Quantifying Eukaryotic Gene Regulation in Hormone Response and Disease.

by

Christopher Michael Vockley

Department of Cell Biology
Duke University

Date:_____

Approved:

_____
Brigid Hogan, Co-Supervisor

_____
Timothy Reddy, Co-Supervisor

_____
Blanche Capel, Chair

_____
Gregory Crawford

_____
Eda Yildirim

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Cell Biology in the Graduate School
of Duke University

2016

ABSTRACT

Quantifying Eukaryotic Gene Regulation in Hormone Response and Disease.

by

Christopher Michael Vockley

Department of Cell Biology
Duke University

Date:_____

Approved:

_____
Brigid Hogan, Co-supervisor

_____
Timothy Reddy, Co-supervisor

_____
Blanche Capel, Chair

_____
Gregory Crawford

_____
Eda Yildirim

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the Department of
Cell Biology in the Graduate School of
Duke University

2016

Abstract

Quantifying the function of mammalian enhancers at the genome or population scale has been longstanding challenge in the field of gene regulation. Studies of individual enhancers have provided anecdotal evidence on which many foundational assumptions in the field are based. Genome-scale studies have revealed that the number of sites bound by a given transcription factor far outnumber the genes that the factor regulates. In this dissertation we describe a new method, chromatin immune-enriched reporter assays (ChIP-reporters), and use that approach to comprehensively test the enhancer activity of genomic loci bound by the glucocorticoid receptor (GR). Integrative genomics analyses of our ChIP-reporter data revealed an unexpected mechanism of glucocorticoid (GC)-induced gene regulation. In that mechanism, only the minority of GR bound sites acts as GC-inducible enhancers. Many non-GC-inducible GR binding sites interact with GC-induced sites via chromatin looping. These interactions can increase the activity of GC-induced enhancers. Finally, we describe a method that enables the detection and characterization of the functional effects of non-coding genetic variation on enhancer activity at the population scale. Taken together, these studies yield both mechanistic and genetic evidence that provides context that informs the understanding of the effects of multiple enhancer variants on gene expression.

Dedication

This dissertation is dedicated to my family. I profoundly appreciate the support of my mother, Kimberly E. Vockley, who has been a perpetual source of optimism even in difficult times. My father, Dr. Joseph G. Vockley, exposed me to biology at an exceptionally young age and dedicated his dissertation on the characterization on gene regulatory elements to me. Our ability to talk about genomics the way that some fathers and sons talk about baseball continues to be a highlight of my intellectual life. I have received tremendous support from my siblings and members of my extended family, who have provided encouragement at every step. My grandparents, George and Mary Vockley provided an example to aspire to and believed in me without conditions. Their memory continues to be a source of strength and inspiration.

Finally, my fiancé, Rebecca Gifford has been the stabilizing force in my life that has made my success in graduate school possible. She is among the most dedicated, kind, and intelligent people I have met. I look forward to our future with gleeful enthusiasm. The best is yet to come.

Contents

Abstract iv

List of Tables ix

List of Figures x

Acknowledgements xiv

1. Introduction 1

1.1 An Introduction to Eukaryotic Gene Regulation 1

1.2 Enhancers are positively acting gene regulatory elements 1

1.3 Enhancers are bound by transcription factors to regulate gene expression 3

1.4 Cis-regulatory modules encode multiple TF binding sites and activate genes cooperatively 4

1.5 Chromatin packaging influences the activity of gene regulatory elements 7

1.6 CRMs contribute to diverse gene expression responses via diverse mechanisms 10

1.7 CRMs can control gene expression from distal regulatory sites 12

1.8 Long-range enhancer CRMs in Locus Control Regions can control multiple genes, concordantly or discordantly 13

1.9 Enhancer CRMs can interact with one another and with promoters via chromatin looping 16

1.10 Chromatin interactions are coordinated within topologically associated chromatin domains 18

1.11 Perturbed gene regulation resulting in Mendelian disease 20

1.12 Variation in regulatory elements as a contributor to complex phenotypes 23

1.13 An integrated view of gene regulation 27

2. GR interaction modules direct the gene expression response to glucocorticoids 29

2.1 Introduction 29

2.2 Results 32

2.2.1 Quantifying GC-induced regulatory element activity 32

2.2.2 DEX-induced GBSs are direct binding sites 37

2.2.3 Cell type-specific DEX-induced enhancers are encoded by direct GBSs 39

2.2.4 Remodeling of GC-induced sites in the endogenous epigenome 41

2.2.5 GBSs cluster in the genome 45

2.2.6 CTCF is depleted within GBS clusters 48

2.2.7 Tethered GBSs cluster around direct GBSs in the genome 50

2.2.8 Direct GBSs recruit AP-1 binding to genomic sites that lack AP-1 recognition motifs 53

2.2.9 Epistatic interactions between GR and AP-1 modulate DEX-responsive regulatory activity 56

2.2.10 Distal binding cluster interactions are a general mechanism of gene regulation 59

2.3 Discussion 61

2.4 Experimental Procedures 65

2.5 Supporting online materials 88

3. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort 91

3.1 Introduction 91

3.2 Results 94

3.2.1 Population-scale reporter assay approach 94

3.2.2 Targeted sequencing of candidate regulatory elements from a GWAS population 96

3.2.3 Quantifying the effects of noncoding variation in a GWAS population 98

3.2.4 Regulatory variants are enriched in active enhancers 102

3.2.5 Effects of haplotypes on regulatory element activity 103

3.2.6 Fine mapping genetic associations with phenotypes 105

3.2.7 Identifying candidate mechanisms of regulatory element activity 107

3.3 Discussion 109

3.4 Methods 112

3.5 Supporting online materials 121

4. Conclusion and Future Outlook 122

Appendix A 129

Appendix B 147

References 159

Biography 184

List of Tables

Table 1: Mendilian position-dependent diseases. Adapted from Kleinjan and Lettice, 2008..... 22

Table 2: Vockley et al., 2015 Supplemental Table 1: Transition:Transversion ratios in 1000 Genomes and Custom Amplicon Sequencing..... 156

Table 3: Vockley et al., 2015 Supplemental Table 2: Proportion of assayed variation in Population STARR-seq..... 156

Table 4: Vockley et al., 2015 Supplemental Table 3: STARR-seq Primers..... 157

Table 5: Vockley et al., 2015 Supplemental Table 4: Luciferase Validation Primers..... 158

List of Figures

Figure 1: Transcription factors bind to degenerate DNA motifs. 4

Figure 2: Regulatory grammar and flexible billboard models of CRM function..... 7

Figure 3: Human β-globin locus. 14

Figure 4: Modulating distal enhancers with dCas9-based synthetic TFs. 16

Figure 5: Topologically Associated Domain Architecture. 19

Figure 6: Chromosomal rearrangements can perturb gene regulation. 21

Figure 7: Genomic evidence of regulatory variation. 24

Figure 8: The multiple enhancer variant hypothesis. 25

Figure 9: ChIP-reporter experimental design. 32

Figure 10: ChIP-reporter data and cross-platform validation. 34

Figure 11: DEX-induced GBSs compared to GR ChIP signal and location of non-DEX-induced sites. 36

Figure 12: DEX responsiveness is predicted by the presence of a GRE. 37

Figure 13: ChIP-exo footprints of DEX-induced and non-DEX-induced GBSs. 38

Figure 14: Cell-type specific DEX-induced GBSs are predicted by the presence of a GRE. 40

Figure 15: Epigenomic remodeling of GBSs after DEX exposure. 42

Figure 16: DEX-induced reporter activity as a function of H3K27 acetylation state..... 43

Figure 17: Epigenomics data improves the performance of models that predict DEX-induced enhancer activity. 45

Figure 18: GBSs bound at low dose DEX treatment occur in coordinated clusters. 46

Figure 19: GBS clusters are less likely to span CTCF insulated domains than expected by chance. 49

Figure 20: DEX induced enhancer function across a GBS cluster at the *NFKBIA* locus. .. 50

Figure 21: GBS clusters consist of DEX-induced GBSs surrounded by tethered GBSs..... 52

Figure 22: Comparisons of various and sequence genomic features among JUND binding sites that are gained, maintained, or lost with DEX treatment..... 55

Figure 23: Evidence for shadows of JUND binding at GBSs. 56

Figure 24: AP-1 and GR can co-activate reporter gene expression beyond the range of direct interactions..... 58

Figure 25: JUN and FOXA1 co-bound ER binding sites are closer to ERE encoded ER binding sites than expected by chance..... 60

Figure 26: ER binding site distal chromatin interactions as a function of ERE motif match..... 61

Figure 27: A revised model of GR-mediated enhancer function. 64

Figure 28: Population STARR-seq assay schematic. 95

Figure 29: Candidate regulatory sites. 97

Figure 30: Population STARR-seq reporter libraries are representative of population diversity..... 100

Figure 31: Identifying regulatory variants. 101

Figure 32: Enhancer activity scores for regulatory variants. 103

Figure 33: Histogram of number of SNPs per assayed element..... 104

Figure 34: Manhattan plot of eQTLs for the long non-coding RNA *LINC00881*. 106

Figure 35: *LINC00881* expression and H3K27 acetylation state as a function of genetic variation..... 107

Figure 36: Regulatory variant disruption of a TEAD4 binding site..... 109

Figure 37: Fragment diversity and composition of GR ChIP-reporter libraries. 129

Figure 38: Fragment size, GBS coverage and DEX-induced activity of GR ChIP-reporters 130

Figure 39:Negative binomial model of sequencing depth and variance, DEX-induced activity as a function of fragment size, assay control and validation data. 131

Figure 40: Nuclear localization of reporter plasmids is not biased by GCs. 132

Figure 41: Addition of GREs increases DEX-induced reporter gene expression from sites bound by the GR but not induced in ChIP-reporter assays. 133

Figure 42: The presence of GRE motifs but not co-factor motifs predicts DEX-induced reporter activity. 134

Figure 43: GR motif strength vs. reporter density. 135

Figure 44: Additive linear regression model of activity in ChIP-reporter assays 136

Figure 45: Additive linear regression model of DEX-responsive activity in ChIP-reporter assays. 137

Figure 46: GR ChIP-exo supporting evidence 138

Figure 47: Epigenetic remodeling at ChIP-reporter assayed sites. 139

Figure 48: GBSs cluster in the genome. 140

Figure 49: Fraction of GBS clusters with at least one DEX-responsive GBS. 141

Figure 50: Negative control motif analysis for JUND-GR interaction experiments. 142

Figure 51: Schematic of GRE/AP-1 epistasis experiments. 143

Figure 52: Functional data from GR/AP-1 epistasis experiments. 144

Figure 53: GR/AP-1 motif spacing experiments. 145

Figure 54: Correlation between reporter activity and endogenous gene regulation. 146

Figure 55: Distribution of TruSeq Custom Amplicon sequencing coverage for 95 individuals. 147

Figure 56: Distribution of allele frequencies. 148

Figure 57: Distribution of median coverage per amplicon in reporter input libraries. .. 149

Figure 58: Distribution of reporter RNA-seq coverage. 150

Figure 59: Allele ratios in plasmid vs. RNA-seq reporter libraries. 151

Figure 60: Allele frequency in reporter libraries vs. allele frequency observed in the population. 152

Figure 61: Minor allele frequency vs. variant effect size. 153

Figure 62: SNP effects vs haplotype effects. 154

Figure 63: eQTLs associated with the expression of *LINC00881*. 155

Acknowledgements

The research contained within this dissertation would not have been possible without the guidance and expert mentorship of my supervisors, Dr. Tim Reddy and Dr. Brigid Hogan. You have provided me with fabulous scientific training and many opportunities, for which I will be eternally grateful. I will spend my career aspiring to emulate your innumerable strengths.

I have received guidance and support from many members of the Duke faculty, including Dr. Blanche Capel, Dr. Greg Crawford, and Dr. Eda Yildirim, who have mentored me as members of my dissertation committee.

I owe a debt of gratitude to my many collaborators, who have fostered my scientific development at Duke. The research presented in this dissertation would be greatly diminished without the scientific contributions of Tony D'Ippolito, Ian McDowell, and Karl Guo. Dewran Koçak has been a friend, collaborator, and sounding board for matters scientific and personal.

My career would not be the same with out the guidance of those who mentored at early stages, especially Dr. Owen Wood, Dr. Francis Collins, and Dr. Richard Maas.

Finally, my career has been most shaped by my father, Dr. Joseph Vockley, who gave me my first job working in a functional genomics lab at age 15. Your wisdom has never failed me.

1. Introduction

1.1 An Introduction to Eukaryotic Gene Regulation

The diploid human genome consists of more than six billion nucleotide bases. The sequence of bases encodes all the information required for cellular homeostasis, the development and organization of complex tissues, and responses to environmental stimuli. It is estimated that three percent of the nucleotides in the genome contain sequences that are translated to proteins. The remaining 97 percent includes elements that regulate the expression of these proteins in various cellular and environmental contexts. Understanding the mechanisms that underlie the activity of non-coding regions of the genome is one of the premier challenges that define the field of functional genomics. Gene regulation studies have revealed that the non-coding genome contains a highly coordinated set of instructions that direct spatiotemporal patterns of gene expression. Understanding these instructions and learning how they are perturbed in the context of disease has been a long-standing challenge in genome biology.

1.2 Enhancers are positively acting gene regulatory elements

The biological task of controlling gene expression, like many other challenges in biology, is met by the coordinated activity and organization of functional subunits. The fundamental subunits of gene regulation are *cis*-regulatory elements and trans-acting transcriptional regulators. DNA sequences known as *cis*-regulatory elements control

gene expression in human cells by recruiting the binding of *trans-* acting proteins known as transcription factors (TFs). Regulatory elements can be classified into different types based on their activity(ENCODE, 2012); three common classifications are promoters, enhancers, and silencers. Promoters are located at the beginning of genes and recruit the enzymatic complexes required for transcribing the downstream gene into RNA. The enzymatic complexes, in turn, include the proteins that together form the transcription pre-initiation complex. The activity of promoters can decreased by silencer elements and increased by enhancer elements. Since positively-acting regulatory elements are the focus of this dissertation I have focused this primer on the biology of enhancers and promoters. Enhancers are bound by complexes of many different proteins and interact physically with protein complexes bound at other regulatory elements in the genome. To describe that complexity, I will first focus on how individual transcription factors bind to enhancers. Then, I will describe how complexes of proteins bind. Finally, I will discuss higher-order interactions between regulatory elements.

Enhancers are DNA elements that induce the expression of their target gene or genes by recruiting TFs and associated machinery that increase transcription from that gene's promoter. The earliest evidence of enhancers came from studies of a 72 nucleotide DNA element contained within the Simian vacuolating virus 40 (SV40) genome (Banerji et al., 1981; Benoist and Chambon, 1981). Unlike promoters, the SV40 enhancer increased gene expression independent of orientation when cloned into a vector containing a

mammalian gene that was introduced into mammalian cells. Based on those initial studies, an enhancer was defined as a DNA element that can increase expression of a reporter gene independent of orientation. This basic operational definition of an enhancer remains in common use today. Enhancers can be located upstream, downstream, and within genes and are thus flexible in terms of their location in the genome (ENCODE, 2012).

1.3 Enhancers are bound by transcription factors to regulate gene expression

Although enhancers that have been isolated from the genome tend to be capable of activating a wide variety of promoters in functional assays, the regulation of RNA transcription is highly controlled. This is accomplished in multiple ways. One mechanism that governs enhancer activity is the frequency of interactions between transcription factors and the enhancers that they target. Recent estimates approximate that there are nearly 1,400 sequence-specific TFs in the human genome (**Figure 1A**) (Vaquerizas et al., 2009). Such TFs contain DNA binding domains that allow them bind to an enhancer in a DNA sequence-directed manner. These sites, called TF binding motifs, are often on the scale of 5-25 nucleotides in length (Sandelin et al., 2004). TF binding motifs are frequently degenerate. Thus, motifs are described by the frequency that a given base is present at a given location within a binding site (**Figure 1B**). TF binding affinity has been proposed to correlate with the significance of the statistical match of a site to the consensus motif (Stormo, 1990). Thus, the affinity of a TF for a

given enhancer is influenced by the primary sequence of the TF binding site that the enhancer contains. Concordantly, many TFs are differentially expressed between tissues. The concentration of a TF in the nucleus directly impacts the number of sites that the TF binds (Reddy et al., 2012a). Together, these attributes enable increased cell-type-specific and gene-specific control of gene expression.

