the difference between them varying between 11% and 30%. This difference

is significant (Fisher Exact test p-value < 0.05) for all but one thresholds where it was marginally significant with p-value = 0.06. This suggests that gene targets of correlated enhancer pairs tend to be functionally related.

Targets of correlated enhancer clusters have correlated expression and shared function

We extended our analyses previous sections to 'clusters' of correlated enhancers. We identified clusters of five or more enhancers that were mutually correlated (various thresholds from 0.2 to 0.5 were used), while enriched for at least one of the previously identified significantly enriched motif cluster. For each enhancer cluster a control cluster was created from non-correlated enhancers that mirrored the former's size and genomic footprint (i.e. intra-cluster genomic distances). As was true for correlated enhancer pairs, putative targets of correlated clusters (*i.e.*, the set of genes nearest to each enhancer), were more highly correlated in their normalized RNA-seq transcript counts than were background clusters. For each triplet of thresholds for (i) minimum cluster size (5-20), (ii) minimum pairwise *I* (0.2-0.5) within a cluster, and (iii) minimum fraction of cluster members (0.7-0.8) harboring the most enriched meta-motif, the genes targeted by enhancers in clusters had higher Spearman correlation of transcription levels than the matching set of background enhancer clusters. For each parameter triplet, we compared the foreground and background for mean pair-wise correlation of expression within clusters. For the entire range of parameters, mean expression correlation within foreground clusters was consistently greater that for corresponding expression correlations within background clusters. Due to the variability in cluster counts for different parameters, p-values ranged from 0.02 to 4.1e-15 (Wilcoxon rank-sum test).

These results suggest that gene targets of correlated enhancer clusters with shared motifs are co-expressed and presumably co-regulated.

Next we assessed enrichment of GO biological processes amongst the targets of an enhancer cluster using R's GOstats package. Enhancer clusters also revealed consistently greater GO functional enrichment than the background clusters. Across 10 parameter settings, the ratio of enriched GO terms (at FDR 0.01) per foreground cluster to enriched GO terms per background cluster ranges from 1.3-fold to 4.8-fold. On average, there is almost 3-fold higher GO term enrichment in the foreground (19.1 terms per cluster). When the FDR threshold is set to ~0 (i.e., $p < 1e-8$), there is 5-fold higher enrichment, on average, in the foreground (7.5 terms per cluster). As an example, for the parameter setting with the greatest fold enrichment of GO terms, the enriched terms are shown, separated by cluster, in Appendix Table 3. These terms are consistently revealed across all parameters settings. Together, the GO enrichment and gene expression results illustrate that co-expression of genes with shared function is coordinately regulated across tissues by enhancers that share motifs and are epigenetically correlated across the same tissues.

Concordant cell type specificity of enhancer clusters and their target genes

Enhancers are believed to regulate cell type specific gene expression. We tested whether there is cell-specificity among the gene targets of correlated enhancers. For identifying cell type specificity of gene expression, we used the online tool CTen (35), which compares input genes to a database of highly expressed cell-specific genes found in public microarray databases, and reports any significant overlaps. Enhancer clusters and associated target genes were identified with three parameter settings resulting in 42, 122, and 182 clusters, with average cluster sizes 64, 31, and 19 genes respectively.

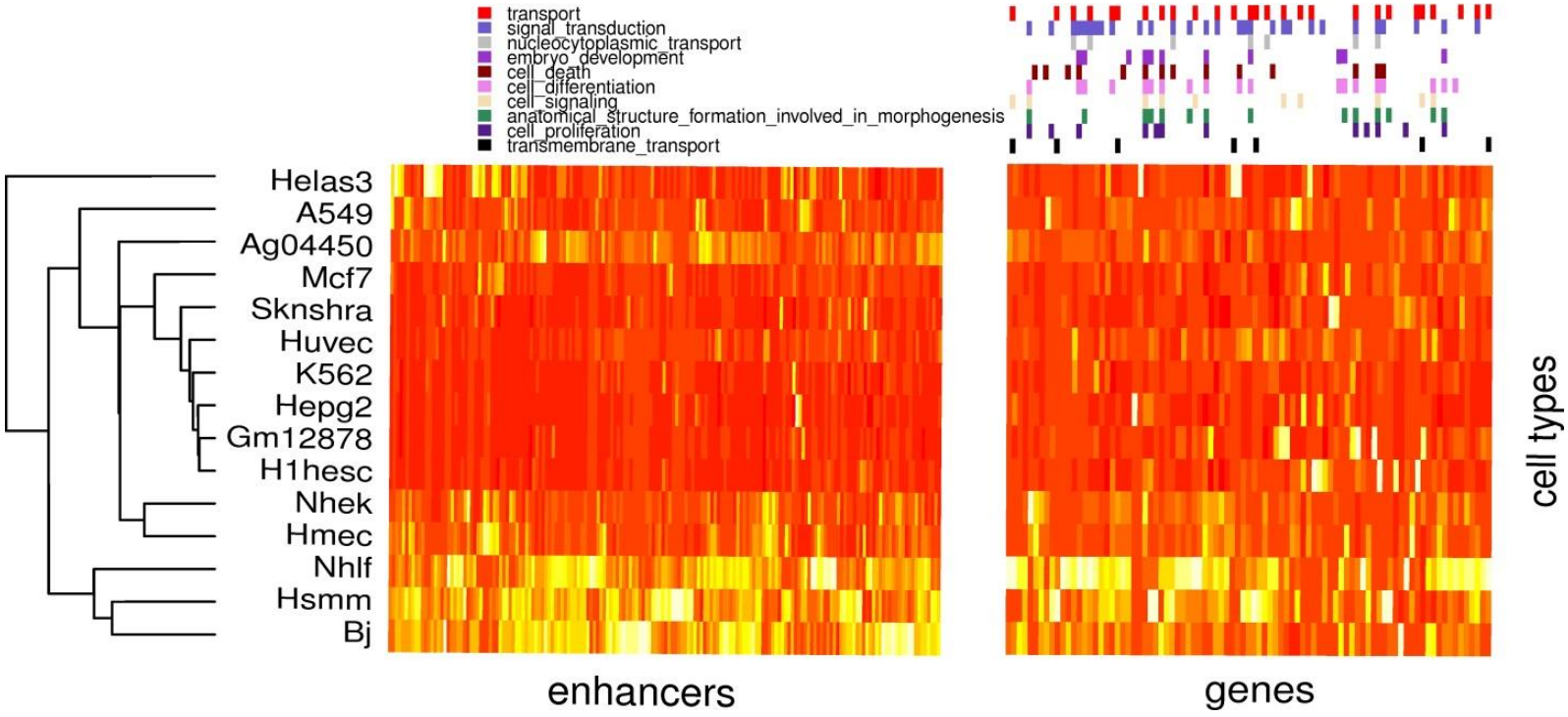
Background gene sets were obtained as in previous section. Our results indicated high tissue enrichment in the gene targets of correlated enhancer clusters. For instance, with 42 clusters, we found enrichment (FDR = 1%) for 23 tissue-specific gene sets involving 16 clusters while no enrichment was detected in the corresponding background clusters; results are qualitatively similar for other parameter settings.

Next we hypothesized that if the genes targeted by an enhancer cluster are expressed in specific cell types then the enhancers in the cluster should have high DHS in the same cell type(s). We determined the average DHS of an enhancer cluster in ENCODE cell types and obtained the DHS-based rank of the cell type in which the corresponding gene cluster was specifically expressed according to CTen; mapping between CTen tissue types and ENCODE cell types was manually determined and organized into classes (Appendix Table 4). For a clustering parameter, we obtained the median rank for the resulting enhancer clusters as well as median rank for an equivalent set of background clusters. We found that across 8 different clusterings the median ranks of enhancer clusters ranged from 4 to 8 with a mean of 6, whereas the expected median rank is 11.5. Overall, this result suggests that there is, indeed, concordance between enhancer clusters and targeted gene clusters in their tissue-specific activity.

Figure 2-9 shows an illustrative example of an enhancer cluster (179 enhancers) and corresponding gene cluster (98 genes) with tissue specific activities across 15 cell types. The DHS profiles of the enhancers (Figure 2-9, left panel) mirror the expression profiles of the genes (Figure 2-9, right panel). These genes are highly expressed in a number of cancer cell lines and an embryonic stem cell line, combined with markedly lower expression in normal adult somatic cells and are highly enriched for terms related

to intra- and inter-cellular signal processing, and regulation of transcription (Appendix Table 5).

Figure 2-9: Tissue activity profile of an enhancer cluster and the corresponding target genes. Left Panel: The tissue-specific DHS activity for 179 coordinately activated enhancers. The data is show only for 15 cell types for which RNA-seq data is also available. Rows (enhancers) and hierarchically clustered. **Right Panel:** Corresponding expression of the 98 target genes in the same 15 cell types. The gene symbol and a representative GO term for the gene are given to the right of each row. Gene rows have been clustered independently, however, column order is preserved from the enhancer heatmap above. In both maps, deeper shades of color indicate higher values.



Discussion

Based on a systematic analysis of correlated enhancer activities across 72 cell types we found a broad range of evidence that support coordinated enhancer activities, potentially mediated by transcription factors, chromatin modification enzymes, and spatial chromatin structure. Our analyses are based on stringent controls at various stages to maximize the robustness of our conclusions. First, we explicitly control for observed autocorrelation along the genome in DHS levels, which would otherwise inappropriately make neighboring enhancers seem correlated. Second, when appropriate we remove transitive correlations between enhancers. Third, when analyzing a group of enhancer pairs we create an appropriate negative control by selecting uncorrelated enhancer pairs with similar inter-enhancer distances. Fourth, to control for cell type similarities, 37 representative cell types were selected from 72 cell types. Fifth, significantly co-occurring motifs in enhancer pairs were screened for high likelihood of active tissue-specific TF binding. Sixth, dependencies due to motif similarity were addressed by clustering motifs. Seventh, clustering parameters settings that included cutoff for mutual information, minimum size, and minimum level of motif enrichment, were varied to ensure robustness of pattern discovery at the network level. For individual analyses additional controls were employed to ensure robustness of our conclusions.

P300 binding has been shown to be an accurate marker of tissue relevant enhancers (5). The base set of 98,000 enhancers was identified based on P300 binding in one of the 4 cell types. P300 binding is a reasonable marker of candidate enhancer for the intended aim of our work, namely, to investigate coordinated enhancer activities and test hypotheses concerning its functional underpinning and consequences. Although there are

alternative ways of identifying the candidate enhancers, such as ChromHMM (31), the combination of DHS and 5C (34), and other epigenomic marks (7), they all can have false positives. Moreover, using DHS as a proxy for an enhancer's tissue-specific activity allowed us to take advantage of the many tissues for which DHS data is currently available, without introducing circular dependence. Even though individual enhancers may be false positives, we infer correlated activity based on highly significant DHS correlation across 37 independent cell types after controlling for potential autocorrelation. Despite noise at the level of individual enhancers, we observe significant patterns when comparing enhancers with coordinated activities with background enhancer pairs, which notably are derived from the same set of enhancers. Approximately 53% of our enhancers overlap with those predicted by ChromHMM. To further ensure the robustness of our conclusions, we repeated some of our analyses separately on the subset of enhancers supported by ChromHMM and the ones not predicted by ChromHMM. In both disjoint datasets, we still observed that correlated enhancers had significant motif co-occurrence, and that the potential targets of correlated enhancers were significantly correlated in their expression and function.

The goal of identifying the full complement of enhancers that drive transcriptional regulation in a specific context remains largely unmet. This work suggests a useful paradigm for organizing enhancers into clusters of coordinated activities. These clusters of enhancers, given their high cross-tissue concordance in epigenetic state, are likely to participate in coordinate transcription regulation of specific genes, or more likely, pathways. Presently, researchers treat enhancers and their gene targets predominantly as

independent edges in a graph. By leveraging prior knowledge of these clusters, searches for enhancer-target genes will benefit from both greater sensitivity and greater specificity.

In addition to finding clusters of enhancers ostensibly involved in coordinate regulation of gene transcription, we also examined the nature of the clusters. We asked, for example, whether there was a pattern in clusters with regard to enhancer strength, as manifest in the expression level of target genes. We found that strong enhancers are much more likely to function in isolation than are weak enhancers. Moreover, strong and weak enhancers assort with enhancers of the same kind: strong (weak) enhancers prefer to interact with strong (weak) enhancers.

TF binding motifs can exert influence on enhancer activity. We found that shared motifs can predict correlated activities of a pair of enhancers. Even though, there is no qualitative difference in density and composition of motifs between enhancers that are involved in coordinate regulation and enhancers that are not, certain motifs preferentially co-occur in correlated enhancers. This could be explained if enhancers with shared motifs respond in unison to a common modulator, such as an allosterically regulated TF, or a pioneer TF that can interact with and recruit CMEs. Indeed, we found that co-occurring motifs do preferentially interact with a subset of CMEs.

We found that correlated enhancers that are in genomic proximity share fewer significantly co-occurring motifs relative to those that are far apart (Table 3b). This, in conjunction with a greater propensity for coordinated activity for nearby enhancers (Figure 2-3), suggests alternative mechanisms for proximal and distal enhancer pairs' coordinated activities. Greater motif sharing between distant enhancer pairs is consistent

with a more active role of motifs in establishing coordinated activity, with or without influencing spatial proximity.

Overall, our analysis suggests that mirroring the known organization of genes into functionally linked co-expressed modules, distal enhancers regulating such genes are also organized into modules of correlated activity across cell types. Strong and weak enhancers exhibit differential correlated activity and assortativity with strong and weak enhancers, respectively. The observed organization of mammalian enhancers into correlated networks is likely mediated by the joint action of TFs through shared motifs, chromatin modification enzymes, and spatial chromatin structure.

Material and Methods

P300 and DHS Data overview:

P300 binding has been shown to be a reliable marker of tissue specific enhancers (Visel et al., 2009). As a starting set of candidate enhancers we extracted from Gene Expression Omnibus (GEO) (Barrett et al., 2010) the genomic regions bound by P300 in at least one of the 4 cell types – HepG2 (GEO accession Id GSM758575), GM12878 (GEO Id GSM803387), H1-HESC (GEO Id GSM803542) and SK-N-SH_RA (GEO Id GSM803495). For each of the 4 datasets, we extracted the P300 peaks and, in case of overlaps, used the center of merged overlapping regions. We thus obtained 98,353 enhancer regions, with an average length of 500 bps centered at the center of the P300 peaks, less than 5% (7%) of which overlap with 2kb (5kb) upstream of annotated ENSEMBL transcripts. From the ENCODE database (Bernstein et al., 2012), we extracted the genome-wide DHS broad peak data for each of the 72 tissue types

represented; for tissue types with more than one data set available, we chose the set with the greatest number of peaks. For each enhancer, with respect to each tissue, DHS was set to 1 if the 500 bp enhancer region overlapped a DHS peak; otherwise it was set to 0. This procedure yielded a 98,353 x 72 binary matrix, with rows corresponding to enhancers, columns to tissue (or cell) types, and matrix entries reflecting the 'activity state' of an enhancer in a tissue. In order to minimize dependencies, tissues were clustered based on similarity, into 37 clusters, including 25 singletons (Appendix Table 1) and only the most representative tissue from each cluster was retained for further analyses. Accordingly, the DHS matrix was reduced from 72 columns to 37.

Mutual Information:

Mutual information between two binary vectors X and Y is defined as

$$MI(X,Y) = \sum_{x\in 0,1} \sum_{y\in 0,1} p(x,y) log \frac{p(x,y)}{p(x)p(y)},$$

where p(x) is the probability of x in X, p(y) is probability of y in Y and p(x,y) is the joint probability that x and y co-occur in vectors X and Y. Informally, mutual information quantifies how much knowing one of the two vectors helps determine the other. Relative advantages of using mutual information over other measures such as correlation have been discussed previously, e.g., (21).

Controlling for DHS autocorrelation:

We controlled for the observed cell type-specific DHS autocorrelation to detect significantly correlated enhancer pairs (Figure 2-1).Separately for each of the 37 cell types, based on 100,000 random genomic segments, we estimated the autocorrelation probability of DHS at a location conditional on DHS at another location at specific

distance-range (or, distance-bin). In particular, given a cell type, enhancer **X** and enhancer **Y** at distance-bin *d* from **X**, we estimate the probability that **Y** is DHS conditional on DHS status of X. This tissue-specific and distance-specific autocorrelation probability was then used to create a 'synthetic' enhancer pair corresponding to each of the actual enhancer pairs. Each synthetic pair consists of the DHS vector for one member of the actual pair and a randomly generated vector of 37 binary DHS values replacing the other member (Figure 2-1). The autocorrelation conditional probabilities estimated above are used to generate the synthetic vector, conditioned on cell type and distance bin. As a consequence, DHS data for synthetic pairs preserves for each tissue type both the mean DHS and extent of autocorrelation observed in the real genome, resulting in a *MI* profile that is virtually identical to that of random genomic segment pairs (Figure 2-2).

TF binding site identification:

For each enhancer sequence and each of the 981 positional weight matrix (PWM) for vertebrate transcription factors in TRANSFAC database (Matys, 2003), we used our previously published tool (Levy and Hannenhalli, 2002) to identify binding sites based on a score threshold of 95th percentile. For each enhancer only presence/absence of a motif was noted.

Motif co-occurrence score:

We quantified the tendency of each motif to co-occur in correlated pairs of enhancers relative to its expected co-occurrence frequency, assuming independent occurrence of motifs among enhancers. If p represents the fraction of enhancers in which a motif occurs then assuming independence the motif is expected to co-occur in p^2 of the enhancer pairs.

The motif co-occurrence score is defined as the ratio of the observed co-occurrence frequency and the expected frequency p^2.

Removing dependencies among pairs:

In both the foreground and the background, transitive dependencies were removed; enhancer pairs were excluded if either of the enhancers was part of a previously included pair. In addition, we ensured that the distribution of inter-enhancer distances was identical for the foreground and the background.

Motif clustering:

Motifs were clustered based on similarity due to structural similarities between the corresponding TFs. All pairwise motif similarity scores for the 981 vertebrate motifs were obtained from the author of STAMP too (Mahony et al., 2005). Using pairwise similarity, the motifs were hierarchically clustered using the *'hierarchy'* module in SciPy's *cluster* package (www.scipy.org) for Python based on Euclidean distance and complete linkage. The resulting tree was trimmed using the module's 'fcluster' function with a maximum co-phenetic distance criterion that produced 142 disjoint clusters.

Tissue clustering:

We computed the pairwise similarity between tissues based on their genome-wide DHS profiles for all enhancers. We used the *linkage* method in Scipy's *hierarchy.cluster* class to perform hierarchical clustering based on average linkage in combination with Russell-Rao pairwise distance (i.e., the fraction of enhancers with a DHS state of 1 in the two tissues). The resulting tree was trimmed using the class's *fcluster* method and with an inconsistency criterion that resulted in 37 clusters, including 25 singletons. In each cluster of size 3 or larger, the tissue with the lowest mean distance to other cluster members was

retained, while in clusters of size 2, it was the tissue with the greatest mean separation from all other tissues in the sample.

Determination of concordance between enhancer cluster's and target gene cluster's tissue-specific activity:

We clustered the 84 tissue types in the CTen database and the 72 cell/tissue types in the DHS database into 34 and 23 cytologically motivated classes, respectively. (Class sizes ranged from 1 to 19 (brain) for CTen tissues and 1 to 15 (endothelium and blood) for DHS cell types). Agreement in tissue specific activity was assessed based on the 17 classes shared between the two domains; tissues falling outside of these classes were not considered. For each target gene cluster we first identified the tissue in which the genes exhibit tissue-specific activity according to CTen (FDR 0.01). Then we obtained the corresponding tissue class in the DHS dataset and determined the rank of that tissue class for the corresponding enhancer cluster activity as follows. For an enhancer cluster, and for each tissue class, we determine the ratio between (i) the fraction of enhancers in the particular cluster having DHS in that tissue class and (ii) the fraction of 'all' enhancers with DHS in that tissue class. We then use this tissue-specific fold enrichment to rank all 23 tissue classes. We are interested in the rank of the specific tissue class in which the corresponding genes had robust and specific activity according to CTen. We thus obtain a rank for each cluster and we determined the median rank among all clusters in a clustering. We applied 8 different clustering parameters and for each clustering obtained the median rank for the actual clusters as well as for randomly generated background clusters with same size. Finally we compared the median ranks for the foreground and background clusters using paired Wilcoxon test.

Chapter 3: **Crowdsourcing: spatial clustering of low-affinity binding sites amplifies *in vivo* transcription factor occupancy**

Abstract

To predict *in vivo* occupancy of a transcription factor (TF), current models consider only the immediate genomic context of a putative binding site (BS) – impact of the site's spatial chromatin context is not known. Using clusters of spatially proximal enhancers, or archipelagos, and DNase footprints and ChIP-Seq to quantify TF occupancy, we report for the first time an emergent group-level effect on occupancy, whereby BS within an archipelago experience greater *in vivo* occupancy than rigorously matched BS outside archipelagos. A TF's occupancy boost in an archipelago is tissue-specific and scales robustly with the total number of archipelago BS for the TF. We explain these results through biophysical modelling, which suggests that a collective of spatially proximal homotypic BS briefly 'trap' a TF inside an archipelago, thereby inducing boosts in local TF concentration and occupancy. Together, we demonstrate for the first time, consistent with a facilitated TF diffusion model, synergism among genomically remote but spatially proximal homotypic BS. We propose that by leveraging three-dimensional chromatin structure and TF availability, weak yet abundant archipelago binding sites *crowdsource* their own occupancy context-specifically.

Introduction

Eukaryotic transcriptional regulation is critically mediated by the binding of specific transcription factors (TF) to their cognate DNA binding sites in the genome (Spitz and Furlong, 2012). A TF's *in vivo* DNA binding varies dramatically over developmental time and across tissues (Plank and Dean, 2014; Yáñez-Cuna et al., 2012), and as such, a TF's *in vitro* binding preference, or motif, does not accurately predict its *in vivo* binding (Yáñez-Cuna et al., 2012; Zinzen et al., 2009). Thus, a TF's DNA binding motif suffers from being, both insufficiently informative to precisely specify binding in the large genomic substrate and insensitive to the *in vivo* environment, making it essential to characterize additional determinants of *in vivo* TF-DNA binding (Heinz et al., 2013; Moses et al., 2004).

Spatio-temporal variation in TF binding has been shown to be, in part, mediated by the local chromatin state of a binding site (BS) (Hesselberth et al., 2009). High nucleosomal density is typically unfavorable to TF binding (Jiang and Pugh, 2009). Recent work has highlighted three additional features of *in vivo* binding: (1) GC content in the flanking region that resembles the GC content of the putative target site (Dror et al., 2015; White et al., 2013), (2) cooperative binding (Smith et al., 2013; Yáñez-Cuna et al., 2012) and (3) genomic clusters of homotypic BS for a common TF, or HCTs (Ezer et al., 2014a; Gotea et al., 2010). These three features have been shown to be enriched in gene promoters and distal enhancers and to contribute to functional *in vivo* binding leading to transcriptional activation (Arvey et al., 2012; Gotea et al., 2010; Sharon et al., 2012; White et al.,

2013). Still, most BS predicted by current models are not bound *in vivo* (Arvey et al., 2012; Moses et al., 2004; Slattery et al., 2014).

To date, research on determinants of functional TF binding have focused on a putative BS and its proximal genomic context, as described above. In parallel, the three-dimensional organization of the genome has emerged as an important mediator of transcriptional regulation, where, as opposed to genomic proximity, spatial proximity is determinative (Babaei et al., 2015; Filippova et al., 2013; Fullwood et al., 2009; Ing-simmons et al., 2014). Chromatin looping can bring into proximity functionally related genes and their genomically distal regulatory regions (Fraser, 2006; Fullwood et al., 2009; Li et al., 2012; Lieberman-Aiden et al., 2009; Schwarzer and Spitz, 2014). In vertebrates, for example, Hox genes, globin genes, and olfactory receptors, along with their distal enhancers, adopt a spatially clustered conformation, termed as 'regulatory Archipelago' (AP), as a prerequisite for robust transcriptional activation (Markenscoff-Papadimitriou et al., 2014; Montavon and Duboule, 2012; Schoenfelder et al., 2010a; Schwarzer and Spitz, 2014; Vernimmen, 2014). Despite mounting evidence supporting functional criticality of chromatin interactions in context-specific transcriptional regulation, the potential impact of spatial clustering of BS on their individual TF occupancy has not been investigated. Recent findings that spatially clustered enhancers (we borrow the term 'archipelago' to refer to such spatially clustered enhancers) often share BS for the same TF, i.e., homotypic sites (Taher et al., 2013; Malin et al., 2013) make such enquiry even more compelling. Notably, these findings echo observations in

enhancer-rich regions of the genome known as super enhancers where BS cognate to key lineage determining TFs have been found to be enriched (Whyte et al., 2013), and three-dimensional interactions among the constituent enhancers unusually frequent (Heinz et al., 2015). Interestingly, super enhancers display extremely high cell type-specific occupancy of certain TFs (Whyte et al., 2013), however the mechanism underlying this is not well characterized (Andersson et al., 2015).

In what follows, it's crucial to distinguish binding affinity of a TF for a BS, which is typically assessed *in vitro,* from TF occupancy at a BS, which is an *in vivo* state and depends on additional factors – most directly, TF concentration (Foat et al., 2006). Importantly, TF concentration and, hence, TF occupancy, may be distributed non-uniformly in the nuclear space (Chakalova and Fraser, 2010; Schoenfelder et al., 2010a). Indeed, as described by facilitated TF diffusion, BS for a common TF in a HCT may act together to briefly 'trap' a TF into diffusing back and forth amongst themselves along the chromatin (Brackley et al., 2012; Ezer et al., 2014a, 2014b), resulting in higher-than-expected occupancy in the HCT. This explains how a genomic HCT synergistically impacts *in vivo* binding at individual BS within the cluster (Ezer et al., 2014a; He et al., 2012). Critically, here, we generalize the notion of "genomic" HCT to investigate the impact of "*spatial"* HCT – that is, spatially clustered but genomically distant BS for a mutual TF – on the *in vivo* occupancy at individual BS in the cluster.

Based on clusters of spatially proximal enhancers, or *AP*s (Malin et al 2013, Sheffield et al 2013), and using nucleotide-resolution DNase footprints as

well as ChIP-Seq data to quantify context-specific *in vivo* TF occupancy (Neph et al., 2012b), we demonstrate a strong group-level effect on TF occupancy whereby individual BS within an AP experience greater *in vivo* occupancy than their counterparts outside APs, i.e., enhancers that are not in spatial proximity with other enhancers, although their local genomic contexts have been carefully matched to their AP counterparts for motif composition and chromatin accessibility. We refer to the differential occupancy in AP enhancer BS relative to the controlled non-AP enhancer BS as '*occupancy boost*'. Strikingly, occupancy boost for a TF in an AP scales robustly with the number of putative BS in the AP, suggesting a strong synergistic impact of spatial HCT on TF occupancy. TFs with degenerate motifs, which are expected to have abundant putative BS, are consistently among the TFs experiencing the greatest occupancy boosts; in large APs, mean occupancy boosts for homotypic BS corresponding to degenerate motifs are between 2 and 3-fold.

Based on these results, we propose that *in vivo* occupancy at particular BS in an AP is amplified by the presence of homotypic BS in spatial proximity, i.e., BS '*crowdsource'* their own occupancy boost along with other homotypic BS in their spatial proximity. We extend the previous biophysical model of facilitated diffusion of TFs explaining the occupancy boost in a genomic HCT to explain spatial HCTs. Our model shows, with striking concordance, that the observed occupancy boost in spatial HCTs can result from TFs briefly 'trapped' into diffusing among multiple spatially proximal BS. In sum, our study shows, for the first time, how hundreds of weak BS, spanning megabases, can leverage

chromatin structure to dramatically boost their own occupancy context-specifically, and in turn, induce higher-order transcriptional changes.

Results

Data and Analysis overview

Archipelagos. Our analysis is based on previously identified enhancer clusters (Malin et al., 2013) comprising ~1600 enhancers in 40 clusters. Enhancers were clustered based on correlated DNase hypersensitivity (DHS) profiles across 37 cell lines (representing 82 cell lines). Enhancers in the same cluster were shown to (i) have functionally related gene neighbors with correlated expression, indicative of coordinated regulation, (ii) share BS for several TFs, and (iii) be spatially proximal to one another. We will refer to such enhancer clusters as 'archipelagos' (APs) borrowing from (Spitz and Furlong, 2012). We refined the APs identified in (Malin et al., 2013) to ensure tight spatial proximity among AP enhancers (see Methods). Note that properties (ii) and (iii) above together imply a higher spatial density of homotypic BS within an AP, particularly for TFs with degenerate motifs, which typically have abundant putative BS (Figures 3-1, 3-2); we quantify a motif's degeneracy by its *relative entropy* (*RE*) (see Methods). For additional validation, key tests were repeated using an alternative set of previously published APs (Sheffield et al., 2013).

Figure 3-1. Spatial homotypic clusters. The combination of spatial proximity and genomic homotypic clusters of TFBS produce high homotypic TF BS concentration. As illustrated, low-RE (degenerate) motif BS have a higher expected frequency in the genome than high-RE motif BS, including more frequent HCTs. In a spatially proximal chromatin context, effective homotypic BS concentrations are particularly elevated for low-RE motif BS. This effect is further accentuated in archipelagos of enhancers, which have been shown to be enriched for HCTs for shared TFs. High effective homotypic BS

concentration is likely a pre-requisite for the crowdsourcing effect. Large ovals denote archipelagos of functionally related enhancers and target genes. Darkness of background color approximates the maximum expected homotypic BS concentration. Not drawn to scale. Green: DNA. Black: BS. BS=binding site; RE=relative entropy; HCT = homotypic (genomic) cluster of TFBS.

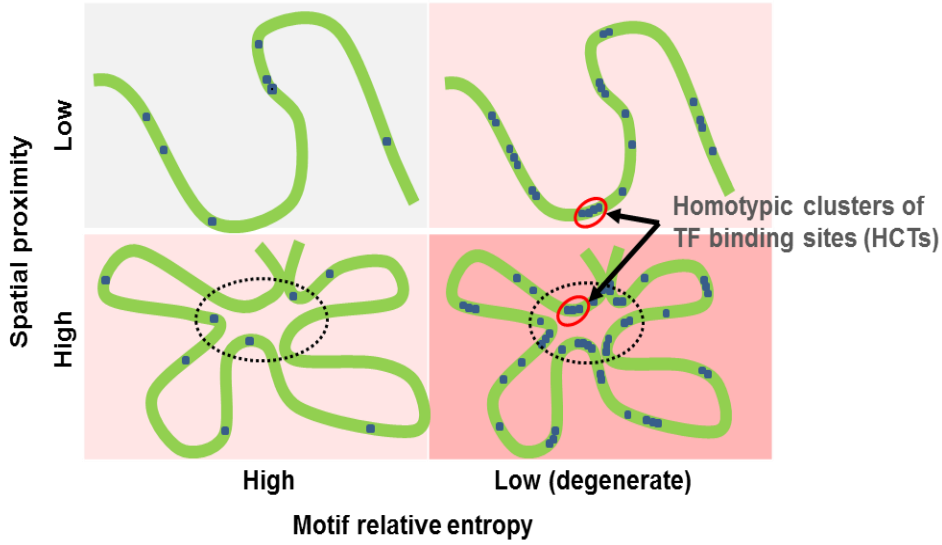
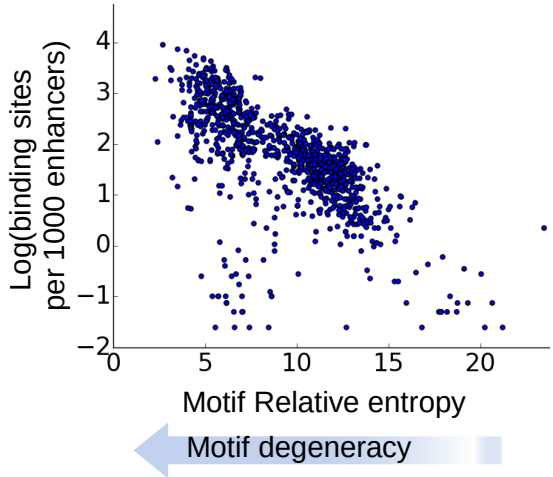


Figure 3-2. TF motif degeneracy is positively associated with frequency of its putative BS in the genome. Degeneracy for each of ~2500 TRANSFAC TF motifs (i.e. position weight matrices) was estimated by its RE (x-axis). Putative BS were identified and tallied in ~40K background (non-AP) enhancers, having mean length ~ 500bp. Putative BS for a TF mapping to multiple motifs were pooled and plotted against the RE of the motif with the lowest. BS: binding site(s), AP: archipelago, RE: relative entropy, TF: transcription factor.



In vivo occupancy. For a putative BS, an initial estimate of its *in vivo* occupancy was determined using high-resolution curated cell type-specific DNase footprint data (Neph et al., 2012b) as well as, data permitting, ChIP-Seq data from ENCODE (see Methods). When using footprint data, we applied highly stringent criteria to ascribe the footprint to a specific TF, similar to (Neph et al., 2012b), while accounting for multiple motifs mapped to a TF (Methods).

Non-archipelago control enhancers. Recognizing the inherent technical challenges in inferring occupancy, especially from footprint data, "raw" estimated AP occupancies were not compared directly with each other. Instead, we quantified occupancy in each AP enhancer, for a given TF, in relation to occupancy in a stringently matched '*non-AP*' enhancer, in the same tissue. The non-AP control enhancers are not spatially clustered (Malin et al 2013), but are otherwise carefully matched with the AP enhancers for each TF in terms of motif composition (motif number and kind) and chromatin accessibility (see Methods). All our results, therefore, marginalize out the contribution of genomic homotypic clusters, while also preempting technical biases due to motif-specific differences in occupancy detection. Additional analyses obviate the need for the non-AP background by comparing an AP BS's occupancy across cell types.

Organization of the Results. We have organized our results into four sections as follows. **(1)** We first establish our central hypothesis - a TF's *in vivo* DNA occupancy in an AP is 'boosted', relative to '*non-AP*' control enhancers, and the occupancy boost robustly scales with the number of BS for the TF in the AP. **(2)** Given the apparent similarities between APs and super enhancers, we compare the

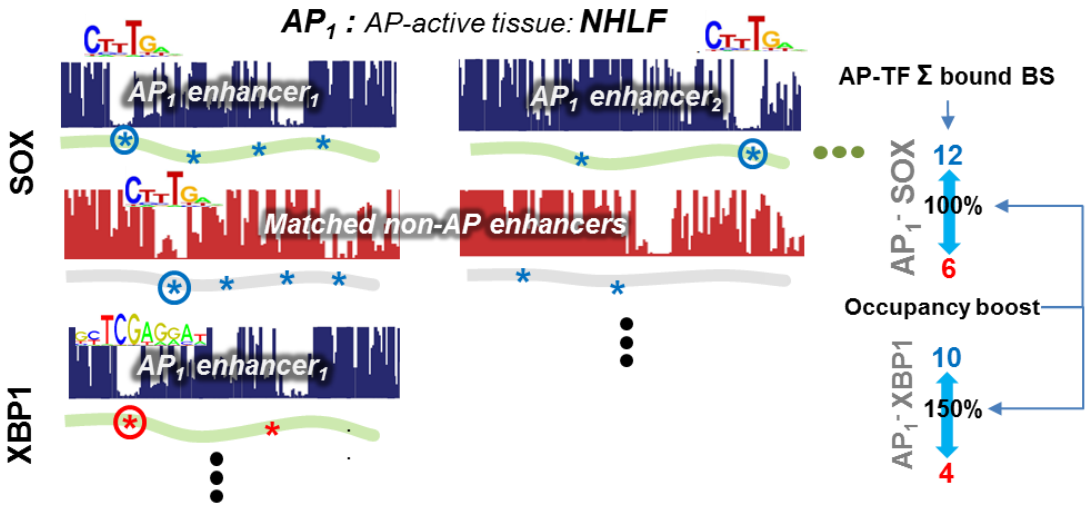
two and show that the occupancy boost in APs can act independently of super-enhancers, as well as independently of protein-protein cooperative binding **(3)** We show, via a biophysical model based on facilitated TF diffusion that the observed occupancy boost can be explained by "trapping" of the TF in a restricted nuclear space. **(4)** Thus far, *in vivo* occupancy of a TF and other functional analyses were primarily rendered in each AP's most active cell line – its so-called 'AP-active' tissue (see Methods). Here, we establish context-specificity of the occupancy boost by comparing the boosts in AP-active tissue with those in 'AP-inactive' tissues.

Occupancy boost at AP BS increases with homotypic BS density within AP, supporting crowdsourcing of *in vivo* TF occupancy

We tested our central hypothesis at the level of a TF-AP pair, in the AP-active cell line (Methods). For a given TF and AP, we calculated the TF's *coverage* as the total number of its cognate BS in the AP, and calculated its *occupancy boost* as the difference in occupancy between AP BS and BS in matched non-AP enhancers, normalized by the latter (Methods; Figure 3-3); for instance, an occupancy boost of 100% corresponds to a 2-fold difference. In comparing mean occupancy boosts of distinct TF-AP pair classes, then, we effectively compare means of AP occupancies normalized by matched non-AP pairs. Because background levels of BS occupancy in the genome are generally low (3-5%), the occupancy is zero in both AP and control non-AP enhancer sets for a majority (65%) of the ~25k TF-AP pairs; these pairs were excluded for this analysis. Of the remaining TF-AP pairs, ~3.6k have non-zero occupancy in both AP and non-AP, encompassing ~95K enhancer-TF pairs and ~205K BS (we call this the *reciprocal* set), and

additional ~5k TF-AP have non-zero occupancy in either AP or in matched non-AP BS (*non-reciprocal* set). We analyze the two sets of TF-AP pairs separately.

Figure 3-3. Calculating differential TF occupancy boost based on curated digital DNase footprint data. Shown is the procedure for calculating occupancy boost for each (AP, TF) pair. For each enhancer in an AP, and each TF with one or more putative BS in the enhancer, a non-AP enhancer is chosen (with replacement) after controlling for mean enhancer-wide chromatin accessibility (DHS) in the AP's most active tissue, and for the number of putative BS. For each TF-AP pair, then, occupancy boost is calculated as the percent difference in the number of putatively bound BS, where binding is determined in a binary manner: 1, if a curated footprint tightly overlaps a given motif instance, 0, otherwise. If multiple TF motifs tightly overlap a given footprint, conservatively, all are classified as bound. Putative BS are indicated by a '1', or '2', respectively, for example TFs SOX and XBP1. A circle around a BS signifies it is imputed as bound by its cognate TF. Note that the toy calculation of occupancy boost does not correspond to the data displayed. AP = archipelago; TF = transcription factor, BS = binding site. DNase digital footprint scans from Neph et al 2012.



We stratified the reciprocally occupied TF-APs into 8 bins with exponentially increasing coverage cutoffs and calculated the overall occupancy boost for each

bin as the mean occupancy boost among member TF-APs. As shown in Figure 3-4A, the occupancy boost robustly increases with the TF coverage in the AP. Specifically, we found a substantial difference in occupancy boost between TF-APs with the highest and lowest 50% coverage (mean of 77.7 % versus 2.1 %; Wilcoxon p-value = 1.4e-5). This trend also holds when coverage was alternatively quantified as the number of enhancers in an AP with at least one BS for the TF (Figure 3-4B), suggesting that the boost is not due to disproportionate contribution from a few enhancers, but instead relies on widely dispersed BS across the AP's enhancers. Interestingly, the boosts for high coverage TF-APs increase when the digital footprint binding criterion for assessing occupancy is made more stringent (Figure 3-5). This highlights the robustness of occupancy estimation, as well as the fidelity of our experimental design. As an alternative measure of occupancy, at the enhancer-wide scale, we used ChIP-Seq data in an independent set of 9 tissues. Despite drastically fewer potentially bound sites analyzed (on average, ~30 fold fewer TFs per cell type), we observed a highly consistent and significant trend (Figures 3-6A, 3-6B).