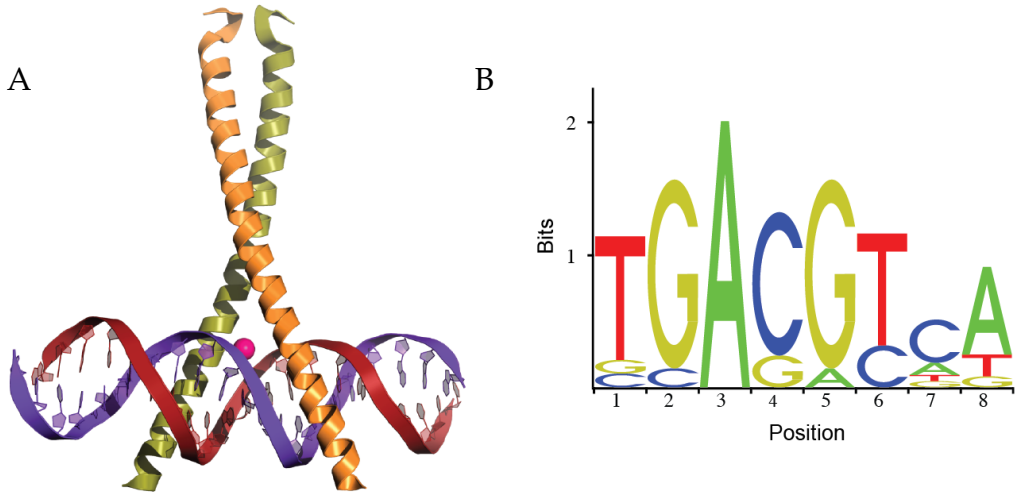

Figure 1: Transcription factors bind to degenerate DNA motifs.

(A) The ribbon structure of CREB1, a leucine zipper transcription factor, bound to DNA (Yikrazuul, 2009). **(B)** A sequence motif logo generated from the position weight matrix that describes the frequency of nucleotides in the CREB1 consensus binding motif (Sandelin et al., 2004).

1.4 Cis-regulatory modules encode multiple TF binding sites and activate genes cooperatively

Enhancers do not act alone in the genome. First, they must interact with promoters to control the expression of target genes. Second, they are also known to

interact with other enhancers that modulate their activity. Regions that contain adjacent TF binding sites that together control an enhancer are called *cis*-regulatory modules (CRMs) (Butz and Hoppe-Seyler, 1993). CRMs have been defined on varying scales. Current literature defines CRMs on the sale of hundreds of nucleotides to 3,000 nucleotides in length (ENCODE, 2012; Gotea et al., 2010). Early evidence for combinatory regulation of gene expression was observed at enhancers that govern the expression of the Serum Amyloid A gene (*SAA1*), which is induced more than 200-fold in response to inflammation (Li and Liao, 1992). The TFs C/EBP and NFκB bind to adjacent regulatory elements that control *SAA1* transcription in response to inflammatory cytokines. Although each DNA element is capable of modestly increasing gene expression independently of the other, the combined effects of both DNA elements synergistically amplify the transcriptional response to cytokine exposure. Subsequent studies have demonstrated that both homotypic and heterotypic *cis*-regulatory modules are prevalent features, occupying nearly two percent of the human genome (Gotea et al., 2010).

There are currently two leading hypotheses about how combinations of TFs interact to regulate gene expression (**Figure 2**). The "regulatory grammar hypothesis" proposes that although individual elements within a CRM can act in an orientation independent manner, in order to act in concert elements must follow specific "grammar rules". In this model, a specific gene regulatory element syntax, involving element

orientation, ordering, and density, is required for a CRM to act as a functional enhancer. These ideas were first formalized in studies of microbial gene regulation that employed linguistic principals (Collado-Vides, 1991, 1992).

A "billboard model" was proposed as an alternative to syntax-dependent languages of gene regulation. The billboard model of CRM organization posits that the specific orientation of elements is less relevant to CRM function than the ability to recruit the appropriate compliment of non-motif driven regulatory factors to a locus (Kulkarni and Arnosti, 2003; Rastegar et al., 2008).

The grammar and billboard models were tested in a comprehensive way in a study that quantified the regulatory effects of nearly 5,000 combinations of synthetic CRMs in which binding site heteromer composition, orientation and spacing were varied (Smith et al., 2013). This study confirmed and refined the billboard model by demonstrating that while orientation and spacing had little effect on regulatory function, the heterogeneity of a CRM correlates with enhancer activity. Heterotypic CRMs consisting of binding motifs for more than one TF had greater activity than homotypic CRMs with the same number of binding motifs for a single TF. Together, these studies demonstrate that combinatory activity of multiple TF-bound enhancers within a CRM is a substantial source of the heterogeneity and specificity of gene expression.

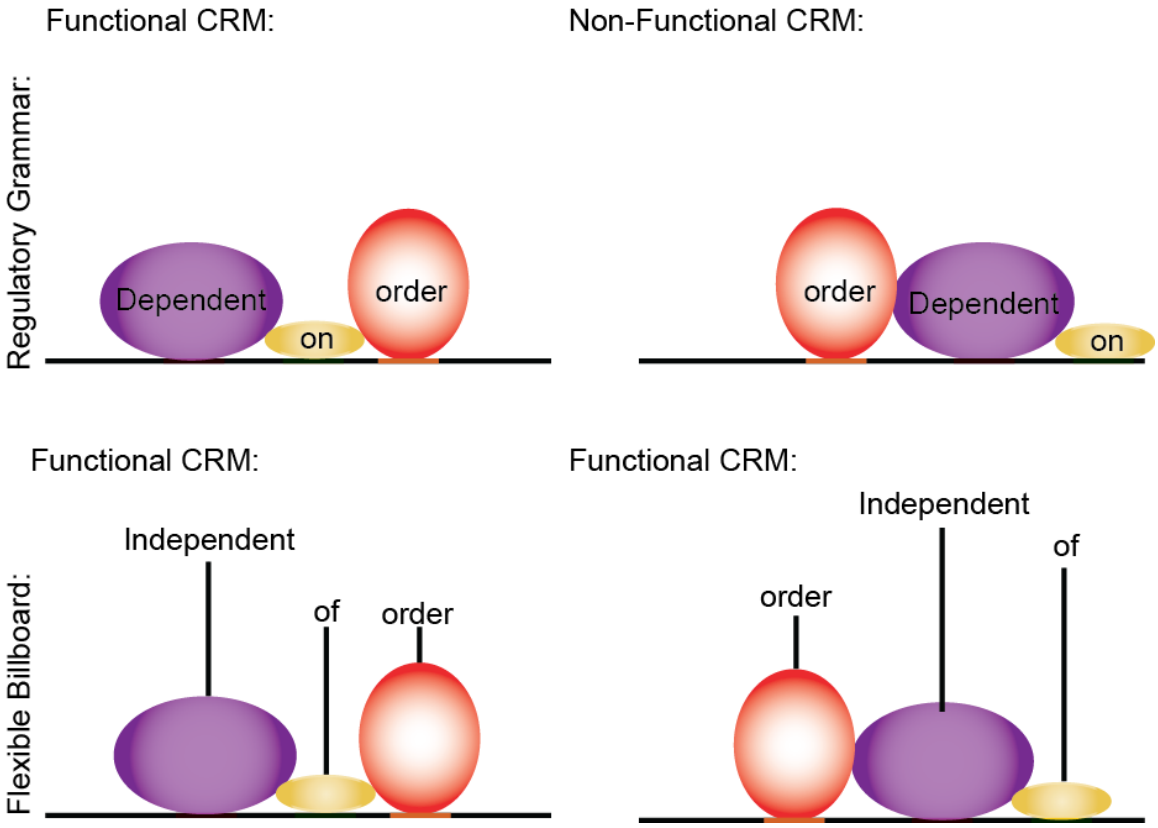

Figure 2: Regulatory grammar and flexible billboard models of CRM function.

The regulatory grammar model (top) hypothesizes that interactions between TFs bound at a CRM relies on specific protein binding syntax. A series of proteins must be arranged in a certain order for a regulatory operation to be specified (top left). When the protein syntax is changed enhancer function is disrupted (top right). The flexible billboard model (bottom) proposes that enhancer function within a CRM does not rely on a specific protein binding order.

1.5 Chromatin packaging influences the activity of gene regulatory elements

Eukaryotic genomes are hierarchically packaged into a nucleic acid and protein complex called chromatin (Levene, 1903). The fundamental subunit of chromatin, the nucleosome, consists of approximately 147 nucleotides of DNA wrapped around an

octamer protein core (Luger et al., 1997). This core consists of two copies of each of four histone proteins. Nucleosomes in turn are linked by histone H1 into a solenoid conformation 30 nM in diameter (Thoma et al., 1979). This fiber is then assembled into successively ordered chromosome structures.

Hierarchical chromatin organization serves both structural and gene regulatory roles. Structurally, chromatin solves the basic challenge of organizing and compacting a two-meter genome into the volume of a nucleus typically ten nanometers in diameter while maintaining an order that permits chromosomal segregation during cell division. Core histone proteins have core domains and tail domains. Core domains interact and scaffold the nucleosome DNA loop. Meanwhile, histone tail domains are localized on the exposed face of the nucleosome and undergo extensive covalent modification (Zheng and Hayes, 2003). These modifications have been correlated with regulatory element activity in the genome (Koch et al., 2007). Methylation of histone 3 lysine 4 and lysine 36 has been correlated with transcriptional activation (Barski et al., 2007). Meanwhile, methylation of histone 3 lysine 27 is associated with facultative gene repression and methylation of histone 3 lysine 9 is associated with constitutive repression (Bernstein et al., 2006). Histone tail acetylation is generally correlated with gene activation (Allfrey et al., 1964). Acetylation of histone 3 lysine 27 has been the focus of intense study and is correlated with enhancer activity (Creyghton et al., 2010). Proteins catalyze the addition and removal of histone modifications with varying

degrees of substrate specificity. For example, KDM6B (JMJD3) specifically removes methyl groups from histone 3 lysine 27 (Hong et al., 2007). However, P300, which catalyzes the addition of acetyl groups to histone 3 lysine 27 is a general acetyl transferase with characterized roles in the acetylation of many proteins (Imhof et al., 1997). *In vitro* biochemical experiments have shown that some histone modifiers can modify purified non-histone proteins. Additional investigation will be required to determine which, if any, histone modifying enzymes function exclusively to modify histones.

The observation that specific histone modifications are enriched among distinct gene regulatory environments has suggested a "histone code hypothesis"(Strahl and Allis, 2000). This hypothesis proposes that the compliment of histone tail modifications present on a nucleosome recruits factors that lack DNA binding domains and potentiates gene regulatory outcomes. This model is supported by evidence that certain modifications can be "read" by specific proteins. For example, BRD4 binds acetylated histones and recruits the positive transcription elongation factor B, which in turn phosphorylates the carboxy terminus of RNA polymerase II, initiating transcription (Itzen et al., 2014). However, the mechanistic relationship between histone modification and gene regulation has been the subject of substantial debate. The degree to which histone modifications cause gene regulatory events, or are the consequence of secondary catalytic activity of proteins that are performing other co-factor functions remains

unclear. For example, the Krupple-associated Box (KRAB) domain containing transcription factors bind to sequence specific sites and recruit the histone 3 lysine 9 (H3K9) methyltransferase SETDB1 and HP1. Consequently, the occupied locus becomes enriched for H3K9 methylation (Schultz et al., 2002). HP1 has been proposed to have a mechanistic role in chromatin condensation and gene silencing. KRAB associated protein 1 (KAP1) contains a HP1 binding domain, and deletion of this domain results in loss of gene silencing function (Nielsen et al., 1999). Meanwhile, studies on synthesized nucleosomes have demonstrated that artificial methylation of H3K9 is required for the association of HP1 with nucleosomes *in vitro* (Hiragami-Hamada et al., 2016). Future studies will be required to investigate this discrepancy and to further clarify the mechanistic relationship between histone modifications and gene regulation.

1.6 CRMs contribute to diverse gene expression responses via diverse mechanisms

Several mechanisms have been proposed to explain the combinatory effects of multi-TF CRMs on enhancer function. Evidence first presented in the study of the TFs SP1 and OTF-1 suggested that protein-protein interactions between TFs could enhance TF binding to adjacent sites in the same CRM (Janson and Pettersson, 1990). These theories were built upon by studies of the relationships between TF binding and chromatin architecture. TF activity and chromatin condensation have long been proposed to act in opposition to control gene regulation (Comings, 1967). Early studies in which DNase foot printing was used to identify transcription factor binding sites

provided evidence that TF-bound regulatory elements reside in chromatin domains that are more accessible to transcription factors (Galas and Schmitz, 1978). The binding of certain TFs displaces nucleosomes. Likewise, heterochromatic regions of the genome are both less accessible to transcription factors and less likely to contain transcribed genes. These observations support a model in which TFs that are capable of increasing the accessibility of adjacent TF binding sites act as "pioneer factors" that potentiate the binding of other TFs in the same CRM (Cirillo et al., 2002). Along with increased accessibility, active CRMs are frequently enriched for nucleosome subunits with covalent histone modifications that are associated with enhancer function. A synthesis of the direct protein-protein interaction model and the pioneer factor model likely explains this observation. Many TFs recruit non-DNA-sequence-specific protein co-factors that are capable of catalyzing the addition or removal of covalent modifications from histone tails. For example, AP-1 family transcription factors both increase chromatin accessibility (Biddie et al., 2011) and recruit P300 to the genome (Brockmann et al., 1999). P300 is a co-factor that catalyzes the addition of an acetyl group to lysine 27 of histone 3—a modification associated with active enhancers. Thus, enhancers that are occupied by AP-1 have both increased chromatin accessibility and are enriched for H3K27 acetylation. The pioneer factor model is confounded by the observation that many factors that require pioneer factor activity at sites that contain a poor match to the factor's binding motif are capable of acting as pioneer factors themselves at sites that

contain highly significant motif matches (John et al., 2011). In chapter two of this dissertation, we provide evidence that supports a model in which sites with a weak TF binding motif, previously proposed to rely on pioneer factor activity, are in fact more likely the result of long range TF-TF interactions. The resulting TF binding observed at such sites can be explained by distal protein tethering rather than by *bona fide* motif driven binding at weak motif matches that have been made accessible by a pioneer factor. Additional studies will be required to further clarify the molecular biology of chromatin condensation-mediated enhancer repression. Together these observations demonstrate that TFs bound at the same CRM cooperate to activate gene expression through diverse mechanisms that include interactions with each other, with the genome, with co-factors, and with the local chromatin environment.

1.7 CRMs can control gene expression from distal regulatory sites

Although many enhancers are capable of recruiting the necessary cellular machinery to initiate transcription, in many cases the genes that enhancers target are located from tens of kilobases to more than a megabase away (ENCODE, 2012). One particularly well-studied example of a distal enhancer is a non-coding functional element contained within an intron of the LMBR1 gene that controls the expression of the gene Sonic Hedgehog (*SHH*). This enhancer is one megabase away from the promoter that it targets. Mutations within this enhancer result in pre-axial polydactyly in humans (Lettice et al., 2002), while deletion of the element in the developing mouse

autopod results in complete limb bud truncation (Sagai et al., 2005). The function of this element is widely conserved across species; a *de novo* mutation of this enhancer has been characterized as the source of the polydactyl phenotype of felines inhabiting Key West, Florida (Lettice et al., 2008). Subsequent studies have demonstrated that this specific enhancer-promoter interaction occurs via chromatin looping (Amano et al., 2009).

TF binding sites frequently act distally. This has been observed in genome-scale studies that map TF occupancy and mRNA expression after perturbing inducible transcriptions factors. One study, in which GRHL2 binding sites were mapped and GRHL function was perturbed using a dominant negative form of GRHL2, revealed that GRHL sites are enriched within forty kilobases of GRHL-regulated genes (Gao et al., 2013). These results were mirrored in our study of the glucocorticoid receptor, a ligand inducible TF. Functionally validated corticosteroid-inducible GR bound enhancers have a median distance of 130 kilobases from the nearest corticosteroid-induced gene and non-corticosteroid-inducible GR binding sites have a median distance of 160 kilobases from the nearest induced gene. Together these studies demonstrate that many TF-bound enhancers are located tens of kilobases or further from the promoter that they target.

1.8 Long-range enhancer CRMs in Locus Control Regions can control multiple genes, concordantly or discordantly

The enhancers that govern the expression of the β-globin genes have emerged as a canonical example of long-range gene regulation (**Figure 3**). The human globin locus is a region containing globin genes that are differentially expressed in a developmental

stage and tissue specific manner. The collection of regulatory elements that control the β-globin locus was first localized in genetic mapping studies in the search for causal alleles for Dutch β-thalassemia. Patients afflicted with Dutch β-thalassemia are frequently heterozygous for a genetic deletion approximately one hundred kilobases in length, but have two intact copies of the β-globin gene (Kioussis et al., 1983). This observation suggested that the regulatory features responsible for β-globin deficiency in these individuals were located within this upstream deleted region rather than in the gene-proximal promoter.

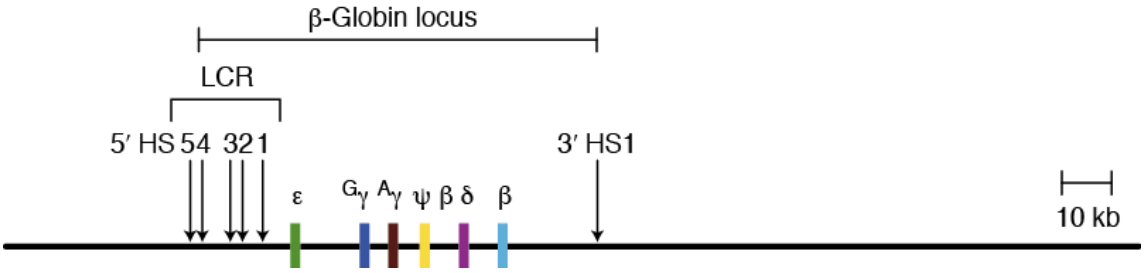

Figure 3: Human β-globin locus.

A schematic of the human β-globin locus. The locus control region is responsible for coordinating the expression of genes in the β-globin gene cluster. Adapted from Harju et al. 2003.

Initial studies of globin regulation using transgenic mice showed that the human β-globin gene and adjacent non-coding sequences could reconstitute partial expression of a human β-globin, but could not reconstitute normal spatiotemporal expression patterns (Kollias et al., 1986; Magram et al., 1985; Townes et al., 1985). A key advance

was the generation of a transgenic mouse that contained a transgene encoding both the human β-globin gene itself and a 50 kilobase region adjacent to the β-globin coding sequence (Grosveld et al., 1987). That study demonstrated that a 17 kilobase region more than 20 kilobases away is responsible for full recapitulation of endogenous patterns of globin gene expression. Additional analyses revealed that a single set of CRMs (the β-globin LCR) control the differential expression of genes encoding fetal and adult-specific β-globin subunits (HBG1 and HBG2). The observation that an enhancer CRM could be shared between multiple genes with mutually exclusive context-dependent expression further demonstrates the flexibility and adaptability of *cis*-regulatory elements.

Recently, nuclease deficient Cas9-based synthetic transcription factors have enabled the targeting of the transcriptional machinery to distal enhancer loci. These technologies include the targeting of synthetic transcriptional activators, transcriptional co-activators and transcriptional co-repressors. Evidence first from the *Grin2C* locus demonstrated that targeting dCas9-VP64 to a distal enhancer could activate gene expression (Frank et al., 2015). Subsequent studies of the β-globin locus demonstrated that targeting enhancers with either the catalytic core of P300 or the KRAB repressive domain can activate and inactivate distal globin genes, respectively (**Figure 4**) (Hilton et al., 2015; Thakore et al., 2015).

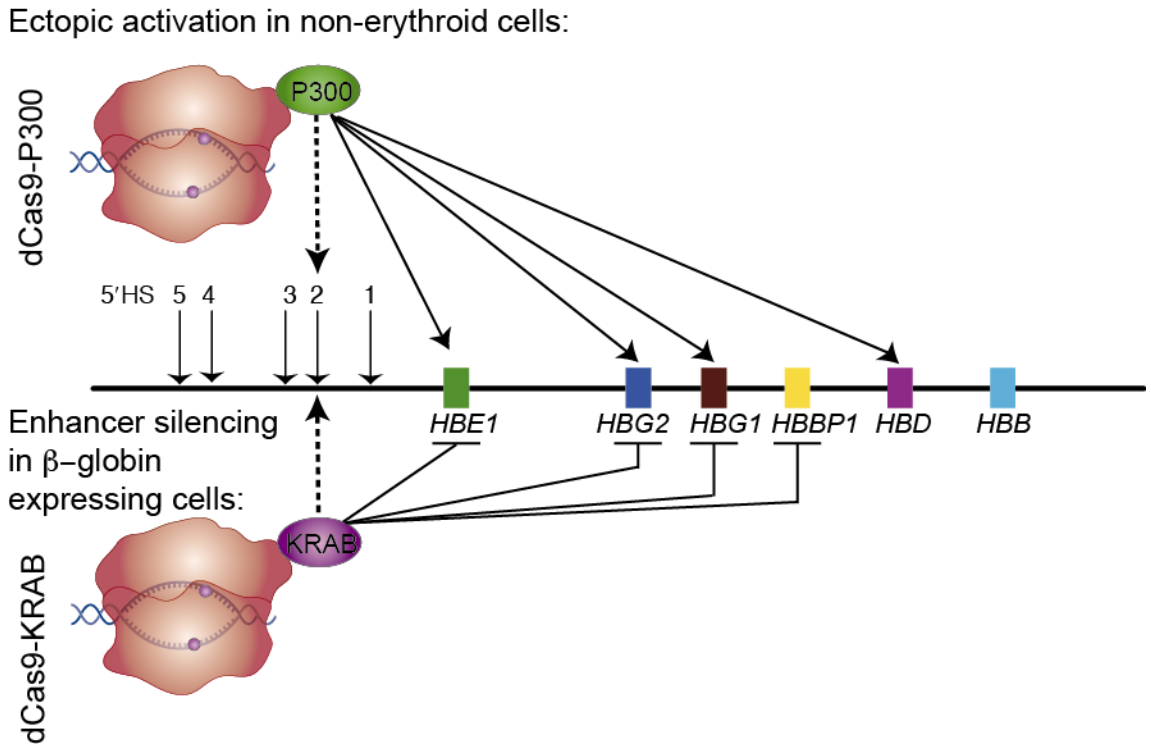

Figure 4: Modulating distal enhancers with dCas9-based synthetic TFs.

A schematic of the β-globin locus indicating distal genes that were activated by nuclease deficient Cas9 fused to the catalytic domain of P300 (top) or inhibited by nuclease deficient Cas9 fused to the KRAB domain. Adapted from Hilton et al., Takore et al., Dominguez et al., and Harju et al.

1.9 Enhancer CRMs can interact with one another and with promoters via chromatin looping

Substantial evidence demonstrates that a common mechanism of long-range enhancer function relies on the formation of chromatin loops. These loops allow for direct interactions between enhancers and promoters. This observation was first made in the context of prokaryotic gene regulation (Ptashne, 1986). Distal interactions have been directly observed and quantified on a locus specific basis, and at the genome-scale.

While chromatin can be visualized via electron microscopy and fluorescence in situ hybridization, high resolution mapping of eukaryotic chromatin interactions was enabled by chromatin conformation capture (3C). This technology was first used to map changes distal interactions between loci in the yeast cell cycle (Dekker et al., 2002). The first definitive evidence of facultative chromatin looping between enhancers and promoters in mammalian gene regulation was obtained in the context of the β-globin LCR (Tolhuis et al., 2002). In that study, 3C was used to quantify the frequency of interactions between individual elements in the β-globin LCR and promoters in the β-globin cluster. Enhancer-promoter interactions were specifically enriched in the embryonic liver, where the globin cluster is expressed. Meanwhile, negative control samples from the developing brain had reduced interaction frequencies. Of particular significance, this study also demonstrated that individual regulatory elements (rather than just enhancer-promoter pairs) within the β-globin LCR can interact with one another in a tissue specific way. A follow up study used transgenic mouse models to demonstrate that deletion of individual sites from the β-globin locus results in decreased interactions between remaining sites (Fang et al., 2007).

Genomic evidence has demonstrated that dynamic distally acting enhancer-promoter and enhancer-enhancer interactions are broadly important for gene regulation. Studies of the estrogen receptor (ER) and glucocorticoid receptor (GR), two ligand-inducible transcription factors, have demonstrated that TF binding sites interact with

one another and the promoter of regulated genes to control expression (Fullwood et al., 2009; Kuznetsova et al., 2015). Understanding the functional implications of those interactions has been a long-standing challenge in genomics. In Chapter 2 of this dissertation, we provide evidence suggesting that enhancers controlled by the GR interact via heterotypic multicomponent clusters. In this model, the GR nucleates clusters of distal chromatin interactions with sites bound by AP-1, a potent co-regulator, to tune gene expression. Evidence from the estrogen receptor suggests that this may be a general mechanism.

1.10 Chromatin interactions are coordinated within topologically associated chromatin domains

Dynamic interactions between promoters and enhancers are coordinated within chromatin structures called topologically associated domains (TADs) (Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012). These domains demarcate loops of chromatin into functionally coordinated sectors. At TAD boundaries, CTCF, a transcription factor with predominantly negative activity on transcription, binds to the DNA insulating the genomic interval within the TAD from external regulatory activity and isolating genes and enhancers within TADs from the rest of the genome (**Figure 5**).

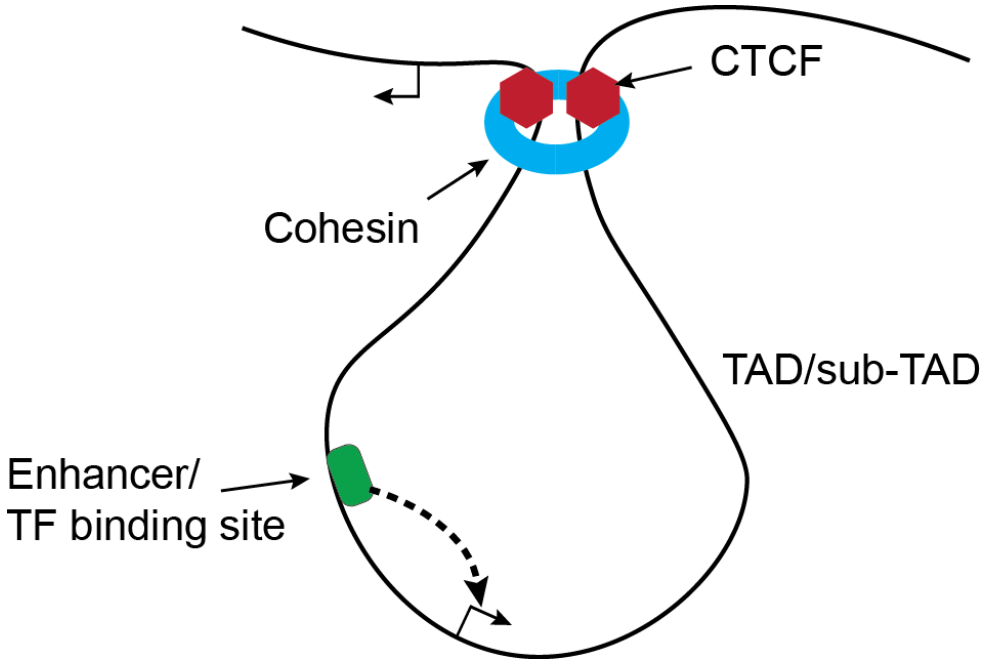

Figure 5: Topologically Associated Domain Architecture.

TADs and sub-TADs are chromatin loops of variable size that are isolated from the rest of the genome by paired CTCF insulation domains that are coupled by cohesion proteins. Sub-TADs are nested within TADs. Enhancer-promoter looping interactions have been described between sites on the same TAD.

It has been proposed that TADs form via a process of loop extrusion, which is governed by cohesins, which associate loop ends, and CTCF, which prevents further extrusion (Fudenberg et al., 2016; Sanborn et al., 2015). The relationship between CTCF boundaries and TAD formation is dependent on CTCF binding site orientation (Guo et al., 2015b). Genome editing studies have revealed that altering the orientation of a CTCF binding site at a TAD boundary can result in ectopic activation of local genes due to altered promoter-enhancer interactions. Accordingly, deletion of CTCF binding sites results in loss of TAD formation (Sanborn et al., 2015). TADs themselves are further

organized into sub-TADs, which contain secondary chromatin interaction domains (Rao et al., 2014). Within these sub-domains, transcription factors distally interact to control gene expression. Thus, the genome is structurally organized in a coordinated series of nested chromatin loops that enable interactions between promoters and the distal regulatory elements that modulate them.

1.11 Perturbed gene regulation resulting in Mendelian disease

Alterations in gene regulation can cause drastic phenotypic changes at the organismal level. The earliest evidence that perturbed gene regulation results in phenotypic changes in humans was gathered in multigenerational studies that attribute genetic disease to "position effects" (Kleinjan and Lettice, 2008). Affected individuals present clinically with phenotypes associated with total or partial loss of gene function, but do not have variants within the putative causal gene's protein coding region. Further investigation has revealed that the genomes of such individuals frequently have chromosomal rearrangements that have either uncoupled a gene from its enhancers, or have resulted in a deletion of those enhancers entirely (**Figure 6**). Many syndromic phenotypes associated with positional effects vary in presentation, consistent with the idea that some deletions and rearrangements only disrupt partial regulatory function of a given gene. Fewer position-dependent Mendelian genetic diseases have been characterized than diseases caused by variants in protein coding regions (**Table 1, (Kleinjan and Lettice, 2008)**). This is likely the result of multiple factors including active

compensation by unaltered alleles in heterozygous individuals, lack of dominant-negative effects induced by mutations, and partial loss of regulatory function dependent on the size of the genetic lesion.

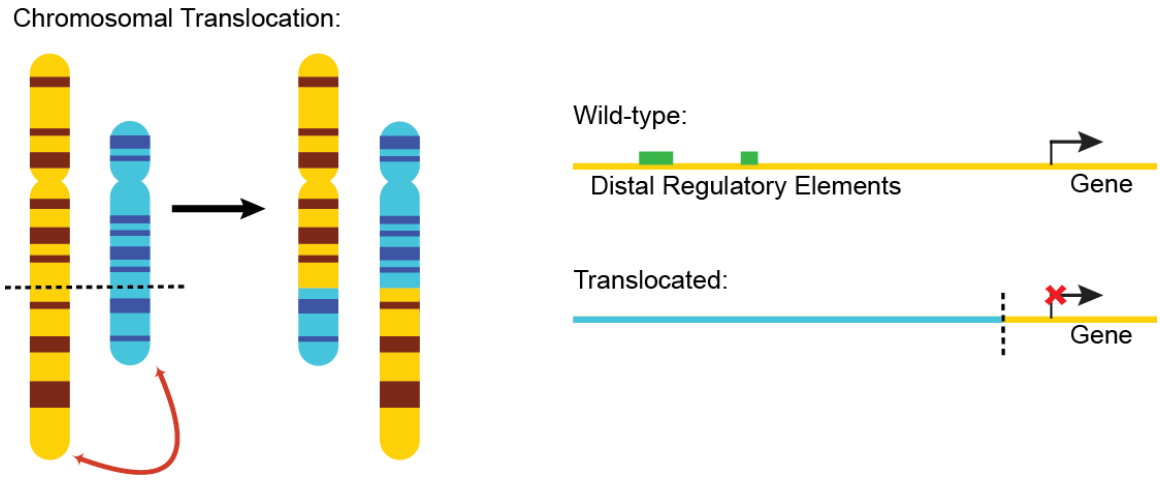

Figure 6: Chromosomal rearrangements can perturb gene regulation.

Chromosomal rearrangements, such as balanced translocations (left) can uncouple regulatory regions from their target gene (right) causing Mendelian disease.

Table 1: Mendilian position-dependent diseases. Adapted from Kleinjan and Lettice, 2008.