Figure 3-4. A, B. Differential AP occupancy 'boost' scales with TF coverage in the AP. TF-AP combinations were sorted on the basis of coverage and mean occupancy boost was determined for each group of TF-APs, where occupancy boost refers to differential occupancy in AP and non-AP enhancers matched 1-to-1 for the TF's motif signature (the number and type of motifs) in a given enhancer, as well as for mean DHS across the AP. Occupancy was calculated based on the overlap of curated DNase digital footprints (Neph et al 2012) with high-confidence TRANSFAC motif instances. TF-APs and their non-AP counterparts were included in this analysis only if they both had non-zero occupancy (See also Figures S1, S2, S3, S4). Coverage was calculated as, alternatively, the number of cognate BS for a given TF in a given AP (A), or the number of enhancers with one or more cognate BS (B).

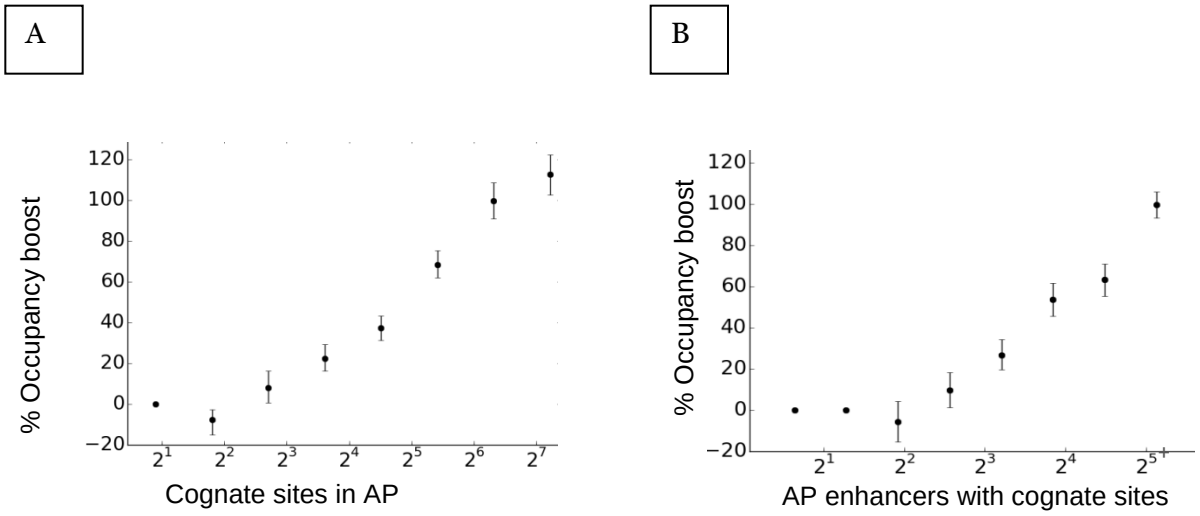


Figure 3-5. Occupancy boost trend improves with a more stringent digital footprint significance threshold, i.e. the 'FOS' (Footprint Occupancy Score) threshold (Neph et al 2012), for curation of high-resolution DNAse hypersensitivity reads. In the top and middle plot, relative lax thresholds of 0.90 and 0.75, respectively, are used, in contrast to the 0.6 threshold (bottom) used for all analyses performed in this work, including Figure 1. TF-AP pairs were binned by 'coverage' (x-axis), i.e. the number of cognate BS in a given AP for a given TF. Occupancy boost with respect to matched non-AP pairs shown on the y-axis. 95% confidence intervals are based on 50K bootstrap samples. . AP: archipelago, BS: binding site(s), TF: transcription factor.

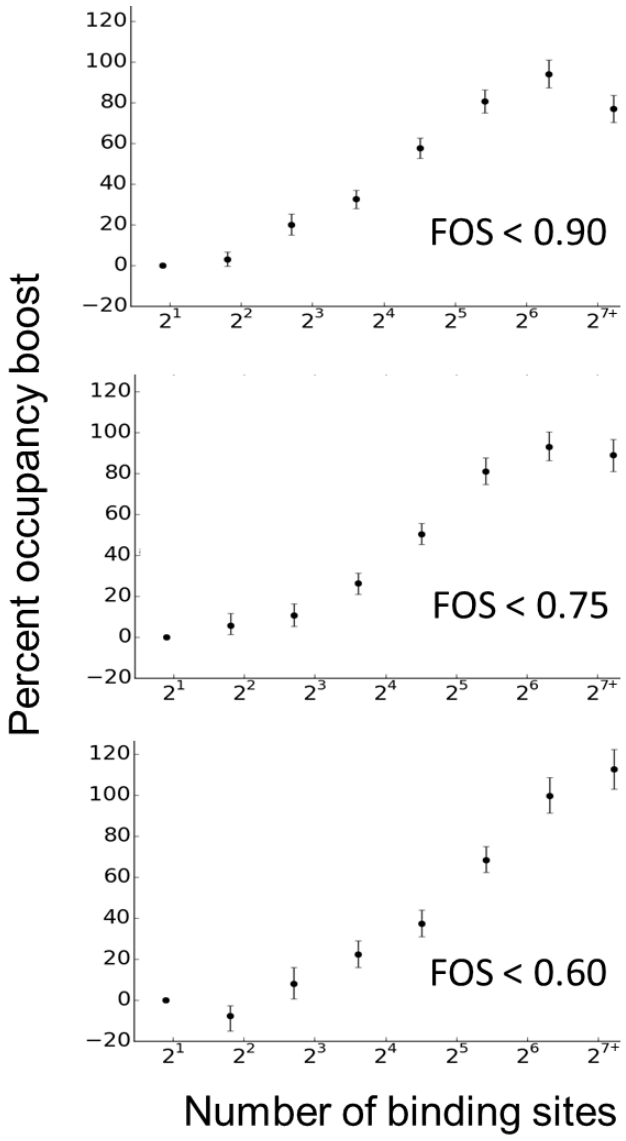
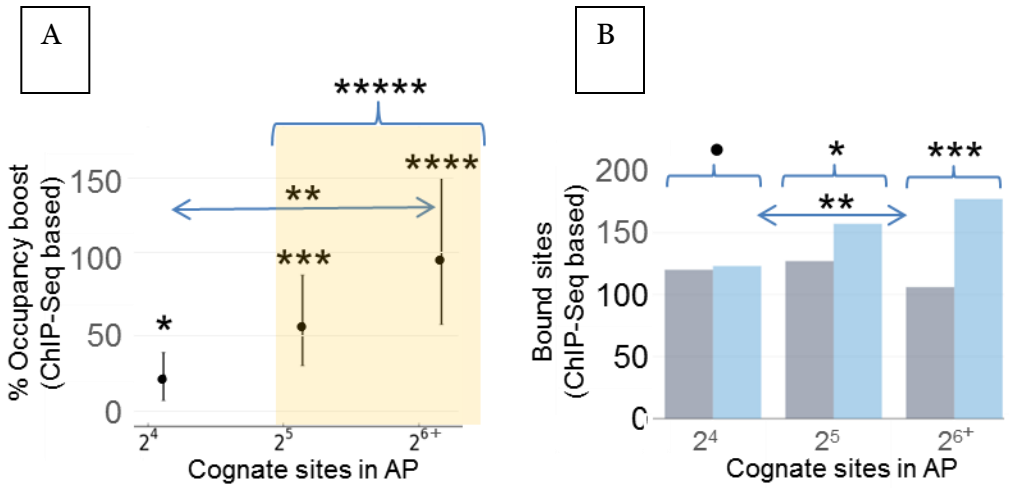
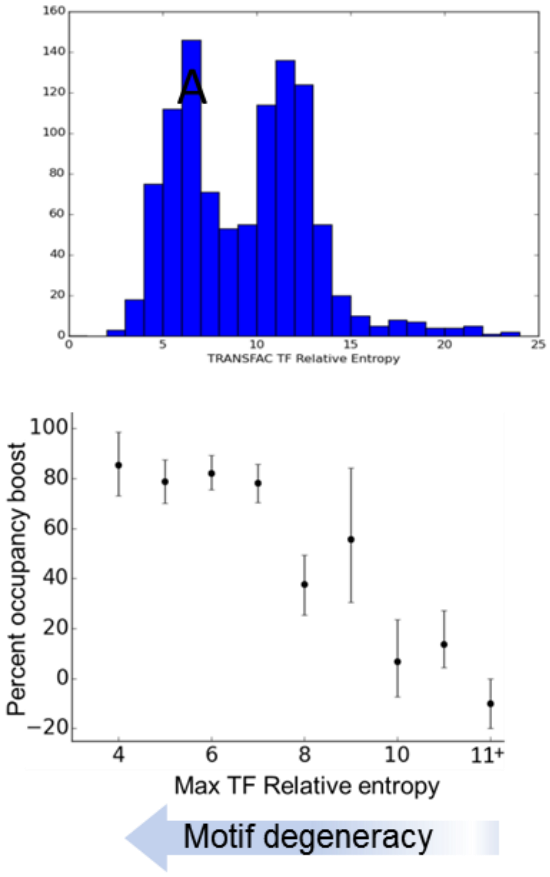


Figure 3-6. (A) Boost in per-enhancer occupancy for reciprocally occupied TF-AP pairs based on 206 ChIP-Seq experiments in 9 cell types. Horizontal arrow represents test comparing first and third coverage bins; remaining tests compare sampled boosts with the null expectation of zero. P-values and 95% confidence intervals computed with bootstrap procedures. ***** $p < 1 \times 10^{-6}$, **** $p < 1 \times 10^{-4}$, *** $p < 5 \times 10^{-4}$, ** $p \leq 1 \times 10^{-2}$, * $p < .05$. (B) Comparison of ChIP-Seq peaks recorded in AP and in non-AP control enhancers of reciprocally occupied TF-AP pairs. Shown are significance levels comparing coverage bins 1 to 3 (horizontal arrow), and bound vs unbound enhancers within each bin. P-values using Fisher Exact tests. *** $p < 5 \times 10^{-6}$, ** $p < 5 \times 10^{-3}$, * $p < 5 \times 10^{-2}$, • $p > 0.1$.



Abundance of a TF's cognate BS is strongly correlated with its motif degeneracy (Figure 3-2). Given this association, we also directly assessed the relationship between TF motif degeneracy and occupancy and found consistent trends (Figure 3-7). Taken together, the above analyses strongly suggest that binding sites for high coverage TFs experience a substantial occupancy boost in AP enhancers relative to BS in comparable non-AP enhancers.

Figure 3-7. Occupancy boost increases with greater TF motif degeneracy. Top: Distribution of RE for vertebrate TF motifs. Counts are shown for the ~1K TRANSFAC vertebrate TFs used in analysis. TFs with more than one identified motif were mapped to that motif having the lowest RE. Bottom: TF RE vs. occupancy boost. TFs partitioned into disjoint RE classes based on RE threshold. For each TF-AP pair, its 'occupancy boost' was estimated as the difference between its occupancy in AP and matched non-AP enhancers, normalized by its occupancy in non-AP enhancers. RE = relative entropy.



The overall TF coverage is affected by both the mean number of BS per AP enhancer ('homotypicity') and the number of enhancers per AP ('AP size'). Next, we assessed the relative contributions of these two constituents of coverage on the occupancy boost. As shown in Figure 3-8, for the reciprocal set, AP size and homotypicity independently and robustly impact the magnitude of occupancy boost (p-value = 4.2E-6). A similar analysis on 5K non-reciprocal TF-AP pairs shows a similar and significant trend (Figure 3-9; p-value 8.1E-5). There was insufficient ChIP-Seq occupancy data to analyze non-reciprocal TF-AP pairs separately, however, we continued to find a highly significant trend for ChIP-Seq-

derived occupancy boost when non-reciprocally and reciprocally bound TF-APs were pooled (p-value = 2.2E-4 Figure 3-10).

Results supported with alternative AP data set

For additional validation, we used alternative sets of AP enhancer clusters reported in (Sheffield et al, 2013). After processing the data to match closely to the data from Malin et al. described in the main text, we obtained 472 AP clusters averaging 15 enhancers per cluster, along with a pool of 18K 'nonAP' enhancers which were then matched to AP enhancers, as described (see Methods). Consistent with the results based on data from Malin et al., we observed a substantial difference ($p=1x10^{-23}$) between high and low coverage occupancy boosts (47% vs 8%, respectively) (Figure 3-11A *bottom*). After Lowess smoothing (with stats.model *Python* package) using default settings, mean boosts exceeded 100% for TF-AP pairs with the highest coverage (Figure 3-11B). Consistent with the crowdsourcing model, we also note that occupancy boosts for AP-TFs with the highest coverage were significantly higher after screening out enhancers in each AP with low mean spatial proximity to fellow AP members, based on Hi-C data from embryonic stem cell (for top 2% coverage, 47% vs. 29% boost $p= 5x10^{-4}$) (see Methods) (Figure 3-11A). This aligns with the established relationship between Hi-C scores and relative spatial distance (Lieberman-Aiden et al., 2009; Mifsud et al., 2015), and with the importance of spatial proximity to occupancy boost.

While these trends based on the Sheffield et al (2013) data are highly significant, the maximum boosts are approximately half of those observed with

the 40 APs from Malin et al. (2013). The most likely explanation centers on a key difference in the two approaches to identify correlated enhancers; Specifically, Malin et al. explicitly controlled for the genome-wide autocorrelation in tissue-specific activity (estimated by DHS), thus screening out many enhancer pairs with high correlation that was nominally due to their genomic proximity. The Sheffield approach did not control for autocorrelation, which results in higher sensitivity for detecting correlated activity, but is also likely to detect a large fraction of enhancer pairs due to their genomic proximity without true coordinate regulation. Nonetheless, their data offers independent evidence of occupancy boost for TFs with degenerate motifs in large APs.

Figure 3-8. Mean occupancy boost versus coverage that has been decomposed along two axes. Each TF-AP pair was binned based on the number of enhancers in an AP (column) and the mean number of BS per AP enhancer (row). Plots to the left of and below the heatplot show mean boost for each row and column, respectively. Red (green) heatmap cells indicate high (low) percentage occupancy boost after Lowess smoothing. Grey cells indicate no data. In the right panel, for all TF-AP pairs in the selected heatmap cell, significant digital DNase hypersensitivity footprints in member AP and matched non-AP enhancers are shown, where the numbers of BS for AP and non-AP enhancer-TF pairs are identical; a blue line indicates a significant footprint overlapping a putative BS. Enhancers are sorted from bottom to top in order of increasing chromatin accessibility. TF: transcription factor, BS: binding site(s), AP: archipelago

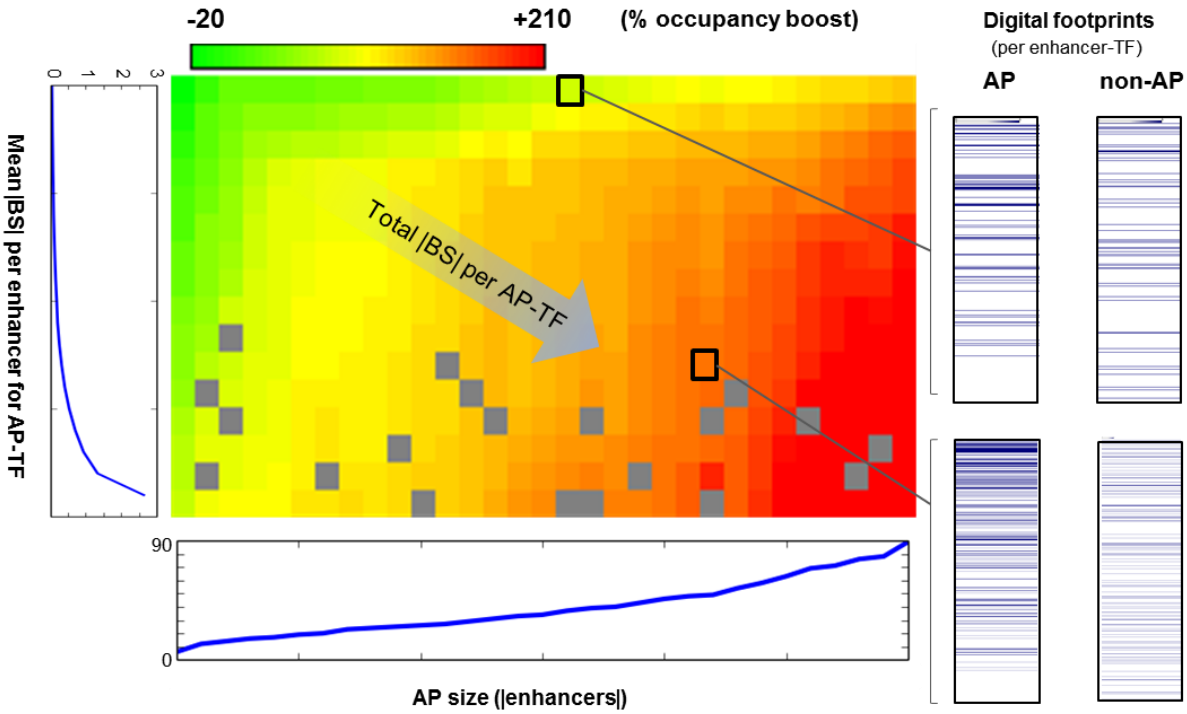


Figure 3-9. Occupancy boost observed in cases of 'non-reciprocal' occupancy -- where exactly one of the AP and matched non-AP enhancer have non-zero occupancy. Given that percent differential occupancy, or 'occupancy boost' , as previously calculated, is not as meaningful in the event that either AP or non-AP occupancy for a given TF-AP = 0, such TF-AP pairs were excluded from the previous calculation and analyzed separately (cases where both occupancy values were zero were excluded). As in the previous analysis, TF-AP pairs were binned based on the combination of AP size (columns) and mean TF homotypicity per enhancer (rows), and for each heatmap cell, the normalized difference was computed between counts of TF-AP pairs exhibiting non-reciprocal AP occupancy (TF-AP occupancy > 0, non-AP synthetic TF-AP occupancy = 0) and counts of pairs exhibiting non-reciprocal non-AP occupancy (TF-AP occupancy = 0, synthetic TF-AP occupancy > 0). This difference was then normalized by |TF-APs| in the cell, and the resulting values Lowess-smoothed along both x- and y-axes using default settings (stats.model Python package). Red hues indicate either 0 or negative differences, while colors spanning orange to green indicate increasingly higher normalized differences, respectively (see scale. Gray indicates no data). P-value based on Wilcoxon test comparing boosts for TF-APs with the lowest and highest 50% coverage.

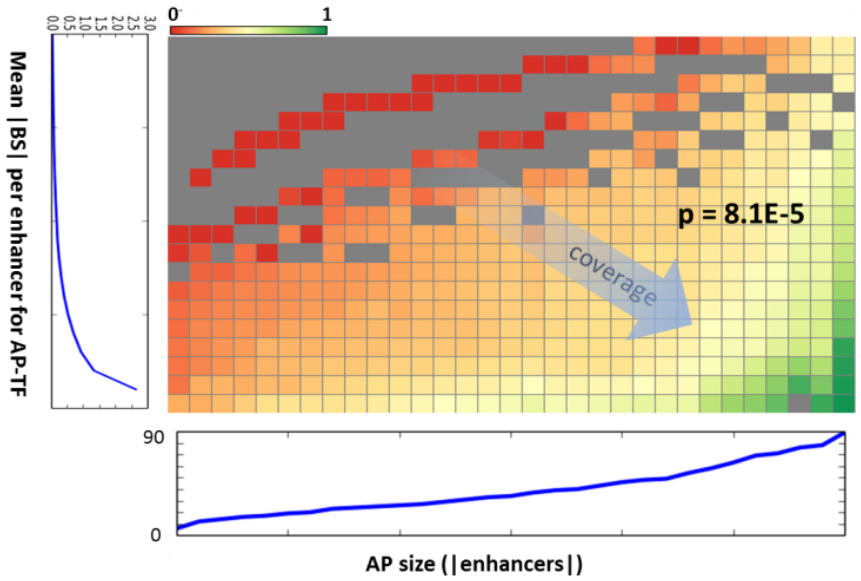


Figure 3-10. Additional validation of occupancy boosts using ChIP-Seq derived occupancy. ENCODE ChIP-Seq data was used for all cell types in which at least one AP was active, that is, for which at least 90% of an AP's enhancers were chromatin accessible. This resulted in 206 Chip-Seq experiments for 89 unique TFs across 9 cell types. (A) Reciprocally and non-reciprocally bound TF-AP combinations were pooled, thereby encompassing TF-AP combinations without mathematically defined occupancy boost. Shown are overall numbers of specifically bound AP (pale blue) and stringently matched non-AP (grey) enhancers, partitioned into coverage bins. Results of two tests are shown: comparing AP/non-AP ratios in the lowest and highest coverage bins (two-sided arrow); and comparing specifically bound and unbound enhancers in the highest coverage bin. P-values are from Fisher Exact tests. Specific occupancy was calculated on a per-enhancer basis, where binding for a given TF was determined based on overlap between a +50bp window surrounding the ChIP-Seq peak and a motif instance in the enhancer. AP: archipelago. (B) Occupancy boost was compared in AP-active cell types to boost in minimally active cell types. Cell type activity for a given AP was computed as the fraction of member enhancers that were DNase hypersensitive. In top plot (same as Fig. 3-6A), minimum AP activity is 90%; in bottom plot, AP activity ranges from 1% to 85%. A Wilcoxon test was used to compare the respective sets of occupancy boosts for TF-AP combinations in the highest coverage bin.

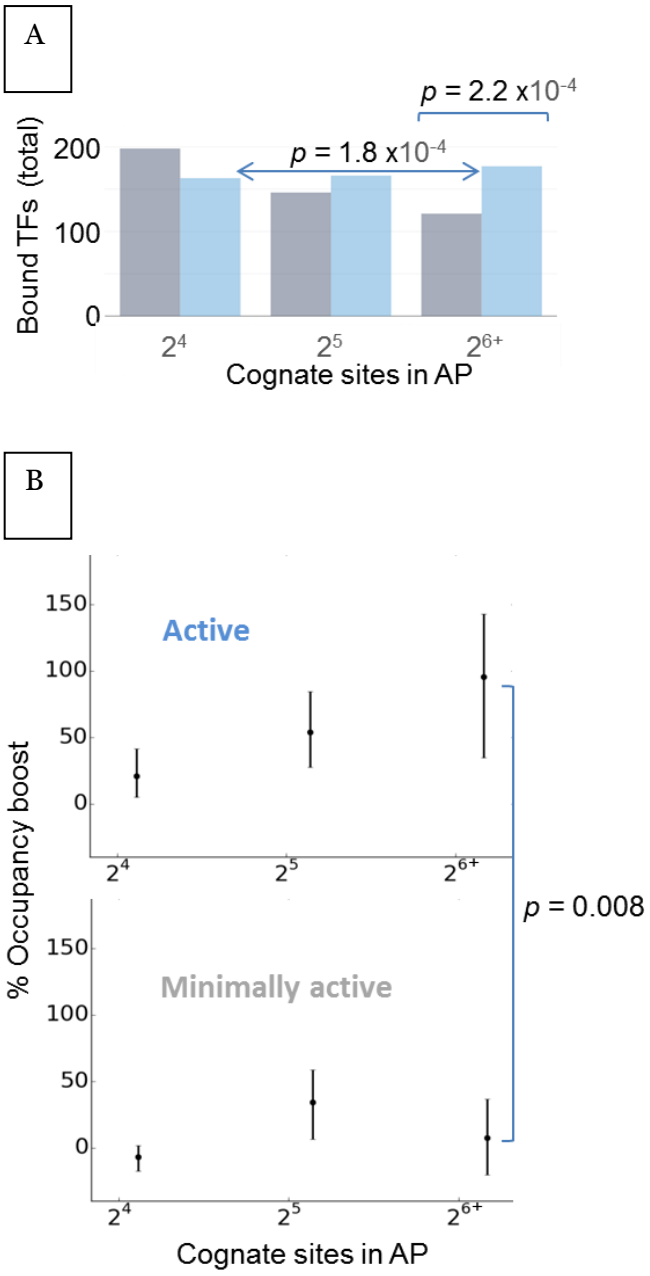
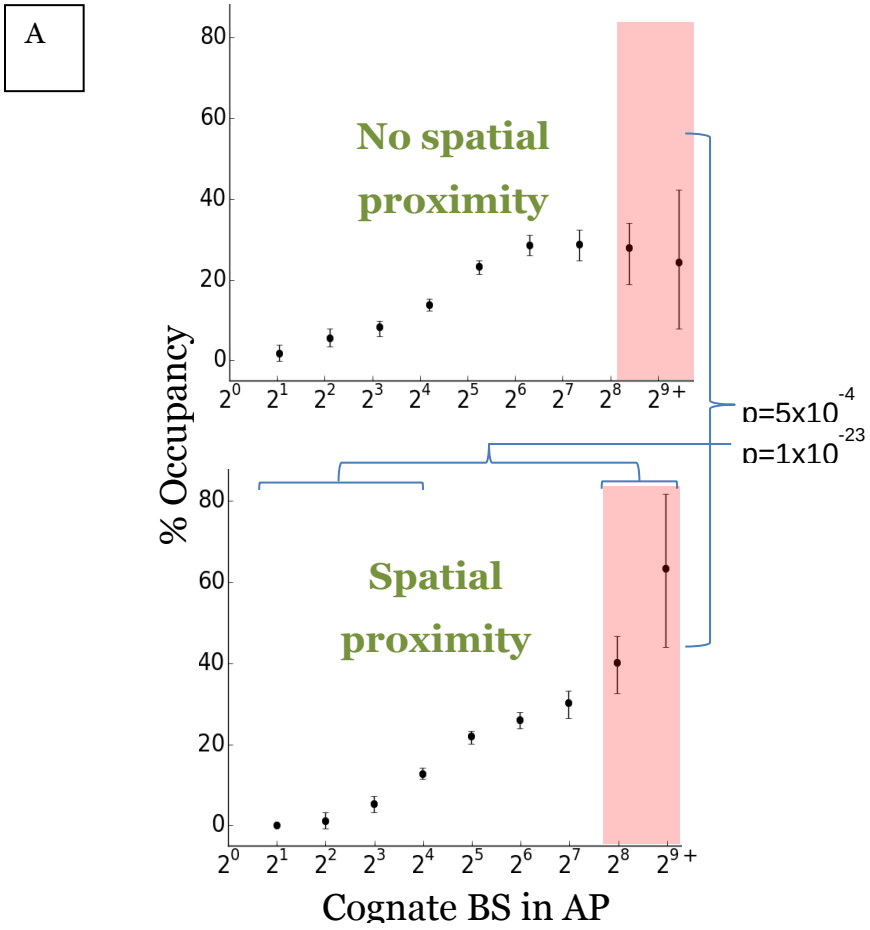
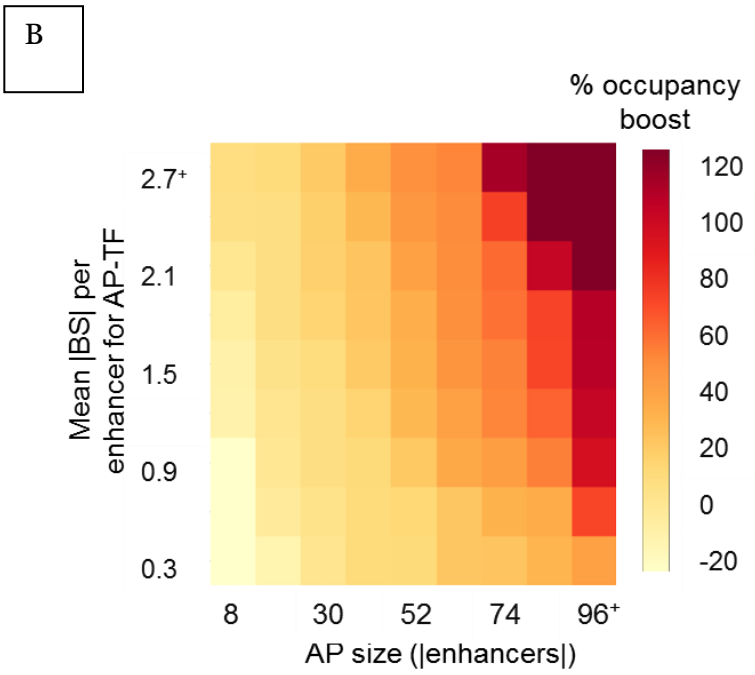


Figure 3-11. Validation of occupancy boosts using alternative archipelago data sets. Occupancy boost was determined for APs comprising sets of coordinately active regions from (Sheffield et al 2013) that were then overlapped with putative enhancers (P300 ChIP-Seq peaks). Boost was calculated with respect to a background of 'non-AP' enhancers, which did not belong to any Sheffield set of co-active regions of size five or greater. (A) Top: Plot shows percentage difference in TF BS occupancy between each TF-AP pair and its matched non-AP enhancer TF BS (y-axis) as a function of coverage – the total number of BS in the AP for a given TF (x-axis). Bottom: In each AP, enhancers with the lowest mean spatial proximity with fellow AP members were

excluded, based on human stem cell Hi-C. Shown are Wilcoxon test p-values from comparing boosts for TF-AP pairs in the bottom four and top two coverage bins (top p-value); comparing boosts in Hi-C screened (bottom plot) and unscreened TF-APs (top plot) with the highest two percent coverage, which approximately corresponds to the top two coverage bins, as indicated by pink shading (bottom p-value). 95% confidence interval shown based on a bootstrap procedure. (B) Coverage for each TF-AP was decomposed into orthogonal components for mean number of BS per enhancer (row) and mean number of enhancers per AP (column). Percentage occupancy boost for cells missing data was interpolated by averaging values in the four or two neighboring cells. Heatmap cells were then Lowess smoothed along both axes.





Occupancy boost can act independently of cooperative binding and of super-enhancers

Higher occupancy boost for cooperatively binding TFs explained by higher coverage

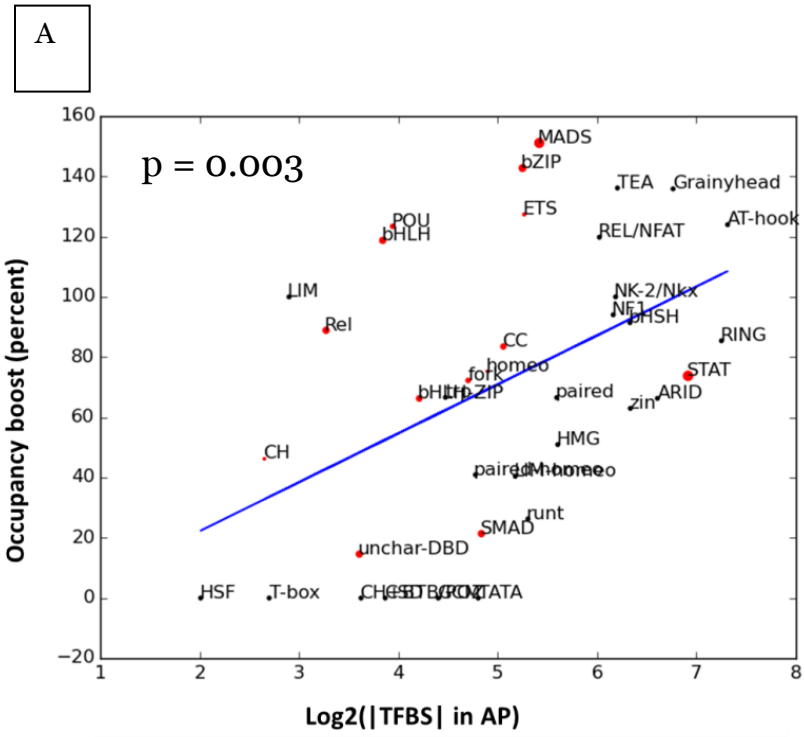
We reasoned that the observed link between spatial BS abundance and occupancy may partly be mediated by cooperativity among the bound TFs within an AP (Martinez and Rao, 2012; Pombo and Dillon, 2015) We therefore assessed whether the occupancy boost varies among TFs in different structural classes. We assigned the analyzed TFs to one of 42 structural families based on the TRANSFAC (version 2013.4). We found that, consistent with the crowdsourcing model, there is an overall significant correlation between family-wise occupancy boost and BS coverage (R^2 = 0.22, p-value = 0.003). Of the 42 families, 14 families included TFs that are known to form heterodimers, often with a member of the same family. Notably, 11 of 14 heterodimerizing (HD) families display

greater-than-expected boost (i.e. above the regression line), with MADS and bZIP families showing the highest boosts (Figures 3-12A). However, many families lacking heterodimer members display robust occupancy boosts and also possess high mean coverage – including Tea, Rel/Nfat, Grainyhead, AT-hook, NF1, bHSH, and Nk2/Nkx.

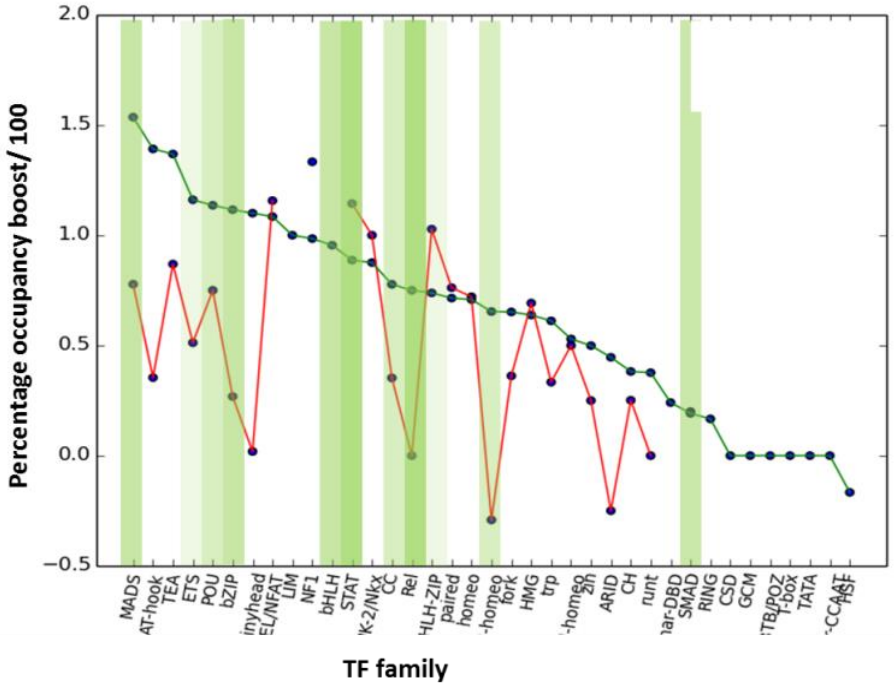
In order to more directly test for a potential link between cooperative binding and occupancy boost, we compared occupancy boost in TF-AP pairs for HD TFs to that in TF-APs that are not HD. Based on boosts for TF-AP pairs among the top 20% (50%) in coverage, HD TFs do, in fact, display higher boost, with a mean of 135% (120%) versus 110% (102%) for all other TFs (p-value = 0.007 (0.0018) Mann-Whitney rank sum test). However this test does not control for differences in coverage between HD and non-HD TFs. Upon closer inspection, we found that indeed, HD TFs have higher coverage than non-HD TFs (159 vs 137 BS per AP, for TF-APs in top 20% by coverage). Within each family however, we found that TFs with higher coverage exhibit higher occupancy boost (Figure 3-12B). Thus, occupancy boost differential between HD and non-HD TFs is largely explained by their inherently different degeneracy and, hence, coverage. Overall, occupancy boosts scale closely with coverage for a majority of TF domain families, and for both cooperatively binding and non-cooperatively binding TFs. Hence, the observed occupancy boost cannot be explained by TF cooperativity alone.

Figure 3-12. Crowdsourcing behavior spans TF domain families with and without strong heterodimerizing tendencies. (A) TF family-wise occupancy boost vs. mean coverage. For a given TF-AP pair, coverage is defined here as the total number of cognate BS in the AP. Plot is based only on TF-APs for APs with > 40 enhancers. Linear regression line (R2 = 0.22, p-value = 0.003) shown in blue. Size of red dot in plot is proportional to the fraction of family members that are heterodimers, as classified by TRANSFAC. Note that 10 of 14 families with heterodimer TFs have occupancy boosts that lie above the regression line, although the fraction of HD members was not significantly associated with the family's mean boost.

(B) Mean occupancy boost stratified by TF family and AP size. TF-APs sorted based on TF domain family were further divided into two classes, based on a cutoff for AP size of 20 enhancers. X-axis shows families sorted by their boost in large APs. Hue of column is proportional to the fraction of TFs in family that are heterodimers – deeper green indicates a larger fraction. Green trace: mean family-wide occupancy boosts in large APs. Red trace: mean boosts in small APs.



B



Non-superenhancer AP enhancers exhibit large occupancy boosts

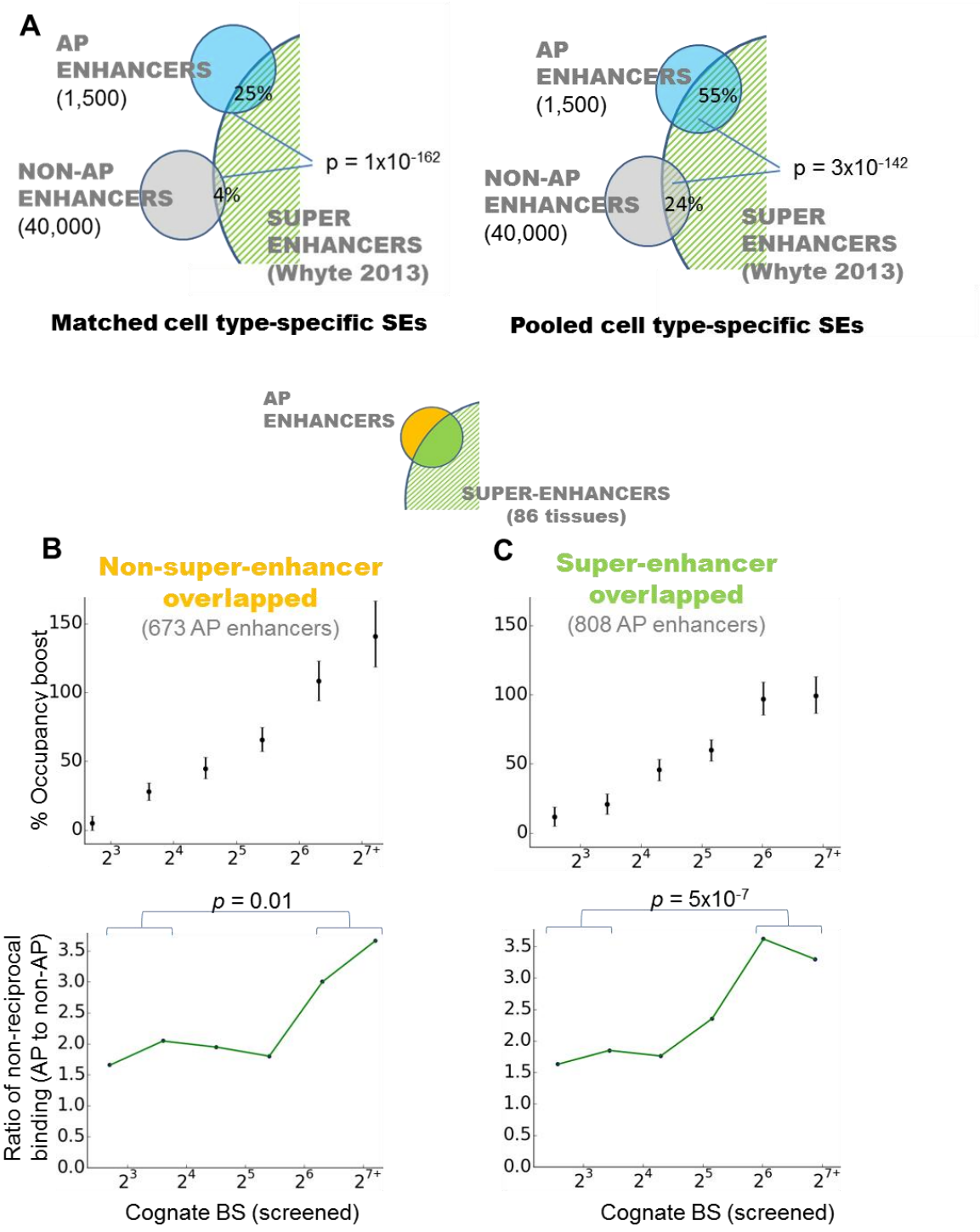
We next probed the potential relationship between occupancy boost in APs and high occupancies of key lineage-determining TFs reported in so-called super enhancers -- compound enhancers extending up to 100Kb or more (Whyte et al., 2013). First, we observed a six-fold greater overlap of cell type-specific super enhancers (downloaded from (Hnisz et al., 2013)) with AP-active enhancers relative to non-AP enhancers, in seven cell types (Figure 3-13A, *left*). This is consistent with the hypothesized association between AP occupancy boost and super enhancer function.