GENE	DISEASE/ SYNDROME	OMIM #	REFERENCE
ALX4	Potocki-Schaffer syndrome	601224	(Wakui et al., 2005)
FOXC1	Glaucoma/autosomal dominant iridogoniodysgenesi	601631	(Davies et al., 1999)
FOXC2	Lymphedema-distichiasis syndrome	153400	(Fang et al., 2000)
FOXL2	Blepharophimosisptosis-epicanthus inversus syndrome	110100	(Crisponi et al., 2004) (Beysen et al., 2005)
FOXP2	Speech-language disorder	602081	(O'Brien et al., 2003)
FSHD	Facioscapulo humeral dystrophy	158900	(Gabellini et al., 2002); (Jiang et al., 2003)
GLI3	Greig cephalopolysyndactyly	175700	(Wild et al., 1997)
HBA	alpha-Thalassemia	141800, 141850	(Viprakasit et al., 2003) (Tufarelli et al., 2003)
HBB	gamma-beta-Thalassemia	141900	(Kioussis et al., 1983) (Driscoll et al., 1989)
HOXD	Lactase persistence	223100	(Enattah et al., 2002) (Tishkoff et al., 2007)
MAF	Cataract, ocular anterior segment dysgenesis, and coloboma	610202	(Jamieson et al., 2002)
PAX6	Aniridia	106210	(Fantes et al., 1995) (Kleinjan et al., 2001)
PITX2	Rieger syndrome	180500	(Trembath et al., 2004)
PLP1	Spastic paraplegia type 2 with axonal neuropathy, Pelizaeus-Merzbacher disease	312920, 312080	(Lee et al., 2006) (Muncke et al., 2004)
POU3F4	X-linked deafness	304400	(de Kok et al., 1996)
REEP3	Increased risk of Hirschsprung disease	142623	(Emison et al., 2005)
RUNX2	Cleidocranial Dysplasia	119600	(Fernandez et al., 2005)

1.12 Variation in regulatory elements as a contributor to complex phenotypes

Evidence from large-cohort genetic studies suggests that regulatory variation plays a critical role in the etiology of complex disease. Genome-wide Association Studies (GWAS) have identified hundreds of loci that are associated with complex phenotypes (Welter et al., 2014). However, relatively few of the protein coding sequences in these loci have contained candidate disease-causing mutations. Thus, there has been increased interest in the possibility that changes in regulatory function and changes in gene expression are likely to contribute to complex phenotypes (Pai et al., 2015). Expression Quantitative Trait Loci (eQTL) studies have associated haplotypes with altered gene expression, and "*cis*-eQTLs" have been identified for many genes, suggesting that the complement of variants contained within a given haplotype block can alter gene expression (Farrall, 2004). However, eQTL studies are limited in their ability to localize potential regulatory variants due to the resolution limits imposed by ancestral haplotype recombination.

Recent advances in genomics have substantially improved the ability to quantify the effects of regulatory variation (**Figure 7**). Allele specific quantification of mRNA transcription has revealed that individual alleles for a gene are frequently differentially expressed relative to one another in a manner that is independent of gene imprinting (Adoue et al., 2014; Battle et al., 2014; Ge et al., 2009; Montgomery et al., 2010). Meanwhile, there is strong evidence that TF occupancy is altered at CRMs that contain

genetic variants within transcription factor binding sites. However, the ability to identify causal regulatory variants is limited by the relative lack of ChIP-seq data available and the high tolerance for motif degeneracy at TF binding sites (Chen et al., 2016; Ni et al., 2012; Reddy et al., 2012b). DNase-seq has been used as a proxy for TF binding in a factor-agnostic manner (Crawford et al., 2004). Likewise, allele specific TAD identification has revealed individual-specific chromatin loop formation (Tang et al., 2015). These studies have revealed that many GWAS associated loci encode a vast number of putative enhancers that can vary in activity by genotype.

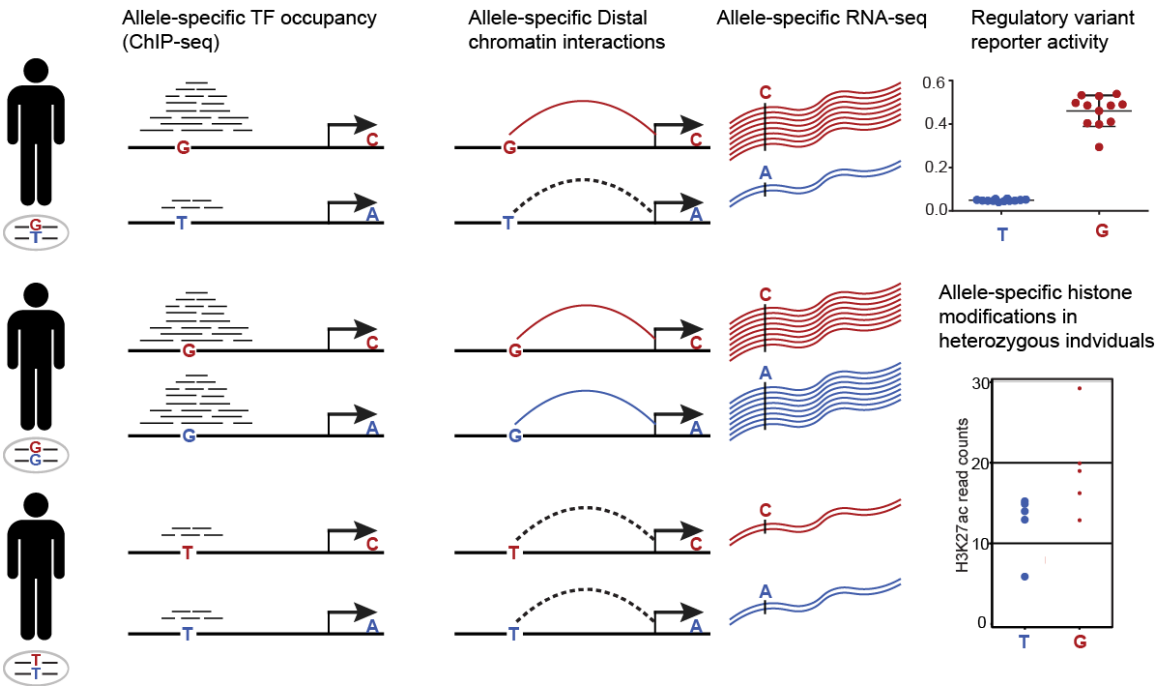

Figure 7: Genomic evidence of regulatory variation.

Evidence for regulatory variation has come from multiple sources including (from right to left): Allele specific TF occupancy, Allele specific distal promoter-enhancer interactions, allele specific gene expression analysis and allele specific enhancer function

(top right), and allele specific histone modifications associated with enhancer function (bottom right). Adapted from Tang et al., 2015 and Vockley et al. 2015.

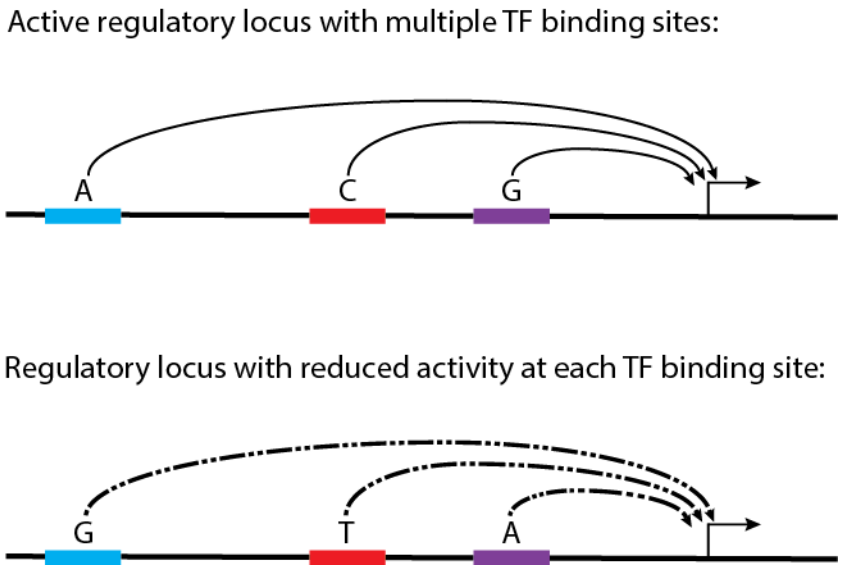

Figure 8: The multiple enhancer variant hypothesis.

The multiple enhancer variant hypothesis states that genetic variants at multiple regulatory sites in linkage disequilibrium can contribute together to alter the expression of a target gene.

These observations led to the "multiple enhancer variant hypothesis" of complex trait etiology (**Figure 8**). The multiple enhancer variant hypothesis was first published in a study of six autoimmune disorders. That study demonstrated that GWAS association for complex traits may be the result of multiple polymorphisms in putative enhancers in linkage disequilibrium (Corradin et al., 2014). Evidence from traditional luciferase reporter vector experiments demonstrates that common variants within DNase hypersensitive sites linked to GWAS loci can alter enhancer function(Guo et al., 2015a).

However, these methods are of limited utility as GWAS follow up approaches due to the large numbers of putative variant enhancers within any given locus. We addressed this concern by pioneering a novel experimental approach that permits the simultaneous quantification of enhancer function at many sites in populations of individuals. Our approach, detailed in chapter 3 of this dissertation, allows for the functional fine mapping of eQTLs to determine which putative variants in linkage disequilibrium with risk loci are responsible for alterations in enhancer function (Vockley et al., 2015). Subsequent studies have used this method to identify 842 regulatory variants from more than 32,000 putative regulatory variants identified using *cis*-eQTL studies (Tewhey et al., 2016). Meanwhile, a survey of nearly 3,000 common variants within 75 GWAS loci identified 32 functional regulatory variants associated with red blood cell related traits (Ulirsch et al., 2016). Thus far, the results of these experiments suggest that perturbations of gene expression are likely the result of the compound effects of mutations within multiple enhancers that together alter gene expression. A complementary study investigated the differences three-dimensional chromatin organization between individuals. That study found that single nucleotide polymorphisms within *cis*-regulatory elements can inhibit the formation of distal chromatin interactions in some individuals. Together, these data support the hypothesis that in some cases complex phenotypes are affected by multiple enhancer variants acting in concert.

1.13 An integrated view of gene regulation

The decades-long effort to characterize how genes are regulated has yielded many insights into the molecules and mechanisms that tune eukaryotic transcription. From the fundamental subunits of gene regulation to the higher order structures that coordinate those sub-units, there is tremendous flexibility. As in many other disciplines of biology, for every rule of gene regulatory logic there are countless exceptions. This flexibility and functional redundancy is essential given the complexity and diversity of tasks for which the genome is used. Despite this intrinsic tolerance of gene regulatory systems, alterations in DNA sequence at regulatory sites has resulted in phenotypic plasticity that drives both evolution and disease.

Recent advances in genomic technology and consortium-led efforts have substantially contributed to the field's understanding of eukaryotic gene regulatory mechanisms (ENCODE, 2012). Integrative genomic analyses have enabled statistical insights into gene regulation that were not obvious in more anecdotal studies. While these analyses have resulted in a rapid expansion of our understanding of enhancer biology, they are also prone to over generalization. Drawing conclusions based on the most frequently observed genomic events is sometimes misleading. For example, studies of the glucocorticoid receptor have employed analyses that disproportionately consider the impact of AP-1-enabled GR binding due to the prevalence of this mechanism (Biddie et al., 2011; John et al., 2011). As detailed in Chapter 2 of this dissertation, a

mechanistically distinct minority of GR binding sites is capable of inducing a gene expression response to glucocorticoids. Understanding the features that distinguish each class of site revealed a substantial revised model of GR function. Thus, while genomic approaches to studying gene regulation have been useful, they are not without their potential pitfalls. Some of these pitfalls can be overcome by employing dynamic experimental systems (e.g. cellular response to hormones). Likewise, the use of population genomic studies to investigate biological mechanisms is an emerging frontier in the field of gene regulation.

2. GR interaction modules direct the gene expression response to glucocorticoids

The text contained in this section has been accepted for publication in the journal *Cell*.

The title an authors of the manuscript that describes these results is:

Direct GR binding sites potentiate clusters of TF binding across the human genome.

Christopher M. Vockley, Anthony M. D'Ippolito, Ian C. McDowell, William H. Majoros, Alexias Safi, Lingyun Song, Gregory E. Crawford, Timothy E. Reddy

2.1 Introduction

Regulation of transcription plays a major role in human health and disease (Maurano et al., 2012; Olansky et al., 1992; Stadhouders et al., 2014; Vockley et al., 2015). The basic mechanism of transcriptional regulation in humans involves transcription factors (TFs) binding to specific genomic regulatory elements. Once bound, TFs influence the recruitment of transcriptional machinery to the promoter of one or more target genes. Several studies have now mapped the location of binding sites across the human genome for many TFs and in many cell types (ENCODE, 2012; Johnson et al., 2007). Those studies have revealed a complex landscape of TF occupancy in which a TF typically binds thousands of locations across the human genome, but only directly regulates hundreds of genes (Gao et al., 2013; Reddy et al., 2009) The discrepancy between TF binding and gene regulation can be explained in part by findings that most

TF binding sites have weak regulatory activity (Kheradpour et al., 2013; Melnikov et al., 2012) and that a TF often binds multiple sites near the same target gene (Gotea et al., 2010). The multiplicity of binding may be the result of functional redundancy between sites (Somma et al., 1991), cooperative assembly of TF complexes (Hertel et al., 1997), or local diffusion of bound factors along the genome (Coleman and Pugh, 1995). Furthermore, numerous studies have shown that the number and diversity of TF binding sites contributes to synergistic rather than additive regulatory activity (Smith et al., 2013; Staller et al., 2015), suggesting a potential relationship between clusters of TF binding sites and the activity of those sites.

Ligand inducible TFs such as the GR are a representative model system for investigating the relationship between TF binding and activity. Once bound by GCs such as the cortisol analogue dexamethasone (DEX) the GR binds at thousands of locations across the human genome and regulates the expression of hundreds of genes (Reddy et al., 2009; So et al., 2007; Wang et al., 2004). The GR binds the genome either directly via DNA-sequence-specific interactions with a GRE or, more often, indirectly via tethering to other proteins such as the AP-1 family of TFs (Chandler et al., 1983; Gertz et al., 2013). Direct binding sites are more often shared across cell types and more likely to occur in genomic regions with closed chromatin prior to induction. Conversely, AP-1 co-occupied sites are more likely to occur at regions of already-accessible chromatin; are more likely to be specific to distinct cell types; and have been

hypothesized to be the basis for differences in the GC responses between tissues (Biddie et al., 2011; Gertz et al., 2013; John et al., 2011).

Here, we propose that the organization of GR binding across the human genome conforms to a model in which GREs recruit GR directly to the DNA, and then those direct sites nucleate clusters of tethered binding nearby. We quantified the activity of GR-bound DNA elements on the genome-scale, assaying approximately 2.9 million unique reporter vectors covering 10,963 GBSs. We found that direct GBSs confer inducible enhancer function, while tethered sites do not. We further provide evidence that GR binding to tethered sites depends on the proximity of the tethered sites to direct sites. The resulting clusters of GBSs modulate the regulatory activity of direct GBSs, potentially contributing to expression levels of cell-specific GC responsive transcription. Together, these results demonstrate that patterns of genomic GR occupancy observed with ChIP-seq reflect locally coordinated and functionally synergistic GR binding events, rather than independent and additive events. We also provide evidence that our enhancer cluster model is general to the estrogen receptor (ER), suggesting that other TFs can act similarly.

2.2 Results

2.2.1 Quantifying GC-induced regulatory element activity

Figure 9: ChIP-reporter experimental design.

For ChIP-reporter assays, we first used PCR to add 15-bp adapters that are complementary to the reporter vector to GR ChIP-seq libraries. Adapted DNA was then ligated into the 3' UTR of the GFP reporter gene of the STARR-seq vector. The resulting plasmid library was transfected into A549 cells. Cells were treated with 100 nM DEX or 0.02% EtOH for 3 h and RNA was collected. cDNA of the reporter GFP was amplified using gene-specific primers. The relative regulatory activity of each fragment was then estimated by counting aligned reads generated by high-throughput paired-end sequencing.

To assess the functional diversity of GBSs, we quantified the regulatory activity of the 10,963 GBSs in A549 cells, a human lung epithelial cell line (**Figure 9**). To do so, we first used ChIP to isolate GBSs from A549 cells after treating the cells for 3 h with 100 nM DEX or with equal-volume ethanol (EtOH) as a vehicle control. High-throughput sequencing of the GR ChIP DNA followed by peak calling and sub-peak splitting identified 27,432 GBSs (**Appx. 1; Table S1**). We then cloned the ChIP-seq library into the STARR-seq reporter assay (Arnold et al., 2013). We estimated that the resulting plasmid library contained 2.9 million unique GR ChIP fragments (**Appx. 1; Figure 37A**). The fragments had a similar genomic distribution and size as the ChIP-seq library, indicating that cloning into the reporter assay did not introduce substantial bias (**Appx. 1; Figure**

37B and 38A). Of the cloned fragments, 6% mapped to the 27,432 called GBS sub-peaks. We transfected the reporter library back into A549 cells and treated the cells with either 100 nM DEX or EtOH to assay DEX-responsive regulatory activity. After filtering to only retain sites with sufficient statistical power to make a determination of regulatory activity (Love et al., 2014a), we quantified the activity and DEX-responsiveness of 10,963 GBSs (**Appx. 1; Figure 38B**). The number of sites assayed was similar to that identified in previous GR ChIP-seq studies (Biddie et al., 2011; Reddy et al., 2012a; Reddy et al., 2009) and included 82% of the sites reported in the same cell line previously (Reddy et al., 2009). We assayed the reporter library in three replicates per condition, and found the approach to be highly reproducible (Spearman $\varrho > 0.8$ between all pairs of replicates, **Figure 10A**). Together, these results indicate that we assayed the activity of nearly all GBSs in A549 cells.

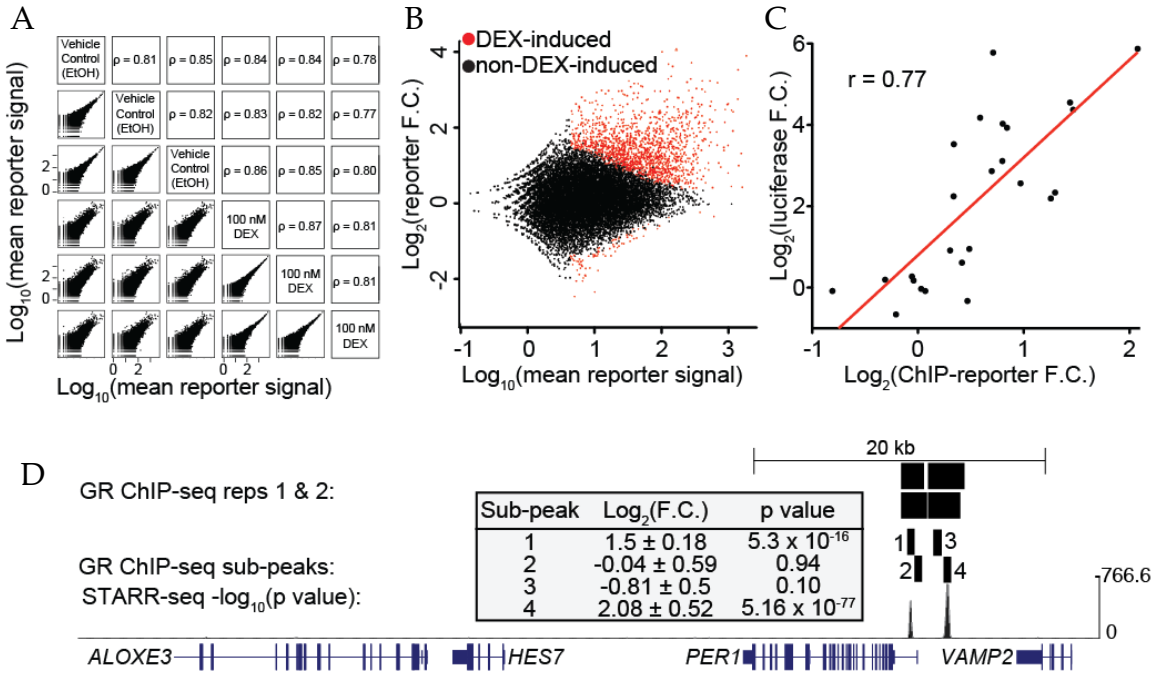

Figure 10: ChIP-reporter data and cross-platform validation.

(A) A comparison of the normalized ChIP-reporter signal for all 27,432 called GBS sub-peaks between all replicates and treatments. Each point represents one GBS, and all sites are included in each plot. **(B)** For each GBS assayed, the fold change in regulatory element activity with DEX treatment is plotted as a function of the mean sequencing coverage at that element. Red points indicate regulatory elements with a statistically significant response (FDR < 5%). **(C)** Correlation of dual luciferase assays with ChIP-reporter results (r = 0.77) **(D)** DEX-induced regulatory features of the *PER1* locus on human chromosome 17. GR ChIP-seq peaks called by MACS (first two rows), GR-ChIP sub-peaks assayed by ChIP-reporters (3^{rd} and 4^{th} rows), -$\log_{10}$ (p value) from BAC-based DEX-induced STARR-seq (5^{th} row). Inset contains ChIP-reporter $\log_2$ (fold change) in response to DEX and associated p value for each sub-peak assayed.

Of the assayed GBSs, 84% had a less than 2-fold difference in activity between treatments (**Appx. 1; Figure 38C**), indicating that most GBSs have modest regulatory activity in our ChIP-reporter assays. That result was consistent with the results from other recent high-throughput reporter assay studies (Arnold et al., 2013; Melnikov et al., 2012). Of the assayed GBSs, 1,376 (13%) had statistically significant regulatory activity at a false discovery rate (FDR) < 5% (Love et al., 2014a) (**Figure 10B, Appx. 1; Figure 39A, Table S2**). Larger fragments were more frequently GC-inducible (**Appx. 1; Figure 39B**). While GCs are known to both activate and repress gene expression (Sakai et al., 1988; Slater et al., 1985) and while STARR-seq can report activation and repression (Arnold et al., 2013) (**Appx. 1; Figure 39C**), 95% (N = 1,330) of the GC-regulated GBSs assayed had increased reporter gene expression in response to DEX. We validated the results with dual luciferase reporter assays (**Figure 10C, Appx. 1; Figure 39D**) and with STARR-seq performed on BACs covering 1 Mb of the human genome that were selected to contain

GBSs near GC-responsive genes (**Figure 10D, Appx. 1; Table S3 and S4**). We observed no enrichment of plasmids containing elements that encode GC-responsive enhancers in the nuclei of cells treated with DEX, suggesting that our results are not due to biased nuclear localization of the reporter vectors (**Appx. 1; Figure 40**). To test if the non-DEX-responsive sites are maximally pre-induced prior to DEX treatment, we performed additional luciferase reporter assays to evaluate if adding one or two GREs to non-DEX-responsive sites increased enhancer activity in response to DEX. The synthetic composite sites had up to 24-fold increased DEX-responsive activity that was lost with targeted mutation of the GRE, suggesting that the non-responsive sites are capable of greater activity with direct GR binding (**Appx. 1; Figure 41**). Together, our results indicate that 13% of GBSs have GC-inducible regulatory activity in a reporter assay, and that those sites overwhelmingly act to increase gene expression. DEX-responsive activity in ChIP-reporters was largely distinct from occupancy as measured by ChIP-seq ($\varrho = 0.22$, **Figure 11A**), indicating that reporters provide complementary information about mechanisms of gene regulation. We used RNA-seq to determine the set of DEX-responsive genes after the same treatment (**Appx.1; Table S5**), and found that the transcription start site (TSS) of DEX-responsive genes were overall closer to DEX-responsive GBSs than to an equal number of randomly sampled non-responsive sites ($p = 0.005$; **Figure 11B**).

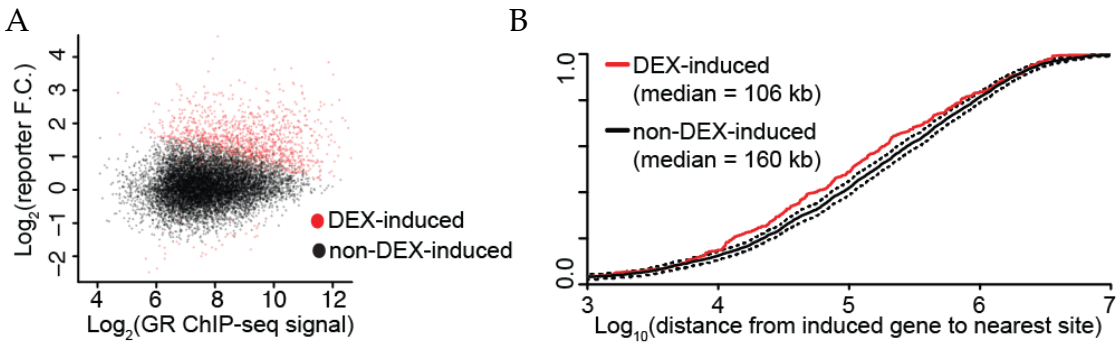

Figure 11: DEX-induced GBSs compared to GR ChIP signal and location of non-DEX-induced sites.

(A) GR reporter signal is plotted against GR ChIP-seq signal for all GBSs assayed with our ChIP-reporter approach. The two datasets correlated with a Spearman's $\rho = 0.22$. **(B)** Fraction of induced genes as a function of the distance from the TSS to the nearest DEX–induced GBS (red line, median distance 106 kb) or an equivalent number of non-DEX-induced GBSs (black line, median distance 160 kb). The non-DEX-induced reporter sites were randomly subsampled to have an equivalent number as for the DEX-induced reporter sites 100 times, and the mean and standard deviation are shown in black solid and black dashed lines, respectively.

2.2.2 DEX-induced GBSs are direct binding sites

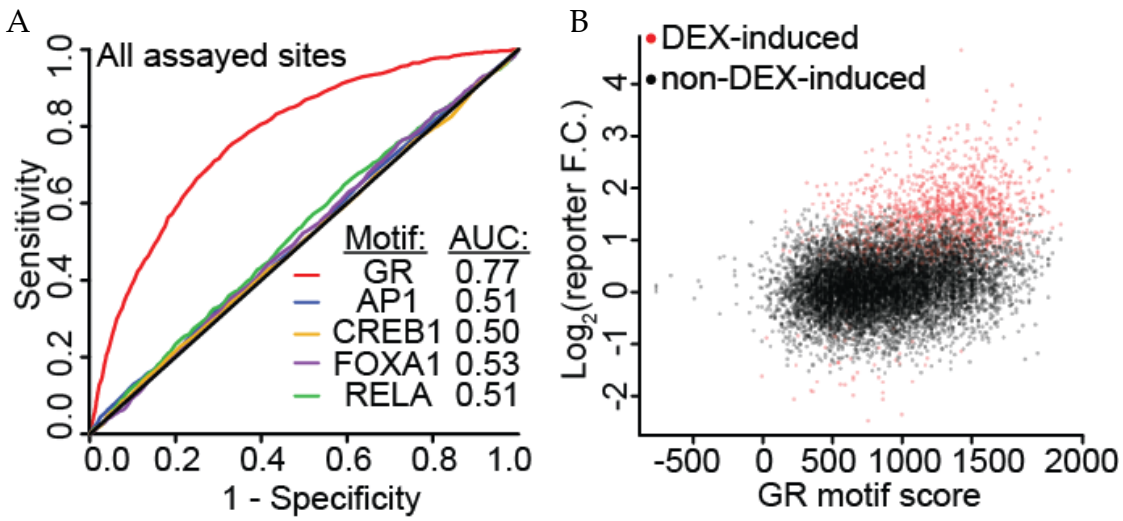

Figure 12: DEX responsiveness is predicted by the presence of a GRE.

(A) ROC analysis of DEX-induced enhancers evaluating DNA sequence motif significance for the GRE and known GR interacting proteins as predictors of enhancer function. **(B)** Scatter plot of GR motif score vs. ChIP-reporter activity. Significant DEX-induced elements are shown in red (r = 0.35).

The GR is known to bind DNA directly at GREs or by tethering to other TFs such as AP-1 (Ratman et al., 2013). We hypothesized that the distinct modes of binding have different regulatory functions. To evaluate that hypothesis, we investigated whether the GRE or the DNA binding motifs of known co-binding TFs [AP-1 (Herrlich, 2001), FOXA1 (Belikov et al., 2009), NFκB (Wang et al., 1997) and CREB1 (Sheppard et al., 1998)] predict DEX-responsive regulatory element activity in reporter assays. Of the binding motifs evaluated, the GRE (Hagerty et al., 2001) was the only one that predicted changes in regulatory element activity after DEX treatment better than random [receiver

operating characteristic (ROC) area under the curve (AUC) = 0.77, **Figure 12A and 12B, Appx. 1; Figure 42A and 42B**]. Using a Gaussian mixture model, we estimated that ~80% of DEX-responsive elements and ~35% of non-responsive elements had a GRE (**Appx. 1; Figure 43**). A *de novo* motif enrichment analysis revealed AP-1 motifs in the DEX-responsive elements with GREs (MEME E-value 3 x 10^{-101}), but not in the corresponding non-responsive elements. In contrast, binding motifs for co-binding TFs were predictive of baseline regulatory activity (**Appx. 1; Figure 44**), but were not predictive of DEX-induced regulatory activity (**Figure 12A, Appx. 1; Figure 45, Table S6**).

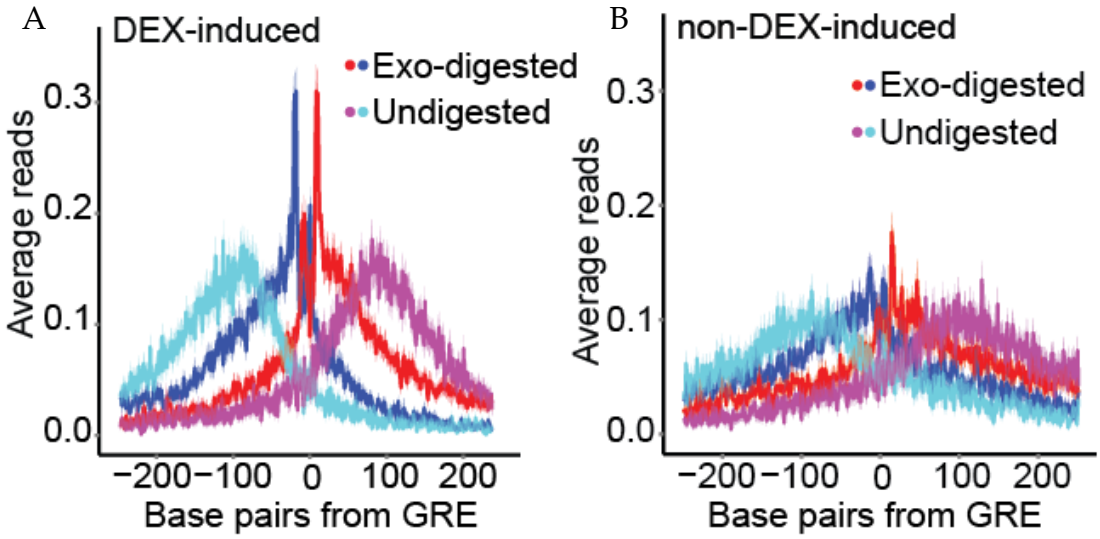

Figure 13: ChIP-exo footprints of DEX-induced and non-DEX-induced GBSs.

(A) Average GR ChIP-exo signal across GC-inducible enhancer GBSs centered on the strongest GRE. Red and blue are the 5' bp of reads aligned to the positive or negative strand, respectively. Cyan and magenta are the 5' bp of reads at non-exonuclease treated fragment end aligned to the positive or negative strand, respectively. Ribbons indicate standard error. **(B)** GR ChIP-exo signal as above, aggregated across an equal number of sub-sampled GBSs that are not induced by DEX in reporter assays.

To distinguish between tethering and GR binding directly to a weak GRE, we mapped the genomic footprints of GR using ChIP-exo. GBSs with significant reporter activity showed a footprint concordant with the known dimer structure of GR (**Figure 13A, Appx. 1; Figure 46A and 46B**) (Starick et al., 2015). In contrast, GBSs matched for ChIP-seq signal but lacking DEX-inducible activity did not have a ChIP-exo footprint consistent with GR:DNA interactions (**Figure 13B, Appx. 1; Figure 46C**). From these results, we conclude that direct GR binding to a GRE is predictive of DEX-inducible reporter activity, whereas tethered binding via other TFs is not.