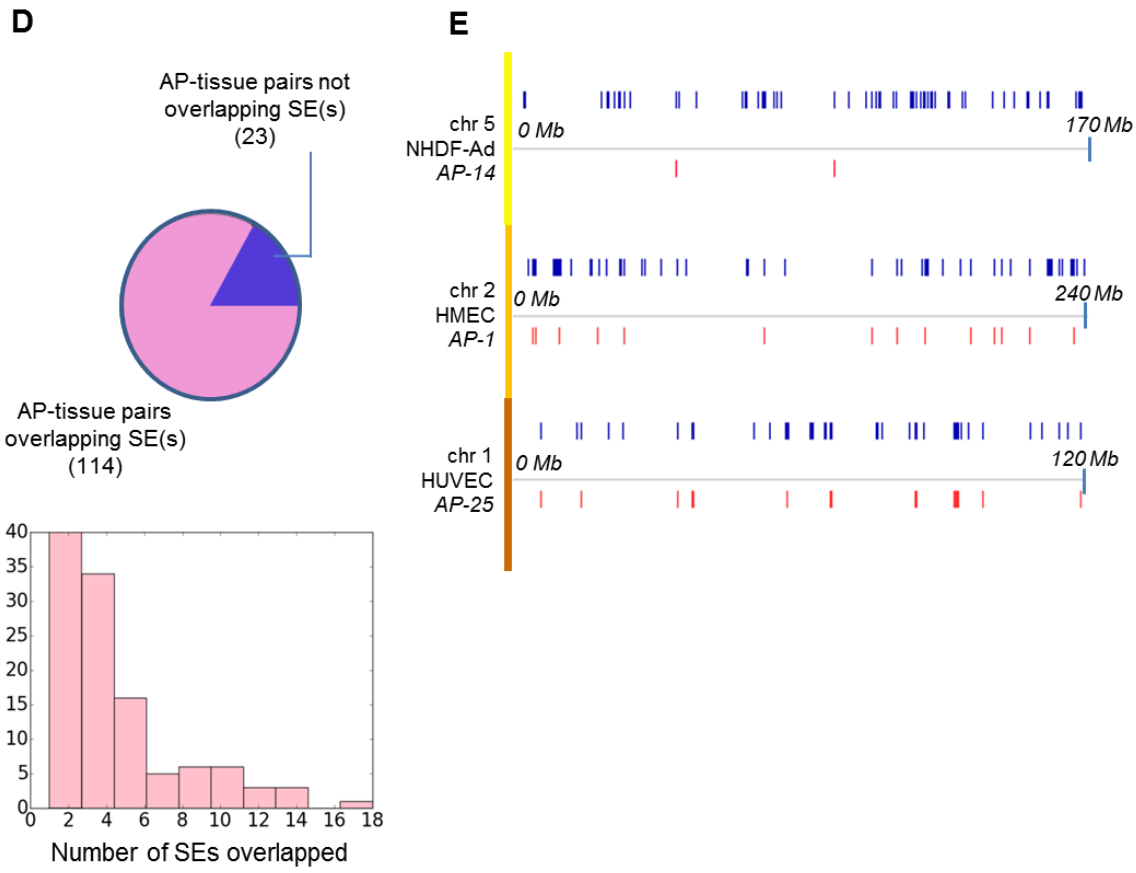
In order to test whether AP occupancy boost is limited to super-enhancers or, conversely, acts more generally, we calculated occupancy boost exclusively at

the 45% of AP enhancers that do not overlap a super-enhancer in any of 86 cell types (Figure 3-13A, *right*). Indeed, the number of cognate BS in just these screened AP enhancers is a highly robust predictor of their own occupancy boost, with mean boost exceeding 140% for the TF-AP pairs with the highest screened coverage (Figure 3-13B *top*, 3-13C *top*). The analogous non-reciprocal binding trend for these AP enhancers was less robust, due largely to few non-reciprocal TF-AP pairs with high coverage, but still significant (AP/non-AP ratio > 3.0 in highest coverage bin, p-value = 0.01, Figure 3-13B *top*, 3-13C *bottom*). In sum, the observed occupancy boost appears to be a general phenomenon not limited to super-enhancers.

Taken together, our extensive analyses based on multiple alternative data sources, both for APs and for inferring occupancy, strongly suggest a group-level effect on TF occupancy, whereby in a spatial cluster of homotypic BS for a TF, occupancy at an individual BS is 'crowdsourced' by the collective contribution of myriad homotypic BS across an AP.

Figure 3-13. Super-enhancers appear to be one instance of crowdsourcing. While there is high enrichment for AP enhancers in super-enhancer (SE) regions, occupancy boost is as well-predicted by non-SE-associated as by SE-associated AP enhancer coverage. (A) Left: Genomic overlap was quantified between cell type-matched SE and (i) AP enhancers in which at least 90% of member enhancers were hypersensitive in the given cell type; (ii) a set of ~40K non-AP enhancers. Overlap was considered anywhere in the span of a super-enhancer region, as annotated in (Hnisz et al 2013), and was found in NHDF-Ad, NHLF, HUVEC, MCF7, HMEC, HeLa, and hESC. P-value based on a Fisher exact test. Right: SEs from 86 cell types and tissues (Hnisz et al 2013) were pooled and overlapped with AP and non-AP enhancers, independent of cell type. (B) Occupancy boost was tested for reciprocally (top) and non-reciprocally (bottom) bound TF-AP combinations for screened AP enhancers. Both occupancy and coverage (numbers of cognate BS) were calculated using the subset of AP enhancers that, conservatively, did not overlap an SE from any of 86 cell types (without regard to cell type), along with an AP enhancer's matched non-AP enhancer for a given TF. 5% confidence intervals determined using a bootstrap method. P-value determined with a Fisher Exact test. (C) Occupancy boost as a function of coverage was determined at those AP enhancers not analyzed in (B), namely AP enhancers that do overlap an SE. (D) Top: Genomic overlap was identified for active APs and cell type-matched SEs. Bottom: Histogram showing number of overlapped SEs per AP for APs overlapping at least one SE. (E) Enhancers in three AP-tissue pairs (red) along with their overlapping SEs. (E) Enhancers in three AP-tissue pairs (red) along with their overlapping SEs.


A
AP ENHANCERS (1,500)
25%
NON-AP ENHANCERS (40,000)
4%
p = 1x10-162
SUPER ENHANCERS (Whyte 2013)
Matched cell type-specific SEs
AP ENHANCERS (1,500)
55%
NON-AP ENHANCERS (40,000)
24%
p = 3x10-142
SUPER ENHANCERS (Whyte 2013)
Pooled cell type-specific SEs
AP ENHANCERS
SUPER-ENHANCERS (86 tissues)
B
Non-super-enhancer overlapped
(673 AP enhancers)
% Occupancy boost
p = 0.01
Ratio of non-reciprocal binding (AP to non-AP)
Cognate BS (screened)
C
Super-enhancer overlapped
(808 AP enhancers)
p = 5x10-7
Cognate BS (screened)



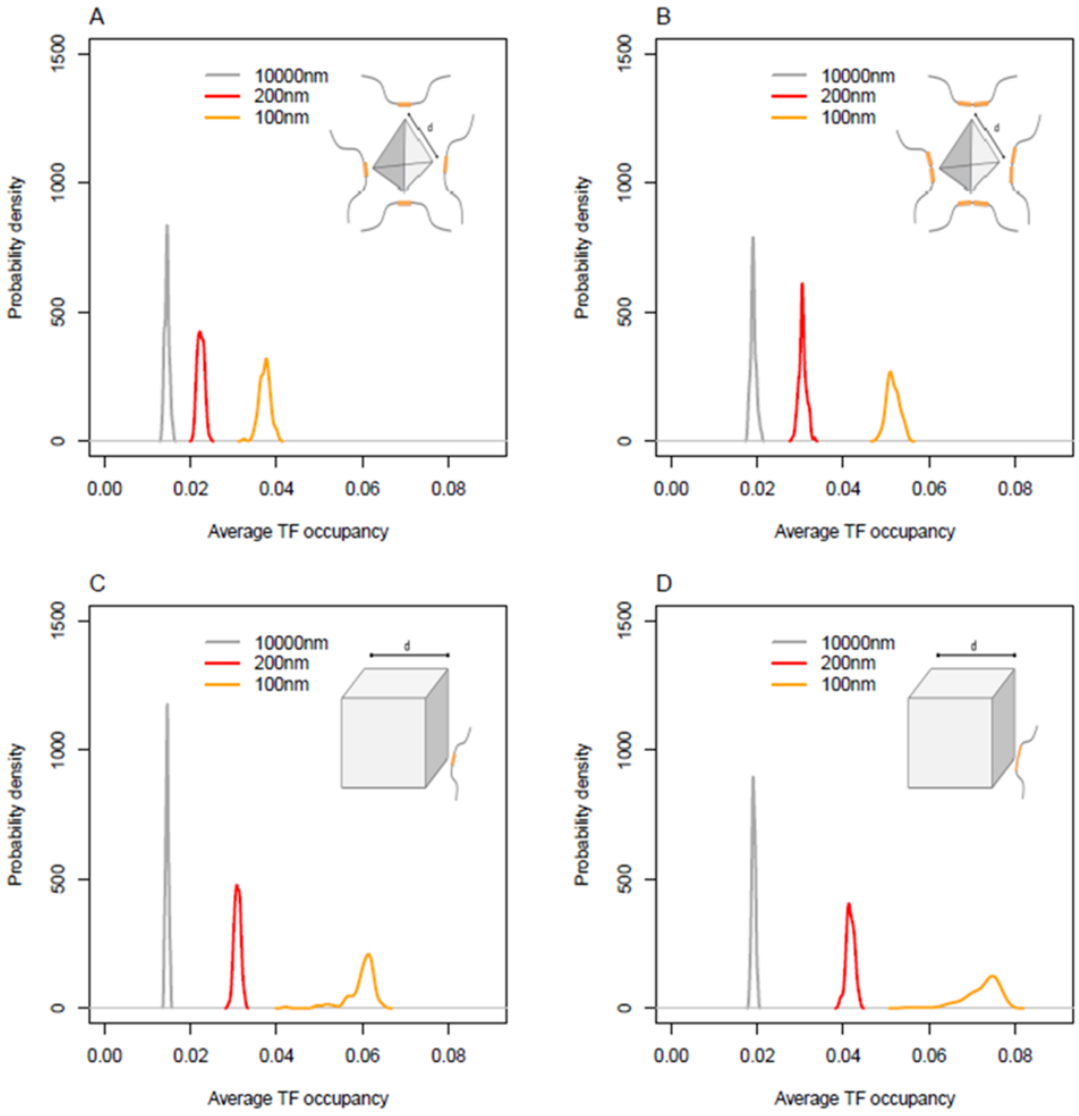
TF occupancy boost in spatial clusters of BS is consistent with a facilitated-diffusion model

Many biophysical simulations and experiments have strongly suggested that facilitated diffusion can have a large influence on TF binding dynamics (Brackley et al., 2012, 2013b; Elf et al., 2007; Hammar et al., 2012; Leith et al., 2012; Mirny et al., 2009; Wunderlich and Mirny, 2008; Zabet and Adryan, 2012). In particular, previous studies have shown that a facilitated diffusion model can explain the greater occupancy in genomic homotypic clusters of TFBS (Brackley et al., 2012). Here we simulated an extended version of the biophysical model for HCTs in isolation in order to determine whether the crowdsourcing effect is sufficient to explain the observed AP-specific

occupancy boost. The crowdsourcing effect was simulated using a modified form of the facilitated diffusion modeling framework fastGRiP (Ezer et al., 2014). While the original implementation of fastGRiP incorporates the influence of the positioning of binding sites along the DNA, it ignores how the 3D organization of the DNA can influence the TF search process. In order to simulate the crowdsourcing effect, TF diffusion between nearby DNA strands was incorporated, by integrating the diffusion equations previously derived by (Carslaw and Jaeger, 1959; Elf et al., 2007; Paramanathan et al., 2014) into the simulation. All details pertaining to the model, algorithms, parameter selection and results are provided in Appendix 'FD-Model'. Centrally, our simulations show that occupancy boost increases with, both, the number of homotypic BS in an HCT (i.e., enhancer) and, novelly, the number of enhancers in an AP. For instance, in the case of four clustered enhancers 100nm to 200nm apart (equivalent to 300 to 600 nucleotide lengths), versus 10000nm apart (approximating non-AP), where each enhancer contained a pair of homotypic binding sites, there was a 60% to 170% increase in TF occupancy, and in the case of eight enhancers containing pairs of homotypic binding sites, there was an 118% to 277% increase in occupancy (Figure 3-14). TF occupancy, consistent with a previous model (Brackley et al., 2012), scaled with the number of BS in an HCT. Less expectedly, the genomic inter-BS distance within an HCT did not significantly impact occupancy – in stark contrast to the large positive effect on occupancy from reduced spatial distance between HCTs (Figures 1C-1E in Appendix FD-Model; Figure 3-14). Together, our simulations demonstrate that inter-strand jumping between HCTs substantially amplifies the occupancy boost experienced at an isolated HCT, and this effect increases robustly with the number of homotypic clusters engaged in 3D

interactions. Indeed, simulation results suggest that the crowdsourcing effect is a biophysically sound strategy for increasing local TF occupancy in APs at a biologically meaningful scale.

Figure 3-14. Biophysically modeling crowdsourcing effect. TF diffusion was simulated for four geometric arrangements of binding sites, and the probability density functions of TF occupancy are shown. The TF occupancy is defined as the average probability that each site is bound. The four simulated scenarios are: a tetrahedron with (A) one binding site or (B) a pair of binding sites in each corner, which contain 4 or 8 binding sites, respectively; a cube with (C) a single binding site or (D) a pair of binding sites in each corner, which contain 8 and 16 binding sites respectively. For an additional figure and details on the simulation, see Appendix.



Cell type-specificity of AP enhancer occupancy boost and activity

Given the link between occupancy boost and spatial clustering of BS, and given the context-specificity of spatial proximity (Ay et al., 2014), we expect the occupancy boost to exhibit cell type specificity. In addition to identifying the cell type where an AP is deemed active (as employed in analyses thus far), we also identified the cell types where an AP is deemed inactive, namely those where less than 40% of the AP enhancers were DNase I hypersensitive. To offset the paucity

of bound sites in inactive tissues, all qualifying inactive tissues for each AP were pooled. We found that for the TF-AP pairs in the highest coverage bin, occupancy boost dropped from ~112% in its AP-active cell type down to 38% in inactive cell types (Figure 3-15A). This trend was also observed when occupancy was computed with ChIP-Seq data (Figure 3-10B). In addition, we estimated tissue specificity of each TF as the cross-tissue dynamic range of its footprint-based occupancy, defined as the ratio of its occupancy in AP-active tissue(s) to that in AP-inactive tissues, calculated over the identical AP BS. Notably, this provides evidence of the occupancy boost's tight association with coverage without the need for non-AP occupancy as a baseline. After controlling for DHS across coverage bins, we find that the TF-APs with top 10% coverage display 135% greater occupancy in active relative to inactive tissues, while in the matched non-AP context it is 38% (Figure 3-15B). Even larger differentials between AP and non-AP contexts were observed for their respective ratios of non-reciprocal binding in active and inactive tissues (Figure 3-16A). Interestingly, we found that high coverage TF-AP pairs for heterodimerizing TFs exhibit substantially higher specificity than other TFs (225% vs. 140%) (Figure 3-16B), particularly TFs in MADS and bZIP domain families, suggesting an augmented level of cooperative binding in APs. This, we suspect, is due to the relatively binary nature of cooperative binding: in response to small increments in TF concentrations, heterodimers exhibit disproportionately large changes in occupancy (Giorgetti et al 2010). These results strongly suggest that occupancy boost in AP enhancers is cell type-specific and leverages context-specific chromatin structure.

Figure 3-15. Occupancy boost is tissue-specific. (A) Occupancy boost in cell types with reduced AP activity. Occupancy was computed as a function of coverage in 'inactive' cell types – those in which fewer than 40% of the AP's enhancers were DNase hypersensitive (bottom). For comparison, the plot for active cell types from 2A is reproduced (top) (B) Tissue specificity of occupancy for TF-AP and matched non-TF-AP pairs as a function of TF-AP coverage. Dynamic range (y-axis) for occupancy was calculated for each TF-AP pair as the percentage difference between mean occupancy in the AP's most active and inactive cell types. Identical BS in active and inactive cell types were tested. Serving as a control, dynamic range was also computed for non-TF-APs that were matched to TF-APs. A TF-AP was required to have non-zero occupancy in both inactive and active cell types. Only TF-APs shared in AP and non-AP contexts were used for analysis. Shown are results for TF-APs and for non-TF-APs each sorted into 4 bins with exponentially increasing coverage cutoffs. Red: AP, Gray: non-AP. 90% and 99% confidence intervals are shown with variable hue. AP-wide DHS as a function of context-specific TF availability.

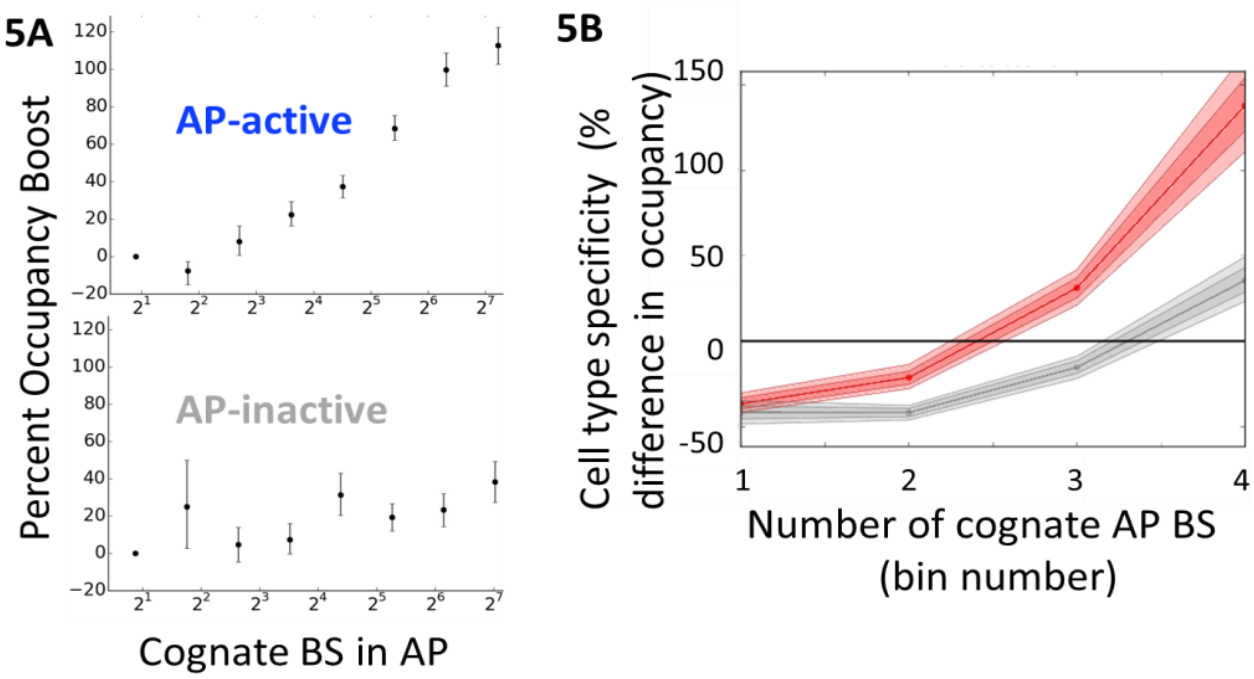
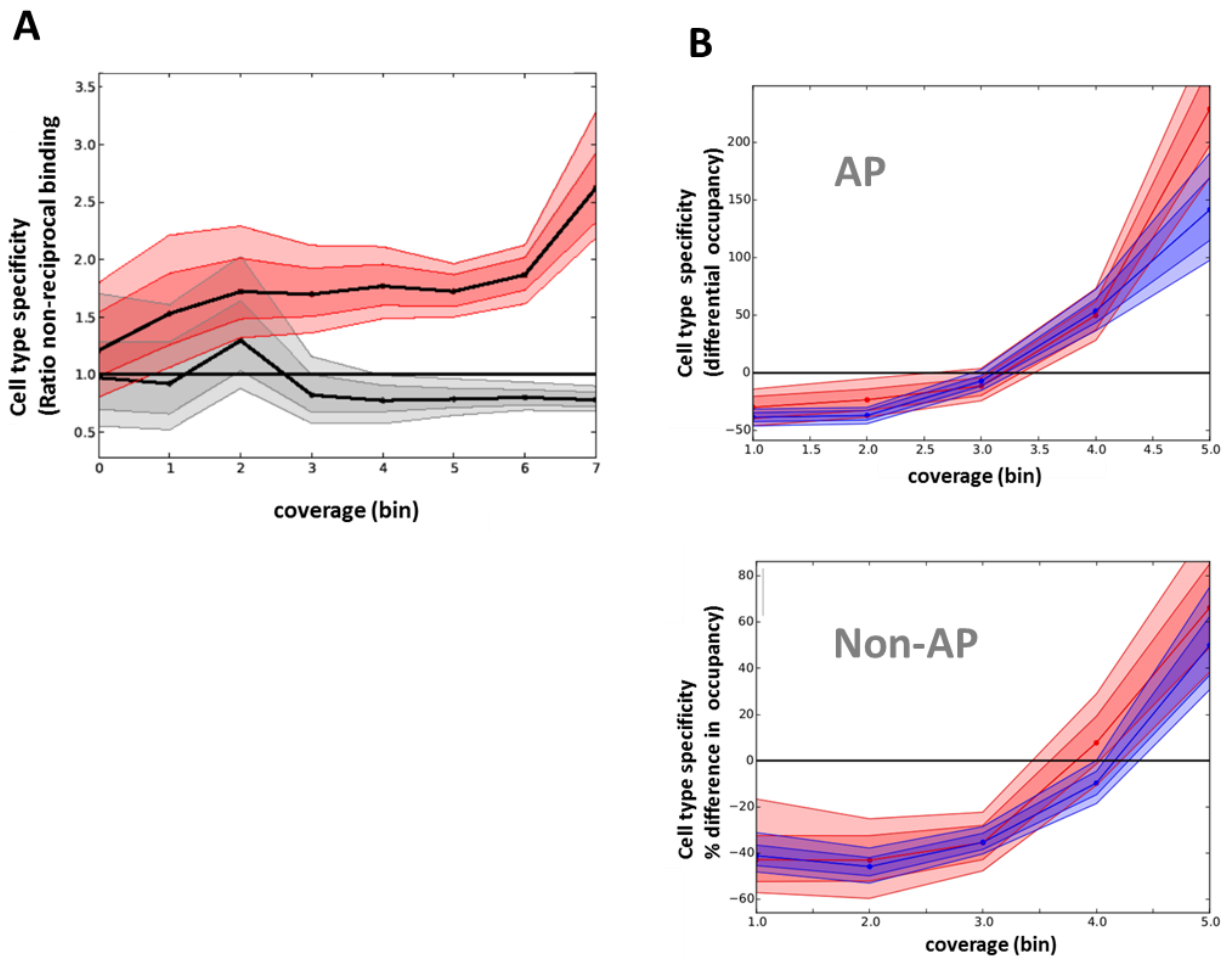


Figure 3-16. Tissue-specificity of occupancy boost. (A) Non-reciprocal tissue-specificity as a function of coverage. The percentage difference between an TF-AP's occupancy in its AP-active tissue and that in AP-inactive tissues was previously computed (Figure 11). TF-APs with zero occupancy in either active or inactive tissues, which were excluded from that analysis, are analyzed here. TF-APs were sorted based on coverage into 8 uniform-sized bins. In each bin, the ratio was computed between the number of TF-AP pairs exhibiting non-reciprocal active-tissue occupancy (active tissue occupancy > 0, inactive tissue occupancy = 0) and the number of pairs exhibiting non-reciprocal inactive-tissue occupancy (active tissue occupancy = 0, inactive tissue occupancy > 0). Unlike Figure 11, where TF-APs are binned based on exponentially increasing coverage cutoff, TF-APs are, instead, binned here uniformly to offset what would otherwise be low sample size in high coverage bins. Red: AP, Gray: non-AP. 90% and 99% confidence intervals are shown with deeper and lighter hue respectively and were computed with bootstrapping. (B) Heterodimers exhibit an elevated trend in cell type specificity, which here, is estimated as the percent difference in occupancy between AP-inactive and AP-active cell types. Left: Differential occupancy (y-axis) between AP-inactive tissues and AP-active tissue was computed for each TF-AP and plotted as a function of TF-AP coverage after partitioning TF-APs based on those with a TF classified as heterodimer (TRANSFAC 2014.3) (red) and all remaining TF-APs (blue). TF-APs were sorted into bins with exponentially increasing coverage cutoffs (x-axis). Right: same as Left except differential occupancy boost between inactive and active tissues was computed for matched non-AP enhancers. Note the different scaling on y-axis compared to in (A). 90% and 99% confidence intervals are shown with deeper and lighter hue respectively.



Discussion

Summary. Here, we have shown that a TF's *in vivo* occupancy at a particular cognate BS is much greater when the BS is in spatial clustered with other homotypic BS (i.e., in an AP) than when it is not. Strikingly, the size of the occupancy boost robustly scales with the number of homotypic BS in the AP, suggesting, for the first time, that the BS in an AP cooperatively crowdsource their own occupancy. To ensure the robustness of our conclusions, we used stringent controls and employed multiple (i) sources for AP enhancers (Malin et al 2013, Sheffield et al 2013), (ii) experimental backgrounds (non-AP enhancers in the AP-active tissue, the same enhancer in AP-inactive tissues), (iii) occupancy scales (per BS, per enhancer), and (iv) types of occupancy data (curated

digital footprints, ChIP-Seq. These observations are not adequately explained by current models, however, they closely agree with a standard biophysical model of facilitated TF diffusion that duly accounts for the augmented diffusion of TFs among spatially proximal homotypic BS. Effectively, a collective of spatial homotypic clusters of TF BS (spatial HCTs) cooperatively alter their microenvironment, raising the local concentration of their cognate TF.

Genomic versus Spatial homotypicity. Our work synthesizes the regulatory roles of HCTs (e.g. Crocker et al 2015), and of stable chromatin structures (e.g. Dowen et al., 2014), by showing that it is precisely the interplay of numerous HCTs mediated by chromatin folding that gives rise to the hitherto undocumented biophysical effect that we have termed *crowdsourcing*. Enhancer-enhancer interactions have been reported in the context of HOX and globin gene regulation as well as in high-throughput ChIA-PET assays, but their functional nature has remained elusive. A notable exception, spatial clustering of enhancers around an olfactory receptor gene have been associated with removal of repressive H3K9me3 (Markenscoff-Papadimitriou et al., 2014); it is plausible that crowdsourcing is an upstream trigger of this change – through a general remodeling of the local chromatin state, or through increased binding of a TF that mediates chromatin remodeling.

Tissue specificity and cooperative binding. We found that crowdsourcing is highly tissue-specific, as high-coverage AP BS exhibit several-fold greater occupancy in AP-active relative to AP-inactive tissues. Such tissue specificity is consistent with the dependence of crowdsourcing on chromatin context and TF availability, where differential TF availability likely acts not only directly but also by influencing higher-

order chromatin conformation (Pombo and Dillon, 2015). Crowdsourcing endows the cell with a high degree of fine-grained regulatory control, as occupancy boost magnitude is shaped by the collective availability of multiple TFs and conditioned on the chromatin-induced spatial proximity of their cognate sites. Fundamentally, crowdsourcing provides an alternative mechanism of cooperativity to direct cooperative binding of heterodimerizing TFs, an established source of tissue-specificity. Indeed, crowdsourcing acts complementarily to cooperative binding.

Differential occupancy as a vehicle for specificity. In contrast to previous work underscoring the functional importance of weak (low occupancy) binding that typical ChIP-Seq processing tends to miss due to stringent cutoffs (Tanay 2006; Biggin 2011), crowdsourcing leverages spatial chromatin context to imbue inherently low-affinity sites with unexpectedly high-occupancy binding. Crowdsourcing may thus explain previous reports linking particular low-affinity sites with context-specific regulation (e.g. (Gaudet, 2002)), or linking unusually robust binding to supposedly individual HCTs, for example . Indeed, occupancy boosts that we observed at spatially clustered HCTs were computed with respect to 'isolated' genomic HCTs. As shown by Crocker et al (2015), occupancy is more robust where degenerate homotypic sites are located in genomic clusters. HCTs, however, are highly abundant in the genome (He et al., 2012) as well as, spatiotemporally invariant, which raises a well-known conundrum, viz. how a TF discriminates among a multitude of candidate BS (Z Wunderlich, 2009). In contrast to the static and relatively low specificity of an individual genomic HCT, a large collective of homotypic low-affinity sites can attain high specificity and spatiotemporal responsiveness precisely by their capacity to configure the local TF environment *en*

masse – in specific favorable chromatin contexts. That is, loci may be coordinately targeted not through a hardwired address on the one-dimensional genome, but as a dynamic nexus dependent on three-dimensional plurality.

Potential implications for transcription factories, superenhancers. An archipelago, as described here, represents a group of spatially clustered enhancers and their likely target genes, which are often functionally related (Malin et al., 2013; Sheffield et al., 2013). Meeting this same general description are subnuclear compartments known as transcription factories (Edelman and Fraser, 2012). Transcription factories have been shown to concentrate resources such as RNA PolII, core components of transcription, as well as some master TF regulators (Schoenfelder et al., 2010b). However, it is unclear precisely how distinct factories achieve specific and differential concentrations of master regulator TFs (Schoenfelder et al., 2010a). Crowdsourcing offers a possible explanation, and is consistent with a speculated role for resident sequences (Andersson et al., 2015; Schoenfelder et al., 2010a). While it is generally assumed that high concentrations of TFs are critical in recruiting genes and their distal regulatory regions to the factory, our work suggests alternative causality, as supported by formal biophysical simulations. Although not confirmed, our characterization of archipelagos suggests their operational overlap with factories.

Our findings are broadly consistent with a mechanistic role for crowdsourcing in super-enhancer (SE) function. Each SE, comprising a contiguous cluster of enhancers, can further form spatial clusters with isolated enhancers (Heinz et al., 2015) as well as with other SEs. We speculate that such spatial clustering of SEs with auxiliary non-SE enhancers may supplement an SE's already-ample BS, thereby further amplifying

occupancy of (typically degenerate) master regulator TFs. Intriguingly, active APs, whose many tens of regulatory elements often span much of a chromosome or potentially many chromosomes (Sheffield et al., 2013), typically overlap multiple active SE (Figure 3-13D, 3-13E). This is consistent with a role for crowdsourcing in coordinating a collective of SEs regulating cell lineage-commitment, interactions currently not well-characterized.

Materials and Methods

Enhancer clusters ('APs'):

In previous work, genomically dispersed clusters of enhancers with correlated activity across cell lines showed evidence of spatial proximity, particularly in tissues in which the enhancers were active, where spatial proximity between two genomic segments was inferred from Hi-C (Malin et al., 2013). Starting with previously published 40 enhancer clusters, we iteratively filtered out the enhancers from each cluster whose mean spatial proximity in stem cell to other enhancers was at least one standard deviation below the original mean across all enhancers in the cluster. This results in 40 APs with a total of 1480 enhancers (Appendix Archipelago enhancers) with ~37 per AP, ranging from 6 to 89 enhancers per AP. Processing of alternative set of APs obtained from Sheffield et al. is described later.

Estimating in vivo occupancy at a BS using digital footprint data:

Putative BS in each enhancer were identified using TRANSFAC vertebrate motifs (Matys et al., 2006) and motif scanning tool PWM_SCAN (Levy and Hannenhalli, 2002) at 95 percentile score cutoff. We estimated *in vivo* TF

occupancy by overlapping putative BS with the high-confidence genome-wide digital DNase hypersensitivity footprints identified in 38 human cell lines (Neph et al., 2012b), using a procedure similar to, but more stringent, than (Neph et al., 2012b). Digital footprints are a single-base-pair resolution readout in which the absence of aligned reads in a particular segment of open chromatin has been shown to predict binding of a protein (Neph et al., 2012b). For a TF, a particular putative BS was considered bound by the cognate TF if there was specific overlap between the BS and a footprint, with further requirement that (i) the midpoint of a footprint must overlap the BS; (ii) the midpoint of the BS must overlap the footprint; and (iii) BS length + 1 > footprint length > BS length - 4. The latter criteria excludes otherwise significant footprints that are either too short or too long to confidently be associated with a given motif instance. When a footprint strongly overlaps sites for multiple TFs, it was included in the analysis for all such TFs; fewer than 25% of the overlapped BS stringently mapped to multiple distinct TFs. These highly stringent criteria were applied identically to AP and to non-AP data.

AP-active and AP-inactive cell lines:

For each AP, we identified the cell line in which it was most active. Cell lines deemed active for a given AP are those in which at least 80% of the AP's enhancers are in open chromatin regions, based on overlap with DHS narrow peaks. In case of more than one such tissue, except where noted, we selected the tissue with the highest percentage of open enhancers (see Figure 1A). Approximately 95 percent of AP enhancers were found to be accessible in an AP's

'most active tissue', which for the 40 APs, span 15 distinct cell types out of 34 tested.

Establishing non-AP control for occupancy boost:

To establish a non-AP control, for each combination of TF and AP enhancer we identified a non-AP enhancer (sampled with replacement) with an identical motif profile, *i.e.* the vector containing the number of instances of each motif mapping to the given TF. This is an important control, as the number of homotypic BS in an enhancer that are cognate to a given TF impacts occupancy (He et al., 2012). We note that AP and non-AP enhancer have very similar distributions of total BS and length. Additionally, for each TF motif and AP, AP enhancers' mean DHS in the AP's most active tissue was matched to within 5% in the corresponding non-AP enhancers' mean DHS in the same tissue. Any TF-AP enhancer pair for which a non-AP could not be found meeting these tight controls was excluded. This procedure yielded 430K AP and non-AP TF-enhancer pairs that harbored 730K BS, of which 31K BS had a DNase footprint suggestive of a binding event.

Determining TF occupancy at enhancer resolution with ChIP-Seq data:

We downloaded ENCODE ChiP-Seq data for 294 experiments in human, including 135 unique TFs in 11 cell types for which there was accompanying DNase hypersensitivity data. This data was then screened to include only cell types in which at least one AP was active, that is, for which at least 90% of an AP's enhancers were found to be chromatin accessible (as per ENCODE DNase hypersensitivity data). This screen resulted in 206 Chip-Seq experiments for 89 unique TFs across 9 cell types – NT2-D1, IMR90, GM12878, Hct-116, MCF-7, Hela-S3, PANC-1, A549, and HUVEC. (Using an AP

activity cutoff of, alternately, 80% or 100% did not change the observed trend). Enhancer occupancy by a given TF was determined based on overlap between a ±50bp window surrounding the ChIP-Seq peak and one or more putative motif instances detected within the enhancer. To mitigate concerns over systematic biases stemming from variability in protocols or labs of origin, we note that all ChIP-Seq data had identical *de facto* weighting for AP and non-AP enhancer-TF pairs, since these were matched by motif for BS counts.

Estimating TF's degeneracy:

A motif's degeneracy was quantified using its relative entropy (RE) (D'haeseleer, 2006). Higher degeneracy corresponds with lower relative entropy. RE was calculated for each TF motif (i.e., position weight matrix) using TRANSFAC (version 2014.3) (Hannenhalli, 2008). In cases where there were multiple motifs associated with a particular TF (coming from different publications etc.), the motif with the lowest RE was chosen, because it is expected to numerically dominate the genome-wide BS for the TF, given its higher degeneracy. Throughout the manuscript the term 'degeneracy' refers to RE and 'degenerate' motif refers to motifs with low RE (at certain RE threshold) and 'specific' motif refers to motifs not deemed to be degenerate, or in some case this whose with RE above certain threshold.

Determining occupancy boost with alternative set of AP enhancers:

We obtained sets of correlated regions generated in (Sheffield et al 2013). Each Sheffield cluster of DNase hypersensitive (HS) regions initially spanned multiple chromosomes. To make them consistent with enhancer clusters from Malin et al

(2013), regions from a single Sheffield cluster located on distinct chromosomes were treated as distinct clusters, and we retained at most the two largest such clusters from each Sheffield cluster. Consistent with previous procedures, we derived enhancer clusters from each Sheffield cluster by only retaining the regions that overlapped a putative enhancer represented by a large pooled set of 98,000 P300 ChIP-Seq peaks used previously (Malin et al 2013).

To further cull the thousands of resulting enhancer clusters, we excluded those with < 10 enhancers or with mean enhancer DHS < 100 in their most active tissue. We further excluded Sheffield clusters in which fewer than 90% of enhancers were DNase hypersensitive in their most active tissue, resulting in 474 clusters – averaging ~16 enhancers each, though ranging to over 100. Similar to above (see 'AP Enhancers') we used Hi-C data to screen enhancers in each AP that were less spatially proximal, on average, to the remaining members. To prevent excessive removal of additional enhancers, given the already modest mean pre-screen AP size, we implemented the Hi-C screen in a single pass, without recursively updating each enhancer's mean Hi-C score after removal of a fellow AP member. This resulted in 472 non-empty APs with an average of ~15 enhancers each.

For background control, we used the complement of P300 ChIP-Seq peaks overlapping any of the screened set of approximately 2.6M Sheffield et al DNase hypersensitive regions. This resulted in too few putative enhancers, and so to this we added back ChIP-Seq peaks overlapping any cluster (on one chromosome) of hypersensitive regions with fewer than five members and with mean DHS > 50 in

its most active cell type; this produced a background pool of ~18K enhancers. Non-AP enhancers from this set were matched with AP enhancers as described above. In order to accommodate the smaller APs in this alternative dataset, we loosened the stringency on DHS control such that at a group level AP and non-AP sets' mean DHS was matched to within 1% while at individual TF-AP combinations, the mean DHS for AP and non-AP enhancers was within 50%.

Chapter 4: **Crowdsourcing: functional impact and gene complex activation**

Abstract

In the previous chapter, we demonstrated an emergent effect among highly spatially clustered BS for the same TF, as may be found in a regulatory archipelago – a cluster of coordinately regulated genes and enhancers. Genomic data and biophysical simulations suggest that such a spatial homotypic cluster of sites may briefly trap a diffusing TF molecule, elevating the TF's observed DNA occupancy within the archipelago. TFs consistently exhibiting the highest occupancy boost were those with degenerate motifs, which tend to have highly abundant cognate sites.