2.2.3 Cell type-specific DEX-induced enhancers are encoded by direct GBSs

GR binding sites can vary dramatically between different cell types. For example, in a recent study comparing GR binding between A549 and Ishikawa cells, 15,220 (75%) of A549 GBSs were specific to A549 cells (Gertz et al., 2013). Most of the differences between cell types occur at sites that lack a GRE, suggesting that changes in the expression or binding of other TFs drive the differences in GR binding between cells. To investigate further, we asked whether DEX-responsive sites are enriched in cell-shared GBSs but depleted in cell-specific GBSs. We assayed 7,951 of the A549 GBSs identified in Gertz et al. Of those 1,088 (14%) were shared between cell types and 6,863 (86%) were A549-specific. Of the cell-shared GBSs, 331 (30%) were DEX-responsive, an enrichment of 2.3-fold over the overall positive rate of 13%. Meanwhile, of A549-specific GBSs, 928 (13.5%) were DEX-responsive, similar to the overall positive rate (**Figure 14A**). Those

results demonstrate that although cell-shared sites but not cell-specific sites are enriched for DEX-responsiveness, most (928 vs. 331) DEX-responsive enhancers are cell-specific. As was observed for all DEX-responsive GBSs, the presence of a GRE and not the motifs of tethering factors explains the DEX-response of A549-specific GBSs (**Figure 14B**). Together, those results demonstrate that both tethered sites and a small fraction of direct sites vary between cell types. Further, A549-specific direct GBSs act as cell type-specific DEX-responsive enhancers.

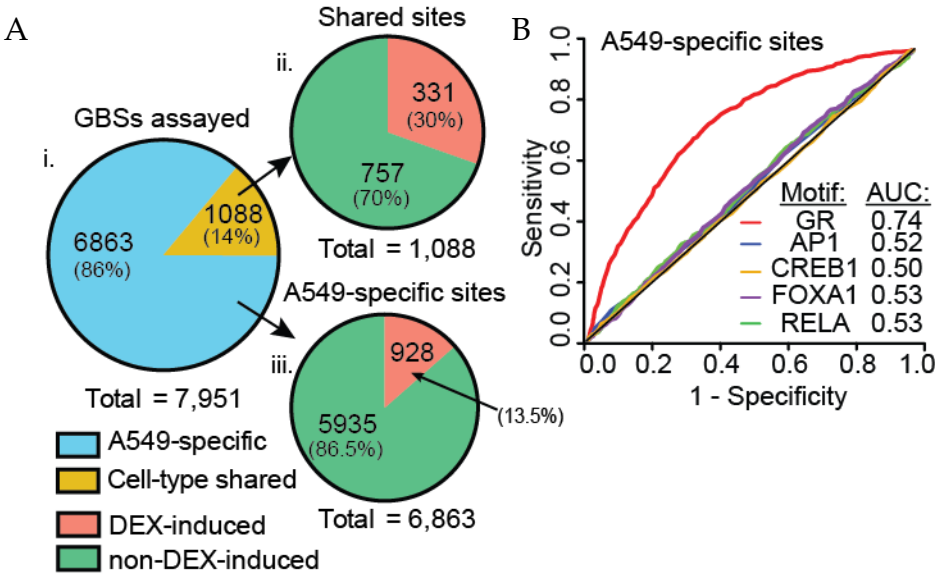

Figure 14: Cell-type specific DEX-induced GBSs are predicted by the presence of a GRE.

(A) i. Fraction of GBSs in a previous study assayed using ChIP-reporters that were specific to A549 cells or common between A549 and Ishikawa cells (N = 7,771). ii. Fraction of A549-Ishikawa shared GBSs with significant DEX-induced reporter activity (N = 1088). iii. Fraction of A549-specific sites with significant DEX-induced reporter activity (N = 6,863). GBSs are categorized as specific to A549 cells or shared between A549 cells and endometrium-derived Ishikawa cells (Gertz et al. 2013). **(B)** ROC analysis of A549-specific DEX-inducible enhancers.

2.2.4 Remodeling of GC-induced sites in the endogenous epigenome

If ChIP-reporters recapitulate endogenous gene regulation, we expect that there will also be differences in the endogenous epigenome between GBSs with and without DEX-responsive reporter activity. We first used DNase-seq to determine if there were differences in chromatin accessibility between the two classes of GBSs (Song and Crawford, 2010). After controlling for differences in GR occupancy between classes of GBSs, the DEX-responsive GBSs had less accessibility than non-DEX-responsive GBSs before and after DEX treatment (T-test, $p < 1 \times 10^{-100}$ and $p = 6 \times 10^{-55}$, respectively), but a significantly greater gain in accessibility after DEX (T-test $p < 1 \times 10^{-100}$) (**Figure 15A**). The increase in accessibility was distributed across a large share of DEX-responsive GBSs, rather than due to a subset of GBSs with larger changes (**Appx. 1; Figure 47A**). Together, these data indicate that the DEX-responsive GBSs undergo more substantial chromatin remodeling after DEX treatment than non-responsive GBSs.

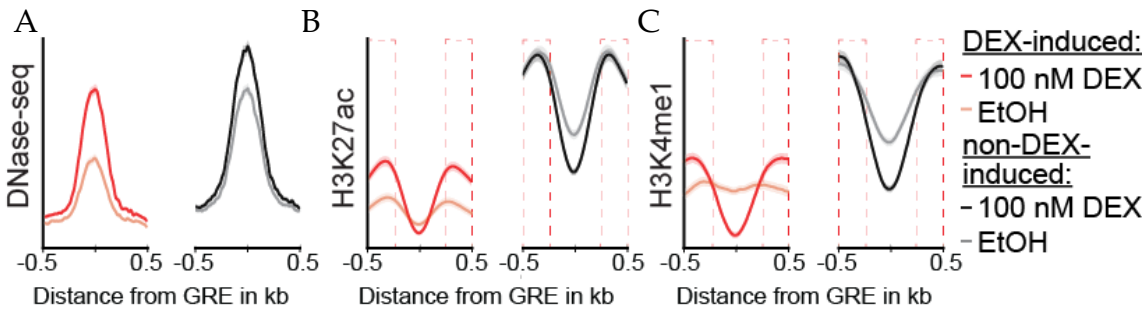

Figure 15: Epigenomic remodeling of GBSs after DEX exposure.

(A) Aggregate profile plot of DNAse hypersensitivity **(B)** H3K27ac and **(C)** H3K4me1 at DEX-inducible and non-DEX-inducible sites before and after 1 h exposure to 100 nM DEX. Flanking 250 bp indicated by red dashed boxes.

We next evaluated whether changes in covalent histone modifications at DEX-responsive and non-DEX-responsive GBSs were consistent with those sites having distinct regulatory activities. Specifically, we evaluated whether there were distinct changes in the enrichment of two covalent histone modifications that are known to be associated with enhancer activity: histone 3 lysine 27 acetylation (H3K27ac), and H3K4 mono-methylation (H3K4me1) (ENCODE, 2012) (**Figure 15B and C, Appx. 1; Table S1**). In control-treated cells, non-DEX-responsive GBSs had greater H3K27ac and H3K4me1 signal than DEX-responsive GBSs before and after DEX induction. Those differences were observed both in the 500 bp window centered on the best match to the GRE, which we term the core of the GBS, (T-test, before DEX: $p = 5 \times 10^{-99}$, $p = 3 \times 10^{-45}$, respectively; after DEX: $p = 1 \times 10^{-38}$, $p = 3 \times 10^{-46}$, respectively), and in the 250 bp flanking that core on either side (T-test, before DEX: $p = 3 \times 10^{-95}$, $p = 10 \times 10^{-70}$, respectively; after DEX: $p = 5 \times$

10^{-21}, p = 5 x 10^{-32}, respectively). DEX-responsive GBSs, however, had greater changes in histone modifications, especially in the flanking regions. H3K27ac and H3K4me1 increased significantly more in the flanks of the DEX-responsive GBSs than in flanks of the non-responsive GBSs (T-test, p < 1 x 10^{-100}). At the core of the GBSs, we observed a greater DEX-dependent decrease in H3K27ac signal in non-DEX-responsive GBSs than in DEX-responsive GBSs (p < 1 x 10^{-100}), but not for H3K4me1 (p = 0.02). We further found that, across DEX-induced GBSs, the sites with the least H3K27ac prior to DEX treatment (**Figure 16**) and with the greatest increase in H3K27ac with DEX treatment (**Appx. 1; Figure 47B**) also had the greatest DEX-responsive reporter activity.

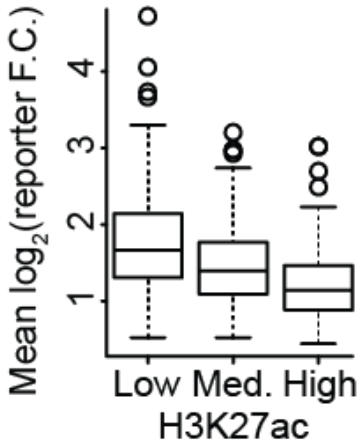

Figure 16: DEX-induced reporter activity as a function of H3K27 acetylation state.

Box plots showing the relative reporter activity of DEX-induced enhancers (FDR < 5%) binned by pre-induction H3K27ac status, where each bin contains an equal number of sites.

The histone acetyltransferase P300 is responsible for establishing H3K27ac across the human genome (Liu et al., 2008). Consistent with our observations of H3K27ac, we observed more ChIP-seq signal for P300 (ENCODE, 2012) at non-DEX-responsive sites prior to hormone treatment than at DEX-responsive sites. Specifically, in control treated cells, there was P300 occupancy at 82% of non-responsive GBSs that were also bound by the AP-1 family member FOSL2. Meanwhile, P300 was bound at 6% of DEX-responsive GBSs that lack evidence of FOSL2 binding in the control treatment (**Appx. 1; Figure 47C**). As a negative control for our histone modification analysis, we examined H3K27 trimethylation (H3K27me3), a histone modification associated with gene repression. We found no significant correlation between endogenous H3K27me3 and reporter assay activity **(Appx. 1; Figure 47D)**. These results show that functionally distinct classes of GBSs have distinct epigenetic states in the genome before DEX treatment, and distinct changes to epigenetic state after DEX treatment.

Integrating the endogenous epigenetic state into our ROC analysis resulted in a statistically significant increase in the area under the curve (DeLong's test, GRE motif and H3K27ac $\log_2$(fold change [F.C.]) ROC vs. GRE motif ROC, $p = 9.8 \times 10^{-3}$; GRE motif, DNase $\log_2$(F.C.), and H3K27ac $\log_2$(F.C.) ROC vs. GRE motif, DNase $\log_2$(F.C.), and H3K27ac ROC, $p = 3.36 \times 10^{-2}$, **Figure 17**). The epigenetic state of both DEX-induced and non-DEX-induced GBSs reflected local chromatin remodeling around the GBS. Together, these data demonstrate that reporter assay output is reflective of the mechanisms that

coordinate steady state expression and the inducible expression response to GCs in the endogenous genome.

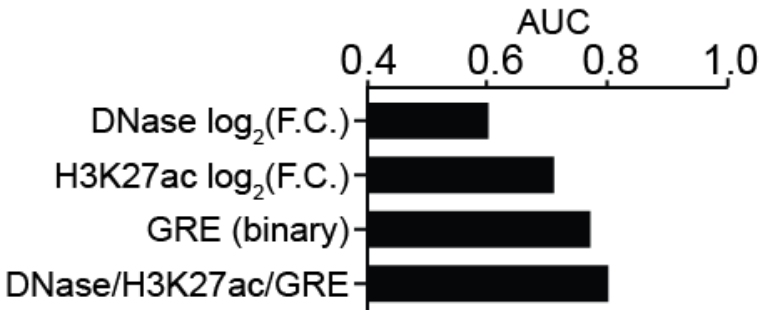

Figure 17: Epigenomics data improves the performance of models that predict DEX-induced enhancer activity.

Area under ROC curves generated by using the presence of enhancer associated epigenetic marks in the endogenous genome independently or with the presence of GR motif as a predictor of enhancer activity.

2.2.5 GBSs cluster in the genome

Several studies have shown that TF binding sites cluster in the human genome (Gotea et al., 2010; Rye et al., 2011). Those clusters may reflect numerous independent binding sites or alternatively, dependencies between sites that may arise, for example, via protein-protein interactions between TFs bound at different sites. One way to differentiate between those two possibilities is to induce the GR to bind a subset of sites, and to then ask whether those sites are a random selection of all GR binding sites, as would be expected if binding sites are independent, or instead enriched within clusters,

as would be expected if there are local dependencies between GR binding sites. To do so, we analyzed data from a previous study in the same cell type in which lower concentrations of DEX (0.5 nM and 5 nM) were used to induce the GR to bind a subset of the sites that are bound at 50 nM DEX (Reddy et al., 2012a). We then asked whether the subset of sites bound at lower concentrations were randomly distributed across all GR binding sites occupied at 50 nM treatment, indicating independent binding events; or were clustered together, indicating local dependencies between sites.

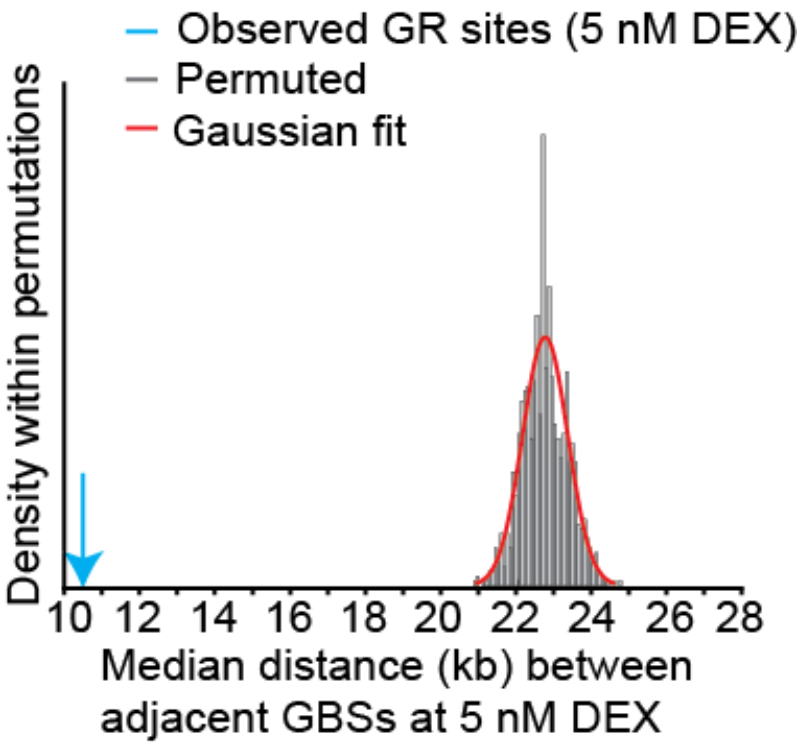

Figure 18: GBSs bound at low dose DEX treatment occur in coordinated clusters.

The median distance between sites bound by GR in A549 cells after 1 h treatment with 5 nM DEX was calculated (blue arrow). As a null model the locations GBSs bound at 5 nM DEX in A549 cells were permuted between the sites bound by the GR at 50 nM DEX in

A549 cells. Sites were permuted 1,000 times and the median distance between sites was calculated after each permutation.

Twenty-four percent of the sites bound by GR in A549 cells after a 1 h treatment with 50 nM DEX are also bound after treating cells for the same period with 5 nM DEX (Reddy et al., 2012a). We hypothesized that if the GR binds to sites independently, then sites bound at low-concentration would be evenly distributed among GBSs bound at higher DEX concentration. Conversely, if clustering reflects binding that is coordinated across a locus, then we expect the sites bound at 5 nM DEX to be closer to each other than expected if an equal number of sites are randomly chosen from the set of sites bound by the GR at 50 nM DEX treatment. The median minimal distance between sites bound by the GR at 5 nM DEX was 11 kb. In contrast, randomly distributing the number of sites that are bound at 5 nM DEX across the larger set of sites bound at 50 nM DEX and recalculating the median minimal distance between bound sites produced a distribution of distances with a median of 23 kb (**Figure 18**). The sites bound by the GR at 5 nM DEX were significantly closer to each other than expected according to that permutation analysis ($z = -20.1$, $p = 1 \times 10^{-90}$). Clustered low dose GR sites should also have fewer stretches of unbound potential GR sites among the possible 50 nM DEX GR sites than expected by chance. We calculated the number of contiguous regions unbound by the GR at 5 nM DEX and compared that to the number of contiguous unbound stretches in our shuffled background model. Consistent with our hypothesis, we found

that there were significantly fewer GR unbound stretches than expected by chance (Z-test, Z = -12.9, p = 1.25 x 10^{-38}, **Appx. 1; Figure 48A**). The same clustering behavior was tested and confirmed for 0.5 nM DEX GR sites in terms of distance to nearest adjacent GR binding site and the number of contiguous unbound regions (Z-test, Z = -8.8, p = 8.95 x 10^{-19}; Z = -11.6, p = 1.16 x 10^{-31}, respectively). Finally, after controlling for GR ChIP-seq counts, we did not find that GBSs bound first at low doses were enriched for DEX-responsive sites (0.5 nM, p = 0.95; 5 nM p = 0.30). These results show that GR binds genomic loci in a dose-specific coordinated manner and not as series of independent binding events.

2.2.6 CTCF is depleted within GBS clusters

The three-dimensional chromatin structure of the genome is organized into topological domains that can functionally separate clusters of regulatory elements from surrounding genomic regions. CTCF is known to bind at and demarcate the boundaries of those topological domains (Botta et al., 2010; Dixon et al., 2012). In that manner, CTCF can insulate the activity of adjacent regulatory elements in the genome. If clustered GR binding results from physically interacting sites, we expect CTCF to be depleted between adjacent pairs of GBSs bound at 5 nM DEX. Indeed, 33% percent of the pairs of GBSs that were both occupied at 5 nM DEX had an intervening CTCF binding site in A549 cells. As a control, we permuted which pairs of all possible 50 nM GBSs had an intervening CTCF site. That model resulted in a greater percentage of sites (47%) with an

intervening CTCF (Z-test, Z = -22.4, $p < 1 \times 10^{-100}$, **Figure 19**). The same result was replicated with 0.5 nM low dose DEX GR sites (Z-test, Z = -2.4, $p = 8 \times 10^{-3}$). Those results show that CTCF is substantially and significantly depleted between coordinately bound GBSs.

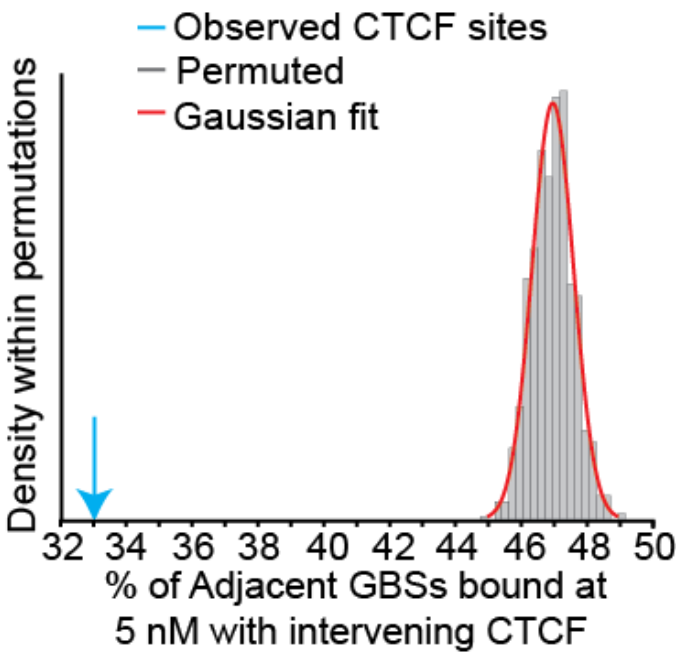

Figure 19: GBS clusters are less likely to span CTCF insulated domains than expected by chance.

Percent of GBSs at 5 nM DEX dose with intervening CTCF sites (blue arrow) and the distribution of a null model in which the location of sites were permuted across possible binding sites occupied across doses and then assayed for the presence of an intervening CTCF binding site.

2.2.7 Tethered GBSs cluster around direct GBSs in the genome

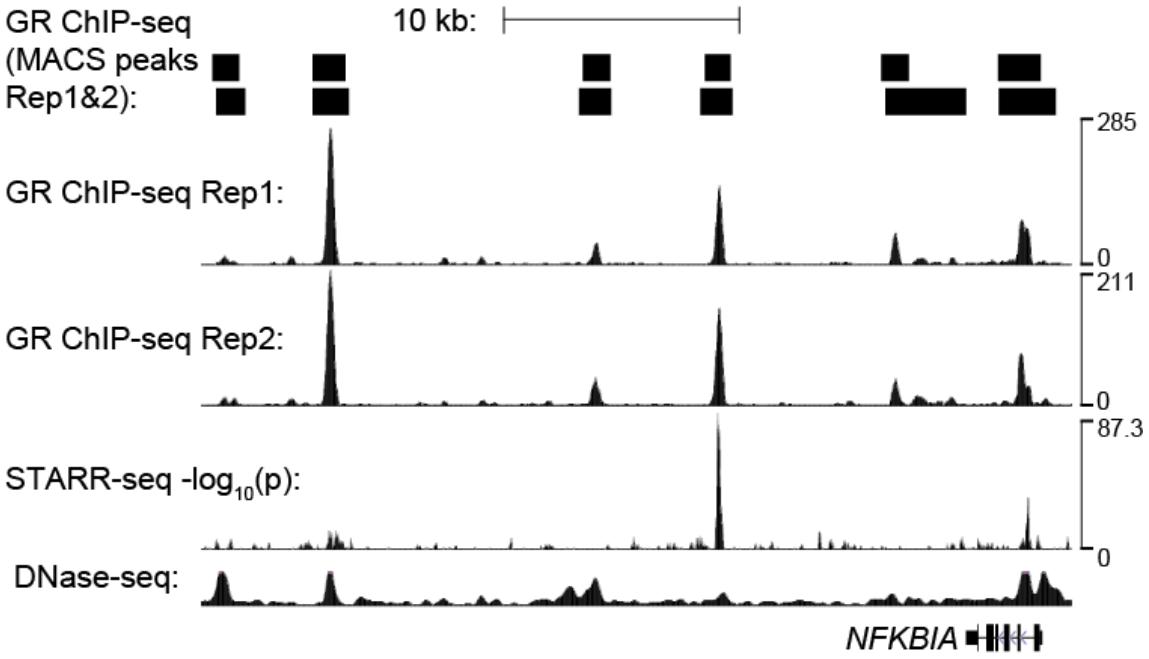

Figure 20: DEX induced enhancer function across a GBS cluster at the *NFKBIA* locus.

An example of a cluster of GBSs in A549 cells near the *NFKBIA* gene that is nucleated by a GC inducible enhancer. A549 DNase-seq under control conditions.

We frequently observed that a small fraction of GBSs in a cluster had reporter activity, suggesting that GBSs in a cluster may serve different functions. For example, there are six GBSs clustered near the *NFKBIA* gene, but only two of the GBSs were DEX-responsive in ChIP-reporter assays (**Figure 20**). We therefore performed an analysis to determine if DEX-responsive GBSs are generally depleted within clusters. We first defined GR binding clusters by identifying groups of GBSs with less than 5 kb between each site. We then quantified the average DEX-responsive activity of the GBSs in each

cluster. If DEX-responsive reporter activity was evenly distributed across all GBSs in the genome, we would expect there to be no relationship between cluster cardinality and average DEX-responsive reporter activity. Instead, there was an inverse relationship such that singular clusters had greater DEX-responsive reporter activity, and the most-populated clusters had the least average reporter activity (**Figure 21A**)**.** Permuting reporter activity across GBSs and repeating the analysis shows that such a trend was unexpected to occur by random. That result was general to the window size used (**Appx. 1; Figure 48B and 48C**). We observed the same trend when evaluating the average number of DEX-responsive GBSs per cluster rather than the average activity of those GBSs (**Appx. 1; Figure 48E**). The fraction of clusters that harbor at least one DEX responsive GBS was enriched for small clusters, another indication that GBSs that are isolated in the genome are more likely to function autonomously in a reporter assay (**Appx. 1; Figure 49**). Finally, to show that the result was not dependent on our approach to define clusters, we found that physically isolated GBSs were more likely to have stronger DEX-responsive activity (**Appx. 1; Figure 48D**). We conclude that clusters of GBSs consist of a mixture of DEX-induced and non-DEX-induced GBSs, that large clusters are depleted for DEX-induced GBSs, and that small clusters or isolated GBSs are enriched for DEX-responsiveness.

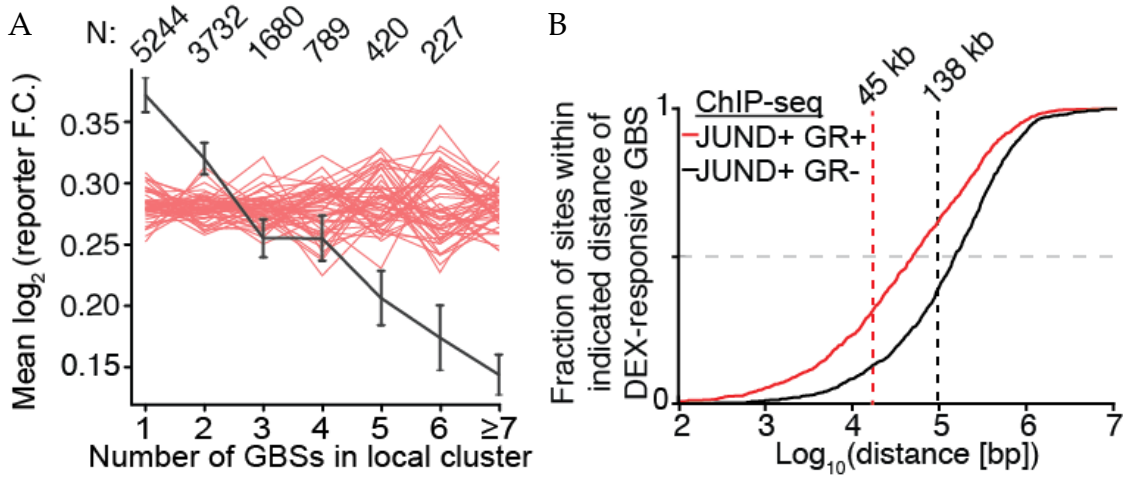

Figure 21: GBS clusters consist of DEX-induced GBSs surrounded by tethered GBSs.

(A) Each GR binding site was assigned to a cluster by grouping together sites within 5 kb of each other. The GBSs were then divided into bins based on the cardinality of their cluster. The mean and standard error of the per-cluster DEX-induced reporter activity was then plotted as a function of cluster cardinality. Red lines are 50 permutations of the reporter activity across GR binding sites. Numbers of clusters of each size indicated at top. **(B)** The cumulative distribution of JUND binding sites that are either bound or unbound by GR (black and red lines, respectively) in A549 cells treated with 100 nM DEX for 1 h as a function of proximity to the nearest motif encoded GR binding site.

Based on those observations, we hypothesized that direct GBSs can nucleate cluster formation by binding to the genome and then interacting with other transcription factors bound beyond the range of immediately proximal TF-TF composite activity, resulting in the appearance of non-GRE driven tethered GBSs. To evaluate that model, we focused on interactions between direct GBSs and sites where GR is tethered to AP-1. As reported previously, we found substantial co-occupancy of GR and the AP-1 subunit JUND after DEX treatment (Reddy et al., 2012a). In our analysis, 2,982 (39%) of the JUND binding sites present prior to DEX treatment were bound by GR after DEX

exposure. Conversely, 1,992 (64%) of the GBSs occurred at sites also bound by JUND in the presence of DEX. If GR binding directly to the genome via a GRE coordinates GR binding at nearby AP-1 sites, we expect that JUND sites that also bind GR after DEX treatment would be closer to direct GBSs. Indeed, we found that JUND binding sites bound by the GR after DEX treatment were substantially closer to direct GBSs than expected by the genomic distribution of JUND binding (median distance 45 kb vs. 138 kb**,** Mann Whitney U-test $p < 7.46 \times 10^{-41}$, **Figure 21B, Table S6**), suggesting that direct GR binding coordinates tethered binding nearby. That finding agrees with a study that used a dominant negative form of AP-1 to show that GR binding at AP-1 co-bound sites was AP-1 dependent, but that GR binding at direct sites was not (Biddie et al., 2011). A recent study demonstrated that sites enriched for distal P300 interactions are enriched for AP-1 binding motifs and that P300 interactions gained after GC-exposure are enriched for the GRE (Kuznetsova et al., 2015). These results suggest that GR binding directly to the genome explains local clustering of AP-1 tethered GR binding sites across the genome.

2.2.8 Direct GBSs recruit AP-1 binding to genomic sites that lack AP-1 recognition motifs

Just as we observe GR occupancy enriched at AP-1 sites that are close to direct GBSs, we also observe gains of AP-1 occupancy enriched at direct GBSs. After DEX treatment, JUND binding was gained at 352 sites, maintained at 2,629 sites and lost at 4,982 sites (**Appx. 1; Table S7**). The majority of gained JUND binding sites (83%) co-

occupied regions bound by the GR. Meanwhile, 56% of maintained JUND binding sites and 25% of lost JUND sites overlapped with GBSs (gained vs. maintained, Fisher's exact test $p = 1.12 \times 10^{-24}$, gained vs. lost, Fisher's exact test $p < 1 \times 10^{-100}$, **Figure 22A**). DEX-induced JUND binding sites were substantially closer to GBSs than maintained JUND binding sites (median distance 0.37 kb vs. 5.23 kb; Mann Whitney U-test $p = 4.83 \times 10^{-15}$) and JUND sites lost after DEX treatment (0.37 kb vs. 21.81 kb; Mann Whitney U-test $p = 5.28 \times 10^{-37}$; **Figure 22B**). The gains and losses of AP-1 were largely determined by whether the GBS had DEX-responsive reporter activity. Of the 352 AP-1 sites that were induced by DEX, 41% overlapped GBSs that were DEX-responsive in reporter assays. Meanwhile, only 6% of the 2,629 sites that maintained JUND occupancy overlapped with reporter-responsive GBSs (Fisher's exact test $p = 2.65 \times 10^{-62}$, **Figure 22C**).

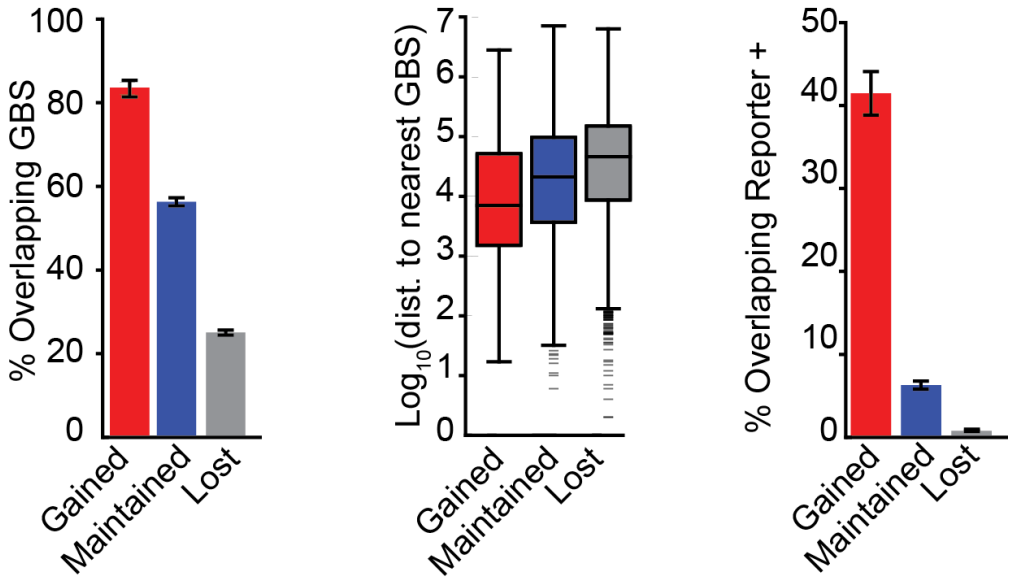

Figure 22: Comparisons of various and sequence genomic features among JUND binding sites that are gained, maintained, or lost with DEX treatment.