In this chapter we scale up and investigate the functional impact of occupancy boosts on an enhancer and on the archipelago, overall. Based on additional analysis, we find that the functional impact, and the magnitude of the boost, itself, strongly diverge among enhancers within an archipelago. Specifically, archipelago enhancers enriched for BS that recognize degenerate motifs exhibit two-fold higher occupancy boost than BS recognizing specific motifs, in addition to far greater overall chromatin accessibility, evolutionary conservation, as well as expression at neighboring gene loci. In order to decouple enhancer chromatin accessibility from enhancer TF occupancy, we tracked accessibility as TF gene expression increased across cell types. Strikingly, archipelago-wide activity scaled with expression of TFs with degenerate motifs, but not TFs with specific motifs. In sum, we find strong evidence suggesting that

the crowdsourcing effect is experienced at a number of scales – binding site, enhancer, and archipelago. At the level of archipelago, crowdsourcing can contribute to switch-like and coordinated activation, mediated by context-specific TF availability and higher-order chromatin structure.

Introduction

Previously we reported a novel group-level phenomenon emergent among homotypic binding sites for the same TF. Mediated by higher-order chromatin structure, spatially concentrated homotypic BS exhibit higher-than-expected TF occupancy. While changes in DNA occupancy at promoters or enhancers appears to frequently precede function (Yáñez-Cuna et al., 2012) – most conspicuously, transcription of coding genes – no obvious function has, to date, been identified for the vast majority of TF binding (Doolittle, 2013; Graur et al., 2013). Hence, for instances of DNA binding, the burden of proof lies in demonstrating their functionality.

The class of TFs for which crowdsourcing is most active, those with degenerate motifs, may have the heaviest burden. Until fairly recently, function was thought to accrue exclusively to stably bound proteins (Chen and Rajewsky, 2007; Spitz and Furlong, 2012). TFs with degenerate motifs tend to be weak binders due to the motif's combination of short length and relatively low levels of adenine and thymine (Pan et al., 2010b), nucleotides that form only single hydrogen bonds with their respective complement. Indeed, it was shown that sites that were bound weakly during fly development were not able to drive a luciferase reporter construct, in contrast to the majority of strongly bound sites tested (Fisher et al., 2012). And among sites bound by the TF RAP1 in a modified yeast strain, those with the highest rates of turnover

(disassociation followed by re-association) were the most likely to incur nucleosome incursion and least likely to induce transcription (Lickwar et al., 2012). Interestingly, RAP1 turnover rate was poorly correlated with occupancy (0.14), as determined by ChIP-Seq, suggesting that high occupancy alone is not a guarantee of function.

Conversely, stereotypically weak binding subject to rapid turnover does not ensure absence of function (Segal et al., 2008). Dynamic binding, by its nature, is often associated with developmentally significant regulation (Cao et al., 2010; Wilczyński and Furlong, 2010). Moreover, there is long-standing evidence that binding affinity does not necessarily correlate with function (Davis et al., 1990). For example, in yeast a significant percentage of sites under purifying selection are of lower predicted affinity than consensus sites that bind the same TF (Tanay, 2006). Similarly, in human T cells, conserved CTCF-bound sites exhibit a wide range of affinities; indeed, the lowest occupancy class is the most strongly identified with cell type-specific function (Essien et al., 2009a)

Putative sites that recognize degenerate motifs, in particular, have gained wider recognition for their importance (Ramos and Barolo, 2013). From an evolutionary perspective, binding sites recognizing small, low-information motifs are critical to maintaining stabilizing selection as the size of cis-regulatory modules has expanded (Stewart and Plotkin 2012; Stewart et al. 2013). In the context of homotypic clusters of BS, such sites are, in fact, unexpectedly prominent in promoters and in enhancers (Gotea et al., 2010). As a likely function of their modest but significant capacity to increase binding robustness as a function of the number of BS they contain (Brackley et al., 2012), homotypic clusters have been implicated in timing of enhancer activation during

development (Rowan et al., 2010); controlling whether TF binding induces activation or repression (Ramos and Barolo, 2013); and shown to be necessary for functional binding at bona-fide sites for the Hox TF Ubx while simultaneously preempting ectopic binding at sites for closely related Hox proteins (Crocker et al., 2015).

In the previous chapter, we demonstrated a substantial boost in occupancy relative to stereotypical homotypic clusters in spatial homotypic clusters. In order to test for functional impact of this occupancy boost, we expanded the scale of observation from binding site to enhancer- and archipelago-wide. Occupancy boost is overwhelmingly centered in BS recognizing degenerate motifs, so we screened archipelago enhancers on the basis of their enrichment for such BS. When the resulting classes were compared, they displayed a striking divergence in character, with AP enhancers enriched in degenerate motifs (‘enriched enhancers’) substantially more affected than AP enhancers depleted for degenerate motifs (‘depleted enhancers’), after being normalized against matched non-archipelago (non-AP) enhancers. Specifically, enriched enhancers exhibited several-fold greater boost in activity, their neighboring genes exhibited several-fold greater expression and, consistent with higher functional significance, they exhibited several-fold greater normalized evolutionary conservation. Finally, we found that tissue-specific AP-wide activity (estimated as chromatin accessibility) scales with the tissue-specific expression of cognate TFs with degenerate – but, not specific – motifs. These results implicate crowdsourcing in: (i) initiating a positive feedback loop whereby greater TF occupancy at enriched enhancer BS increases the overall accessibility at these enhancers, thus facilitating further occupancy; (ii) endowing enriched enhancers with switch-like behavior, activating them in specifically those tissues where chromatin

structure and TF availability together result in sufficient occupancy boost. Together, we find strong evidence for the crowdsourcing occupancy boost's functional role in tissue-specific gene complex activation.

Results

AP enhancers enriched for degenerate motifs have greater occupancy boost

Previous results (Chapter 3) showed that a TF's occupancy boost scales with its BS abundance, or equivalently, its motif degeneracy in an AP. This led us to hypothesize that occupancy boost due to crowdsourcing may not uniformly impact all AP enhancers, but rather predominate in AP enhancers that are enriched for degenerate motif BS. For a specific dichotomous threshold for motif degeneracy, we defined '*enriched*' enhancers as those having significantly greater-than-expected degenerate motif BS; '*depleted*' enhancers are at the other end of the spectrum and, hence, have greater-than-expected non-degenerate motif BS. Note that, *a priori*, enriched enhancers are not expected to have a greater occupancy boost for a given TF compared to a fellow AP member with the same number of cognate sites but which, overall, is depleted for degenerate motifs. Unexpectedly, however, the enriched enhancers displayed boosts of up to 50% higher magnitude than those observed in depleted enhancers for the same mean coverage (for the given TF) despite no significant differences in either total BS per enhancer or chromatin accessibility.

We reasoned that if enriched enhancers were disproportionately larger contributors to the occupancy boost than depleted enhancers, then coverage (number of cognate BS) tallied based on enriched enhancers alone would be a

more direct predictor of occupancy boost. As shown in Figure 4-1 in the highest coverage bin, occupancy boost was two-fold higher in enriched than in depleted enhancers (~160% vs ~80%). We observed a similar trend when occupancy was determined using ChIP-Seq data (135% vs. 70%) (Figure 4-2), and in an alternative set of APs (Sheffield et al., 2013), where there was more than a 2-fold difference in footprint-based occupancy boost between enriched and depleted enhancers (70% vs. 33% occupancy boost for top 5% coverage, $p=3x10^{-3}$, Figure 4-3). To account for these unexpectedly high boosts, we address the potential for higher-order interactions within enriched enhancers among BS for distinct TF (see Discussion).

Figure 4-1. Enhancer enriched for degenerate motifs feature higher occupancy boost than enhancers depleted for degenerate motifs. Percentage occupancy boost is shown as a function of coverage for AP enhancers with the highest 20% enrichment (blue line) and the highest 20% depletion (green line) for low-RE BS, along with their 95% confidence intervals. Coverage for a given TF-AP pair was calculated as the number of cognate BS in the AP among enriched (depleted) enhancers only. A p-value is given for a Wilcoxon test comparing boosts among TF-APs with top 20% coverage. RE cutoff = 5.

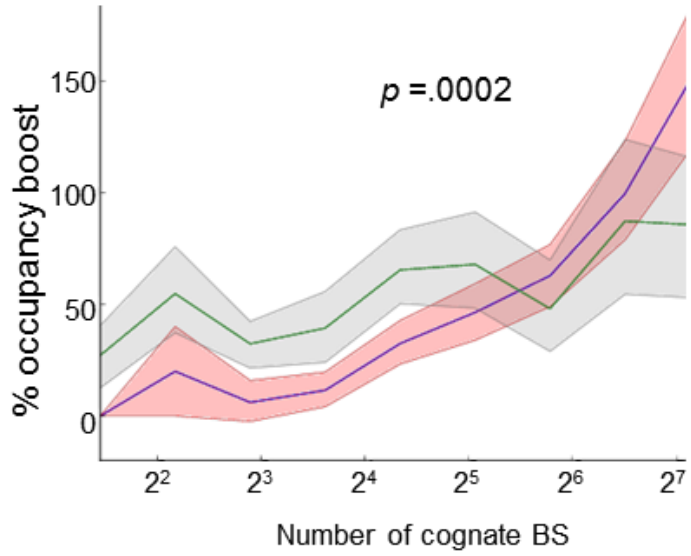


Figure 4-2. Validation using ChIP-Seq derived occupancy of higher boost in degenerate motif-enriched than depleted AP enhancers. ENCODE ChIP-Seq data was used for all cell types in which at least one AP was active, that is, for which at least 90% of an AP's enhancers were chromatin accessible. This resulted in 206 Chip-Seq experiments for 89 unique TFs across 9 cell types.. Enhancers with more degenerate motifs than expected ('enriched' enhancers) have higher occupancy boost than enhancers with fewer than expected ('depleted enhancers'). TF-AP coverage (x-axis) was computed as the number of cognate BS in just their enriched or depleted enhancers, respectively. Test result shown is for comparison of occupancy boosts computed in enhancers with the highest 50% in enrichment (green) to occupancy boosts computed in enhancers with the lowest 50% enrichment (grey), pooled across the three topmost coverage bins. An RE threshold of 5 was used to classify motif instances as degenerate for the purpose of computing enrichment, based on a Fisher Exact test. RE: relative entropy; BS: binding sites.

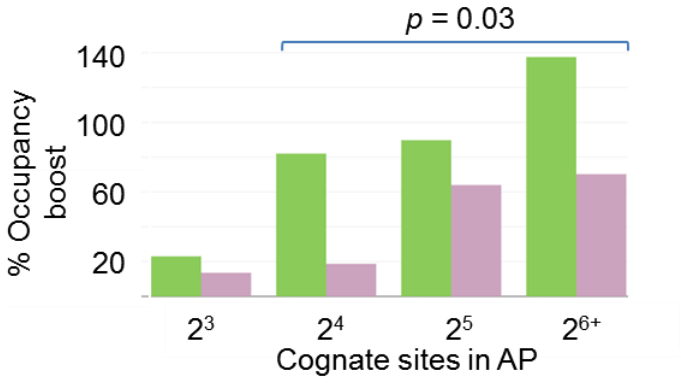
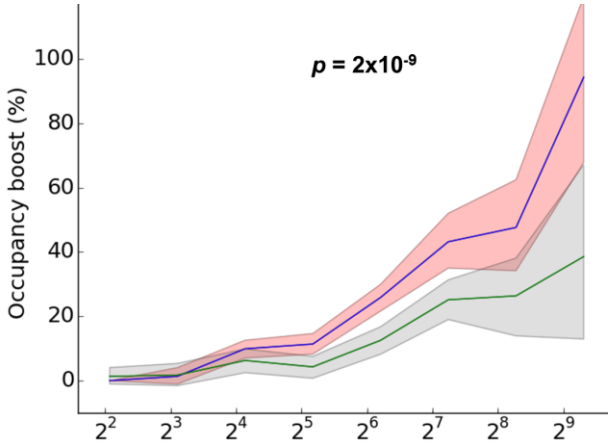


Figure 4-3. Validation of occupancy boosts using alternative archipelago data sets. Occupancy boost was determined for APs comprising sets of coordinately active

regions from (Sheffield et al 2013) that were then overlapped with putative enhancers (P300 ChIP-Seq peaks). Boost was calculated with respect to a background of 'non-AP' enhancers, which did not belong to any Sheffield set of co-active regions of size five or greater. Occupancy boost is more robust in enriched than in depleted AP enhancers, where 'enriched' and 'depleted' refer to the balance of low-RE BS. Percentage occupancy boost is shown for AP enhancers with the highest 20% enrichment (blue line) and the highest 20% depletion (green line) for low-RE BS, along with their 95% confidence intervals. Coverage for a given TF-AP pair was calculated as the number of cognate BS in in the AP among enriched (depleted) enhancers only. A p-value is given for a Wilcoxon test comparing occupancy boosts between enriched and depleted enhancers in TF-APs having top 20% coverage. 95% confidence interval shown based on a bootstrap procedure. RE threshold of 5 was used to calculate enhancer enrichment for low-RE BS. RE: relative entropy; BS: binding sites.



Enriched enhancers exhibit greater activity and evolutionary conservation

Enriched enhancers are more strongly associated with strong neighbor gene expression

Given the elevated occupancy boosts at degenerate motif enriched enhancers, we assessed whether such enhancers are associated with a greater expression of their target genes (Fisher et al., 2012; Smith et al., 2013). For each AP enhancer, assuming its closest gene neighbor to be its putative target (Djebali et al., 2012), we calculated its 'expression boost', as the relative difference in expression between its target gene and the target gene of the control non-AP enhancer. In

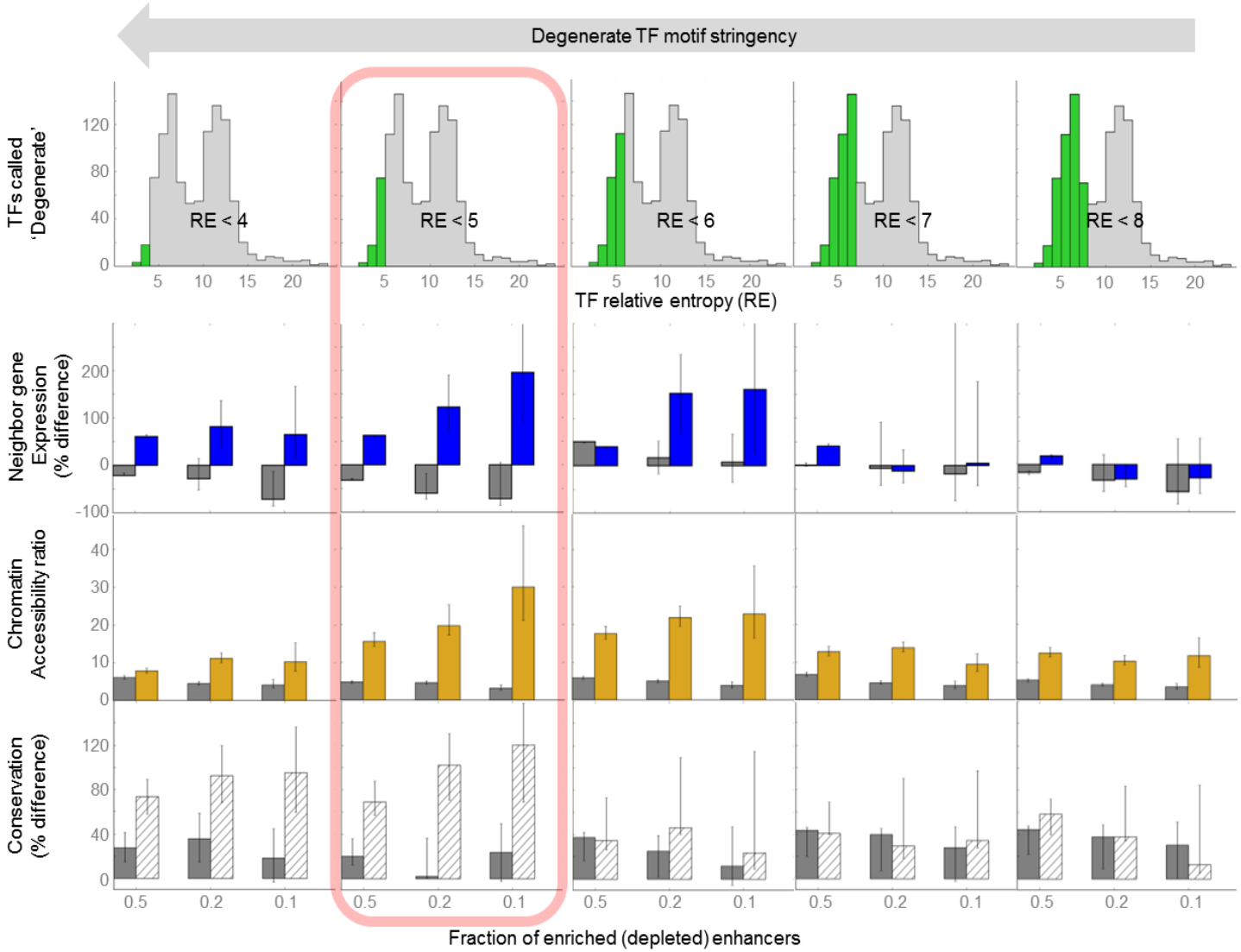
contrast to the determination of occupancy boost (Chapter 3), in the following analyses each enhancer is mapped to exactly one non-AP (control) enhancer on the basis of enhancer-wide distribution of degenerate and non-degenerate sites, using a variable threshold of degeneracy (Methods). We then compared expression boost for *enriched* enhancers to that for *depleted* enhancers (Methods). As shown in Figure 4-4 (row 2), the putative targets of enriched enhancers have much greater expression than their non-AP counterparts, while the depleted enhancers do not. Moreover, as the degree of enrichment increases from top 50% to top 10%, the relative expression boost increases from 62% to 196% (for degeneracy cutoff of 5, indicated by the pink loop); In contrast, genes near depleted AP enhancers have lower expression than their non-AP counterparts (discussed later) GC content differences between enriched and depleted enhancers do not explain these trends, as GC content in non-AP enriched (respectively, depleted) enhancers is on average <10% (respectively, 15%) higher than in corresponding AP enhancers. Note that at higher degeneracy cutoff (being more permissive) the observed effect weakens and eventually disappears.

AP enhancers near highly expressed genes bind a disproportionately high fraction of degenerate motifs.

As a complementary test of our hypothesized link between degenerate site enrichment in an AP enhancer and its target gene expression, we assessed whether degenerate BS are more abundant in AP enhancers driving highly expressed genes than in AP enhancers driving weakly expressed genes. We compared the ratios of bound degenerate sites to bound specific sites in enhancers that were within 50Kb of genes with, alternatively, top and bottom 25% expression. Each BS was

classified as degenerate, specific, or neither based on its putative TF and two variable degeneracy thresholds for degenerate and specific BS. We found that the ratio of degenerate to specific occupancy is consistently greater in enhancers neighboring highly expressed genes by 1.8 to 3-fold compared to enhancers neighboring low-expressed genes, and monotonically increases for more stringent thresholds for specific motifs. In contrast, for a control set of non-AP enhancers chosen based on the same proximity criteria as AP enhancers, the ratio of two classes of TF binding do not deviate significantly from 1.0. Taken together, these results suggest that degenerate binding specifically at enriched AP enhancers has a significant impact on downstream gene expression.

Figure 4-4. Enhancers enriched for degenerate BS are more functional than expected. Enhancers enriched and depleted for low-RE BS were compared in terms of DNase hypersensitivity (row 2), evolutionary constraint (row 3), and neighbor gene expression (row 4). Readouts on y-axes indicate values normalized against carefully matched non-AP enhancers for the given RE cutoff (column). Within each plot, the 10%, 20%, and 50% (x-axis) most enriched enhancers are indicated in non-grey, while the most depleted enhancers are shown in grey. Note that 50% most depleted enhancers for degenerate motifs are synonymous with the 50% most enriched for enhancers specific motifs. The histograms in the top row indicate the fraction (green) of all TFs deemed low-RE for the purpose of calculating each enhancer's low-RE BS enrichment. The pink loop shows the consensus degeneracy (RE) level at which all metrics are most divergent between enriched and degenerate enhancers. RE = relative entropy


Degenerate TF motif stringency
TFs called 'Degenerate'
RE < 4
RE < 5
RE < 6
RE < 7
RE < 8
TF relative entropy (RE)
Neighbor gene Expression (% difference)
Chromatin Accessibility ratio
Conservation (% difference)
Fraction of enriched (depleted) enhancers

Enriched enhancers are more accessible and more highly acetylated than expected

TF binding and chromatin accessibility are intimately connected; higher accessibility typically leads to higher occupancy, while TF binding can help displace a nucleosome and increase accessibility (Teif and Rippe, 2012). We therefore assessed whether enriched enhancers exhibit a greater boost in overall accessibility compared with depleted enhancers. For this analysis, we normalized AP enhancer accessibility by that of stringently matched non-AP enhancers as described previously, except the variable of interest, DHS, was explicitly left uncontrolled for this analysis. As shown in Figure 4-4A (row 3) and Figure 4-5, at the stringent degeneracy threshold of 5 (higher thresholds are more permissive), the most enriched enhancers exhibit ~10-fold greater DHS boost with respect to matched non-AP enhancers than do depleted enhancers. To further resolve the effect degeneracy on enhancer accessibility, we tracked changes in accessibility as we increased the number of degenerate (specific) sites, while holding relatively constant the number of specific (degenerate) sites. As shown in Figure 4-6, increasing the number of degenerate BS has a substantial positive impact on enhancer's accessibility – especially when the number of specific BS is low, while increasing the number of specific BS does not. In addition, we found that histone acetylation level (H3K27Ac), which is associated with active enhancers, also is ~3-fold higher in enriched enhancers than expected (Figure 4-7).

Figure 4-5. Enhancers enriched for degenerate motifs exhibit the largest fold-change in accessibility from non-archipelago to archipelago state. Chromatin accessibility shown for AP and non-AP enhancers within 50KB of a highly expressed gene that were matched one-to-one for motif composition with an AP enhancer. Enhancers sorted by degenerate motif enrichment. Aligned heatplots display one enhancer per row (most enriched at top). (Left) heatplot in which red signifies a low-RE motif and blue a high-RE motif. Low-RE motif enrichment based on Fisher exact test against a background that included all non-AP enhancers. For visualization purposes, enhancer lengths and BS lengths standardized. (Middle, right). Log of non-AP and AP enhancer DHS, respectively. RE cutoff for low-high degeneracy

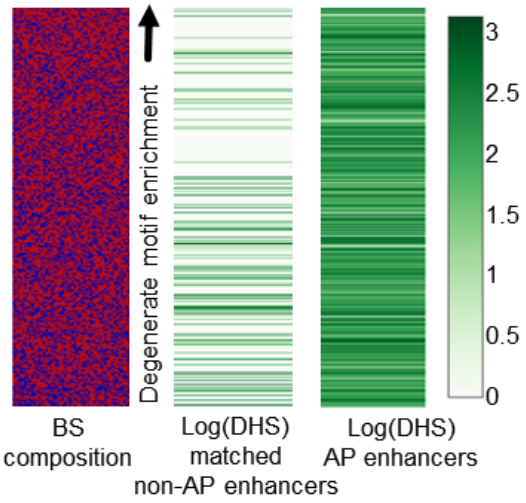


Figure 4-6. The ratio of AP to non-AP enhancer DHS rises with increasing numbers of low-RE BS, but not high-RE BS. Plot indicates trend in DHS ratio as the count of degenerate (low-RE) BS increases (along axis labeled '|Low-RE-sites|'), and the count of non-degenerate (high-RE) motifs is held roughly constant -- or vice versa. AP enhancers were partitioned into equal sized bins along each of two axes based on degenerate and non-degenerate motif counts, as shown. y-axis gives the mean AP DHS normalized by non-AP DHS. The trend remained strong when enhancers were subdivided into a greater number of bins (3 or 4, not shown). Motif degeneracy classification was based on an RE threshold of 5. RE: relative entropy.

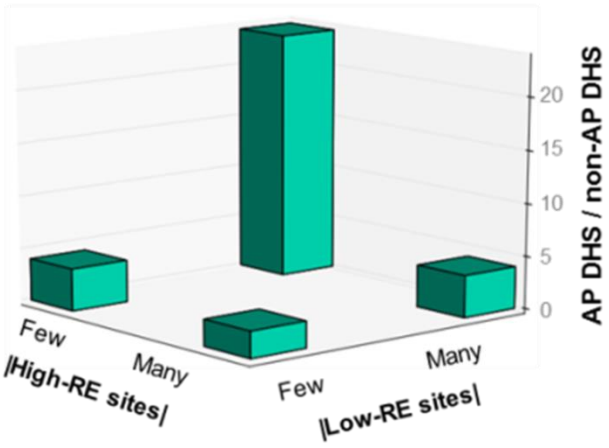
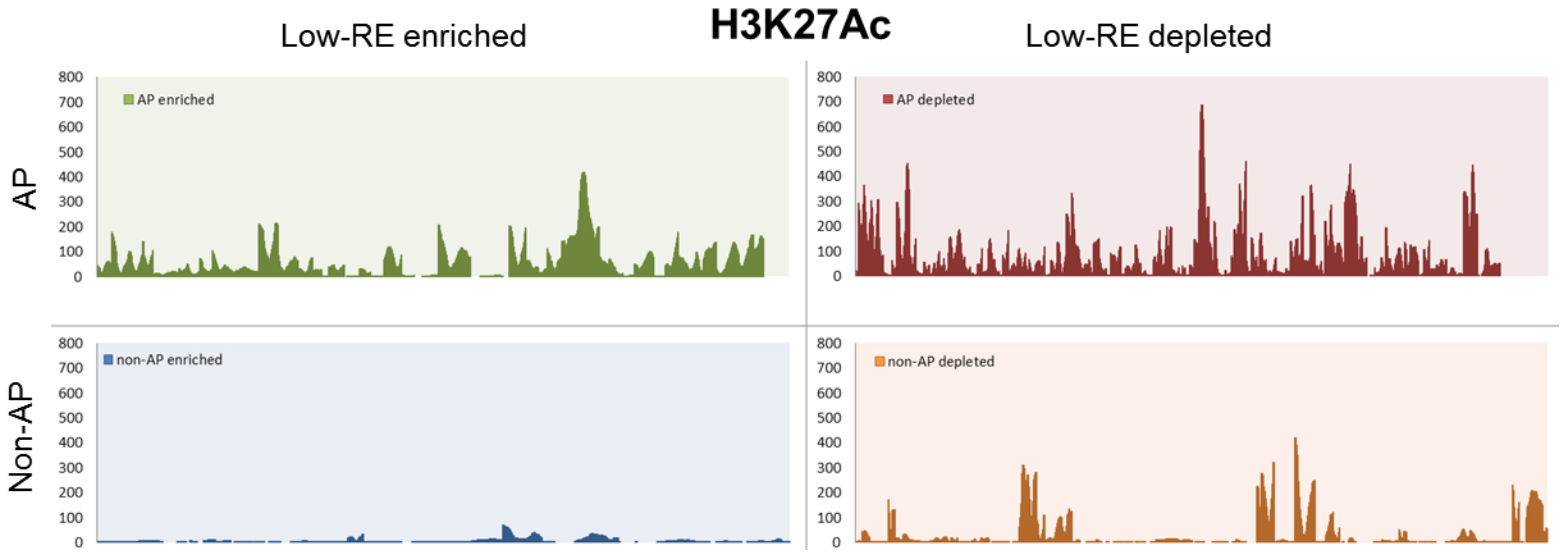

AP DHS / non-AP DHS
20
15
10
5
0
Few
Many
|High-RE sites|
Few
Many
|Low-RE sites|

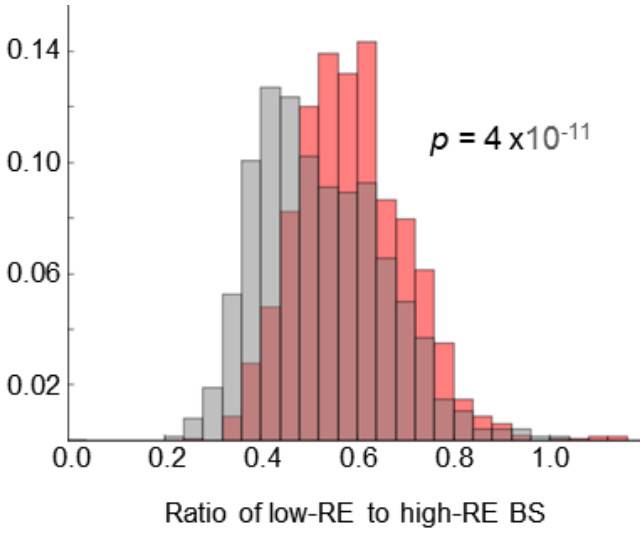
Figure 4-7. Acetylation levels in enriched vs. depleted enhancers. Juxtaposed views of H3K27Ac ChIP-Seq in HUVEC are shown for 40 (100) AP enhancers in the top row that are in an AP that is active and in the top 10% for enrichment (depletion). Shown in the bottom row are views for matched non-AP enhancers. Motif degeneracy classification was based on an RE threshold of 5. RE = Relative entropy



Evolutionary conservation for enriched enhancers greater than expected

As an additional ascertainment of the functional importance of enriched AP enhancers, we found such enhancers to be up to 120% more evolutionarily conserved (using 20-species PhastCons scores (Siepel et al., 2005)) than matched non-AP enhancers; indeed, the greater their enrichment, the greater the evolutionary constraint we observed (Figure 4-4, row 4). Depleted AP enhancers, by contrast, were at most 40% more conserved than their non-AP counterparts. Finally, we observed that there is a substantially higher proportion of enriched enhancers in AP than non-AP (Figure 4-8). These results – the relatively higher occupancy boosts, chromatin accessibility, downstream gene expression, and evolutionary constraint in enriched enhancers, along with greater prevalence of enriched enhancers among AP than non-AP enhancers – strongly suggest a hitherto unreported special functional relevance of AP enhancers that are enriched for degenerate binding sites.

Figure 4-8. Ratio of low-RE to high-RE motifs in AP enhancers vs. non-AP enhancers. AP and non-AP enhancers were matched one-to-one for DHS in each AP enhancer's most active tissue. Putative BS were identified based on 95 percentile motif match threshold. The x-axis shows the ratio of low-RE to high-RE motif sites in each enhancer. Y-axis shows percentage of enhancers analyzed. P-value from a Wilcoxon test comparing ratios in AP and non-AP enhancers.



AP enhancer activity is correlated with availability of TFs with degenerate motifs only

Our results thus far suggest that crowdsourcing may be intimately connected to the regulation of AP enhancer-gene complexes, as it provides a way for the cell to prime or induce activity in multiple genomic elements simultaneously, in a specific spatial and tissue context. As shown above, the boost in overall activity (approximated by DHS) of an AP enhancer is in fact far higher in AP enhancers enriched for degenerate (high coverage) BS. However, the direction of causality is not clear – that is, whether the binding of TFs corresponding to the degenerate motifs increases overall accessibility at enriched enhancers, or alternatively, already increased accessibility at enriched enhancers (by some unknown mechanism) fosters greater occupancy of particular TFs at those enhancer. In order to resolve this circularity, we tracked tissue-specific gene expression of TFs in 9 cell types and studied its relationship with tissue-specific AP enhancer accessibility (Methods). As shown in Figure 4-9A, mean AP enhancer

accessibility increases robustly (up to 400%) with increasing expression of TFs comprising high-coverage (red), but not low-coverage, (gray) TF-APs. In non-AP enhancer sets controlled for degenerate and specific BS counts, no such associations were observed at all (Figure 4-9B). Thus, AP enhancer accessibility and activity is highly responsive to the levels of high-coverage TF-APs as they vary across tissues. Together, these results strongly suggest that crowdsourced boosts in TF occupancy, through the context-specific binding of high coverage TFs, may help drive tissue-specific activation of enhancer networks and their target gene complexes.

Figure 4-9. Mean AP accessibility scales with context-specific availability of TFs with degenerate motifs but not TFs with specific motifs. For each TF-AP, tissue specific DHS was compared across each of 15 tissues for which there was RNA-Seq data available. (TF, AP, tissue) triplets were segregated into lowest-20%-coverage (cyan) and highest-20%-coverage (red) classes based on TF-AP, and then further subdivided into low and high expression based on tissue-specific TF expression. Bar height indicates the percentage increase in DHS level associated with an increase in TF expression from bottom <x> to top <x> percentage levels, where <x> is read off the x-axis. 1% confidence intervals from a bootstrap procedure. (D) same as (C) except matched non-AP triplets were used.

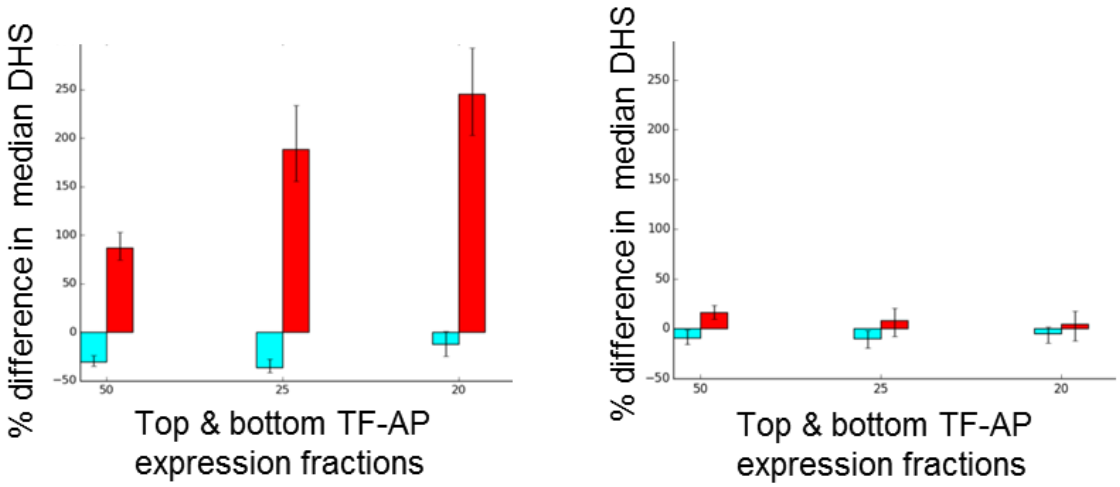
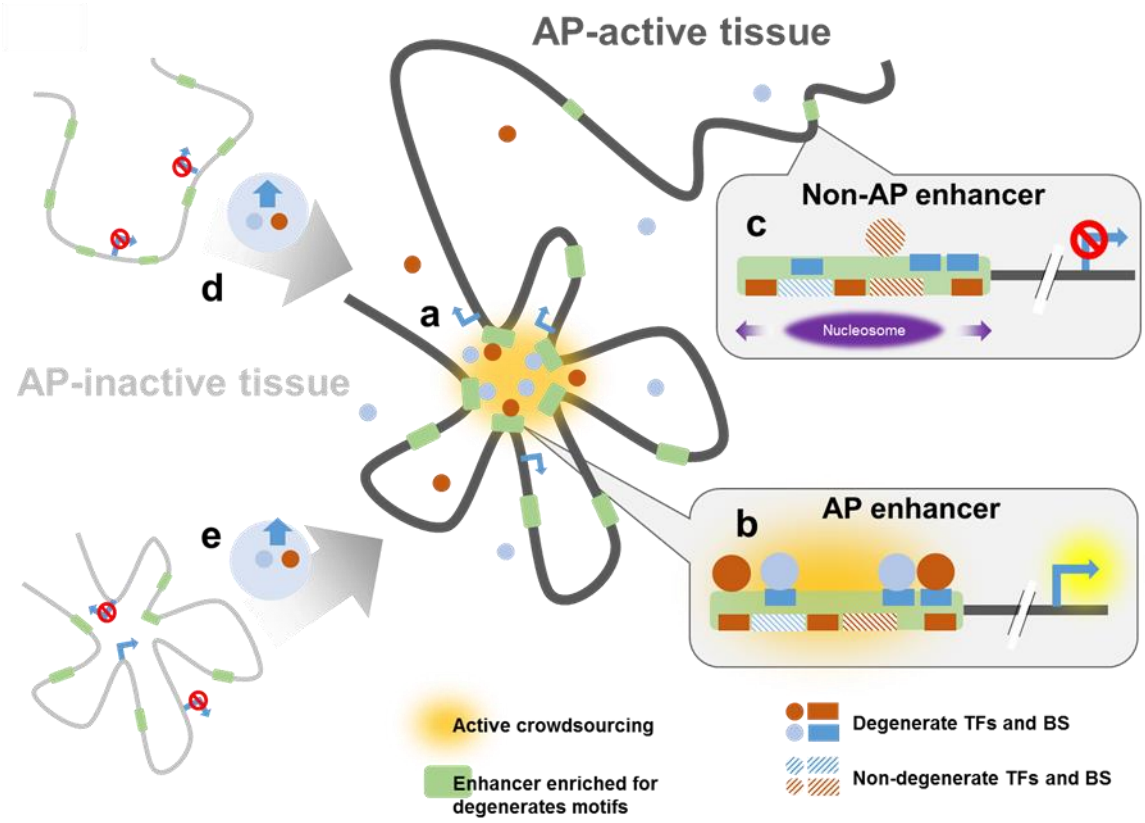


Figure 4-10. Model of crowdsourcing effect. (a) The yellow highlighted region represents a regulatory archipelago (AP) consisting of genes and distal enhancers. Within an AP, spatially proximal binding sites (BS) for a common TF 'crowdsource' an increase in their own occupancy. Facilitated by increased TF diffusion among large numbers of spatially proximal BS, a spatial homotypic BS cluster favorably alters TF protein concentration in its microenvironment. Predictably, TFs with degenerate motifs, and hence pervasive BS, exhibit the highest occupancy boosts. (b) In turn, AP enhancers enriched in degenerate motifs experience switch-like multi-fold boosts in accessibility and target gene expression. Overall, a context-specific increase in availability of TFs with degenerate motifs – but not high-specificity motifs – drives a multi-fold boost in chromatin accessibility, thereby underscoring crowdsourcing's likely role in AP activation. (c) In contrast, a non-AP enhancer does not experience an occupancy boost and activation. The crowdsourcing mechanism integrates well with the two prevailing models of context-specific gene module activation: in a targeted tissue, higher expression of TFs with a degenerate motif may (d) induce chromatin loop formation; or alternatively (e) facilitate release of paused polymerase in pre-formed enhancer-promoter loops. In both cases, crowdsourcing ensures a high degree of context-specificity, mitigating spurious occupancy outside of or AP-active tissue or AP enhancers enriched for degenerate motifs.



Discussion

Summary. In a previous work, we demonstrated a previously undescribed occupancy boost that is broadly emergent in spatially concentrated clusters of homotypic BS – typical of sites cognate to a degenerate motif TF in an active regulatory archipelago. Here, we have probed the functional importance, if any, of this crowdsourcing occupancy boost, by comparing whole enhancers that are, alternatively, enriched or depleted for degenerate motifs. Consistent with functional significance for the observed occupancy boost, we detected at least two-fold higher gene expression of neighbor gene loci, normalized evolutionary conservation, and chromatin accessibility among the enriched enhancers.

Higher order impact of crowdsourcing. Unexpectedly, we observed up to two-fold higher occupancy boost in addition to 10-fold greater normalized chromatin accessibility in AP enhancers enriched for degenerate motifs ('enriched enhancers') than in depleted enhancers. A likely explanation is the emergence of an aggregate occupancy effect among an enriched enhancer's abundant degenerate BS, which serves to remodel the local chromatin state. Under inactive conditions – that is, in AP-inactive tissues or outside of APs – we found that enriched enhancers (which inherently tend toward far lower GC content than depleted enhancers) display substantially higher chromatin accessibility compared to depleted enhancers. This is consistent with previous work suggesting that nucleosomes favor unbound, low GC-content sequence, yet are readily displaced by strongly binding pioneer factors, or, as in the case of crowdsourcing, by an aggregate of distinct TFs (Barozzi et al., 2014; Wasson and Hartemink, 2009).

In an AP-active tissue, enriched AP enhancers experience a widespread surge in binding, thereby displacing the nucleosome and boosting occupancy further, in a positive feedback loop (Figure 4-10). Taken together, the markedly divergent accessibility inside versus outside an active AP confer to enriched enhancers switch-like behavior, where their state is determined by their context: included in an AP replete with degenerate homotypic BS, their accessibility increases – but only in tissues in which the cognate TFs are available. In light of this highly context-specific activation and the rapid evolutionary gain of BS for degenerate motifs, we suggest that enriched AP enhancers can evolve adaptively relatively free of consequences from spurious binding. This is the first work to highlight the special functional significance of AP enhancers enriched for abundant, degenerate motif BS. Intriguingly though, we found that the genes near depleted AP

enhancers are expressed at up to three-fold *lower* levels than their non-AP counterparts (Figure 4-4A row 2). Further work is needed to investigate to what extent depleted enhancers have a unique, perhaps repressive, role. Interestingly, genes controlling cell identity in stabilized chromatin structures were found accompanied by repressed genes that coded for yet other lineage-specifying regulators (Dowen et al., 2014), while certain super-enhancer constituent enhancers, confounded expectation by not inducing transcription when cloned into reporter constructs (Hnisz et al., 2015).

Crowdsourcing integrates well with the two prevailing models of coordinated activation of spatially co-localized gene complexes (Figure 4-10), while providing a missing piece of the puzzle. Whether (a) long-range enhancer-gene loops form *de novo* upon (or along with) activation of a gene cluster (Deng et al., 2012), or (b) the loops are pre-formed and the paused polymerase is released due to a change in TF availability (Ghavi-Helm et al., 2014), the cell requires TFs to functionally bind and activate elements specifically in a targeted gene cluster. Crowdsourcing of low-affinity BS is well-suited for such targeting, as it can induce specificity through emergent switch-like binding behavior, discussed in Chapter 3. Interestingly, a recent study showed a strong correlation between pathway-level gene activity and pathway-level spatial proximity across cell types (Karathia, Hannenhalli et al., under review), suggesting that chromatin structure is intimately connected with gene complex activation. In contrast to direct enhancer-gene interactions in the standard model for distal transcriptional regulation, crowdsourcing and its downstream impact are not observable at the level of single enhancer-gene interaction, but instead emerges only at higher levels of chromatin organization and co-regulated gene modules.

Materials and Methods

Enhancer clusters ('APs'). *See Chapter 3 Methods*

Estimating in vivo occupancy at a BS using digital footprint data. *See Chapter 3 Methods*

AP-active and AP-inactive cell lines. *See Chapter 3 Methods*

Establishing non-AP control for occupancy boost. *See Chapter 3 Methods*

Determining TF occupancy at enhancer resolution with ChIP-Seq data. *See Chapter 3 Methods*

Estimating TF's degeneracy. *See Chapter 3 Methods*

Determining occupancy boost with alternative set of AP enhancers. *See Chapter 3 Methods*

Identifying degenerate motif enriched and depleted AP enhancers. For a specific RE cutoff each putative BS in an enhancer was classified as either degenerate or specific (complement of degenerate). This cutoff was varied from RE = 4 (classifies ~2% enhancers as low-RE) to RE = 9 (classifies > 50% enhancers as low-RE). For each AP enhancer, after tallying the number of degenerate and specific motif BS, an enrichment p-value was generated by applying a Fisher Exact test comparing the numbers of BS in each class in the enhancer to those in the pooled set of control (non-AP) enhancers. Based on this enrichment p-value, enhancers were sorted, and the top (enriched) and bottom (depleted) ranked x% of enhancers compared in subsequent analysis (x in {10, 20, 50}).

Creating a non-AP control for enriched and depleted AP enhancers. We paired each AP enhancer with one of the remaining non-AP enhancers while controlling for DHS peak height (within 2%) and numbers of both degenerate and specific motif sites (within 2%), where degeneracy class is based on a (variable) degeneracy threshold. This yielded ~1200 pairs of AP and matched non-AP enhancers; the exact number varied with the degeneracy threshold.

Comparing neighbor gene expression between AP and non-AP enhancers. As a proxy for an enhancer's target gene, following the convention (Djebali et al., 2012), we used the gene closest to the enhancer. As an extra measure of stringency, in case of non-AP enhancer, we excluded those enhancers that were farther than 50kb from the nearest gene promoter. For gene expression, five cell types were used for which overall AP activity, calculated as described above, was at or near its maximum as observed in 15 cell types for which we had digital DNase footprint and RNA-Seq data (www.encodeproject.org/ENCODE). These were HSMM, A549, NHLF, Ag04450, and Bj.

Calculating a normalized conservation score. To compare evolutionary conservation of degenerate BS enriched AP enhancers to depleted AP enhancers, we used PhastCons scores, based on 20 mammalian species (Siepel et al., 2005), which are resolved to the individual base. Mean scores across the two classes of enhancers were normalized with respect to non-AP enhancers matched one-to-one with an AP enhancer, as elsewhere in the manuscript. Additionally, we ensured that non-AP enhancers were within 50Kb of the promoter of a highly expressed

gene (fpkm > 1.0), which includes approximately the ten percent most highly expressed genes.

TF expression-AP activity correlation. This analysis used data from each of 15 cell lines for every AP, encompassing ~2.4 million BS. Each (TF, AP enhancer, cell line) triplet was assigned (i) a DHS value, corresponding to AP enhancer and cell line; (ii) a coverage score, corresponding to AP enhancer and TF; (iii) a normalized RNA-Seq value corresponding to TF and cell line. Analysis was limited to triplets with a coverage score in the top and bottom 20%. In each of these coverage classes, triplets were further sorted based on the TF's expression in the given cell line and screened to include only triplets with top or, alternatively, bottom 20 (or 25 or 50) percent TF expression. For each coverage class, the percentage difference in mean cell-type specific DHS between the low TF expression and high TF expression cohorts was plotted. Confidence intervals for each percentage difference were computed on the basis of 50K bootstrap replicates.

H3H27Ac levels. We downloaded Encode ChIP-Seq peaks from human umbilical vein cells (HUVEC) for histone mark H3K27Ac, known to be associated with active enhancer states (Calo and Wysocka, 2013). This cell line was chosen for its combination of available data and a large number of enhancers in APs that are active in the cell line. We compared the ratio in mean ChIP-Seq levels between top 10% enriched and top 10% depleted AP enhancers to the same ratio for non-AP enhancers, matched one-to-one with the AP enhancers as described above. An AP enhancer and its matched non-AP enhancer were included only if the AP

enhancer belonged to an AP that was ‘active’ in HUVEC (>80% of its enhancers was DNase hypersensitive). This resulted in ~40 enriched and ~100 depleted AP enhancers, and the same number of non-AP enhancers.

Chapter 5: **Crowdsourcing fosters archipelago compaction**

Abstract

Chromatin's three dimensional topology has emerged as a critical facilitator of transcriptional regulation. In particular, spatial proximity among genes and distal enhancers in a co-regulated complex appears to be a prerequisite for strong expression. Little is known, however, about the extent to which spatial proximity, itself, is functionally regulated across tissues, let alone the mechanisms responsible.

In our previous work, using known chromatin hubs, or 'archipelagos', of spatially colocalized enhancers, we demonstrated that spatially concentrated binding sites (BS) for a shared, typically degenerate motif transcription factor (TF), can reshape the TF's local micro-environment and 'crowdsource' higher TF concentration and BS occupancy. Here, we test whether this crowdsourced increase in local TF concentration, through a positive feedback loop, itself augments chromatin looping and, consequently, spatial proximity among archipelago BS. Specifically, we seek evidence for two complementary mechanisms: (1) increased interactions between non-contiguously DNA-bound heterodimer TFs, which create anchor points for chromatin loops; and (2) increased recruitment of proteins implicated in chromatin looping – cohesin and Mediator complex – and various chromatin modifying enzymes (CME).

Based on high resolution Hi-C data and consistent with previous reports limited to a few isolated systems, we find that, indeed, there is a generalized tendency for archipelagos to significantly compact in 'active' cell types relative to less active cell types; this is achieved through increased formation of chromatin loops. As predicted by

the crowd-sourcing effect, the increased looping that accompanies transition from an inactive to active cellular context occurs at a several-fold higher rate between enhancers enriched for degenerate BS than between enhancers across different archipelago, generally. To test whether degenerate motif TF binding is a principal driver of the observed compaction, we next assayed changes in looping as a function of increasing TF availability across cell types, as estimated by TF gene expression. Consistent with the crowdsourcing effect, looping increased more in lockstep with increased expression of degenerate TFs than of specific TFs. In turn, supporting TF-TF interactions as a contributing mechanism, we found that as cell type-specific expression of heterodimer TFs with degenerate motifs increases, their involvement in indirect ChIP-Seq interactions grows – in contrast to non-heterodimer TFs.

While more work remains, our preliminary findings suggest that the crowdsourcing effect exerts a positive feedback loop between BS concentration and TF concentration. This manifests as an increased local abundance of chromatin loops and thus greater spatial proximity among co-regulated archipelago enhancers. As such, this work reveals how DNA sequence, mediated by tissue-specific TF availability, contributes to the higher-order chromatin structure that underlies coordinate gene complex activation.

Introduction

Chromatin's spatial component has proven indispensable to understanding regulation of gene transcription. Genes encoding developmental regulators, for example, must integrate a complex set of regulatory inputs, many quite distal. By displacing intervening chromatin, chromatin loops foster spatial proximity, and interaction, between regulators and the transcriptional unit (Mukherjee et al. 1988; Montavon and Duboule 2012). In the

comprehensive picture of higher-order chromatin structure during interphase offered by the chromosomal conformation capture technique Hi-C, thousands of 'topological domains' ranging up to several megabases and demarcated by loops have been identified that are largely maintained across cell types – suggesting such domains represent a fundamental organizing unit of chromatin (Dixon et al., 2012; Shen et al., 2012; Vietri Rudan and Hadjur, 2015). Chromatin loops have also been found to aggregate into chromatin hubs or 'archipelagos' in a variety of species and model systems, including HOXD, olfactory receptor, and alpha-globin (Markenscoff-Papadimitriou et al., 2014; Montavon et al., 2013; Vernimmen, 2014). An important insight into the source of tissue-specific regulation comes from the most resolved examination of chromatin structure to date (Rao et al., 2014), in which so-called topologically associating domains (TAD) colocalizing in a nuclear compartment broadly share chromatin state, indicative of co-regulation, but colocalize with a changing cast of TADs across cell types. To understand cell type-specific regulation, therefore, particularly of coordinately regulated complexes, greater insights are required into the forces which contribute to the higher-order chromatin organization (Schwarzer and Spitz, 2014).

There is a long tradition of seeking the roots of fine-grained DNA structure in DNA sequence (Burge et al. 2006) and in sequence-protein interactions (Schultz et al., 1991). Indeed, DNA-protein interaction is thought to largely account for chromatin looping. Specifically, the ring-forming cohesin complex, best known for tethering sister chromatids after DNA replication (Nasmyth and Haering 2009) has more recently been shown to anchor loops critical to transcription, in complex with Mediator, when recruited by DNA-bound proteins, particularly CCCTC binding factor (CTCF) (Hadjur et al. 2009;

Mifsud et al. 2015; Kagey et al. 2010). Intriguingly, cohesin has recently been shown to bind tissue-specifically to non-CTCF TFs (Schmidt et al., 2010). However, at the level of archipelago and the context-specific coordination of multiple loops, mechanistic results are missing.

Notwithstanding, a consensus is emerging that protein binding likely holds the key to higher-order DNA organization (Bickmore and van Steensel, 2013; Dekker et al., 2013; Feuerborn and Cook, 2015a; Pombo and Dillon, 2015; Sexton and Cavalli, 2015). Recent biophysical simulations have successfully modeled broad patterns of DNA folding based on generic interactions between a polymer representing DNA and a collection of protein-like particles, each with two 'sticky' ends for binding DNA. In the 'String and binders switch' model, Barbieri et al (2012) found through Monte Carlo simulations that as the concentration of particles is increased, a threshold is reached where DNA exhibits a switch-like compaction into a structure replete with loops, recapitulating the power law that describes DNA contact probabilities observed *in vivo.* The non-specific TF-bridging model (Brackley et al., 2013a), based on molecular dynamical simulations, also produced compact DNA folding. Crucially, neither these nor related biophysical models, to our knowledge have been tested *in vivo*, perhaps due to the abstracted quality of their predictions. This stands in contrast to the concrete predictions made by the model we offer in this work.

Previously, we proposed and found functional genomic evidence for a novel group-level biophysical effect we named 'crowdsourcing', which augments local TF occupancy and TF concentration levels tissue-specifically, ultimately inducing gene complex activation. Briefly, we showed that such occupancy boosts are the likely

consequence for a TF in an archipelago that features a large number of cognate sites: TF proteins are briefly 'trapped' as they sequentially disassociate and re-associate amid the many spatially proximal, if genomically distal, BS. Here we ask, 'Does the accompanying local boost in TF concentration, mediated by tissue-specific chromatin conformation, result in further compaction of the chromatin structure? As the concentration boost is both spatially local and tissue-specific, it is a logical candidate for providing the needed coordination among archipelago enhancers to increase compaction. Based on observation in several distinct systems, such compaction appears to characterize the main difference in regulatory configurations of active and inactive gene complexes (Markenscoff-Papadimitriou et al., 2014; Montavon et al., 2011).

Testing this model leverages the expected divergence in behavior between, on the one hand, TFs with degenerate BS and their cognate BS, and on the other hand, TFs with higher information content and more specifically-binding motifs. Degenerate TFs stereotypically have very high abundances of cognate sites and hence, per the crowdsourcing model, should exhibit far higher effect size than TFs with specific motifs. Similarly, enhancers enriched for degenerate motifs are expected to be more intimately associated with tissue-specific changes in looping.

Additionally, we propose two complementary mechanisms through which increase in concentration and DNA occupancy for TFs with degenerate motifs, specifically, leads to increased archipelago compaction and, ultimately, transcriptional activation: (i) increased interactions between non-contiguously DNA-bound heterodimer TFs, consistent with the String and binders switch model; (ii) increased recruitment of the chromatin proteins cohesin and Mediator complex, as well as chromatin modifying

enzymes (CMEs) and co-factors, such as P300, critical to complex activation. Together with previous findings, this work aims to show that archipelago BS for degenerate motif TFs experience a virtuous cycle that drives increased TF concentration and cognate binding site concentration, in the form of increased archipelago compaction. We suggest this is mediated by tissue-specific TF availability, and premised on a nominal degree of spatial proximity in the 'ground state', i.e. in inactive cell types.

Indeed, preliminary results, as laid out in this chapter, are consistent with the described sequence-based mechanism for context-specific archipelago compaction. Specifically, we have found that (i) degenerate BS-enriched enhancers exhibit several-fold higher ratio of active-to-inactive state archipelago compaction (estimated by Hi-C interaction frequency) than archipelago enhancers, generally; (ii) archipelago loop formation occurs in closer lockstep with expression of degenerate TFs than of specific TFs; and (iii) heterodimers appear to form bridges between distal chromatin far more often within APs than outside of APs, and at rates that scale with the availability of degenerate TFs alone. Pending work, described below, aims to provide additional support for crowdsourcing's role in archipelago compaction, generally, and in boosting local recruitment of cohesin and chromatin modifying enzymes, specifically.

Results

In Chapters 3 and 4, we found spatial proximity among archipelago enhancers, as estimated by Hi-C, to be a pre-requisite for TF occupancy amplification. In this chapter this spatial proximity, or compactness, is treated as a dependent variable. Specifically, we hypothesize and test for a positive feedback mechanism in which increased TF concentration represents one half of a virtuous cycle, and increased BS concentration,

estimated by chromatin compactness, represents the other. To estimate compactness of an AP in a given cell type, we use its Hi-C based *edge fraction* – the fraction of all possible pairs of AP enhancers with evidence of significant interaction.

AP adopts more compact conformation in active tissues than in inactive tissues

Previous reports have shown that spatial proximity among an archipelago of enhancer elements is required for stereotypical gene activation (Markenscoff-Papadimitriou et al., 2014; Montavon et al., 2011). To test whether this is exhibited more generally across the genome, we analyzed 40 previously identified archipelagos (AP) using published 5-Kb resolution Hi-C data for 6 cell types: HUVEC, HMEC, IMR90, NHEK, K562, and GM12878 (Rao et al., 2014). Among the set of enhancers in each AP, we compared the combinatorial interaction frequency in active cell types to interaction frequency in inactive cell types, where AP activity was defined as the fraction of member enhancers that are DNase hypersensitive (DHS > 0). As in previous chapters, we used thresholds of >0.90 and <0.50, respectively. APs without at least one active and one inactive cell type were excluded, leaving 25 APs (~900 enhancers in total) for analysis. Within an AP, individual enhancer pairings within 100Kb of one another were excluded. Finally, we used a paired Wilcoxon test across the aggregate of ~21K enhancer pairs to compare interaction presence in active and inactive cell types. Surprisingly, interaction frequency was significantly lower in active than in inactive AP-cell type combinations – 0.046 vs. 0.067 (p-value = 4.5e-65). This can putatively be explained by the far greater presence in the inactive archipelagos of heterochromatic enhancers (Figure 5.1). Heterochromatin not only adopts a highly condensed configuration, locally but, critically, also co-localizes to a common nuclear compartment (Dixon et al., 2012, 2015; Rao et al., 2014). Genomic

regions in such a shared compartment often appear in Hi-C assays to have high contact frequencies (Dixon et al., 2012; Ulianov et al., 2015).

In Chapter 4, we observed that as cell type-specific AP activity fell below 90%, occupancy boost quickly dropped; at AP activity = 50%, crowdsourcing occupancy boost is nearly dormant (data not shown). We therefore repeated the above analysis with the same 'active' AP threshold of 90%, but with an inactive window now ranging from 50% to 90% AP activity. We now observe a highly significant increase in enhancer-enhancer interactions in active cell types with respect to inactive (0.051 vs. 0.037, p-value 1.6e-17) (Figure 5-2), with compaction occurring in 15 of the 19 APs with sufficient data. Moreover, and consistent with crowdsourcing mechanism, when enhancer-enhancer interaction are screened to include only enriched enhancers, the average compaction from inactive to active cell types for any given AP monotonically increases to more than two-fold for top-20% enriched enhancers (0.058 vs. 0.027, p=0.006) (Figure 5-2, *top*). As a control, we compared these results to those for inter-AP interactions, albeit on the same chromosome. Here, we see that, in contrast, and consistent with the absence of crowdsourcing effect, the compaction ratio does not increase among enhancers enriched for degenerate BS (Figure 5-2, *bottom*)

Figure 5-1. Percentage of enhancers in a given AP-tissue combination that are heterochromatic. Determination based on overlap between an enhancer and a region classified by ChromHMM as 'heterochromatin' (processed ChromHMM genome segmentation data for 5 cell types downloaded from ENCODE).

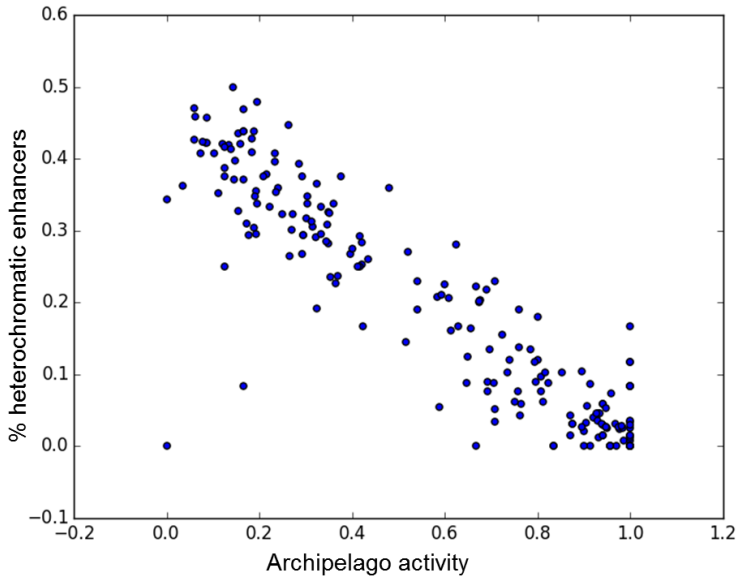
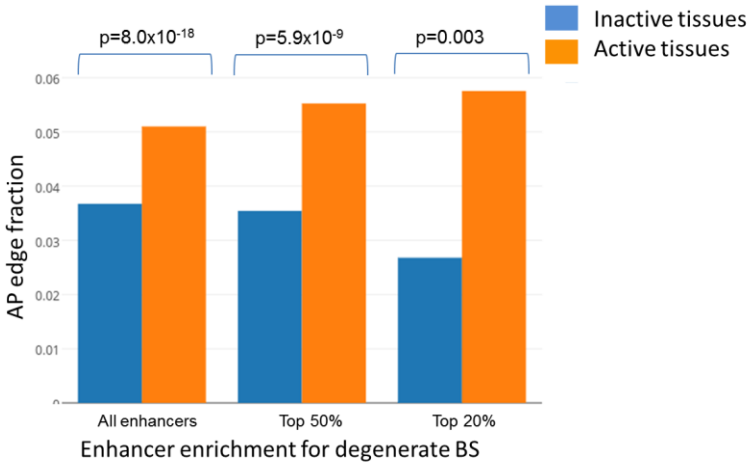
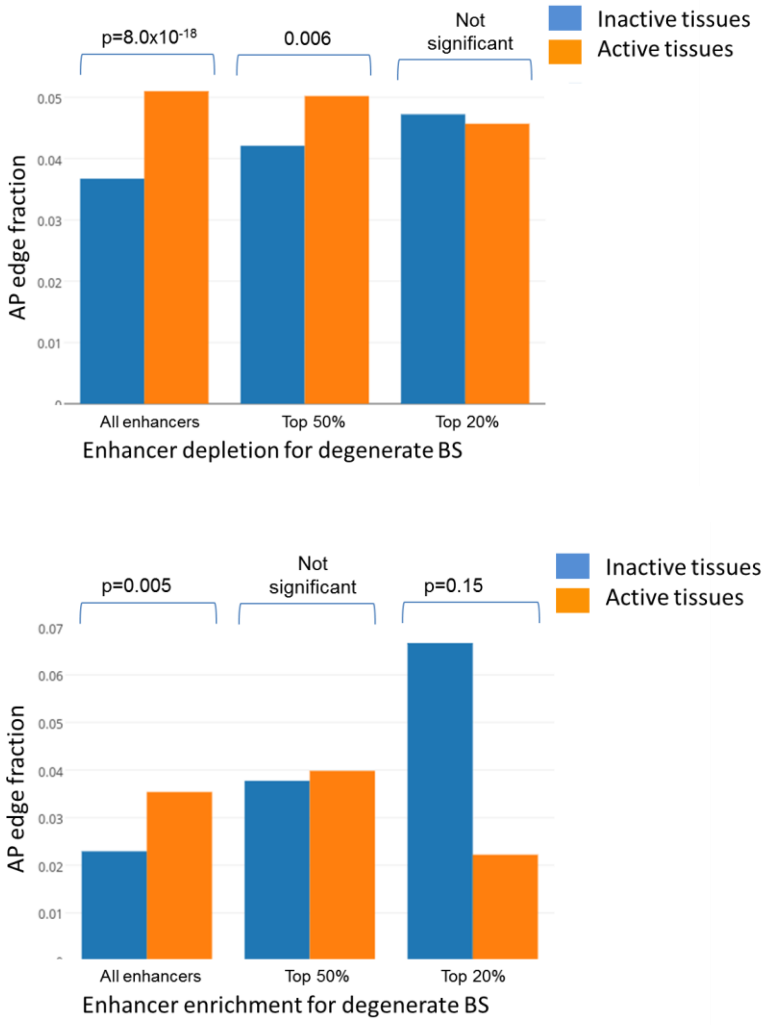


Figure 5-2. (Top) Fraction of interactions among AP enhancers in active vs. inactive cell types. Results shown for all enhancer pairs, as well after screening for enhancers enriched for degenerate BS at two enrichment levels. One-sided p-values based on paired Wilcoxon test across relevant enhancer pairs. (Middle) Same as Top but with enhancer depleted for degenerate motifs. (Bottom) Same as Top, except interactions are between enhancers residing in different APs, rather than enhancers within the same AP, and p-values are two-sided.



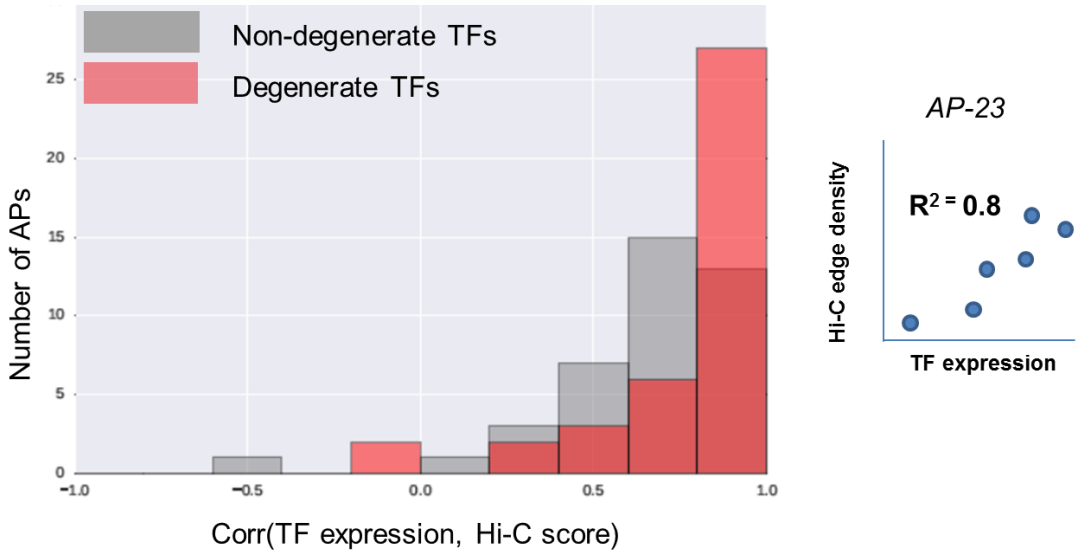


AP compactness scales more closely with expression of degenerate than specific TFs

If crowdsourcing does indeed help drive AP compaction, then we expect that TFs with degenerate BS contributed more to the compaction than non-degenerate TFs. To test this, we tracked AP compactness as a function of TF expression for ~400 TFs whose motif placed them in the top or bottom 20% of TFs for degeneracy. For each AP-cell type combo, mean TF expression was calculated separately for degenerate and specific TFs, while tailored to include only those TFs recognized by at least one enhancer in the AP.

Finally, for each of the two degeneracy classes and for each AP, we computed the correlation across all 6 cell types between cell type-specific mean TF fpkm and mean AP compactness (edge fraction). As seen in Figure 5-3, correlations for degenerate TFs trend higher than those for specific TFs, consistent with our hypothesis.

Figure 5-3. (Left) tallies of correlation values (x-axis) computed across 6 cell types between mean TF expression and AP compactness (estimated by its edge fraction) for each of 40 APs, where TFs were segregated into degenerate and non-degenerate. (*Right*) An illustrative AP. AP = archipelago

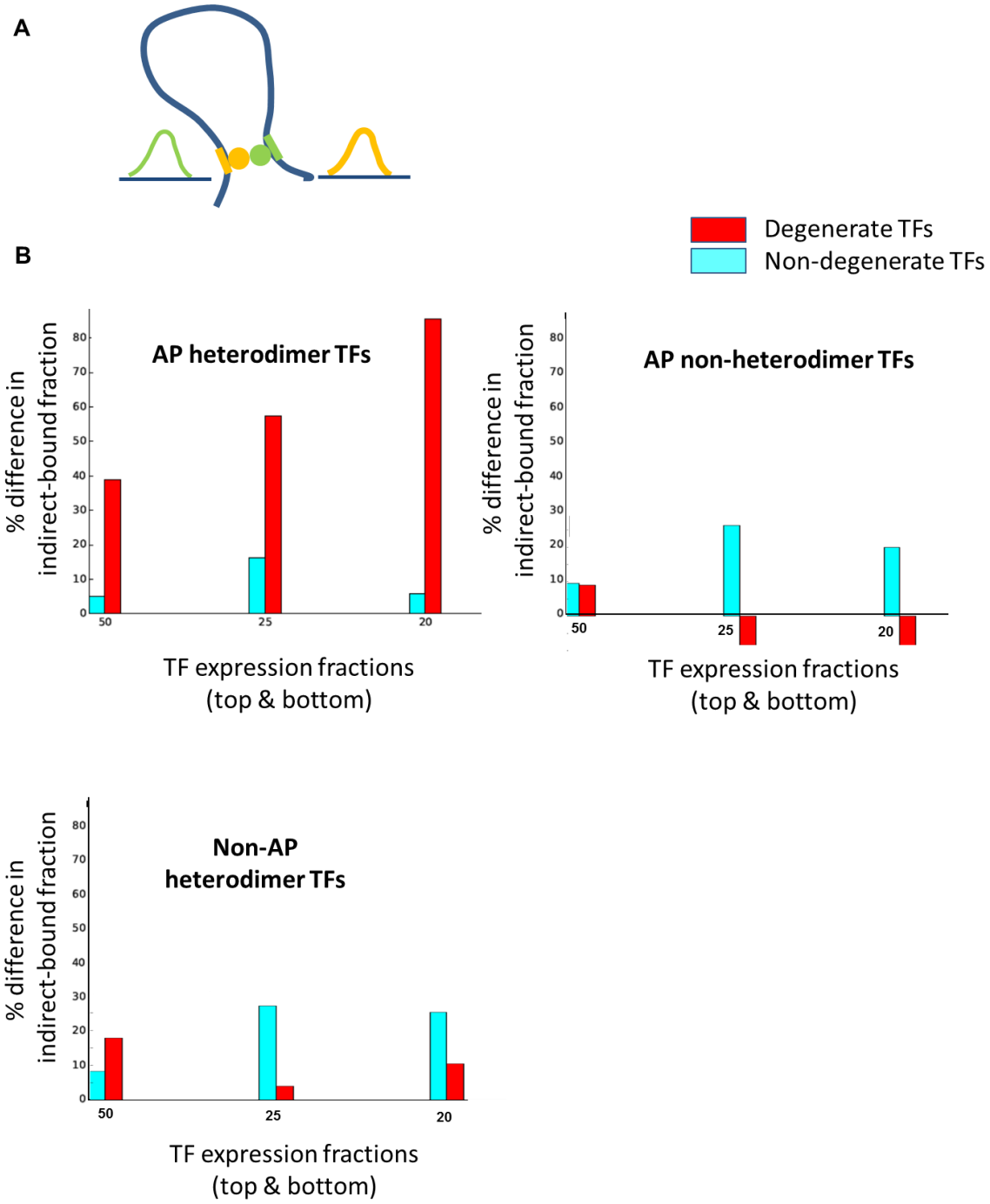


There is greater heterodimer-induced DNA-bridging than expected in active APs

We propose two mechanisms to account for the greater correlation between TF expression and AP compactness for degenerate TFs. Here we test the first: elevated AP TF concentrations boost the rate of protein-protein bridges formed between degenerate heterodimers bound at non-contiguous DNA loci. Such bridges form de facto chromatin loops, as described in the Strings and Binders model, described above. To estimate the

relative change in heterodimer bridges formed, we used fraction of binding that was indirect, that is, binding to another TF which is, itself, bound to the DNA. Indirect binding is frequently inferred through identification of ChIP-Seq peaks for a TF in the absence of any corresponding motif instances under the peak (Jothi et al., 2008). As above, we tracked the level of indirect binding against TF expression. We did this for heterodimers and non-heterodimers in both AP and non-AP enhancers. As can be seen in Figure 5-4, results support the hypothesized increase in heterodimer-induced TF bridging in APs, as indirect binding scales robustly with expression of degenerate motifs TFs, but not non-degenerate motifs.

Figure 5-4. TF chromatin bridging depends on TF motif degeneracy, TF expression, and AP context. (A) Cartoon of the experimental proxy used to detect heterodimer –anchored TF chromatin bridging. **(B)** Plots comparing fraction of indirect ChIP-Seq peaks in AP-cell type combinations with high TF expression to the corresponding fraction in AP-cell type combinations with low TF expression**.** The percent difference is plotted on the y-axis as a function of cutoffs for high/low TF expression (x-axis) for heterodimers in APs (top-left); non-dimers in APs (top-right); heterodimers in matched non-AP enhancers (bottom-left).



Pending work

To further address concerns about high Hi-C edge fractions for the least active AP-cell type combinations (activity from 0% to 50%), we will try to determine whether these 3D contacts can be explained by a common heterochromatic compartment(s). Specifically,

we will examine Hi-C interactions between APs on distinct chromosomes and (i) test whether these interactions are disproportionately frequent in the least active AP-cell type combinations; and (ii) cluster these cross-chromosomal interactions to show an unexpectedly high proportion colocalize in the least active AP-cell type combos. Both observations would best be explained by incidental contact in a shared, highly compacted compartment, typical of heterochromatin.

We will repeat the analysis in Figure 5-3 that assays the correlation between 3-D interactions and expression of TFs classified on the basis of degeneracy, however, at the level of enhancer instead of AP in order to increase statistical power.

We will use ENCODE histone modification data to mimic published results in the context of our APs and show that spatially proximate enhancer pairs (within an AP) have more similar chromatin/histone state than non-AP pairs matched with AP pairs for pairwise genomic distance and chromatin accessibility. To test whether this observation can be explained by crowdsourcing-boosted recruitment of cohesin and chromatin proteins, we will use available ChIP-Seq data to ascertain whether cohesin, relevant co-factors, and CMEs are disproportionately recruited in APs by the most degenerate TF recruiters, as predicted by crowdsourcing. We will also test whether CME-recruiting TFs are more degenerate than expected by chance, and the extent to which the degenerate recruiters have deeper evolutionary conservation than the non-degenerate recruiters. Finally, we will compare impacts on a given CME's presence when degenerate TF recruiters are knocked down compared to when a specific TF recruiter is knocked down, using published experimental data.

Complementing these genomic results, our collaborators Daphne Ezer and Xiaoyan Ma at the University of Cambridge will provide results from a biophysical simulation whose goal is to model the crowdsourcing-induced boost in local TF concentration as a function, effectively, of BS concentration.

Chapter 6: **Perspective and future work**

The contribution of chromatin structure to regulating cellular processes, such as gene expression, is an active area of research (Pombo and Dillon 2015; Zhang et al. 2013). There is also keen interest by the community in uncovering factors that, conversely, shape chromatin and can account for its structural variation across cell (Bickmore and van Steensel, 2013; Sexton and Cavalli, 2015). In this work, we describe and offer the first evidence of a general biophysical mechanism that provides insights on both fronts, whereby a high spatial concentration of genomically remote binding sites for a given transcription factor serves to remodel the protein's microenvironment, increasing its concentration. This, in turn, further compacts chromatin, elevating the spatial concentration of binding sites, and likely setting up a feedback loop. As a direct consequence, occupancy is boosted for TFs (typically degenerate) with abundant archipelago BS, and expression multiplies for genes near enhancers enriched in such BS. Crowdsourcing, then, mechanistically bridges effects at two starkly different scales – single BS versus chromatin structure spanning megabases – through 'mass action' of tens to hundreds of binding sites, with the resulting dialog between TF binding and chromatin structure contributing to gene complex activation.

TF occupancy, specificity, and superenhancers

To date, there has not been work synthesizing the flanking region perspective that dominates modeling of TF-DNA binding, on the one hand, and the spatial perspective now common in study of coordinate regulation, on the other. Our work suggests that occupancy is much more accurately modeled, and false positives mitigated, when spatial

context is accounted for – namely, the quantity of spatially proximal homotypic sites – particularly for TFs with degenerate motifs.

The impact of crowdsourcing on occupancy also holds surprising implications for binding specificity. There is a well-documented dearth of binding information encoded in the transcription factor motifs of higher eukaryotes, which tend to have, at once, larger genomes yet shorter motifs (Stewart and Plotkin, 2013). The challenge for a TF to discriminate between its bona fide sites and the many inevitable duplicate but non-functional sites is only exacerbated for degenerate motifs – which are recognized by up to millions of putative sites (Mirny et al., 2009). One solution to this conundrum, as employed by the cell, is to require added information in the form of cooperative binders, which must recognize their own binding site nearby. A 3-dimensional and more diffuse version of this approach, as suggested by our results, is to require a spatial plurality of similar sites in a given regulatory compartment. In its absence, binding is too weak to induce complex-wide activation. Hence, a degenerate TF discriminates on the basis of genomic sequence in addition to higher-order chromatin structure. The result is, effectively, mobile and context-specific area codes within the nucleus.

Degenerate motifs, and weak binding, more generally, have gained notice for their unexpectedly high contribution to cell- and condition-specific regulation of gene (Essien et al., 2009b; Segal et al., 2008; Tanay, 2006) Master regulator TFs, which are hierarchically situated at the beginnings of regulatory cascades, tend to be degenerate (Heinz et al., 2015). Master regulators bind in strikingly high occupancy in super-enhancers at levels, interestingly, that scale with superenhancer size (typically from 10-100Kb) (Whyte et al., 2013). It has also been recently learned that superenhancers feature

a dense thicket of chromatin contacts within their borders (Heinz et al., 2015). A mechanistic explanation connecting these observations, however, has not been offered (Andersson et al., 2015). Here we have shown that superenhancer function is likely informed by crowdsourcing, even as crowdsourcing also acts widely as a general mechanism among standard enhancers. Additionally, we observed that small superenhancers were several times more likely to co-inhabit the same archipelago than large enhancers (data not shown), which echoes our results showing enhances classified as 'weak' (chromHMM – Ernst and Kellis 2012) interacted with significantly larger networks of correlated enhancers than 'strong' enhancers. Future work could explore a potential role for crowdsourcing in coordinating the many and widely-dispersed superenhancers that collectively govern lineage

determination.

Archipelagos, transcription factories, and meta-enhancers

The dependency of coordinate regulation and co-expression of functionally related genes on the activity of regulatory archipelagos has now been demonstrated in model systems such as HOX, alpha- and beta-globin (Montavon et al. 2013; Fang et al. 2009). In olfactory neural receptors, archipelagos are critical for expression of even a single coding gene (Markenscoff-Papadimitriou et al. 2014). While enhancer-promoter proximity is well-established to be critical for transfer of information encoded in TFs (Krivega and Dean, 2012), there has been an absence of well-elaborated mechanisms explaining the adaptive role, if any, of observed enhancer-enhancer contacts (Li et al., 2012; Sandhu et al., 2012) of the type that lie at the heart of crowdsourcing. Moreover, our findings suggest a refinement of the widely-held view of coordinate gene regulation wherein

enhancers and genes are recruited by high concentrations of (master) TFs. Rather, the truth appears to be more circular, as enhancers also concentrate TFs. In the literature on transcription factories – nuclear sub-compartments that concentrate transcriptional resources and feature high transcriptional output – several have proposed similar ideas (Eskiw et al., 2010; Feuerborn and Cook, 2015b) , but without empirical support.

Transcription factories to date have been explored primarily with bench science and microscopy, with minimal if any sequence-related results. This raises a challenge in applying factory-gleaned results to archipelagos, and vice versa, two research tracks that have advanced with negligible crossing despite indications suggesting 'archipelago' and 'factory' are two descriptions of the same biological phenomenon. A first step toward integration might be to expand use of fluorescent labeling in factories to include enhancer elements identified in computational archipelago analyses, including 3C. Transcription factories are often subject to proximal promoter pausing of polymerase elongation (Buckley and Lis, 2014). By applying ChIP-Seq data with a polII antibody to the area near transcription start sites, high levels of stalled polymerase would ostensibly show up in archipelagos, helping cement archipelagos' 'hidden identity' as transcription factories.

Factories show evidence of being customized to particular transcriptional outputs, with the distribution of resources, such as TFs, similarly customized by factory (Babu et al., 2008; Bulger and Groudine, 2010). The source of this customization, however, has not been resolved (Sutherland and Bickmore, 2009). Crowdsourcing is a good candidate mechanism, as it ostensibly recruits TFs in proportions similar to motif instances in their member enhancers (and less numerous promoters). In this work, we treated APs, except for their size, as generic. But, in fact, we observed large variation among APs in their

relative TF-specific coverage levels not predicted by degeneracy (data not shown). It would be straightforward to test whether differences in binding site enrichment in a given AP are consistent with the functional enrichment of the AP's genes. Further predictions could be tested through fluorescent labeling and bench experimentation.

Interestingly, we also observed across archipelagos, generally, an unexpected dearth of motif instances for degenerate binding sites, relative to non-archipelago regions. To be sure, there were relatively more AP sites, overall, that recognized degenerate motifs, however there were fewer sites per given TF in each enhancer (ie, each genomic homotypic cluster). This could be explained by the limitations of evolutionary selection to functionally preserve a regulatory region from the vagaries of mutations if it is too large (Stewart and Plotkin, 2013). Evolution appears to have leveraged the (spatial) proximity of enhancers and their collective abundance of sites for a TF to reduce the quantity of its sites in any *individual* enhancer. Instead, we find, this expensive real estate accommodates sites for a wider variety of TFs – consistent with the more complex regulatory demands of archipelagos compared to non-AP transcription. As further evidence for this unique example of group level purifying selection, we observed, counter-intuitively, far greater sharing of binding motifs among AP enhancers separated by megabases than by AP enhancers separated by 20Kb or less; sharing of motifs climbs monotonically as inter-enhancer grows. This finding is consistent with the inevitably high spatial proximity of enhancers separated by relatively negligible genomic distance. Unique to archipelagos, genomically proximal enhancers thus have motif composition suggestive of their membership in larger, 'meta-enhancers'. These may, interestingly,

turn out to consist largely of super-enhancer regions, although on the other hand, super-enhancers have, in fact, been shown to be *enriched* for motifs of master regulator TFs.

Higher **higher order transcriptional regulation**

In the domain of spatial chromatin structure, scale may be defining. At the scale of 1-3 Mb (or 200Kb-500Kb, as per Rao et al), topologically associated domains (TAD) remain intact across cell types (Dixon et al., 2012). Rao found that TADs came together in a nuclear compartment tissue-specifically, where they shared histone marks, a strong indicator of co-regulation, only to co-localize with a different set of TADs in other cell types. Based on this description, these co-regulated domains appear to be related to the archipelagos we identified. (Shared histone state among enhancers is a predicted consequence of crowdsourcing and compartment-wide recruitment by bound degenerate TFs of chromatin modifying enzymes). Interestingly, this view of regulation suggests that enhancers are repurposed under different cellular contexts – a view that dovetails with the prevailing view of evolution as endlessly resourceful, evidenced by the numerous genomic structures coopted over time for new or added functions. Indeed, in (Sheffield et al., 2013) where archipelago enhancers clusters were identified without requiring they be disjoint, many enhancers appear in multiple such clusters.

Hence, enhancers appear to be frequently subject to reuse, rather than constrained to a single archipelago/transcription factory and function. This could account for enhancers' strikingly high abundance of putative binding sites, typically numbering in the hundreds. But this raises the question of how enhancers ensure a binding regime specific to a given factory. In principle, crowdsourcing can account for this. Depending on fellow factory members, and consistent with the factory's function, only required TFs would be

raised to functional concentrations, while remaining TFs would not. Cognate sites for such TFs would hence remain effectively dormant and unbound. Confirming this model requires showing specialized occupancy patterns as a function of cell-type specific factory activity. Importantly, in addition to testing enhancer clusters identified based on shared (correlated) activity across cell types as done in Sheffield and Malin – representing, in essence, constitutive regulatory archipelagos – clusters identified based on *single*-tissue activity should be identified and tested. If the model is verified, it would highlight crowdsourcing's role in organizing this highest level of transcriptional coordination, while providing direct evidence for the Rao and Dixon models of modular TAD function.

Appendices

Appendix 1: Author contributions

Chapter 2:

Conceived: SH

Designed analysis: SH, JM

Performed analysis: JM with help from Radhouane Aniba

Wrote manuscript from which chapter taken: SH, JM

Chapter 3, 4:

Crowdsourcing mechanism and functional implications: JM

Designed genomic analysis JM with help from SH, Steve Mount

Biophysical Modeling with Simulations: Daphne Ezer, Xiaoyan Ma

Performed genomic analysis: JM with help from Hiren Karathia

Wrote manuscript from which chapter taken: JM, SH, DE

Help with illustrations: Seung Gu Park

Chapter 5:

Conceived: JM

Designed computational analysis JM with help from Hiren Karathia, SH, Kan Cao

Performed genomic analysis: JM, HK

Hi-C data processing: Hiren Karathia

Appendix 2: Tables for correlated enhancer analysis

Appendix Table 1. 73 cell types sorted into 37 clusters.

One cell type from each cluster (first in row) was used as the representative for the cluster. See text (Chapter 2) for how the representative was selected.

Cluster	Representative Cell Type	Cluster Members						
1	A549							
2	Aoaf	M059j						
3	Be2c							
4	Cd20ro01778							
5	Gm04503							
6	Gm04504							
7	Hah							
8	Hasp	Nt2d1						
9	Hbmec	Hff						
10	Hipe							
11	Hmf							
12	Hmvecdad							
13	Hmvecdblneo							
14	Hmvecdlyneo							
15	Hmveclly							
16	Hpaf	Hsmmt						
17	Hrgec	Th1wb54553204						
18	Hs5							
19	Hsmm							
20	Huvec	Hbvp	Hct116	Hmec	Hmvecdblad	Hmvecdneo	Hpdlf	Nhek

21	Jurkat							
22	Mcf7	Lhcnm2						
23	Monocd14ro1746							
24	Msc							
25	Nha							
26	Nhbera							
27	Nhdfad	Hmveclbl	Hpaec					
28	Prec							
29	Gm12864	Hac	Hcfaa	Hconf	Rptec	Th17		
30	Sknmc							
31	Cd34mobilized	T47d						
32	Th1wb33676984							
33	Th2							
34	Th2wb33676984							
35	Cd4naivewb78495824	Th2wb54553204						
36	Cd4naivewb11970640	H7es	Hbvsmc	Hffmyc	Hmvecdlyad	Hpf	Hs27a	Hvm
37	Werirb1							

Appendix Table 2. 153 significantly co-occurring motifs sorted into 51 disjoint clusters based on motif similarity.

Cluster	Motifs
1	M00762
2	M00497
3	M00431 M00428 M00940 M00427 M00430 M00919 M00425 M00736 M00920 M00739 M00426 M00738
4	M01240
5	M01199 M01253 M01593
6	M00646
7	M01721 M01598
8	M01298
9	M00925 M01267 M00199 M00174 M00926 M00821 M00188 M00173
10	M00801
11	M01201
12	M01747 M01798
13	M01756 M00789 M00347
14	M00644 M00175 M00277 M01288 M00176 M00804 M00927 M01716 M01287
15	M01072
16	M01147 M01016
17	M01292 M00471 M00980 M00216
18	M00100
19	M01275
20	M00145

21	M01759 M01658 M00722
22	M01117
23	M00775
24	M00075
25	M00332
26	M01177
27	M00641 M01023
28	M00648 M00032 M01258 M01197 M00743 M00771
29	M01020
30	M01653
31	M01118 M00649 M01783 M00008 M01219 M01100 M01175 M00695 M00255 M00716 M00196 M00803 M00807 M00931 M00933 M00932 M01303 M00720 M01273 M01837 M01104 M01816 M01597 M00982 M01714 M00706 M00491 M01231 M00333 M01835 M01587 M01122
32	M01733 M00083
33	M01028
34	M01220
35	M00466
36	M00615 M00322 M00976 M00799 M00055 M00217 M01249 M01116 M00726
37	M01482 M00468
38	M00967
39	M00470 M00469 M00915 M00189 M00800 M01045 M01047
40	M01742 M00652
41	M01243
42	M00986

43	M01113 M01588 M00378 M01657 M01042 M00749
44	M01318
45	M01261 M01599 M00724 M01765
46	M00492
47	M01162 M01654
48	M01294
49	M00076
50	M01230 M00489
51	M01169

Appendix Table 3. GO enrichment of enhancer cluster target genes

Gene Ontology (GO) annotation terms for the clusters of target genes corresponding to correlated enhancer clustering with the highest ratio of enrichment terms between itself and a background gene cluster. In this list are GO terms separated by targeted gene cluster with adjusted p-values < 0.0005 and that are supported by three or more genes in the cluster. 7 of 52 clusters were enriched for at least one term that met this highly stringent standard. There were 149 separate instances of enrichment. This enhancer cluster was identified using the following parameters: min mean mutual information = 0.2, minimum cluster size = 20, minimum percent occupancy for most enriched motif = 0.0. Background clusters are matched for chromosome, the number of enhancers and signature of inter-enhancer distances, but consist of otherwise random enhancers. GO enrichment analysis performed with R's GOstats package. Adjusted p-value = 0.05*p-value/ q-value.

---

cluster size: 65 genes

Enriched term	#genes	Adjusted p-value	Description
GO:0009790	4	3.3e-04	embryo development
GO:0007411	3	3.3e-04	axon guidance
GO:0051179	10	3.3e-04	localization
GO:0009605	5	3.3e-04	response to external stimulus
GO:0051093	3	3.5e-04	negative regulation of developmental process
GO:0048519	8	3.5e-04	negative regulation of biological process
GO:0045597	3	3.5e-04	positive regulation of cell differentiation
GO:0016337	3	3.5e-04	cell-cell adhesion
GO:0001775	4	3.5e-04	cell activation
GO:0060284	3	3.6e-04	regulation of cell development
GO:0051960	3	4.0e-04	regulation of nervous system development
GO:0048523	8	4.2e-04	negative regulation of cellular process
GO:0065008	8	4.2e-04	regulation of biological quality
GO:0072358	4	4.2e-04	cardiovascular system development
GO:0072359	4	4.2e-04	circulatory system development
GO:0045596	3	4.2e-04	negative regulation of cell differentiation
	3	4.3e-04	negative regulation of cellular component organization
GO:0048568	3	4.3e-04	embryonic organ development
GO:0071845	3	4.3e-04	cellular component disassembly at cellular level
GO:0051239	6	4.3e-04	regulation of multicellular organismal process
GO:0022411	3	4.3e-04	cellular component disassembly
GO:0050767	3	4.3e-04	regulation of neurogenesis
GO:0007155	5	4.3e-04	cell adhesion
GO:0022610	5	4.3e-04	biological adhesion
GO:0007507	3	4.3e-04	heart development
GO:0050793	5	4.3e-04	regulation of developmental process
GO:0030182	5	4.5e-04	neuron differentiation

GO:2000026	5	5.0e-04	regulation of multicellular organismal development

cluster size: 141 genes

GO:0048812	4	2.3e-04	neuron projection morphogenesis
GO:0048667	4	2.3e-04	cell morphogenesis involved in neuron differentiation
GO:0001525	3	2.3e-04	angiogenesis
GO:0051172	5	2.3e-04	negative regulation of nitrogen compound metabolic process
GO:0048585	4	2.3e-04	negative regulation of response to stimulus
GO:0048568	3	2.3e-04	embryonic organ development
GO:0007409	4	2.3e-04	axonogenesis
GO:0001558	3	2.3e-04	regulation of cell growth
GO:0051090	3	2.3e-04	regulation of transcription factor activity
GO:0048468	6	2.3e-04	cell development
GO:0002009	3	2.3e-04	morphogenesis of an epithelium
GO:0050767	3	2.3e-04	regulation of neurogenesis
GO:0090046	3	2.3e-04	regulation of transcription regulator activity
GO:0010629	5	2.3e-04	negative regulation of gene expression
GO:0007507	3	2.3e-04	heart development
GO:0007399	7	2.3e-04	nervous system development
GO:0045934	5	2.3e-04	negative regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
GO:0016481	5	2.4e-04	negative regulation of transcription
GO:0032501	16	2.5e-04	multicellular organismal process
GO:0001503	3	2.7e-04	ossification
GO:0035239	3	2.7e-04	tube morphogenesis
GO:0006357	6	2.8e-04	regulation of transcription from RNA polymerase II promoter
GO:0042127	6	2.9e-04	regulation of cell proliferation
GO:0032582	3	3.0e-04	negative regulation of gene-specific transcription
GO:0008283	7	3.0e-04	cell proliferation
GO:0050673	3	3.3e-04	epithelial cell proliferation
GO:0009887	5	3.3e-04	organ morphogenesis
GO:0042692	3	3.3e-04	muscle cell differentiation
GO:0007411	4	3.7e-04	axon guidance
GO:0009890	6	3.7e-04	negative regulation of biosynthetic process
GO:0002697	3	3.8e-04	regulation of immune effector process
GO:0048869	10	3.9e-04	cellular developmental process
GO:0031327	6	3.9e-04	negative regulation of cellular biosynthetic process
GO:0010553	3	3.9e-04	negative regulation of gene-specific transcription from RNA polymerase II promoter
GO:0009892	7	4.1e-04	negative regulation of metabolic process
GO:0031324	7	4.2e-04	negative regulation of cellular metabolic process
GO:0035295	4	4.2e-04	tube development
GO:0010605	7	4.2e-04	negative regulation of macromolecule metabolic process
GO:0045892	5	4.2e-04	negative regulation of transcription, DNA-dependent
GO:0051253	5	4.2e-04	negative regulation of RNA metabolic process
GO:0009605	7	4.2e-04	response to external stimulus

GO:2000113	6	4.2e-04	negative regulation of cellular macromolecule biosynthetic process
GO:0022603	4	4.2e-04	regulation of anatomical structure morphogenesis
GO:0060284	4	4.2e-04	regulation of cell development
GO:0001763	3	4.2e-04	morphogenesis of a branching structure
GO:0019216	3	4.2e-04	regulation of lipid metabolic process
GO:0030154	10	4.2e-04	cell differentiation
GO:0050678	3	4.2e-04	regulation of epithelial cell proliferation
GO:0031347	4	4.2e-04	regulation of defense response
GO:0006935	5	4.2e-04	chemotaxis
GO:0042330	5	4.2e-04	taxis
GO:0072358	5	4.2e-04	cardiovascular system development
GO:0072359	5	4.2e-04	circulatory system development
GO:0061061	4	4.2e-04	muscle structure development
GO:0010558	6	4.2e-04	negative regulation of macromolecule biosynthetic process
GO:0061138	3	4.2e-04	morphogenesis of a branching epithelium
GO:0050727	3	4.9e-04	regulation of inflammatory response
GO:0051146	3	4.9e-04	striated muscle cell differentiation
GO:0048754	3	5.0e-04	branching morphogenesis of a tube

---

cluster size: 33 genes

GO:0006936	3	0	muscle contraction
GO:0051259	3	0	protein oligomerization
GO:0003012	3	0	muscle system process
GO:0003013	3	0	circulatory system process
GO:0008015	3	0	blood circulation
GO:0061061	3	0	muscle structure development
GO:0035556	6	0	intracellular signal transduction
GO:0022607	5	0	cellular component assembly
GO:0010627	3	0	regulation of intracellular protein kinase cascade
GO:0050794	13	0	regulation of cellular process
GO:0044085	5	0	cellular component biogenesis

---

cluster size: 6 genes

GO:0007268	4	0	synaptic transmission
GO:0019226	4	0	transmission of nerve impulse
GO:0035637	4	0	multicellular organismal signaling

---

cluster size: 27 genes

GO:0001775	5	0	cell activation
GO:0001568	4	0	blood vessel development
GO:0001944	4	0	vasculature development
GO:0051716	11	0	cellular response to stimulus
GO:0007265	3	0	Ras protein signal transduction

GO:0007166	7	0	cell surface receptor linked signaling pathway
GO:0072358	4	0	cardiovascular system development
GO:0072359	4	0	circulatory system development
GO:0007165	9	0	signal transduction
GO:0006928	4	0	cellular component movement
GO:0007167	4	0	enzyme linked receptor protein signaling pathway
GO:0023052	9	0	signaling

---

cluster size: 53 genes

GO:0045785	3	0	positive regulation of cell adhesion
GO:0007167	6	0	enzyme linked receptor protein signaling pathway
GO:0071844	6	0	cellular component assembly at cellular level
GO:0040007	5	0	growth
GO:0030155	3	0	regulation of cell adhesion
GO:0032268	6	0	regulation of cellular protein metabolic process
GO:0051246	6	0	regulation of protein metabolic process
GO:0048589	3	0	developmental growth
GO:0031399	5	0	regulation of protein modification process
GO:0034622	4	0	cellular macromolecular complex assembly
GO:0071845	3	0	cellular component disassembly at cellular level
GO:0022411	3	0	cellular component disassembly
GO:0009967	4	0	positive regulation of signal transduction
GO:0048584	5	0	positive regulation of response to stimulus
GO:0043623	3	0	cellular protein complex assembly
GO:0022607	6	0	cellular component assembly
GO:0010647	4	0	positive regulation of cell communication
GO:0023056	4	0	positive regulation of signaling
GO:0031401	3	0	positive regulation of protein modification process
GO:0007169	4	0	transmembrane receptor protein tyrosine kinase signaling pathway
GO:0044085	6	0	cellular component biogenesis
GO:0042060	4	0	wound healing
GO:0001932	4	0	regulation of protein phosphorylation

---

clusterID 38 cluster size: 53 genes

GO:0048583	4	3.5e-04 regulation of response to stimulus
GO:0007267	3	3.5e-04 cell-cell signaling
GO:0022008	3	3.6e-04 neurogenesis
GO:0050793	3	3.6e-04 regulation of developmental process
GO:0048731	5	3.7e-04 system development
GO:0048699	3	3.8e-04 generation of neurons
GO:0030182	3	4.0e-04 neuron differentiation
GO:0048518	5	4.0e-04 positive regulation of biological process
GO:0051128	3	4.0e-04 regulation of cellular component organization

GO:2000026	3	4.1e-04 regulation of multicellular organismal development
GO:0030030	3	4.2e-04 cell projection organization
GO:0048522	5	4.4e-04 positive regulation of cellular process
GO:0045595	3	4.8e-04 regulation of cell differentiation
GO:0048666	3	4.9e-04 neuron development

Appendix Table 4. Mapping of tissues between CTen and ENCODE databases.

We clustered the 84 tissue types in the CTen database and the 72 types in the ENCODE DHS database into 34 and 23 cytologically motivated classes, respectively. Agreement in tissue enrichment was assessed based on the 17 classes, shown below, that are shared between CTen and ENCODE.

Encode cell type (enhancer domain)	**Tissue class**	***Cten* cell type (gene domain)**
Cd20ro01778	blood	721 b lymphoblasts
Cd34mobilized		bdca4+ dentritic cells
Cd4naivewb11970640		cd19+ b cells
Cd4naivewb78495824		cd33+ myeloid
Gm12864		cd34+
Jurkat		cd4+ t cells
Th1		cd56+ nk cells
Th17		cd71+ early erythroid
Th1wb33676984		cd8+ t cells
Th1wb54553204		leukemia chronic myelogenous k-562
Th2		leukemia lymphoblastic (molt-4)
Th2wb33676984		leukemia promyelocytic hl-60
Th2wb54553204		lymph node
Tregwb78495824		lymphoma burkitts (daudi)
Tregwb83319432		lymphoma burkitts (raji)
		whole blood

Nhbera	bronchial epithelium	bronchial epithelial cells
Hs27a	bone marrow	bone marrow
Hs5		
Be2c	brain	amygdala
Hah		caudate nucleus
M059j		cingulate cortex
Nha		globus pallidus
Sknmc		hypothalamus
		medulla oblongata
		occipital lobe
		olfactory bulb
		parietal lobe
		pineal day
		pineal night
		pituitary
		pons
		prefrontal cortex
		subthalamic nucleus
		temporal lobe
		thalamus
		whole brain
Hac	cerebellum	cerebellum
		cerebellum peduncles
Hct116	colon	colorectal adenocarcinoma
		colon
Aoaf	endothelium	cd105+ endothelium
Hbmec		
Hbvp		

Hbvsmc		
Hmvecdad		
Hmvecdblad		
Hmvecdblneo		
Hmvecdlyad		
Hmvecdlyneo		
Hmvecdneo		
Hmveclbl		
Hmveclly		
Hpaec		
Hpaf		
Huvec		
Hconf	eye	ciliary ganglion
Werirb1		retina
Hcfaa	heart	atrioventricular node
		cardiac myocytes
		heart
Hrgec	kidney	kidney
Hpf	lung	fetal lung
Nhbera		lung
Wi38		
Monocd14ro1746	monocytes	cd14+ monocytes
Hsmm	muscle	skeletal muscle
Hsmmt		smooth muscle
Lhcnm2		
Lncap	prostate	prostate
Prec		
Gm04503	skin	skin

Gm04504		
Nhdfad		
Nhek		
Rpmi7951		
Hasp	spine	dorsal root ganglion
		spine
		superior cervical ganglion
		trigeminal ganglion
Nt2d1	testis	testis
		testis germ cell
		testis intersitial
		testis leydig cell
		testis seminiferous tubule

Appendix Table 5. Genes targeted by the illustrative enhancer cluster

(see legend in Chapter 2 Figure 2-6 for more information).

- transport
- signal transduction
- nucleocytoplasmic transport
- embryo development
- cell death
- cell differentiation
- cell signaling
- anatomical structure formation involved In morphogenesis
- cell proliferation
- transmembrane support

Gene Symbol	Gene Description	GO Slim Terms								
STK38L	protein_kinase_activity		*							
SSPN	cell_junction									
DIP2B	transcription_factor_binding									
PPM1H	catalytic_activity									
KITLG	signal_transduction		*		*	*	*			*	
KCNA5	transmembrane_transport	*						*			*
PLEKHA5	phospholipid_binding									
WNK1	protein_kinase_activity	*	*				*				
ADAMTS20	proteolysis		*			*	*				
SRGAP1	signal_transduction		*				*		*		
CCDC91	protein_transport	*									
IFNG	cytokine-mediated_signaling_pathway	*	*	*		*	*	*	*	*	
BTBD11	DNA_binding										
TMTC2	endoplasmic_reticulum										
E2F7	regulation_of_transcription_DNA-dependent		*		*		*		*	*	
CDK17	protein_kinase_activity										
PPTC7	metal_ion_binding										
ETNK1	ATP_binding										
VEZT	cell_junction				*						
PRICKLE1	transcription_factor_binding	*	*	*	*		*		*		
CALCOCO1	signal_transduction		*								
LIMA1	cell_junction										
IFT81	cell_differentiation						*				
SYT1	cell_junction	*						*			
PTPRQ	receptor_activity				*		*		*		
CACNA1C	transmembrane_transport	*					*	*	*		
ERC1	Golgi_membrane	*	*								
KRR1	RNA_binding										
TMEM117	integral_to_membrane										
AEBP2	regulation_of_transcription_DNA-dependent										
DRAM1	apoptotic_process					*					
NUDT4	intracellular_signal_transduction	*	*	*							
EPS8	signal_transduction		*							*	
IFLTD1	cell_proliferation									*	
ANO6	ion_transport	*									*
DDX47	ATP_binding					*					
SLC6A15	transmembrane_transport	*									*
HPD	Golgi_membrane										
PTHLH	Golgi_apparatus		*				*	*	*	*	
IGF1	signal_transduction	*	*	*		*	*		*	*	
STAB2	receptor_activity	*							*		
EEA1	membrane_fraction	*						*			
C1R	proteolysis										

TMEM119	integral_to_membrane										
TSPAN11	membrane										
PPFIA2	receptor_activity										
NCOR2	negative_regulation_of_transcription_		*								
ATP2B1	transmembrane_transport	*									*
MLXIP	regulation_of_transcription_DNA-dependent	*		*							
GLIPR1L2	integral_to_membrane										
EPYC	extracellular_region										
PPP1R12A	signal_transducer_activity	*	*	*							
AMIGO2	cell_adhesion					*					
FAR2	endoplasmic_reticulum_membrane										
BICD1	transport	*	*								
NUAK1	protein_kinase_activity									*	
SLC38A2	transmembrane_transport	*						*			*
CRADD	signal_transduction		*			*					
EP400	nucleotide_binding										
DYRK2	protein_kinase_activity	*	*	*		*					
DCN	extracellular_space										
ZNF664	regulation_of_transcription_DNA-dependent										
SLC41A2	transmembrane_transport	*									*
HMGA2	negative_regulation_of_transcription_	*			*	*	*	*	*	*	
PDE3A	signal_transduction		*			*	*				
CHST11	transferase_activity		*		*	*	*		*	*	
PLEKHG6	phospholipid_binding		*								
TMTC3	integral_to_membrane										
ANO4	ion_transport										
NAV3	ATP_binding										
SLC38A4	transmembrane_transport	*									*
ANKS1B	cell_junction										
C12orf70	integral_to_membrane										
PLCZ1	intracellular_signal_transduction	*	*								
HCAR1	G-protein_coupled_receptor_activity										
CKAP4	perinuclear_region_of_cytoplasm										
USP15	proteolysis		*								
ITPR2	transmembrane_transport	*	*					*			
TBX3	negative_regulation_of_transcription_	*			*	*	*	*	*	*	
PTPRR	receptor_activity				*		*				
WNT5B	receptor_binding		*		*		*		*		
TSPAN8	signal_transducer_activity		*								
ST8SIA1	Golgi_membrane									*	
RASSF9	signal_transduction	*	*								
TSFM	intracellular										
TEAD4	regulation_of_transcription_from_RNA_pol_II_pro		*		*		*		*		
TMEM132B	integral_to_membrane										
PHLDA1	regulation_of_transcription_from_RNA_polII_prom					*					

Appendix 3: Enhancer coordinates for 40 archipelagos

AP ID	chrm	start	stop
24	chr1	95164417	95164990
24	chr1	64292181	64292559
24	chr1	68444790	68445268
24	chr1	60169236	60169669
24	chr1	115732459	115733152
24	chr1	60731878	60732134
24	chr1	115973287	115973736
24	chr1	60598553	60599033
24	chr1	78622648	78623135
24	chr1	120490104	120490512
24	chr1	95500965	95501522
24	chr1	112005979	112006315
24	chr1	64636945	64637982
24	chr1	115872698	115873416
24	chr1	51536339	51536786
24	chr1	56038885	56039096
24	chr1	64504984	64505702
24	chr1	77836420	77837040
24	chr1	56184343	56185108
24	chr1	56097743	56098133
24	chr1	59840890	59841404
24	chr1	55909167	55909671
24	chr1	112106280	112106659
24	chr1	98623601	98623908
24	chr1	78005294	78005892
25	chr1	94263880	94264343
25	chr1	94134742	94135281
25	chr1	109739985	109740365
25	chr1	94269939	94270369
25	chr1	94791950	94792244
25	chr1	94510972	94511623
25	chr1	94087786	94088194
25	chr1	15683492	15683671
25	chr1	94736296	94736703
25	chr1	58861009	58861375
25	chr1	68355587	68355924
25	chr1	68306107	68306399
25	chr1	87691978	87692472
25	chr1	78957353	78957707
25	chr1	84831150	84831766
25	chr1	67090077	67090486
25	chr1	117635800	117636209
25	chr1	77980799	77981121
25	chr1	85794343	85795136
25	chr1	85796497	85797013
25	chr1	85779637	85780145
25	chr1	94725107	94725515
25	chr1	94724323	94724845
25	chr1	59083976	59084404
25	chr1	59639399	59639827
25	chr1	85809925	85810291
25	chr1	77977045	77977566
25	chr1	8262119	8262695
25	chr1	67029766	67030435
25	chr1	25051303	25052018
25	chr1	8121237	8121658

25	chr1	22091187	22091852
25	chr1	16471562	16471984
25	chr1	16472673	16473256
25	chr1	120188458	120188846
25	chr1	36602196	36602457
25	chr1	39576554	39576986
25	chr1	94791092	94791836
25	chr1	100056376	100056822
25	chr1	94790696	94791081
25	chr1	96828051	96828577
25	chr1	95552663	95553087
25	chr1	85832159	85832465
25	chr1	64240629	64241296
25	chr1	67029084	67029542
25	chr1	59228798	59229247
25	chr1	59058194	59058572
25	chr1	112275837	112276365
25	chr1	67410299	67410672
25	chr1	52364034	52364332
25	chr1	39857734	39858075
25	chr1	85755184	85755475
25	chr1	68639710	68640111
25	chr1	59229797	59230355
25	chr1	77939451	77939710
25	chr1	16508297	16508896
25	chr1	68190504	68191046
25	chr1	55776276	55776744
25	chr1	55776756	55777308
25	chr1	64508093	64508795
25	chr1	86072787	86073459
25	chr1	77787887	77788256
25	chr1	95329248	95329903
25	chr1	8197745	8198274
25	chr1	115721930	115722556
25	chr1	39513392	39513831
25	chr1	39644671	39644936
25	chr1	86044065	86044579
25	chr1	64196797	64197513
26	chr10	24498248	24498721
26	chr10	4812873	4814515
26	chr10	3269102	3269837
26	chr10	34722168	34722650
26	chr10	64342468	64343190
26	chr10	93100545	93101157
26	chr10	4414792	4415283
26	chr10	63222803	63223399
26	chr10	63223409	63223776
26	chr10	17104013	17104336
26	chr10	17007974	17008508
26	chr10	34612169	34612443
26	chr10	117696736	117697026
26	chr10	62587627	62588496
26	chr10	23113868	23114211
26	chr10	14003025	14003546
26	chr10	13923903	13924675
26	chr10	65426709	65427074
26	chr10	92690690	92691602
26	chr10	44352738	44353448
26	chr10	14032923	14033335
26	chr10	123551892	123552204
26	chr10	17029319	17029655
26	chr10	116629249	116629562

26	chr10	13966391	13966755
26	chr10	63854889	63855896
26	chr10	63853361	63854104
26	chr10	13726283	13727166
26	chr10	63991160	63991461
26	chr10	31109193	31109561
26	chr10	4285810	4286241
27	chr10	4746234	4746828
27	chr10	98783923	98784227
27	chr10	121439596	121439934
27	chr10	21581629	21582206
27	chr10	21623903	21624386
27	chr10	25155478	25155739
27	chr10	73631049	73631487
27	chr10	116398514	116399072
27	chr10	116031245	116031933
27	chr10	93347751	93348406
27	chr10	97067477	97067675
27	chr10	75647443	75647808
27	chr10	97068397	97068947
27	chr10	59783398	59783804
27	chr10	116383467	116384052
27	chr10	123886826	123887178
27	chr10	14116361	14116840
27	chr10	30073196	30073703
27	chr10	123900607	123900999
27	chr10	73526408	73526879
27	chr10	29824077	29824717
27	chr10	74075275	74075698
27	chr10	33552379	33553678
27	chr10	34413618	34413984
27	chr10	80917233	80917781
27	chr10	73015123	73015880
27	chr10	95228178	95228675
27	chr10	33595684	33596055
27	chr10	5986399	5986699
27	chr10	34815575	34816449
27	chr10	123942594	123943026
27	chr10	21625800	21626171
27	chr10	21655033	21655223
27	chr10	124060437	124060876
27	chr10	78801584	78801958
27	chr10	3290968	3291459
27	chr10	124067375	124067819
27	chr10	3581246	3581651
27	chr10	80720185	80720528
27	chr10	33274531	33274986
27	chr10	84741766	84742210
27	chr10	76951864	76952130
27	chr10	12887168	12887547
27	chr10	6763386	6763848
27	chr10	6764097	6764479
27	chr10	97033145	97033719
27	chr10	124264142	124264479
27	chr10	93364940	93365398
27	chr10	29273323	29273656
27	chr10	45297365	45297672
27	chr10	103699043	103699341
27	chr10	65498159	65498614
27	chr10	104364195	104364505
27	chr10	95226051	95226409
27	chr10	95225536	95225917

27	chr10	62240686	62241203
27	chr10	102650334	102650814
27	chr10	3928641	3929481
27	chr10	33626346	33626723
27	chr10	80872843	80873226
27	chr10	95218492	95218942
27	chr10	3966518	3966864
27	chr10	103127361	103127700
20	chr8	40929580	40929758
20	chr8	126340768	126341065
20	chr8	98194112	98194386
20	chr8	38770183	38770543
20	chr8	106998269	106998592
20	chr8	40381741	40382508
20	chr8	130492481	130492818
20	chr8	98446369	98446792
20	chr8	98102693	98103262
20	chr8	58663405	58663757
20	chr8	58503247	58503675
20	chr8	96817838	96818268
20	chr8	122101585	122102096
20	chr8	123330843	123331309
20	chr8	117587096	117587507
20	chr8	131245061	131245387
20	chr8	123199833	123200264
20	chr8	126082325	126082992
20	chr8	98102189	98102667
20	chr8	102300197	102300515
20	chr8	118632011	118632240
20	chr8	118631640	118632006
20	chr8	41053940	41054490
20	chr8	82106037	82106662
20	chr8	51052146	51052586
20	chr8	41092900	41093745
20	chr8	49236833	49237057
20	chr8	96820364	96820911
20	chr8	51096075	51096421
20	chr8	75690270	75690631
20	chr8	50968879	50969932
20	chr8	41228936	41229246
20	chr8	95232609	95233011
20	chr8	90962952	90963307
20	chr8	98995989	98996350
20	chr8	76661626	76662229
20	chr8	129912815	129913283
21	chr8	40032030	40032971
21	chr8	143757421	143757799
21	chr8	141489755	141490062
21	chr8	61911690	61912123
21	chr8	22131646	22132000
21	chr8	118922271	118922581
21	chr8	39916949	39917474
21	chr8	49320307	49320695
21	chr8	27474519	27475113
21	chr8	141655904	141656283
21	chr8	49541191	49541440
21	chr8	8870858	8871599
21	chr8	119023688	119023960
21	chr8	26122968	26123156
21	chr8	49321857	49322146
21	chr8	8395193	8395619
21	chr8	49320986	49321351

21	chr8	8167941	8168532
21	chr8	23268946	23269278
21	chr8	8153392	8154059
21	chr8	141001417	141001762
21	chr8	128961833	128962243
21	chr8	129188725	129189659
22	chr12	67040776	67041367
22	chr12	88865439	88865738
22	chr12	68110798	68111166
22	chr12	104571218	104572214
22	chr12	80427224	80427581
22	chr12	78019180	78019476
22	chr12	47392783	47393174
22	chr12	67928487	67928985
22	chr12	47353232	47353683
22	chr12	47315105	47315834
22	chr12	71950389	71950990
22	chr12	91417114	91417466
22	chr12	116966861	116967211
22	chr12	91492679	91493114
22	chr12	65858732	65859103
22	chr12	77257716	77258254
22	chr12	18853240	18853589
22	chr12	106316343	106316796
22	chr12	25340802	25341089
22	chr12	105651118	105651637
22	chr12	26522578	26522916
22	chr12	102961042	102961282
22	chr12	106381183	106381483
22	chr12	66430077	66430498
22	chr12	102933336	102933572
22	chr12	26392242	26393260
22	chr12	78836263	78836891
22	chr12	65721267	65721889
22	chr12	2378210	2378467
22	chr12	2353225	2353611
22	chr12	75709356	75709526
22	chr12	103954367	103954871
22	chr12	80377098	80377862
22	chr12	80378010	80378437
22	chr12	89018766	89019032
22	chr12	88956772	88957193
22	chr12	95720244	95720527
22	chr12	66008969	66009276
22	chr12	89058934	89059103
22	chr12	64495324	64495603
22	chr12	58881417	58881855
22	chr12	58923581	58924292
22	chr12	77583034	77584168
22	chr12	102921260	102921583
22	chr12	20131073	20131541
22	chr12	66158809	66159172
23	chr12	66286471	66287125
23	chr12	66285650	66286336
23	chr12	66284926	66285347
23	chr12	75979145	75979454
23	chr12	66050013	66051167
23	chr12	86531658	86532089
23	chr12	92955377	92955681
23	chr12	65930503	65931457
23	chr12	93139664	93140297
23	chr12	63126817	63127123

23	chr12	65824728	65824963
23	chr12	66089200	66089982
23	chr12	47408269	47408623
23	chr12	66329839	66331604
23	chr12	14989145	14989552
23	chr12	95512175	95512489
23	chr12	58808160	58808497
23	chr12	89340081	89340739
23	chr12	80662703	80663038
23	chr12	15815520	15815885
23	chr12	79837967	79838230
23	chr12	15781075	15781371
23	chr12	15781885	15782423
23	chr12	75840514	75840899
23	chr12	15490049	15490481
23	chr12	26150497	26150832
23	chr12	71055578	71055900
23	chr12	71039175	71040088
23	chr12	78333363	78334239
23	chr12	93349796	93350106
23	chr12	27726923	27727311
23	chr12	89845098	89845403
23	chr12	65997673	65998085
23	chr12	58951212	58951519
23	chr12	26939888	26940477
23	chr12	18615331	18615773
23	chr12	66220552	66221018
23	chr12	26164101	26164532
23	chr12	101412258	101412647
23	chr12	86020150	86020607
23	chr12	15933242	15933881
23	chr12	89767970	89768291
23	chr12	12550910	12551680
28	chr1	184807733	184808184
28	chr1	246168304	246168825
28	chr1	215130501	215131558
28	chr1	240405802	240406327
28	chr1	246755984	246756332
28	chr1	222252370	222252775
28	chr1	244275964	244276315
28	chr1	232246236	232246943
28	chr1	240562711	240563023
28	chr1	240549400	240550041
28	chr1	221290827	221291475
28	chr1	168457262	168457754
28	chr1	183681182	183681552
28	chr1	169843429	169843742
28	chr1	240422157	240422736
28	chr1	244229397	244229694
28	chr1	201665103	201665977
28	chr1	215959040	215959335
28	chr1	219214582	219214869
28	chr1	203526490	203526910
28	chr1	232613479	232613852
28	chr1	246035118	246035596
28	chr1	164613874	164614415
28	chr1	245951510	245951982
28	chr1	221441653	221441990
28	chr1	217735707	217736261
28	chr1	243368510	243368844
28	chr1	170514633	170514988
28	chr1	201735128	201735631

28	chr1	170278626	170279163
28	chr1	224780872	224781242
28	chr1	164573663	164574003
28	chr1	168475747	168476097
28	chr1	203527106	203527902
28	chr1	149958310	149958832
28	chr1	202474960	202475384
28	chr1	229099854	229100610
28	chr1	202548932	202549449
28	chr1	168622712	168623147
28	chr1	161972916	161973223
29	chr1	167597232	167597674
29	chr1	171407591	171407910
29	chr1	214724825	214725519
29	chr1	201683595	201684174
29	chr1	218879266	218880097
29	chr1	157989302	157989643
29	chr1	218553764	218554262
29	chr1	244524901	244525537
29	chr1	157979524	157979783
29	chr1	234767309	234767897
29	chr1	201430254	201430632
29	chr1	183239788	183240270
29	chr1	183203569	183204008
29	chr1	165868074	165868308
29	chr1	178207776	178208362
29	chr1	178021982	178022484
29	chr1	177980579	177980846
1	chr2	18595019	18595328
1	chr2	180325308	180325601
1	chr2	73291030	73291314
1	chr2	20714624	20714933
1	chr2	98950812	98951154
1	chr2	33329066	33329365
1	chr2	173096649	173097083
1	chr2	216576113	216576415
1	chr2	7198041	7198409
1	chr2	69525395	69525778
1	chr2	56114969	56115400
1	chr2	98485776	98486183
1	chr2	150982264	150983184
1	chr2	10715997	10716376
1	chr2	150518437	150518662
1	chr2	56089838	56090252
1	chr2	173647772	173648337
1	chr2	216394998	216395286
1	chr2	18480691	18481157
1	chr2	239552687	239553175
1	chr2	239553238	239553778
1	chr2	33650740	33651195
1	chr2	202013780	202014285
1	chr2	106020583	106020915
1	chr2	36665599	36666069
1	chr2	69365641	69365943
1	chr2	45355813	45356255
1	chr2	223709024	223709343
1	chr2	9319179	9319669
1	chr2	216708278	216708609
1	chr2	225086318	225086718
1	chr2	224330490	224330884
1	chr2	20001154	20001508
1	chr2	19911185	19911634

1	chr2	162949264	162949999
1	chr2	235158366	235158914
1	chr2	114575978	114576372
1	chr2	204549696	204550394
1	chr2	180136813	180137649
1	chr2	171382011	171382381
1	chr2	62805872	62806456
1	chr2	47082188	47083021
1	chr2	161084843	161085558
1	chr2	208258765	208260127
1	chr2	46165902	46166339
1	chr2	20348844	20349308
1	chr2	19340465	19340896
1	chr2	9778636	9778987
1	chr2	225133127	225133504
1	chr2	187755850	187756256
1	chr2	54860418	54860786
1	chr2	17824406	17824850
1	chr2	33294769	33295284
1	chr2	10423384	10423788
1	chr2	230309617	230310043
1	chr2	228682666	228684045
1	chr2	9427687	9428218
1	chr2	192030694	192031158
1	chr2	163100948	163101429
1	chr2	228727418	228727650
1	chr2	9450450	9450875
1	chr2	47077904	47078237
1	chr2	69273280	69273754
1	chr2	36788598	36788962
1	chr2	235160498	235160976
1	chr2	36683586	36684030
1	chr2	236239698	236239988
1	chr2	39721844	39722340
1	chr2	25039885	25040164
1	chr2	36599176	36599514
1	chr2	225965612	225966076
1	chr2	234395659	234396089
1	chr2	191624424	191624889
1	chr2	173860476	173860890
1	chr2	47182018	47182504
1	chr2	233853179	233853457
0	chr2	119496057	119496468
0	chr2	54893814	54894475
0	chr2	216393241	216393619
0	chr2	101383304	101384162
0	chr2	216396105	216396327
0	chr2	159992450	159992870
0	chr2	159991824	159992319
0	chr2	216565607	216566010
0	chr2	216558058	216558674
0	chr2	202519818	202520176
0	chr2	55390350	55390661
0	chr2	207986522	207986910
0	chr2	201642772	201643038
0	chr2	192722164	192722525
0	chr2	45543054	45543382
0	chr2	190212131	190212651
0	chr2	190075476	190075817
0	chr2	159793466	159793905
0	chr2	181388664	181389057
0	chr2	203182092	203182561

0	chr2	141778701	141779040
0	chr2	102051839	102052268
0	chr2	114518018	114518332
0	chr2	216405355	216405743
0	chr2	192743367	192743643
0	chr2	216529526	216529916
0	chr2	36713703	36714277
0	chr2	36677985	36678424
0	chr2	191700068	191700290
0	chr2	161785243	161785723
0	chr2	178261307	178261874
0	chr2	159887086	159887889
0	chr2	36598133	36598766
0	chr2	33516376	33516781
0	chr2	109788336	109788837
0	chr2	114519999	114520267
0	chr2	191656338	191656606
0	chr2	106005498	106005850
0	chr2	109901311	109901644
3	chr2	121692321	121692720
3	chr2	43861656	43862320
3	chr2	162941597	162942114
3	chr2	217447885	217448207
3	chr2	101358773	101359734
3	chr2	33378165	33378458
3	chr2	45405113	45405393
3	chr2	202662838	202663503
3	chr2	147739837	147740320
3	chr2	19238349	19238857
3	chr2	19455857	19456426
3	chr2	227354523	227354848
3	chr2	227292236	227292629
3	chr2	147784136	147784573
3	chr2	38241888	38242156
3	chr2	67516874	67517865
3	chr2	221089869	221090264
3	chr2	19237763	19238307
3	chr2	181486687	181487133
3	chr2	38373282	38374826
3	chr2	19106690	19106907
3	chr2	45346982	45347361
3	chr2	191702039	191702513
3	chr2	33523188	33523586
3	chr2	19807579	19808286
3	chr2	162847107	162847619
3	chr2	203000056	203000605
3	chr2	37699078	37699477
3	chr2	121487329	121487607
3	chr2	189064365	189064677
3	chr2	187321741	187322098
3	chr2	28453880	28454328
3	chr2	19376661	19377044
3	chr2	208929209	208929569
3	chr2	19738011	19738767
3	chr2	235150656	235151107
3	chr2	163089803	163090131
3	chr2	221144444	221145007
3	chr2	227050272	227051041
3	chr2	12461104	12461800
3	chr2	67487766	67488446
3	chr2	40630539	40631015
3	chr2	190399471	190399962

3	chr2	227291142	227291494
3	chr2	221143950	221144336
3	chr2	234397003	234397361
3	chr2	36565861	36566178
3	chr2	238208502	238208925
2	chr2	189488958	189489403
2	chr2	189879376	189879984
2	chr2	198774035	198774458
2	chr2	220941495	220941931
2	chr2	180904765	180905596
2	chr2	232878378	232879085
2	chr2	189844603	189845120
2	chr2	223958194	223958584
2	chr2	33370486	33370889
2	chr2	196763183	196763625
2	chr2	180881105	180881481
2	chr2	19157453	19158237
2	chr2	149950765	149951116
2	chr2	158089996	158090417
2	chr2	216518856	216519209
2	chr2	39643287	39643665
2	chr2	216763948	216764194
2	chr2	55087435	55087758
2	chr2	72642527	72642876
2	chr2	226943942	226944944
2	chr2	161725809	161726291
2	chr2	161693224	161693716
2	chr2	39608041	39608359
2	chr2	152213825	152214309
2	chr2	162962606	162962890
2	chr2	146997137	146997550
2	chr2	216592453	216592900
2	chr2	189718052	189718438
2	chr2	189833978	189834379
2	chr2	189673603	189674061
2	chr2	159927985	159928349
2	chr2	45639755	45640302
2	chr2	227658201	227658644
2	chr2	192575078	192575459
2	chr2	139466070	139466611
2	chr2	139464924	139465729
2	chr2	216266564	216267190
2	chr2	190133028	190133282
2	chr2	238144417	238144717
2	chr2	163113103	163113571
2	chr2	238131848	238132270
2	chr2	238130767	238131048
2	chr2	238117000	238117805
2	chr2	72505184	72505643
2	chr2	191723625	191723935
2	chr2	227565897	227566389
2	chr2	19015512	19016031
2	chr2	151466318	151466698
5	chr3	134082770	134083454
5	chr3	160881073	160881379
5	chr3	130682817	130683083
5	chr3	189782006	189782236
5	chr3	189949552	189950022
5	chr3	149104437	149105244
5	chr3	134046478	134046792
5	chr3	37986847	37987828
5	chr3	189729977	189730363

5	chr3	160984754	160985035
5	chr3	193523236	193523619
5	chr3	64671378	64671863
4	chr3	55223453	55223758
4	chr3	18140251	18140851
4	chr3	64250068	64250431
4	chr3	55201662	55202080
4	chr3	79216714	79217079
4	chr3	79346344	79346908
4	chr3	147892960	147893327
4	chr3	20663677	20664209
4	chr3	12478109	12478406
4	chr3	16482889	16483450
4	chr3	115553025	115553243
4	chr3	61912787	61913538
4	chr3	155108878	155109264
4	chr3	155109279	155109809
4	chr3	22286578	22286906
4	chr3	123535453	123535836
4	chr3	61941927	61942463
4	chr3	55203447	55204258
4	chr3	60565471	60565820
4	chr3	55525256	55525817
4	chr3	24237962	24238365
4	chr3	73649159	73650055
4	chr3	45126696	45127118
4	chr3	61616138	61616499
4	chr3	146831031	146831400
4	chr3	73651264	73651748
4	chr3	112530341	112530761
4	chr3	61623369	61623735
4	chr3	114225752	114226210
4	chr3	146685914	146686514
4	chr3	8597329	8597802
4	chr3	37115860	37116096
4	chr3	64329171	64330026
4	chr3	61770012	61770384
4	chr3	64331053	64331446
4	chr3	104078987	104079208
4	chr3	64429410	64429849
4	chr3	115719053	115719546
4	chr3	12264171	12264572
4	chr3	16024069	16024687
4	chr3	105231879	105232171
4	chr3	54988015	54988421
4	chr3	36705444	36705740
4	chr3	16165994	16166629
4	chr3	62095304	62095722
4	chr3	54904429	54904790
4	chr3	112967942	112968421
4	chr3	21880993	21881420
4	chr3	105192366	105192647
4	chr3	21662869	21663110
4	chr3	63710599	63710992
4	chr3	55195925	55196496
4	chr3	27208439	27208728
4	chr3	64447908	64448283
4	chr3	24296184	24296529
4	chr3	136069011	136069295
4	chr3	73776644	73776942
4	chr3	25891222	25891547
4	chr3	59545390	59545863

4	chr3	55034584	55034889
4	chr3	147724297	147724674
4	chr3	21488311	21489402
4	chr3	16781309	16781664
4	chr3	114482956	114483380
4	chr3	25078474	25078951
4	chr3	43736991	43737299
4	chr3	9256194	9256560
4	chr3	8531719	8532330
4	chr3	73815609	73816031
4	chr3	154497365	154497717
4	chr3	37240578	37240856
7	chr3	191130547	191130894
7	chr3	123407050	123407563
7	chr3	123377012	123377243
7	chr3	143021455	143022064
7	chr3	127453514	127453956
7	chr3	171024458	171024843
7	chr3	106446995	106447637
7	chr3	23727075	23727750
7	chr3	123967276	123967833
7	chr3	105077852	105078377
7	chr3	194931007	194931341
7	chr3	58614720	58615178
7	chr3	110133758	110134032
7	chr3	187990156	187991092
7	chr3	14493567	14494020
7	chr3	14513908	14514277
7	chr3	134092385	134092837
7	chr3	177650707	177651075
7	chr3	187980630	187981349
7	chr3	98827317	98827898
7	chr3	18799577	18799965
7	chr3	171591969	171592827
7	chr3	126191107	126191475
7	chr3	129370444	129370971
7	chr3	15676654	15677266
7	chr3	158443047	158443815
7	chr3	114958456	114959003
7	chr3	158420899	158421250
7	chr3	114271163	114271573
7	chr3	11494930	11495323
7	chr3	99759204	99759601
7	chr3	114343520	114344064
7	chr3	149057687	149058506
7	chr3	149864850	149865167
7	chr3	150166893	150167547
7	chr3	170715581	170716201
7	chr3	129107670	129108047
7	chr3	129213860	129214224
7	chr3	123469501	123470000
7	chr3	70898991	70899349
7	chr3	11550538	11550973
7	chr3	11609901	11610389
7	chr3	106787760	106788097
7	chr3	70881958	70882356
7	chr3	176843783	176844128
7	chr3	99764389	99764866
7	chr3	99760107	99760653
7	chr3	123976812	123977246
7	chr3	126645479	126645906
7	chr3	187785770	187786123

7	chr3	159493264	159493688
7	chr3	11590216	11590511
7	chr3	61834576	61835034
7	chr3	124277824	124278543
7	chr3	16101607	16102350
7	chr3	194085390	194086103
7	chr3	11555704	11556365
7	chr3	170520160	170520844
6	chr3	23710054	23710376
6	chr3	111593237	111593569
6	chr3	58147395	58147867
6	chr3	57015347	57015902
6	chr3	56960519	56960990
6	chr3	111458004	111458500
6	chr3	30327268	30327885
6	chr3	159590143	159590553
6	chr3	45677084	45677605
6	chr3	16181918	16182279
6	chr3	111583891	111584491
6	chr3	110245599	110245949
6	chr3	67660933	67661161
6	chr3	71586115	71586865
6	chr3	29800952	29801280
6	chr3	71160995	71161223
6	chr3	31316434	31317108
6	chr3	158486113	158486437
6	chr3	101645435	101645866
6	chr3	29373475	29373868
6	chr3	45175223	45175678
6	chr3	43911561	43911936
6	chr3	53271995	53272502
6	chr3	141086041	141086572
6	chr3	12791663	12791929
6	chr3	188003721	188004325
6	chr3	40546104	40546433
6	chr3	5036531	5036998
6	chr3	189721468	189721817
6	chr3	29281727	29282124
6	chr3	39680698	39681187
6	chr3	183088656	183088936
6	chr3	43795243	43795544
6	chr3	170444372	170444794
9	chr6	106043764	106044079
9	chr6	106044106	106044681
9	chr6	122021300	122021563
9	chr6	136990364	136990732
9	chr6	136931295	136931838
9	chr6	80310589	80310940
9	chr6	148588929	148589313
9	chr6	47002714	47003088
9	chr6	128828500	128829393
9	chr6	2688667	2688943
9	chr6	25193851	25194164
9	chr6	25192704	25193798
9	chr6	82732280	82732641
9	chr6	105876319	105876953
9	chr6	81317471	81318131
9	chr6	2630565	2630983
9	chr6	116879025	116879366
8	chr6	140382946	140383514
8	chr6	140383567	140383963
8	chr6	139908617	139909024

8	chr6	132511976	132512330
8	chr6	126860255	126860514
8	chr6	140399645	140400290
8	chr6	45618210	45618787
8	chr6	132509950	132510406
8	chr6	131174544	131175102
8	chr6	113697361	113698108
8	chr6	54482514	54483010
8	chr6	45558426	45558843
8	chr6	163950612	163951061
8	chr6	9482208	9482683
8	chr6	102254044	102254481
8	chr6	132384781	132385326
8	chr6	55955840	55956306
8	chr6	148829497	148829937
8	chr6	148822215	148822630
8	chr6	118872907	118873208
8	chr6	52405344	52405651
8	chr6	132303871	132304401
8	chr6	126304658	126305083
8	chr6	132474335	132474703
8	chr6	9523718	9524040
8	chr6	56393136	56393577
8	chr6	142737328	142737646
8	chr6	113706146	113706578
8	chr6	148593224	148593518
13	chr5	38783213	38783747
13	chr5	38845714	38846305
13	chr5	73611781	73612442
13	chr5	14185691	14186032
13	chr5	150017103	150017485
13	chr5	72661048	72661662
13	chr5	72592583	72592906
13	chr5	172330830	172331139
13	chr5	148608543	148608851
13	chr5	148865207	148866182
13	chr5	148941370	148941763
13	chr5	149318549	149318863
13	chr5	15500068	15500299
13	chr5	148825380	148825867
13	chr5	148352627	148352891
13	chr5	68816869	68817255
13	chr5	149896781	149897547
13	chr5	173191368	173191852
13	chr5	148442664	148443001
13	chr5	172882348	172882665
12	chr5	52658217	52658850
12	chr5	92608881	92609399
12	chr5	92839370	92839754
12	chr5	124429305	124429858
12	chr5	157841800	157842025
12	chr5	124470320	124470737
12	chr5	92498696	92499187
12	chr5	109201217	109202139
12	chr5	89226586	89226988
12	chr5	102786093	102786518
12	chr5	148515082	148515485
12	chr5	13985829	13986180
12	chr5	72666346	72666861
12	chr5	89226278	89226566
12	chr5	72666876	72667513
12	chr5	97803845	97804186

12	chr5	169122788	169123619
12	chr5	169053807	169054104
12	chr5	57137911	57138214
12	chr5	131599387	131600157
12	chr5	169133809	169134138
12	chr5	14038441	14038859
12	chr5	140900640	140901272
12	chr5	159510143	159510484
12	chr5	71683906	71684149
12	chr5	89316973	89317361
12	chr5	34608803	34609806
12	chr5	60870936	60871282
12	chr5	56737228	56737831
12	chr5	121464967	121465422
12	chr5	108791196	108791947
12	chr5	39501508	39501888
12	chr5	33807134	33807433
12	chr5	33343840	33344069
12	chr5	52674486	52674788
12	chr5	58375731	58376106
12	chr5	33321062	33321600
12	chr5	38601930	38602341
12	chr5	144875329	144875633
12	chr5	144858831	144859119
12	chr5	139365299	139365675
12	chr5	129885494	129886013
12	chr5	60347088	60347399
12	chr5	141825733	141826078
12	chr5	124822155	124822607
12	chr5	159276598	159276873
12	chr5	53358833	53359174
12	chr5	38468366	38468658
12	chr5	34587710	34588272
12	chr5	75996980	75997534
12	chr5	157809492	157809891
12	chr5	123043737	123044247
12	chr5	102294785	102295134
12	chr5	74936075	74936602
11	chr5	158419401	158419625
11	chr5	156814563	156814994
11	chr5	111083832	111084327
11	chr5	125338065	125338581
11	chr5	57031531	57032020
11	chr5	149849215	149849549
11	chr5	156815020	156815584
11	chr5	125338620	125338939
11	chr5	123120483	123121200
11	chr5	135329431	135330641
11	chr5	158444886	158445233
11	chr5	57111930	57112418
11	chr5	146933767	146934228
11	chr5	146915087	146915665
11	chr5	157811849	157812517
11	chr5	17000380	17000745
11	chr5	156990345	156990742
10	chr6	151766622	151766953
10	chr6	54200611	54201066
10	chr6	134742007	134742381
10	chr6	53848908	53849406
10	chr6	54155915	54156210
10	chr6	8083578	8084088
10	chr6	113672319	113672827

10	chr6	144736158	144736504
10	chr6	132511175	132511542
10	chr6	86161573	86162069
10	chr6	161678789	161679155
10	chr6	150176091	150176821
10	chr6	147384586	147385220
10	chr6	56614405	56614609
10	chr6	82723662	82723971
10	chr6	4358754	4359331
10	chr6	12490904	12491644
10	chr6	82853609	82854996
10	chr6	54546852	54547075
10	chr6	112537872	112538805
10	chr6	132405957	132406263
10	chr6	121803218	121803508
10	chr6	76463734	76463970
10	chr6	8108455	8108948
10	chr6	121844061	121844472
10	chr6	154830010	154830424
10	chr6	72188476	72188833
10	chr6	140887816	140888358
10	chr6	79316912	79317319
10	chr6	90021701	90022097
10	chr6	112525064	112525607
10	chr6	112526617	112527067
10	chr6	112527176	112527571
10	chr6	71791997	71792219
10	chr6	52794042	52794481
10	chr6	154800572	154800939
10	chr6	86173814	86174957
10	chr6	86070764	86071344
10	chr6	110114959	110115330
10	chr6	145000166	145000480
10	chr6	142618990	142619709
10	chr6	151381018	151381768
10	chr6	121759273	121759815
10	chr6	56235236	56236240
10	chr6	11650404	11650923
10	chr6	151390016	151390415
10	chr6	151388735	151389923
10	chr6	151384388	151384860
10	chr6	100748276	100748697
10	chr6	56715915	56716241
10	chr6	56728195	56728558
10	chr6	56234833	56235196
10	chr6	54058890	54059288
10	chr6	56579624	56580240
10	chr6	113735715	113736041
10	chr6	82575371	82575725
10	chr6	4607401	4608042
10	chr6	4600135	4600552
10	chr6	149884689	149885013
10	chr6	1821918	1822477
10	chr6	81167003	81167503
10	chr6	153488069	153488794
10	chr6	113880116	113880549
10	chr6	17865437	17865792
10	chr6	57130440	57130880
10	chr6	17865951	17866326
10	chr6	148685192	148685491
10	chr6	132301303	132301596
10	chr6	148735969	148736379

10	chr6	3797696	3798414
10	chr6	110111313	110111595
10	chr6	132272187	132272914
10	chr6	117819292	117819788
10	chr6	153413686	153414180
10	chr6	85333637	85333981
10	chr6	113123361	113123740
10	chr6	143718339	143719152
10	chr6	113882108	113882468
39	chr20	4206396	4206782
39	chr20	36796352	36796806
39	chr20	19973624	19974353
39	chr20	45887592	45888066
39	chr20	45944280	45944803
39	chr20	10903545	10904041
39	chr20	19766991	19767281
39	chr20	11098470	11098810
39	chr20	30300251	30301082
39	chr20	19716130	19716477
39	chr20	11247824	11248466
39	chr20	10579322	10579940
39	chr20	10585040	10585667
39	chr20	4493571	4493848
39	chr20	4487952	4488370
39	chr20	19955119	19955624
39	chr20	11435635	11435910
39	chr20	1789803	1790118
39	chr20	1793703	1794062
39	chr20	10844325	10844597
39	chr20	1810949	1811314
39	chr20	46196652	46197032
39	chr20	10829707	10830085
38	chr16	19201820	19202177
38	chr16	75278704	75279718
38	chr16	84565631	84566341
38	chr16	11295373	11296122
38	chr16	24681874	24682557
38	chr16	14493141	14493846
15	chr4	125099064	125099477
15	chr4	13922525	13922971
15	chr4	79567515	79567914
15	chr4	86932350	86932699
15	chr4	13908906	13909414
15	chr4	125861549	125861945
15	chr4	74982759	74983075
15	chr4	177714977	177715664
15	chr4	107504614	107505097
15	chr4	125825344	125825618
15	chr4	75133977	75134359
15	chr4	177909377	177909786
15	chr4	169445155	169445412
15	chr4	126310900	126311276
14	chr5	124678829	124679253
14	chr5	77973477	77974155
14	chr5	77189205	77189883
14	chr5	110918881	110920064
14	chr5	143267794	143268329
14	chr5	53625696	53626238
14	chr5	58429846	58430044
14	chr5	35047012	35047267
14	chr5	120114205	120114616
14	chr5	31364795	31365782

14	chr5	130639285	130639586
14	chr5	123679081	123679592
14	chr5	123789358	123789792
14	chr5	159354947	159355523
14	chr5	111290521	111290817
14	chr5	78102240	78102874
14	chr5	80718525	80718868
14	chr5	107148225	107148662
14	chr5	98098902	98099360
14	chr5	36502526	36503136
14	chr5	121501227	121501598
14	chr5	126706829	126707341
14	chr5	121485846	121486289
14	chr5	135352539	135353374
14	chr5	134726845	134727206
14	chr5	39400725	39401260
14	chr5	119790095	119790407
14	chr5	72606986	72607712
14	chr5	52721519	52721781
14	chr5	81628787	81629106
14	chr5	143300997	143301406
14	chr5	159310666	159310946
14	chr5	33310709	33310954
14	chr5	52541812	52542281
14	chr5	9456509	9456844
14	chr5	135393207	135394733
14	chr5	52630465	52630826
14	chr5	97967317	97967833
14	chr5	9440532	9440840
14	chr5	77974166	77974508
14	chr5	125807794	125808300
14	chr5	82771293	82772258
14	chr5	31361982	31362508
14	chr5	123529011	123529514
14	chr5	168078723	168078992
14	chr5	125720652	125721047
14	chr5	114750058	114750507
14	chr5	114734404	114734790
14	chr5	36402773	36403840
14	chr5	158878204	158878641
14	chr5	76111968	76112244
14	chr5	111265506	111266029
14	chr5	111266324	111267226
14	chr5	153656403	153656841
14	chr5	149416041	149416474
14	chr5	120534892	120535215
14	chr5	97793691	97794066
14	chr5	168588074	168588752
14	chr5	9055236	9055521
14	chr5	37771685	37772161
14	chr5	9350345	9350799
14	chr5	9468016	9468584
14	chr5	109814609	109814929
14	chr5	167170770	167171040
14	chr5	77929597	77929897
14	chr5	156942816	156943503
14	chr5	108665411	108665864
14	chr5	82617380	82617796
14	chr5	167851575	167852033
14	chr5	71520634	71521243
14	chr5	130958328	130958602
14	chr5	81709542	81710223

14	chr5	111333100	111333929
14	chr5	111334551	111334986
14	chr5	153280690	153280881
14	chr5	115790056	115790693
14	chr5	111312380	111312882
14	chr5	72670243	72671018
14	chr5	72669543	72670139
14	chr5	72672476	72672806
14	chr5	32811051	32811343
14	chr5	32790611	32791133
14	chr5	146548844	146549568
14	chr5	119640392	119640788
14	chr5	102014042	102014346
14	chr5	119629480	119629810
14	chr5	122493333	122493787
14	chr5	122492507	122493009
14	chr5	168777656	168778296
17	chr4	48604273	48604595
17	chr4	151052160	151052429
17	chr4	169559122	169559321
17	chr4	147363987	147364297
17	chr4	127837745	127838295
17	chr4	174365361	174365693
17	chr4	28773670	28773964
17	chr4	177468249	177468460
17	chr4	101906086	101906404
17	chr4	24107416	24107672
17	chr4	182889523	182889758
17	chr4	26328382	26328739
17	chr4	54721507	54722428
17	chr4	54728321	54728768
17	chr4	15453263	15454026
17	chr4	38121289	38121678
17	chr4	38950039	38950509
17	chr4	169472384	169472757
17	chr4	125829114	125829370
17	chr4	153511455	153511896
17	chr4	87260181	87260490
17	chr4	47553662	47553960
17	chr4	177191713	177192277
17	chr4	138231792	138232059
17	chr4	138228945	138229304
17	chr4	87013799	87014163
17	chr4	173772000	173772363
17	chr4	33838657	33839007
17	chr4	156525829	156526267
17	chr4	177688249	177688571
17	chr4	157898510	157898939
17	chr4	54345253	54345556
17	chr4	169785893	169786387
17	chr4	54629896	54630386
17	chr4	177761268	177761692
17	chr4	54554260	54554920
17	chr4	169059265	169059681
17	chr4	138882249	138882623
17	chr4	182582247	182582510
17	chr4	15236338	15236795
17	chr4	53719592	53719874
17	chr4	183110582	183111070
17	chr4	94315599	94315918
17	chr4	169505673	169505940
17	chr4	157607192	157607671

17	chr4	38152366	38152740
17	chr4	138077171	138077523
17	chr4	178040140	178040498
17	chr4	78500985	78501237
17	chr4	28289336	28289887
17	chr4	138445950	138446558
17	chr4	170304827	170305255
17	chr4	158972523	158972822
17	chr4	66378562	66378961
17	chr4	13981916	13982371
17	chr4	129490720	129491026
17	chr4	157243757	157244115
17	chr4	138676076	138676336
16	chr4	111462101	111462300
16	chr4	111462338	111462720
16	chr4	107286687	107287163
16	chr4	27007168	27007421
16	chr4	107462388	107462640
16	chr4	169525671	169526123
16	chr4	54600621	54601082
16	chr4	169724914	169725399
16	chr4	123704268	123704905
16	chr4	77905860	77906462
16	chr4	53967015	53967259
16	chr4	17143889	17144364
16	chr4	107508694	107509052
16	chr4	109513305	109513580
16	chr4	126242997	126243429
16	chr4	126243562	126243830
16	chr4	154434717	154435122
16	chr4	154435266	154435688
16	chr4	170175108	170175415
16	chr4	16628560	16628926
16	chr4	124620894	124621238
16	chr4	114388741	114389084
16	chr4	186713826	186714266
16	chr4	158941888	158942306
16	chr4	158941178	158941824
16	chr4	126354530	126354855
16	chr4	114357040	114357415
16	chr4	114365515	114365912
16	chr4	126289710	126290140
16	chr4	115008698	115008984
16	chr4	114304242	114304684
16	chr4	186760237	186760908
16	chr4	126289320	126289667
16	chr4	126656309	126656606
19	chr7	46851069	46851476
19	chr7	46949119	46949765
19	chr7	20263040	20263364
19	chr7	151006544	151006786
19	chr7	93977149	93977811
19	chr7	80412872	80413246
19	chr7	18801819	18802148
19	chr7	40611312	40611911
19	chr7	43575428	43575819
19	chr7	13985191	13986047
19	chr7	18818135	18818733
19	chr7	41960893	41961344
19	chr7	34135725	34136247
19	chr7	33893746	33894145
19	chr7	33914540	33914872

19	chr7	41068600	41069132
19	chr7	41136444	41136814
19	chr7	132422972	132423463
19	chr7	132424188	132424709
19	chr7	13914436	13914791
19	chr7	43824941	43825386
19	chr7	18793478	18794173
19	chr7	93926656	93927133
19	chr7	80456116	80456550
19	chr7	46948306	46948801
19	chr7	73406375	73406711
19	chr7	30264259	30264616
18	chr7	116141463	116141789
18	chr7	43609017	43609255
18	chr7	115911951	115912412
18	chr7	17639852	17640308
18	chr7	115994860	115995583
18	chr7	116083029	116083506
18	chr7	32627155	32627612
18	chr7	112124718	112125163
18	chr7	123273572	123274007
18	chr7	22600868	22601307
18	chr7	16168936	16169418
18	chr7	7900793	7901167
18	chr7	16779805	16780151
18	chr7	129995915	129996582
18	chr7	22626501	22626705
18	chr7	98048148	98048781
18	chr7	55200296	55200959
18	chr7	30843999	30844374
18	chr7	47644314	47644980
18	chr7	116346693	116346954
18	chr7	99684608	99685074
18	chr7	116356686	116357127
18	chr7	73693828	73694341
18	chr7	55132673	55133174
18	chr7	55133225	55133859
18	chr7	7478571	7478935
18	chr7	47492944	47493402
18	chr7	7532048	7532504
18	chr7	23374088	23374733
18	chr7	30315721	30316314
18	chr7	130576201	130576633
18	chr7	43733526	43733868
18	chr7	130571896	130572452
31	chr11	95846461	95846821
31	chr11	33394047	33394408
31	chr11	86976309	86976496
31	chr11	12221857	12222269
31	chr11	12204078	12205174
31	chr11	86448756	86449194
31	chr11	101981981	101983096
31	chr11	86171012	86171324
31	chr11	44787687	44788177
31	chr11	12714123	12714654
31	chr11	95895912	95896350
31	chr11	19617850	19618362
31	chr11	12222368	12223160
31	chr11	11994944	11995289
31	chr11	122059893	122060397
31	chr11	29328105	29328378
31	chr11	11998717	11999215

31	chr11	19617432	19617799
31	chr11	122080630	122081001
31	chr11	121955982	121956411
31	chr11	27739857	27740298
31	chr11	12011593	12012104
31	chr11	73045519	73046053
31	chr11	36094975	36095380
31	chr11	130347930	130348478
31	chr11	130767139	130767514
31	chr11	73032675	73033295
31	chr11	73034508	73034858
31	chr11	69311905	69312450
31	chr11	114166847	114167324
31	chr11	114165456	114166206
31	chr11	114178156	114178616
31	chr11	86451615	86451977
31	chr11	130392077	130392522
31	chr11	12000673	12001706
31	chr11	96044212	96044803
31	chr11	36033928	36034318
31	chr11	86235008	86236037
31	chr11	44008900	44009525
31	chr11	28855190	28855603
31	chr11	122051032	122051659
31	chr11	122067571	122068243
31	chr11	102866942	102867334
31	chr11	27955628	27956183
31	chr11	122007612	122008127
31	chr11	19736501	19737395
30	chr11	101737308	101737744
30	chr11	77056649	77056981
30	chr11	128287481	128287924
30	chr11	86719052	86719270
30	chr11	128351072	128351376
30	chr11	129119352	129119573
30	chr11	26864051	26864435
30	chr11	123045318	123045844
30	chr11	123043970	123044580
30	chr11	35551403	35552148
30	chr11	102473661	102474196
30	chr11	102107767	102108035
30	chr11	111428230	111428970
30	chr11	12527976	12528344
30	chr11	35310727	35311029
30	chr11	111506400	111506832
30	chr11	122214912	122215367
30	chr11	119438909	119439360
30	chr11	12455227	12455589
30	chr11	121807782	121808120
30	chr11	122011258	122011755
30	chr11	121806846	121807403
30	chr11	26842274	26842614
30	chr11	12419489	12420023
30	chr11	130668133	130668524
30	chr11	106911891	106912141
37	chr18	56246822	56247346
37	chr18	42596440	42596988
37	chr18	56248444	56248844
37	chr18	41242656	41242979
37	chr18	18697326	18697624
37	chr18	42406392	42406854
37	chr18	74157211	74157510

37	chr18	65450694	65451156
37	chr18	68086788	68087133
37	chr18	39518397	39518786
37	chr18	42181451	42181778
37	chr18	42182282	42182661
37	chr18	67713465	67713877
37	chr18	65092834	65093423
37	chr18	42630823	42631279
37	chr18	42835364	42835789
36	chr14	62031332	62031701
36	chr14	57849129	57849381
36	chr14	59204993	59205218
36	chr14	100223815	100224196
36	chr14	29705959	29706504
36	chr14	55263777	55263984
36	chr14	58549227	58549531
36	chr14	69161905	69162434
36	chr14	85881611	85882041
36	chr14	106465433	106465944
36	chr14	62087241	62087688
36	chr14	69010524	69011121
36	chr14	85996280	85996574
36	chr14	55981467	55981868
36	chr14	50441894	50442306
36	chr14	85982552	85983453
35	chr15	67417701	67418424
35	chr15	99440009	99440597
35	chr15	71385714	71385930
35	chr15	44394798	44395482
35	chr15	74532210	74532799
35	chr15	71587230	71588219
35	chr15	71149065	71149546
35	chr15	99270795	99271215
35	chr15	44205452	44205874
35	chr15	62405020	62405471
35	chr15	63189088	63189673
35	chr15	33571057	33571370
35	chr15	91229682	91230150
35	chr15	63311600	63311901
35	chr15	71588239	71588621
35	chr15	33116922	33117288
35	chr15	71570995	71571844
35	chr15	67175624	67176199
35	chr15	67224339	67224833
34	chr13	47789875	47790272
34	chr13	94725224	94725634
34	chr13	48235237	48235604
34	chr13	91137773	91138094
34	chr13	48247161	48247505
34	chr13	48246694	48247083
34	chr13	44892386	44892651
34	chr13	31293051	31294030
34	chr13	49349917	49350395
34	chr13	49349276	49349892
34	chr13	32324045	32324468
34	chr13	30096118	30096458
34	chr13	47613279	47613589
34	chr13	48432662	48432915
34	chr13	51163225	51164403
34	chr13	51149153	51149616
34	chr13	94764513	94764818
34	chr13	45629484	45630309

34	chr13	76207429	76207828
34	chr13	110527262	110527756
34	chr13	75346086	75346376
34	chr13	75335841	75336220
34	chr13	31424886	31425388
34	chr13	74825029	74825371
33	chr9	81835994	81836356
33	chr9	81839492	81839995
33	chr9	118135643	118136139
33	chr9	84946838	84947428
33	chr9	118377922	118378475
33	chr9	84635960	84636483
33	chr9	118434879	118435798
33	chr9	118863810	118864191
33	chr9	104344814	104345127
33	chr9	118789820	118790076
33	chr9	117878243	117878577
33	chr9	117821225	117821521
33	chr9	89808686	89809707
33	chr9	117797104	117797393
33	chr9	106838599	106839018
33	chr9	84521551	84521858
33	chr9	112533385	112533772
33	chr9	112555404	112555917
33	chr9	110470169	110470620
33	chr9	110469352	110470130
33	chr9	113531939	113532277
33	chr9	112578611	112579270
33	chr9	89598180	89598441
33	chr9	89598470	89599243
33	chr9	89816299	89816543
33	chr9	118131821	118132311
33	chr9	110014075	110014646
33	chr9	117996344	117996744
33	chr9	118012928	118013371
33	chr9	84738973	84739295
33	chr9	119311609	119312025
33	chr9	113205732	113206072
33	chr9	118760205	118760508
33	chr9	95324928	95325333
33	chr9	118453912	118454328
33	chr9	118452931	118453876
33	chr9	85104816	85105172
33	chr9	118701260	118702131
33	chr9	106838031	106838415
33	chr9	119038260	119038519
33	chr9	89409224	89409685
33	chr9	117974113	117974471
33	chr9	113412596	113412983
33	chr9	111149238	111149679
33	chr9	118293704	118294100
33	chr9	118367326	118367905
33	chr9	112562080	112562349
33	chr9	111313567	111314014
33	chr9	117908872	117909262
32	chr9	133837918	133838360
32	chr9	116382700	116383161
32	chr9	114812784	114813267
32	chr9	114714880	114715226
32	chr9	133712495	133712974
32	chr9	116383756	116384099

Appendix 4: Facilitated Diffusion Model

Biophysical model of simulations of the crowdsourcing effect

We used simulations of the transcription factor (TF) target finding process to evaluate the hypothesis that the crowdsourcing effect has a measurable impact on TF occupancy. This is largely uncharacterized territory, so we derive biologically plausible biophysical parameters, given the scarce experimental data. We implemented an extension of fastGRiP [Ezer et al., 2014], which allows for compute-efficient simulation of the facilitated diffusion process, with an additional translocation mode in which TFs can jump between DNA strands. In the original fastGRiP implementation, there is an interval surrounding each binding site along the DNA called the *sliding window* (as in: allowing for sliding), and any TF that binds to the DNA within this range will almost certainly reach the binding site by 1D diffusion. In the improved simulation, we introduced an absorbing sphere around each binding site, and if a TF enters this sphere it will almost certainly reach the binding site (See Figure 1AI). The equation that describes the probability of a TF distance r away reaching the absorbing sphere at time t has been previously derived [Carslaw and Jaeger, 1959] [Paramanathan et al., 2014]. In this equation, s is the radius of the absorbing sphere, and D_{eff} is the effective diffusion rate.

$$\phi(r,t) = \frac{s(r-s)}{2r\sqrt{\pi D_{eff} t^3}} \exp(-\frac{(r-s)^2}{4D_{eff}t}) \tag{1}$$

In analogy to our sliding window, we adjust the diameter of the absorbing sphere s to 30nm for absorbing TFs from outside of the clusters. In the case of internal jumps within homotypic clusters, the absorbing sphere is set to be 2nm [Wunderlich and Mirny, 2008] (representing directly reaching the binding site from 3D diffusion), because fastGRiP already incorporates the TFs' sliding between nearby binding sites, and we must be careful not to double-count this effect. We calculate the diffusion coefficient D_{eff} using the following equation, as previously described [Elf et al., 2007].

$$D_{eff} = (1-a)D + a\frac{D_1}{3} \tag{2}$$

where D_{eff} is the effective diffusion coefficient, a is the proportion of time the TF spends sliding on the DNA non-specifically, D is the 3-dimensional diffusion coefficient, and D_1 is the 1-dimensional diffusion coefficient of the

TF on DNA. In the following analysis, we choose D to be $3\mu m^2/s$, D_1 to be $0.046\mu m^2/s$,and a to be 90%, as estimated by single molecule tracking of LacI in live *E.coli*. [Elf et al., 2007]. In the absence of experimental data, we are acting under the assumption that TFs in eukaryotic nuclei have similar diffusion parameters.

In the original simulation, we assumed that any unbound TF is equally likely to bind to any other binding site, so when a TF dissociates from a binding site it enters a *pool of TFs*. In the updated simulation, a recently dissociated TF is more likely to bind to a nearby site than a far away site, as described by the probability density functions depicted in Figure 1B. Figure 1B illustrates that even after 0.1 seconds, the probability density functions for TFs jumping between DNA strands that are 100nm, 200nm, and 2000nm apart nearly converge. After 10 seconds, the probability of the TF binding to a DNA strand 100nm, 200nm, or 2000nm away is less than 1% for all cases, so we replace an TF that is still free floating after 10 seconds into the TF pool.

The other parameters we used were identical to those described by Ezer et al, 2014; we set $\tau_0 = 3.3$, $cn = 100$, and the distance between binding sites in a cluster to 5 bp, unless otherwise stated.

Modifications to fastGRiP

The fastGRiP simulation tool (available in http://logic.sysbiol.cam.ac.uk/fgrip/) is a stochastic simulation that models TF binding and unbinding using the Gillespie algorithm. It models each unique combination of bound TFs as a state, which can transition to another state through either a TF association, dissociation or (in the case of homotypic cluster) translocation to neighboring sites, but it does not allow TFs to jump between strands. In our updated fastGRiP, we incorporate the jumping probability from one strand to another by combining pre-computed diffusion probability look-up table with Gillespie algorithm.

Given a set of possible reactions, the Gillespie algorithm can (i) randomly select the time when the next reaction will occur (ii) randomly select which reaction will most likely happen next. A core assumption of the Gillespie algorithm is that the distribution of reaction events must approximate an exponential distribution, which implies that the probability of a reaction event is time-independent. However, the diffusion of a TF from one DNA

strand varies with time; for instance, it is impossible for a TF to immediately detach from one strand and attach to another, because the TF must have enough time to travel the distance between the two DNA strands. Equation (1) in the supplements describes how the probability density function for TF jumping varies with time. Selecting the time of the next reaction requires sampling a value from the averaged probability density function of the reaction times for all of the possible reactions, which is easy in the case of exponential functions, but would require time consuming steps such as averaging custom functions and sampling values from this distribution. Selecting the next reaction would be even more time consuming when the probability density functions are not exponential, since it would require a numerical integration step for the custom distribution.

Instead, we modify fastGRiP as follows to allow diffusion between DNA strands to be incorporated without substantially decreasing the runtime of the simulation. In the earlier version of fastGRiP, once a TF became dissociated from the DNA, it enters a pool where the TF is equally likely to bind to any location along the DNA. Now, once a TF dissociates, it enters a second pool of *diffusible TFs*. It samples the time of its next expected jump from a 100,000 element pre-computed lookup table generated in Matlab. All of the possible TF jumps are stored in a PriorityQueue, a data structure that efficiently stores these values in sorted order. When the Gillespie algorithm reaches the step in which it selects the time of the next TF association, dissociation or intra-cluster translocation reaction, it first checks the pool of diffusible TFs to see if any TF jumping events have happened in the meantime, and updates the state of the system accordingly. Sometimes, a TF jump event can no longer occur, because that DNA binding site is already occupied by the time the new TF diffused to it. In these cases, we recomputed a new location for the TF to diffuse to and add it to the PriorityQueue again. If at any time, the sampled TF jump time is greater than 10 seconds, we do not store this TF in the pool of diffusible TFs, because it has nearly equal likelihood of diffusing to any binding site, and we place the TF in the original TF pool. This algorithm modification allows us to model TF jump events, even though the probability density function is not exponential, without substantially increasing the runtime of the algorithm. The code for this modification is available at https://github.com/ezer/DiffusionMarkovModelJumping.

Simple scenarios for occupancy boost in two binding site clusters of spatial proximity

We compare three scenarios 1) First, we look at a pair of homotypic clusters that are on two different strands, as shown in Figure 1AII, and we vary the distance between two DNA strands 2) Then, we take the same scenario and adjust the distance between TF binding sites within the homotypic cluster (Figure 1AIII). 3) Finally, we vary the number of TF binding sites within the homotypic cluster (Figure 1AIV).

In each of these cases, we are interested in determining how these binding site organizations influence TF occupancy, which we define here as the average probability that each TF binding site is bound. For instance, if the TF occupancy is 0.05, it means that (on average) each TF binding site is bound 5% of the time. Of course, if there are 20 binding sites in the simulation, this would mean that on average 1 TF is bound at any given time.

In the first scenario with two binding site clusters located at different distances from each other, we see that the closer these two clusters are in 3D, the higher average occupancy they have, which shows jumping between strands substantially increases the average TF occupancy of the region (Figure 1C).

Next, we vary the distance between binding sites within homotypic clusters, and discover that this only slightly influences overall TF occupancy, at least given the parameters that we simulated (Figure 1D). This result is a reflection that there are two opposite effects influencing TF binding site occupancy. On one hand, there is increased translocation of TFs between two binding sites in a cluster as the distance between binding sites decrease. On the other hand, the absorbing spheres around each of the TF binding sites will intersect if the two sites are very close together in a homotypic cluster, so the overall chance that a TF jumps to another binding site is reduced. This is comparable to playing a game of darts with two dartboards that are partially overlapping - the chance of scoring is higher the less they overlap. Therefore, the distance between binding sites in homotypic clusters might not have very much influence on TF occupancy.

Finally, we consider homotypic clusters with four binding sites (Figure 1E). Homotypic clusters with more binding sites are more greatly impacted by having 3D jumping between strands, with a 58% improvement in TF occupancy when DNA strands are 100nm apart in the quadruple TF binding sites homotypic cluster case as opposed to a 42% improvement in the double

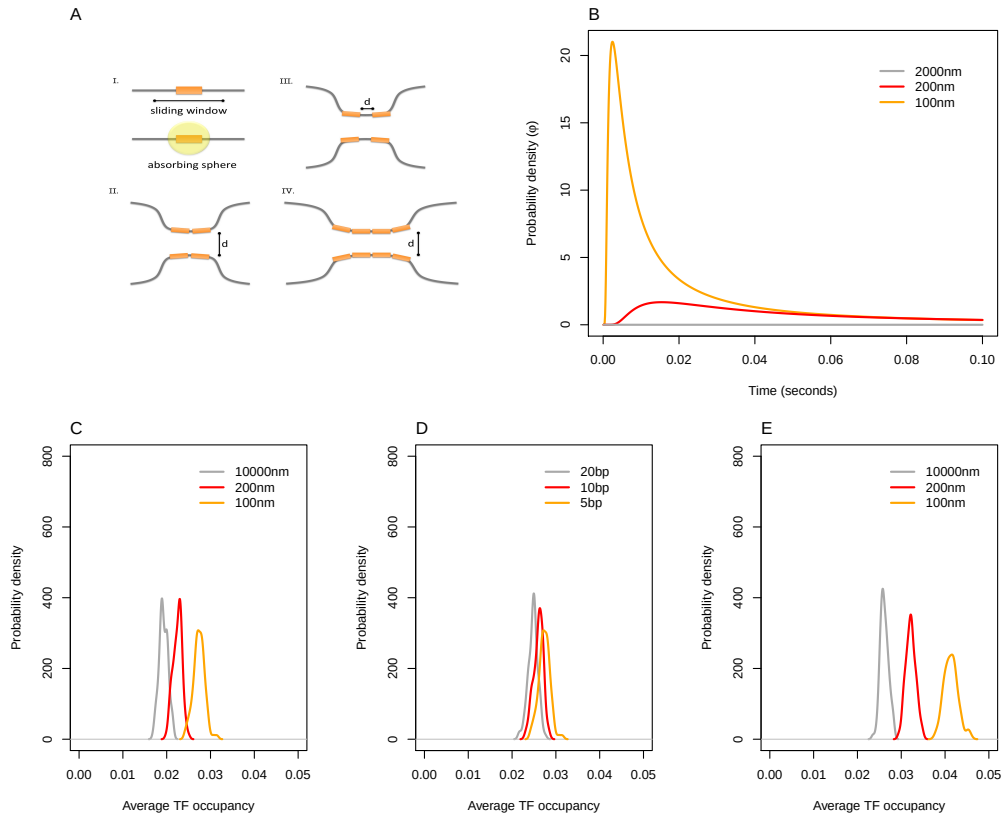


Figure 1: *Biophysical simulations of the crowdsourcing effect.* We assess the biophysical plausibility of the crowdsource effect using fastGRiP simulations. Subfigure AI demonstrates how fastGRiP's sliding length concept is extended to an absorbing sphere as we consider 3D diffusion. AII-AIV illustrate the simulated scenarios that were evaluated. The shape of the probability density function ϕ from equation 2 is shown in B. The results from the simulated scenarios AII-AIV are depicted in C-E, respectively, as probability density plots of the TF occupancy, which is the probability of each TF binding site being bound. Note that the TF occupancy, as defined by fastGRiP, includes not only the time at which a binding site is occupied, but also the time when the TF is within 90bp of the binding site.

TF binding site cluster case.

References

[Carslaw and Jaeger, 1959] Carslaw, H. S. and Jaeger, J. C. (1959). *Conduction of Heat in Solids*. Clarendon Press.

[Elf et al., 2007] Elf, J., Li, G.-W., and Xie, X. S. (2007). Probing transcription factor dynamics at the single-molecule level in a living cell. *Science*, 316(5828):1191–1194.

[Ezer et al., 2014] Ezer, D., Zabet, N. R., and Adryan, B. (2014). Physical constraints determine the logic of bacterial promoter architectures. *Nucleic acids research*, page gku078.

[Paramanathan et al., 2014] Paramanathan, T., Reeves, D., Friedman, L. J., Kondev, J., and Gelles, J. (2014). A general mechanism for competitor-induced dissociation of molecular complexes. *Nature communications*, 5.

[Wunderlich and Mirny, 2008] Wunderlich, Z. and Mirny, L. A. (2008). Spatial effects on the speed and reliability of protein–dna search. *Nucleic acids research*, 36(11):3570–3578.

References

Adelman, K., and Lis, J.T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. Nat. Rev. Genet. *13*, 720–731.

Adli, M., and Bernstein, B.E. (2011). Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. Nat. Protoc. *6*, 1656–1668.

Anderson, G.M., and Freytag, S.O. (1991). Synergistic activation of a human promoter in vivo by transcription factor Sp1. Mol. Cell. Biol. *11*, 1935–1943.

Andersson, R., Sandelin, A., and Danko, C.G. (2015). A unified architecture of transcriptional regulatory elements. Trends Genet. *31*, 426–433.

Arnold, C.D., Gerlach, D., Spies, D., Matts, J. a, Sytnikova, Y. a, Pagani, M., Lau, N.C., and Stark, A. (2014). Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. Nat. Genet. *46*, 685–692.

Arvey, A., Agius, P., Noble, W.S., and Leslie, C. (2012). Sequence and chromatin determinants of cell-type-specific transcription factor binding. Genome Res. *22*, 1723–1734.

Ay, F., Bunnik, E.M., Varoquaux, N., Bol, S.M., Prudhomme, J., Vert, J.P., Noble, W.S., and Le Roch, K.G. (2014). Three-dimensional modeling of the P. falciparum genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. Genome Res. *24*, 974–988.

Babaei, S., Akhtar, W., de Jong, J., Reinders, M., and de Ridder, J. (2015). 3D hotspots of recurrent retroviral insertions reveal long-range interactions with cancer genes. Nat. Commun. *6*, 6381.

Babu, M.M., Janga, S.C., de Santiago, I., and Pombo, A. (2008). Eukaryotic gene regulation in three dimensions and its impact on genome evolution. Curr. Opin. Genet. Dev. *18*, 571–582.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299–308.

Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.-M., Dostie, J., Pombo, a., and Nicodemi, M. (2012). Complexity of chromatin folding is captured by the strings and binders switch model. Proc. Natl. Acad. Sci. *109*, 16173–16178.

Bar-Joseph, Z., Gerber, G.K., Lee, T.I., Rinaldi, N.J., Yoo, J.Y., Robert, F., Gordon, D.B., Fraenkel, E., Jaakkola, T.S., Young, R. a, et al. (2003). Computational discovery of gene modules and regulatory networks. Nat. Biotechnol. *21*, 1337–1342.

Barozzi, I., Simonatto, M., Bonifacio, S., Yang, L., Rohs, R., Ghisletti, S., and Natoli, G. (2014). Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA Features of Mammalian Enhancers. Mol. Cell *54*, 844–857.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., et al. (2010). NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res. *39*, D1005–D1010.

Berman, B.P., Pfeiffer, B.D., Laverty, T.R., Salzberg, S.L., Rubin, G.M., Eisen, M.B., and Celniker, S.E. (2004). Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. Genome Biol *5*, R61.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Bickmore, W.A., and van Steensel, B. (2013). Genome architecture: domain organization of interphase chromosomes. Cell *152*, 1270–1284.

Biggin, M.D. (2011). Animal transcription networks as highly connected, quantitative continua. Dev. Cell *21*, 611–626.

Blow, M.J., McCulley, D.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2010). ChIP-Seq identification of weakly conserved heart enhancers. Nat. Genet. *42*, 806–810.

Boyle, A.P., Song, L., Lee, B.-K., London, D., Keefe, D., Birney, E., Iyer, V.R., Crawford, G.E., and Furey, T.S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. *21*, 456–464.

Brackley, C. a, Taylor, S., Papantonis, A., Cook, P.R., and Marenduzzo, D. (2013a). Nonspecific bridging-induced attraction drives clustering of DNA-binding proteins and genome organization. Proc. Natl. Acad. Sci. U. S. A. *110*, E3605–E3611.

Brackley, C. a., Cates, M.E., and Marenduzzo, D. (2012). Facilitated diffusion on mobile DNA: Configurational traps and sequence heterogeneity. Phys. Rev. Lett. *109*, 1–5.

Brackley, C.A., Cates, M.E., and Marenduzzo, D. (2013b). Intracellular Facilitated Diffusion: Searchers, Crowders, and Blockers. Phys. Rev. Lett. *111*, 108101.

Buckley, M.S., and Lis, J.T. (2014). Imaging RNA Polymerase II transcription sites in living cells. Curr. Opin. Genet. Dev. *25*, 126–130.

Bulger, M., and Groudine, M. (2010). Enhancers: the abundance and function of regulatory sequences beyond promoters. Dev. Biol. *339*, 250–257.

Bulger, M., and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. Cell *144*, 327–339.

Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K., and Neidle, S. (2006). Quadruplex DNA: sequence, topology and structure. Nucleic Acids Res. *34*, 5402–5415.

Calo, E., and Wysocka, J. (2013). Modification of Enhancer Chromatin: What, How, and Why? Mol. Cell *49*, 825–837.

Cao, Y., Yao, Z., Sarkar, D., Lawrence, M., Sanchez, G.J., Parker, M.H., MacQuarrie, K.L., Davison, J., Morgan, M.T., Ruzzo, W.L., et al. (2010). Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. Dev. Cell *18*, 662–674.

Carslaw, H.S., and Jaeger, J.C. (1959). Conduction of heat in solids. Oxford Clarendon Press.

Chakalova, L., and Fraser, P. (2010). Organization of transcription. Cold Spring Harb. Perspect. Biol. *2*, a000729.

Chen, K., and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. Nat. Rev. Genet. *8*, 93–103.

Coleman, R.A., and Pugh, B.F. (1995). Evidence for Functional Binding and Stable Sliding of the TATA Binding Protein on Nonspecific DNA. J. Biol. Chem. *270*, 13850–13859.

Cook, P. (1995). A chromomeric model for nuclear and chromosome structure. J. Cell Sci. *108*, 2927–2935.

Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet. *12*, 628–640.

Crocker, J., Abe, N., Rinaldi, L., McGregor, A.P., Frankel, N., Wang, S., Alsawadi, A., Valenti, P., Plaza, S., Payre, F., et al. (2015). Low Affinity Binding Site Clusters Confer Hox Specificity and Regulatory Robustness. Cell 191–203.

D'haeseleer, P. (2006). What are DNA sequence motifs? Nat. Biotechnol. *24*, 423–425.

Daniel, B., Nagy, G., and Nagy, L. (2014). The intriguing complexities of mammalian gene regulation: how to link enhancers to regulated genes. Are we there yet? FEBS Lett. *588*, 2379–2391.

Davis, R.L., Cheng, P.-F., Lassar, A.B., and Weintraub, H. (1990). The MyoD DNA binding domain contains a recognition code for muscle-specific gene activation. Cell *60*, 733–746.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.

Dekker, J., Marti-Renom, M. a, and Mirny, L. a (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. *14*, 390–403.

DeMare, L.E., Leng, J., Cotney, J., Reilly, S.K., Yin, J., Sarro, R., and Noonan, J.P. (2013). The genomic landscape of cohesin-Associated chromatin interactions. Genome Res. *23*, 1224–1234.

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P.D., Dean, A., and Blobel, G. a (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell *149*, 1233–1244.

Dermitzakis, E.T., Reymond, A., and Antonarakis, S.E. (2005). Conserved non-genic sequences - an unexpected feature of mammalian genomes. Nat. Rev. Genet. *6*, 151–157.

Dewey, F.E., Perez, M. V, Wheeler, M.T., Watt, C., Spin, J., Langfelder, P., Horvath, S., Hannenhalli, S., Cappola, T.P., and Ashley, E. a (2011). Gene coexpression network topology of cardiac development, hypertrophy, and failure. Circ. Cardiovasc. Genet. *4*, 26–35.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dixon, J.R., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J.E., Lee, A.Y., Ye, Z., Kim, A., Rajagopal, N., Xie, W., et al. (2015). Chromatin architecture reorganization during stem cell differentiation. Nature *518*, 331–336.

Djebali, S., Davis, C. a., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101–108.

Doolittle, W.F. (2013). Is junk DNA bunk? A critique of ENCODE. Proc. Natl. Acad. Sci. *2013*, 1–7.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., et al. (2014). Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. Cell *159*, 374–387.

Dror, I., Golan, T., Levy, C., Rohs, R., and Mandel-Gutfreund, Y. (2015). A widespread role of the motif environment on transcription factor binding across diverse protein families. Genome Res. *25*, 1268–1280.

Edelman, L.B., and Fraser, P. (2012). Transcription factories: genetic programming in three dimensions. Curr. Opin. Genet. Dev. *22*, 110–114.

Elf, J., Li, G.-W., and Xie, X.S. (2007). Probing transcription factor dynamics at the single-molecule level in a living cell. Science *316*, 1191–1194.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat. Methods *9*, 215–216.

Eskiw, C.H., Cope, N.F., Clay, I., Schoenfelder, S., Nagano, T., and Fraser, P. (2010). Transcription factories and nuclear organization of the genome. Cold Spring Harb. Symp. Quant. Biol. *75*, 501–506.

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

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

Ezer, D., Zabet, N.R., and Adryan, B. (2014a). Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression. Comput. Struct. Biotechnol. J. *10*, 63–69.

Ezer, D., Zabet, N.R., and Adryan, B. (2014b). Physical constraints determine the logic of bacterial promoter architectures. Nucleic Acids Res. *42*, 4196–4207.

Fang, X., Yin, W., Xiang, P., Han, H., Stamatoyannopoulos, G., and Li, Q. (2009). The Higher Structure of Chromatin in the LCR of the β-Globin Locus Changes during Development. J. Mol. Biol. *394*, 197–208.

Feuerborn, A., and Cook, P.R. (2015a). Why the activity of a gene depends on its neighbors. Trends Genet. 1–8.

Feuerborn, A., and Cook, P.R. (2015b). Why the activity of a gene depends on its neighbors. Trends Genet. *31*, 483–490.

Filippova, D., Patro, R., Duggal, G., and Kingsford, C. (2013). Multiscale identification of topological domains in chromatin. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) *8126 LNBI*, 300–312.

Fisher, W.W., Li, J.J., Hammonds, A.S., Brown, J.B., Pfeiffer, B.D., Weiszmann, R., MacArthur, S., Thomas, S., Stamatoyannopoulos, J. a, Eisen, M.B., et al. (2012). DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. Proc. Natl. Acad. Sci. *109*, 21330–21335.

Foat, B.C., Morozov, A. V, and Bussemaker, H.J. (2006). Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. Bioinformatics *22*, e141–e149.

Fraser, P. (2006). Transcriptional control thrown for a loop. Curr. Opin. Genet. Dev. *16*, 490–495.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y. Bin, Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58–64.

Furey, T.S. (2012). ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nat. Rev. Genet. *13*, 840–852.

Gaudet, J. (2002). Regulation of Organogenesis by the Caenorhabditis elegans FoxA Protein PHA-4. Science (80-.). *295*, 821–825.

Ghavi-Helm, Y., Klein, F. a., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E.E.M. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. Nature *512*, 96–100.

Giniger, E., and Ptashne, M. (1988). Cooperative DNA binding of the yeast transcriptional activator GAL4. Proc. Natl. Acad. Sci. *85*, 382–386.

Gotea, V., Visel, A., Westlund, J.M., Nobrega, M. a, Pennacchio, L. a, and Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res. *20*, 565–577.

Graur, D., Zheng, Y., Price, N., Azevedo, R.B.R., Zufall, R.A., and Elhaik, E. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. Genome Biol. Evol. *5*, 578–590.

Hadjur, S., Williams, L.M., Ryan, N.K., Cobb, B.S., Sexton, T., Fraser, P., Fisher, A.G., and Merkenschlager, M. (2009). Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature *460*, 410–413.

Hammar, P., Leroy, P., Mahmutovic, A., Marklund, E.G., Berg, O.G., and Elf, J. (2012). The lac repressor displays facilitated diffusion in living cells. Science *336*, 1595–1598.

Hannenhalli, S. (2008). Eukaryotic transcription factor binding sites - Modeling and integrative search methods. Bioinformatics *24*, 1325–1331.

Hatzis, P., and Talianidis, I. (2002). Dynamics of Enhancer-Promoter Communication during Differentiation-Induced Gene Activation. Mol. Cell *10*, 1467–1477.

He, X., Duque, T.S.P.C., and Sinha, S. (2011). Evolutionary Origins of Transcription Factor Binding Site Clusters. Mol. Biol. Evol. *29*, 1059–1070.

He, X., Duque, T.S.P.C., and Sinha, S. (2012). Evolutionary origins of transcription factor binding site clusters. Mol. Biol. Evol. *29*, 1059–1070.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009a). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009b). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Heinz, S., Romanoski, C.E., Benner, C., Allison, K. a, Kaikkonen, M.U., Orozco, L.D., and Glass, C.K. (2013). Effect of natural genetic variation on enhancer selection and function. Nature *503*, 487–492.

Heinz, S., Romanoski, C.E., Benner, C., and Glass, C.K. (2015). The selection and function of cell type-specific enhancers. Nat. Rev. Mol. Cell Biol. *16*, 144–154.

Henikoff, S., and Shilatifard, A. (2011). Histone modification: cause or cog? Trends Genet. *27*, 389–396.

Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., Reynolds, A.P., Thurman, R.E., Neph, S., Kuehn, M.S., Noble, W.S., et al. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *6*, 283–289.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A. a, Hoke, H. a, and Young, R. a (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Hnisz, D., Schuijers, J., Bradner, J.E., Young, R.A., Hnisz, D., Schuijers, J., Lin, C.Y., Weintraub, A.S., Abraham, B.J., and Lee, T.I. (2015). Short Article Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super- Short Article Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers. Mol. Cell *58*, 362–370.

Ing-simmons, E., Seitan, V.C., Faure, A.J., Flicek, P., Dekker, J., Fisher, A.G., Lenhard, B., and Merkenschlager, M. (2014). Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. *3*, 318–356.

Jaenisch, R., and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat. Genet. *33 Suppl*, 245–254.

Jiang, C., and Pugh, B.F. (2009). Nucleosome positioning and gene regulation: advances through genomics. Nat. Rev. Genet. *10*, 161–172.

Jothi, R., Cuddapah, S., Barski, A., Cui, K., and Zhao, K. (2008). Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res. *36*, 5221–5231.

Junier, I., Martin, O., and Képès, F. (2010). Spatial and topological organization of DNA chains induced by gene co-localization. PLoS Comput. Biol. *6*, e1000678.

Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E.H., Birney, E., and Furlong, E.E.M. (2012). A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell *148*, 473–486.

Kagey, M.H., 1, *, 1, 2, 3, Ebmeier, C.C., 4, 2, and Young, & R.A. (2010a). Mediator and cohesin connect gene expression and chromatin architecture. Nature.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D. a, van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010b). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kazemian, M., Pham, H., Wolfe, S. a, Brodsky, M.H., and Sinha, S. (2013). Widespread evidence of cooperative DNA binding by transcription factors in Drosophila development. Nucleic Acids Res. *41*, 8237–8252.

Krivega, I., and Dean, A. (2012). Enhancer and promoter interactions-long distance calls. Curr. Opin. Genet. Dev. *22*, 79–85.

Leith, J.S., Tafvizi, A., Huang, F., Uspal, W.E., Doyle, P.S., Fersht, A.R., Mirny, L.A., and van Oijen, A.M. (2012). Sequence-dependent sliding kinetics of p53. Proc. Natl. Acad. Sci. U. S. A. *109*, 16552–16557.

Lelli, K.M., Slattery, M., and Mann, R.S. (2012). Disentangling the many layers of eukaryotic transcriptional regulation. Annu. Rev. Genet. *46*, 43–68.

Lettice, L.A. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. *12*, 1725–1735.

Levine, M. (2010). Transcriptional enhancers in animal development and evolution. Curr. Biol. *20*, R754–R763.

Levo, M., and Segal, E. (2014). In pursuit of design principles of regulatory sequences. Nat. Rev. Genet. *15*, 453–468.

Levy, S., and Hannenhalli, S. (2002). Identi ® cation of transcription factor binding sites in the human genome sequence. *514*, 510–514.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell *148*, 84–98.

Li, Q., Ritter, D., Yang, N., Dong, Z., Li, H., Chuang, J.H., and Guo, S. (2010). A systematic approach to identify functional motifs within vertebrate developmental enhancers. Dev. Biol. *337*, 484–495.

Li, X.-Y., Thomas, S., Sabo, P.J., Eisen, M.B., Stamatoyannopoulos, J. a, and Biggin, M.D. (2011). The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. Genome Biol. *12*, R34.

Lickwar, C.R., Mueller, F., Hanlon, S.E., McNally, J.G., and Lieb, J.D. (2012). Genome-wide protein–DNA binding dynamics suggest a molecular clutch for transcription factor function. Nature *484*, 251–255.

Lieberman-Aiden, E., Berkum, N.L. Van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). of the Human Genome. *33292*, 289–294.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. Nature *478*, 476–482.

Liu, R., Hannenhalli, S., and Bucan, M. (2009). Motifs and cis-regulatory modules mediating the expression of genes co-expressed in presynaptic neurons. Genome Biol *10*, R72.

Ludwig, M.Z., Kittler, R., White, K.P., and Kreitman, M. (2011). Consequences of Eukaryotic Enhancer Architecture for Gene Expression Dynamics, Development, and Fitness. PLoS Genet. *7*, e1002364.

Lutz, B., Lu, H.C., Eichele, G., Miller, D., and Kaufman, T.C. (1996). Rescue of Drosophila labial null mutant by the chicken ortholog Hoxb-1 demonstrates that the function of Hox genes is phylogenetically conserved. Genes Dev. *10*, 176–184.

Mahony, S., Auron, P., and Benos, P. (Takis) V (2005). DNA Familial Binding Profiles Made Easy: Comparison of Various Motif Alignment and Clustering Strategies. PLoS Comput Biol *preprint*, e61.

Maier, T., Güell, M., and Serrano, L. (2009). Correlation of mRNA and protein in complex biological samples. FEBS Lett. *583*, 3966–3973.

Malin, J., Aniba, M.R., and Hannenhalli, S. (2013). Enhancer networks revealed by correlated DNAse hypersensitivity states of enhancers. Nucleic Acids Res. *41*, 6828–6838.

Markenscoff-Papadimitriou, E., Allen, W.E., Colquitt, B.M., Goh, T., Murphy, K.K., Monahan, K., Mosley, C.P., Ahituv, N., and Lomvardas, S. (2014). Enhancer Interaction Networks as a Means for Singular Olfactory Receptor Expression. Cell *159*, 543–557.

Martinez, G.J., and Rao, A. (2012). Immunology. Cooperative transcription factor complexes in control. Science *338*, 891–892.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional Regulatory Elements in the Human Genome. Annu. Rev. Genomics Hum. Genet. *7*, 29–59.

Matys, V. (2003). TRANSFAC(R): transcriptional regulation, from patterns to profiles. Nucleic Acids Res. *31*, 374–378.

Matys, V., Kel-Margoulis, O. V, Fricke, E., Liebich, I., Land, S., Barre-Dirrie, a, Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. *34*, D108–D110.

May, D., Blow, M.J., Kaplan, T., McCulley, D.J., Jensen, B.C., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., et al. (2011). Large-scale discovery of enhancers from human heart tissue. Nat. Genet. *44*, 89–93.

Mercer, T.R., and Mattick, J.S. (2013). Understanding the regulatory and transcriptional complexity of the genome through structure. Genome Res. *23*, 1081–1088.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., Wingett, S.W., Andrews, S., Grey, W., Ewels, P. a, et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. *47*, 598–606.

Mirny, L., Slutsky, M., Wunderlich, Z., Tafvizi, A., Leith, J., and Kosmrlj, A. (2009). How a protein searches for its site on DNA: the mechanism of facilitated diffusion. J. Phys. A Math. Theor. *42*, 434013.

Moggs, J.G., and Orphanides, G. (2001). Estrogen receptors: orchestrators of pleiotropic cellular responses. EMBO Rep. *2*, 775–781.

Montavon, T., and Duboule, D. (2012). Landscapes and archipelagos: spatial organization of gene regulation in vertebrates. Trends Cell Biol. *22*, 347–354.

Montavon, T., Soshnikova, N., Mascrez, B., Joye, E., Thevenet, L., Splinter, E., de Laat, W., Spitz, F., and Duboule, D. (2011). A regulatory archipelago controls Hox genes transcription in digits. Cell *147*, 1132–1145.

Montavon, T., Duboule, D., and B, P.T.R.S. (2013). Chromatin organization and global regulation of Hox gene clusters Chromatin organization and global regulation of Hox gene clusters.

Moses, A.M., Chiang, D.Y., Pollard, D. a, Iyer, V.N., and Eisen, M.B. (2004). MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol. *5*, R98.

Mukherjee, S., Erickson, H., and Bastia, D. (1988). Enhancer-origin interaction in plasmid R6K involves a DNA loop mediated by initiator protein. Cell *52*, 375–383.

Naranjo, S., Voesenek, K., de la Calle-Mustienes, E., Robert-Moreno, A., Kokotas, H., Grigoriadou, M., Economides, J., Van Camp, G., Hilgert, N., Moreno, F., et al. (2010). Multiple enhancers located in a 1-Mb region upstream of POU3F4 promote expression during inner ear development and may be required for hearing. Hum Genet *128*, 411–419.

Narlikar, L., Sakabe, N.J., Blanski, A.A., Arimura, F.E., Westlund, J.M., Nobrega, M.A., and Ovcharenko, I. (2010). Genome-wide discovery of human heart enhancers. Genome Res. *20*, 381–392.

Nasmyth, K., and Haering, C.H. (2009). Cohesin: Its Roles and Mechanisms.

Nelson, A.C., and Wardle, F.C. (2013). Conserved non-coding elements and cis regulation: actions speak louder than words. Development *140*, 1385–1395.

Neph, S., Stergachis, A.B., Reynolds, A., Sandstrom, R., Borenstein, E., and Stamatoyannopoulos, J. a (2012a). Circuitry and dynamics of human transcription factor regulatory networks. Cell *150*, 1274–1286.

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

Nobrega, M.A., Ovcharenko, I., Afzal, V., and Rubin, E.M. (2003). Scanning human gene deserts for long-range enhancers. Science *302*, 413.

Orphanides, G., Lagrange, T., and Reinberg, D. (1996). The general transcription factors of RNA polymerase II. Genes Dev. *10*, 2657–2683.

Pagel, M., and Johnstone, R.A. (1992). Variation across Species in the Size of the Nuclear Genome Supports the Junk-DNA Explanation for the C-Value Paradox. Proc. R. Soc. B Biol. Sci. *249*, 119–124.

Pan, Y., Tsai, C.-J., Ma, B., and Nussinov, R. (2010a). Mechanisms of transcription factor selectivity. Trends Genet. *26*, 75–83.

Pan, Y., Tsai, C.-J., Ma, B., and Nussinov, R. (2010b). Mechanisms of transcription factor selectivity. Trends Genet. *26*, 75–83.

Paramanathan, T., Reeves, D., Friedman, L.J., Kondev, J., and Gelles, J. (2014). A general mechanism for competitor-induced dissociation of molecular complexes. Nat. Commun. *5*, 5207.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*, 499–502.

Perry, M.W., Boettiger, A.N., Bothma, J.P., and Levine, M. (2010). Shadow enhancers foster robustness of Drosophila gastrulation. Curr. Biol. *20*, 1562–1567.

Pique-Regi, R., Degner, J.F., Pai, A. a, Gaffney, D.J., Gilad, Y., and Pritchard, J.K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res. *21*, 447–455.

Plank, J.L., and Dean, A. (2014). Enhancer Function: Mechanistic and Genome-Wide Insights Come Together. Mol. Cell *55*, 5–14.

Pombo, A., and Dillon, N. (2015). Three-dimensional genome architecture: players and mechanisms. Nat. Rev. Mol. Cell Biol. *16*, 245–257.

Ptashne, M. (1986). Gene regulation by proteins acting nearby and at a distance. Nature *322*, 697–701.

Ramos, A.I., and Barolo, S. (2013). Low-affinity transcription factor binding sites shape morphogen responses and enhancer evolution. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *368*, 20130018.

Rao, S.S.P.S.P., Huntley, M.H.H., Durand, N.C.C., Stamenova, E.K.K., Bochkov, I.D.D., Robinson, J.T.T., Sanborn, A.L.L., Machol, I., Omer, A.D.D., Lander, E.S.S., et al. (2014). A 3D

Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. Cell *159*, 1665–1680.

Reiner, A., Yekutieli, D., and Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics *19*, 368–375.

Rowan, S., Siggers, T., Lachke, S.A., Yue, Y., Bulyk, M.L., and Maas, R.L. (2010). Precise temporal control of the eye regulatory gene Pax6 via enhancer-binding site affinity. Genes Dev. *24*, 980–985.

Sandhu, K.S., Li, G., Poh, H.M., Quek, Y.L.K., Sia, Y.Y., Peh, S.Q., Mulawadi, F.H., Lim, J., Sikic, M., Menghi, F., et al. (2012). Large-scale functional organization of long-range chromatin interaction networks. Cell Rep. *2*, 1207–1219.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Schmidt, D., Schwalie, P.C., Ross-Innes, C.S., Hurtado, A., Brown, G.D., Carroll, J.S., Flicek, P., and Odom, D.T. (2010). A CTCF-independent role for cohesin in tissue-specific transcription. Genome Res. *20*, 578–588.

Schoenfelder, S., Clay, I., and Fraser, P. (2010a). The transcriptional interactome: gene expression in 3D. Curr. Opin. Genet. Dev. *20*, 127–133.

Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J. a, Umlauf, D., Dimitrova, D.S., et al. (2010b). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat. Genet. *42*, 53–61.

Schultz, S., Shields, G., and Steitz, T. (1991). Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. Science (80-.). *253*, 1001–1007.

Schwarzer, W., and Spitz, F. (2014). The architecture of gene expression: integrating dispersed cis-regulatory modules into coherent regulatory domains. Curr. Opin. Genet. Dev. *27*, 74–82.

Segal, E., Raveh-Sadka, T., Schroeder, M., Unnerstall, U., and Gaul, U. (2008). Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature *451*, 535–540.

Seitan, V.C., Faure, A.J., Zhan, Y., McCord, R.P., Lajoie, B.R., Ing-Simmons, E., Lenhard, B., Giorgetti, L., Heard, E., Fisher, A.G., et al. (2013). Cohesin-Based chromatin interactions enable regulated gene expression within preexisting architectural compartments. Genome Res. *23*, 2066–2077.

Sexton, T., and Cavalli, G. (2015). The Role of Chromosome Domains in Shaping the Functional Genome. Cell *160*, 1049–1059.

Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. Nat. Biotechnol. *30*, 521–530.

Sheffield, N.C., Thurman, R.E., Song, L., Safi, A., Stamatoyannopoulos, J. a, Lenhard, B., Crawford, G.E., and Furey, T.S. (2013). Patterns of regulatory activity across diverse human cell types predict tissue identity, transcription factor binding, and long-range interactions. Genome Res. *23*, 777–788.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V, et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. *15*, 1034–1050.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends Biochem. Sci. *39*, 381–399.

Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat. Genet. *45*, 1021–1028.

Song, S.-H., Kim, A., Ragoczy, T., Bender, M.A., Groudine, M., and Dean, A. (2010). Multiple functions of Ldb1 required for beta-globin activation during erythroid differentiation. Blood *116*, 2356–2364.

Spitz, F., and Furlong, E.E.M. (2012). Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. *13*, 613–626.

Stergachis, A.B., Neph, S., Sandstrom, R., Haugen, E., Reynolds, A.P., Zhang, M., Byron, R., Canfield, T., Stelhing-Sun, S., Lee, K., et al. (2014). Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature *515*, 365–370.

Stewart, A.J., and Plotkin, J.B. (2013). The evolution of complex gene regulation by low-specificity binding sites. Proc. Biol. Sci. *280*, 20131313.

Stewart, A.J., Hannenhalli, S., and Plotkin, J.B. (2012). Why transcription factor binding sites are ten nucleotides long. Genetics *192*, 973–985.

Stormo, G.D., and Zhao, Y. (2010). Determining the specificity of protein–DNA interactions. Nat. Rev. Genet. *11*, 751–760.

Stuart, J.M. (2003). A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science (80-.). *302*, 249–255.

Sutherland, H., and Bickmore, W. a (2009). Transcription factories: gene expression in unions? Nat. Rev. Genet. *10*, 457–466.

Taher, L., Smith, R.P., Kim, M.J., Ahituv, N., and Ovcharenko, I. (2013). Sequence signatures extracted from proximal promoters can be used to predict distal enhancers. Genome Biol. *14*, R117.

Tanay, A. (2006). Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res. *16*, 962–972.

Teif, V.B., and Rippe, K. (2012). Calculating transcription factor binding maps for chromatin. Brief. Bioinform. *13*, 187–201.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Tolhuis, B., Palstra, R., Splinter, E., Grosveld, F., and Laat, W. De (2002). Looping and Interaction between Hypersensitive Sites in the Active [illegible] -globin Locus. *10*, 1453–1465.

Ulianov, S. V, Khrameeva, E.E., Gavrilov, A. a, Flyamer, I.M., Kos, P., Mikhaleva, E. a, Penin, A. a, Logacheva, M.D., Imakaev, M. V, Chertovich, A., et al. (2015). Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. Genome Res. *7*.

Vakoc, C.R., Letting, D.L., Gheldof, N., Sawado, T., Bender, M. a, Groudine, M., Weiss, M.J., Dekker, J., and Blobel, G. a (2005). Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. Mol. Cell *17*, 453–462.

Vernimmen, D. (2014). Uncovering Enhancer Functions Using the α-Globin Locus. PLoS Genet. *10*, e1004668.

Vietri Rudan, M., and Hadjur, S. (2015). Genetic Tailors: CTCF and Cohesin Shape the Genome During Evolution. Trends Genet. *xx*, 1–10.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.

Vogel, C., and Marcotte, E.M. (2012). Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat. Rev. Genet. *13*, 227–232.

Wallace, J.A., and Felsenfeld, G. (2007). We gather together: insulators and genome organization. Curr. Opin. Genet. Dev. *17*, 400–407.

Wang, X., Bai, L., Bryant, G.O., and Ptashne, M. (2011). Nucleosomes and the accessibility problem. Trends Genet. *27*, 487–492.

Wasserman, W.W., and Fickett, J.W. (1998). Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol *278*, 167–181.

Wasson, T., and Hartemink, A.J. (2009). An ensemble model of competitive multi-factor binding of the genome. Genome Res. *19*, 2101–2112.

White, R.J. (2011). Transcription by RNA polymerase III: more complex than we thought. Nat Rev Genet *12*, 459–463.

White, M.A., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis -regulatory function of ChIP-seq peaks.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell *153*, 307–319.

Wilczyński, B., and Furlong, E.E.M. (2010). Dynamic CRM occupancy reflects a temporal map of developmental progression. Mol. Syst. Biol. *6*, 383.

Wunderlich, Z., and Mirny, L.A. (2008). Spatial effects on the speed and reliability of protein-DNA search. Nucleic Acids Res. *36*, 3570–3578.

Yáñez-Cuna, J.O., Dinh, H.Q., Kvon, E.Z., Shlyueva, D., and Stark, A. (2012). Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding. Genome Res. *22*, 2018–2030.

Yáñez-Cuna, J.O., Kvon, E.Z., and Stark, A. (2013). Deciphering the transcriptional cis-regulatory code. Trends Genet. *29*, 11–22.

Z Wunderlich, L.M. (2009). Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet. *25*, 429–434.

Zabet, N.R., and Adryan, B. (2012). A comprehensive computational model of facilitated diffusion in prokaryotes. Bioinformatics *28*, 1517–1524.

Zentner, G.E., Tesar, P.J., and Scacheri, P.C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome Res. *21*, 1273–1283.

Zhang, Y., Wong, C.-H., Birnbaum, R.Y., Li, G., Favaro, R., Ngan, C.Y., Lim, J., Tai, E., Poh, H.M., Wong, E., et al. (2013). Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. Nature.

Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M., and Furlong, E.E.M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature *462*, 65–70.