(A) Percent that overlap with GBSs. **(B)** Distance between to the nearest non-overlapping GBS. **(C)** Percent that overlap DEX-induced GBSs.

We investigated whether GR and AP-1 DNA binding motifs explain the gains and losses of AP-1 with DEX treatment. We computed the distribution of motif scores for the AP-1 binding motif within gained, maintained, and lost JUND sites. The distribution of motif scores in each class was bimodal, suggesting distinct classes of AP-1 motif driven and non-AP-1 motif driven JUND binding (**Figure 23A, Appx. 1; Figure 50A**). Next, we computed GRE motif scores at each class of JUND binding site. While the majority of JUND binding sites had GREs consistent with a background model in which sequences were shuffled prior to motif finding, many novel JUND binding sites contained strong matches to the GRE (**Figure 23B**). When two-component Gaussian

mixture models were fit separately to the GR motif scores of JUND gained, JUND maintained, and JUND lost sites, only JUND gained sites had a mixture component with a significantly greater GR motif score than that of matched dinucleotide shuffled sequences (Z-test, $p = 3 \times 10^{-3}$) (**Appx. 1; Figure 50B**). We estimate that 47% of the gained JUND sites belonged to the class of sites with greater GR motif score. These results suggest that, just as GR binding to the genome at tethered sites corresponds with enrichment for AP-1 motifs, AP-1 can be observed binding at strong GREs after DEX exposure.

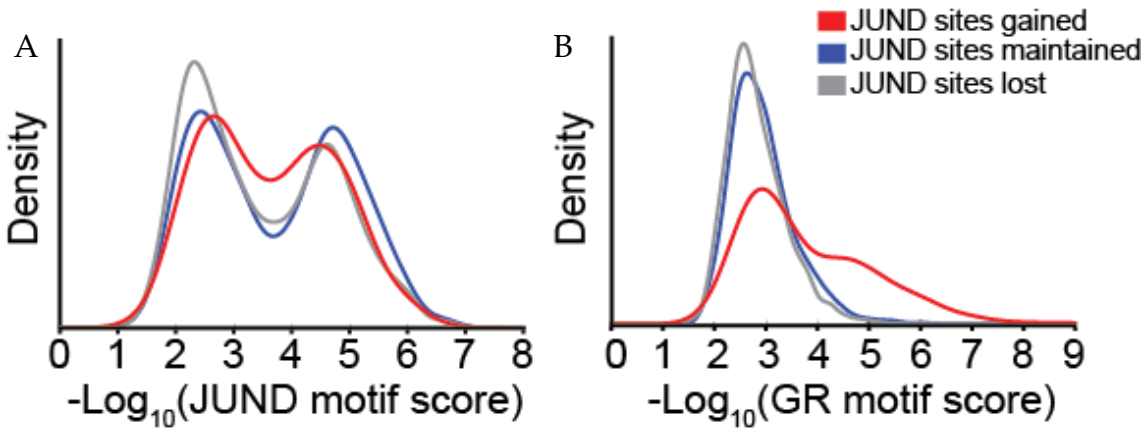

Figure 23: Evidence for shadows of JUND binding at GBSs.

(A) Distribution of AP-1 motif scores at JUND binding sites that overlap DEX-induced GBSs. **(B)** GRE motif scores of JUND binding sites that overlap DEX-induced GBSs.

2.2.9 Epistatic interactions between GR and AP-1 modulate DEX-responsive regulatory activity

It is well established that GR and AP-1 act together to regulate gene expression [e.g. (Diamond et al., 1990; Herrlich, 2001; Ratman et al., 2013)], and that AP-1 plays a

major role in determining where GR binds in different cell types (Biddie et al., 2011; Gertz et al., 2013). Our GR ChIP-reporter experiments and subsequent analyses have shown that AP-1 sites are not sufficient for DEX-responsive regulatory activity in a reporter assay, yet are clustered around direct GR binding sites in the genome. Previous studies have demonstrated that GR and AP-1 binding immediately adjacent to each other can interact and synergistically amplify the transcriptional response to glucocorticoids. We confirm those results here (**Appx. 1; Figure 51A-C, 52A and B; Table S3**) (Mittal et al., 1994; Pearce et al., 1998). Our clustering analysis also suggests that GR and AP-1 interactions occur over tens of kilobases (**Appx. 1; Figure 4**). One possible explanation is that direct GBSs interact with such AP-1 binding sites via DNA looping, and that those interactions also alter the GC response. To distinguish between immediately adjacent binding and looping, we varied the spacing between a DEX-induced GBS and a canonical AP-1 response element, and again tested for synergy between the sites (**Figure 24 Appx. 1; Figure 53A-C**). Consistent with previous studies of GR and AP-1 binding immediately adjacent to each other (Pearce et al., 1998), AP-1 amplified the effects of GC-mediated gene activation ~20-fold when the AP-1 binding motif was between 23 and 123 nt away from the GRE. The extent of synergistic activation was decreased to a minimum of 3.5 fold induction when the AP-1 response element was between 143 and 183 nucleotides from the GRE. That decrease agrees with previous observations (Pearce et al., 1998), and is consistent with the axial stiffness of

DNA inhibiting interactions between GR and AP-1 (Lee and Schleif, 1989). Finally, reporter gene expression increased to a maximum induction of 58-fold when the interval between the GRE and the AP-1 binding site was between 203 and 243 nucleotides (**Figure 24, Appx. 1; Figure 53C**). That interval is substantially more than the estimated 150 nt persistence length of DNA, and is therefore consistent with DNA looping between the GR and AP-1 site (Lee and Schleif, 1989). These data provide evidence that AP-1 binding sites could interact distally with GC-inducible GR occupied enhancers to increase the potency of GC-responsive regulatory elements.

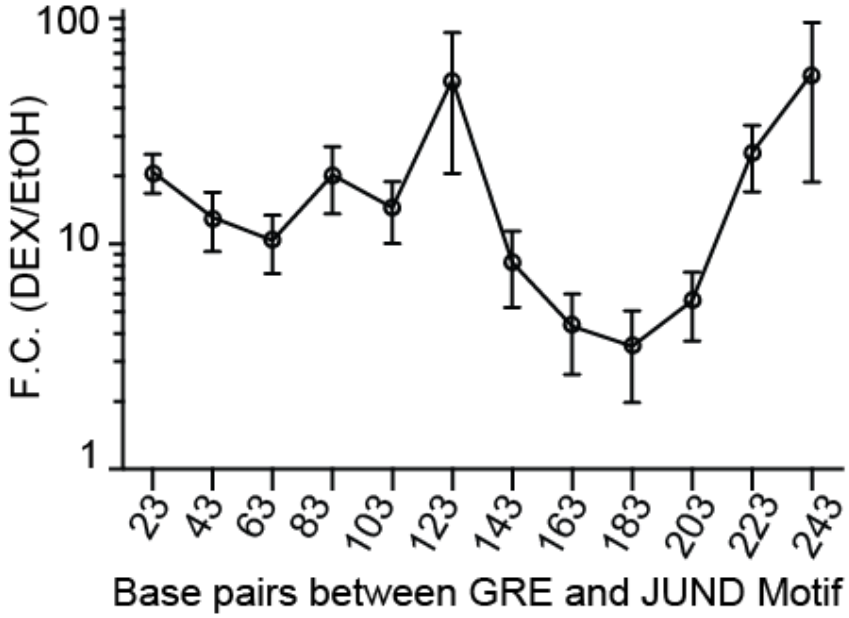

Figure 24: AP-1 and GR can co-activate reporter gene expression beyond the range of direct interactions.

Dual luciferase assays in A549 cells treated with 100 nM DEX or vehicle control using plasmids with increasing stretches of non-DEX-responsive DNA between the *TSC22D3* DEX-inducible enhancer and a canonical AP-1 binding motif.

2.2.10 Distal binding cluster interactions are a general mechanism of gene regulation

Many TFs bind to the genome in clusters, raising the possibility that our model of TF clustering via interactions between direct and tethering sites may be a general phenomenon. To address this possibility we analyzed published genomic datasets generated to study the ER (Hurtado et al., 2011; Joseph et al., 2010). Previous studies have demonstrated that the ER engages in cooperative interactions with both AP-1 and FOXA1. In those studies, AP-1 and FOXA1 have been described as either pioneer factors that increase chromatin accessibility prior to ER binding at composite sites or, as tethering factors that indirectly bind ER to the genome. Alternatively, we hypothesized that like the GR, the ER binds to the genome in clusters that reflect interactions between direct ER binding sites and nearby AP-1 or FOXA1 binding sites. As we observed in the case of the GR, we found that direct ER binding sites were substantially closer to JUN binding sites that become ER bound than the genomic distribution of JUN sites (**Figure 25A**, 19 kb vs. 51 kb, Mann Whitney U-test $p = 3.97 \times 10^{-14}$).

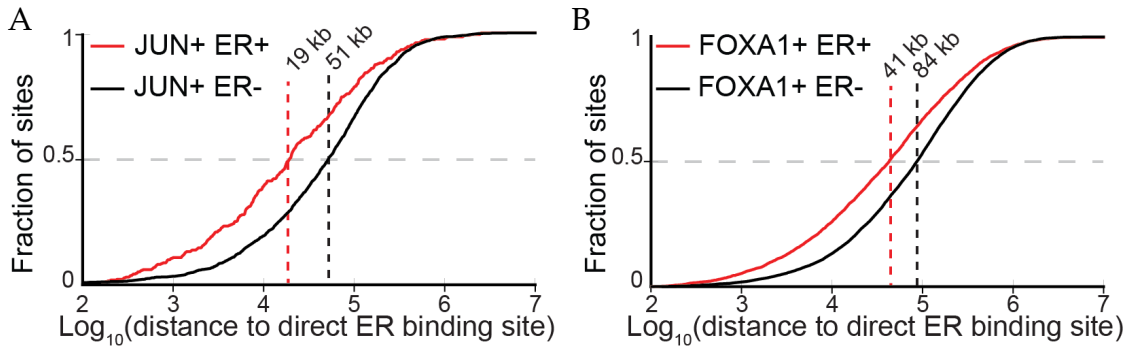

Figure 25: JUN and FOXA1 co-bound ER binding sites are closer to ERE encoded ER binding sites than expected by chance.

(A) The cumulative distribution of JUN binding sites that are either bound or unbound by ER (black and red lines, respectively) as a function of proximity to the nearest ERE motif encoded ER binding site. **(B)** The cumulative distribution FOXA1 binding sites that are either bound or unbound by ER (black and red lines, respectively) as a function of proximity to the nearest ERE motif encoded ER binding site.

Likewise, we found that direct ER binding sites were substantially closer to FOXA1 binding sites that became bound by the ER after estrogen treatment than the genomic distribution of FOXA1 binding sites (**Figure 25B,** 41 kb vs. 84 kb, Mann Whitney U-test $p < 2.06 \times 10^{-71}$). Analysis of published ER ChIA-PET data revealed that at sites bound by the ER and FOXA1 or sites bound by the ER alone, the percentage of sites with distal chromatin interactions was directly correlated with ERE strength (**Figure 26A, 26B**) (Fullwood et al., 2009). These analyses demonstrate that analogous to the GR, the ER participates in distal interactions with tethering associated proteins nucleated by the presence of an ER binding site with a high scoring ERE match.

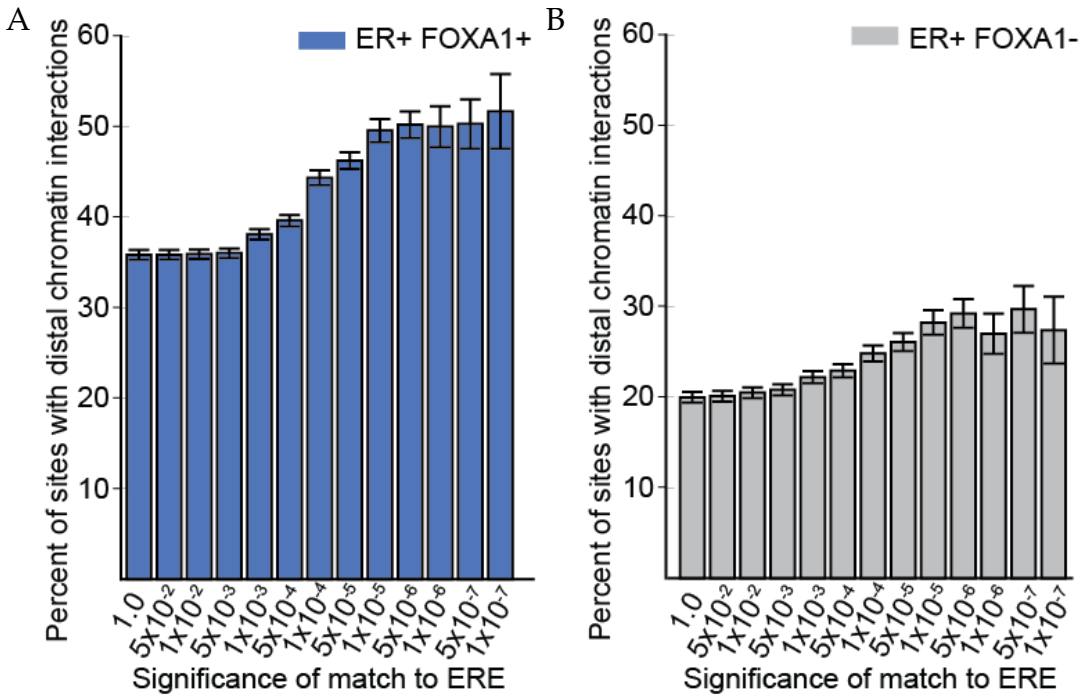

Figure 26: ER binding site distal chromatin interactions as a function of ERE motif match.

(A) Prevalence of distal chromatin interactions at ER binding sites co-bound by FOXA1 binned by the significance of the best scoring match to the ERE contained in the binding site. **(B)** Distal ER interactions as in panel C, for sites with no evidence of FOXA1 binding.

2.3 Discussion

Genome-wide mapping of TF-DNA interactions at the whole-genome scale has revealed a vast excess of TF binding sites relative to the number of regulated genes (Gao et al., 2013; Reddy et al., 2009). Using a comprehensive empirical analysis of the regulatory activity of each GBS, we have developed an enhancer-cluster model in which direct GBSs interact with co-binding TFs kilobases away via protein-protein interactions (**Figure 27**). We have predominantly focused on the well-known interactions that occur

between GR and AP-1. Several models have been proposed to explain those interactions such as AP-1 recruiting GR to sites that lack a GRE (Teurich and Angel, 1995) or AP-1 enabling GR to bind weak versions of a GRE (Biddie et al., 2011; John et al., 2011). Our model expands on those possibilities by suggesting that GR also binds AP-1 sites via chromatin loops forming between AP-1 sites and distinct direct GR binding sites. We also show that such interactions are necessary for the DEX-responsiveness of some GBSs observed with ChIP-seq. Our model predicts the reciprocal effect of AP-1 binding at direct GREs upon DEX treatment, which we observe (**Figure 23B**). Similarly, a recent study showed recruitment of FOXA1 to GR and ER binding sites (Swinstead et al., 2016). Our interaction model helps to explain the following observations: (i) the discrepancy between the number of TF binding sites and the number of regulated genes, (ii) the fact that only a small fraction of the binding sites for a given TF have a DNA binding motif for that TF (Gertz et al., 2013), and (iii) the distinct patterns of epigenetic remodeling at direct and tethered GBSs. Genetic studies have also shown that it is common for single nucleotide variants to disrupt TF binding without direct binding motifs for that TF (Reddy et al., 2012b; Soccio et al., 2015). According to our model, an explanation is that the variant disrupts interactions between TFs, thus leading to distal regulatory perturbations.

Previous studies have demonstrated that AP-1 tethered GBSs vastly outnumber direct GBSs and that tethered sites are enriched in cell-type-specific GR binding. Based on those two findings, tethered sites are thought to play a major role in cell type-specific transcriptional GC responses (Gertz et al., 2013). Meanwhile, our results show that while many tethered GBSs act as steady-state enhancers, only direct GBSs initiate GC-responsive enhancer function. A potential resolution to that discrepancy is that distal interactions between direct GBSs and tethered sites tune the activity of both cell-type-specific and cell-type-shared direct GBSs. Each direct GBS can interact with multiple nearby tethering sites. Thus, cell-type-specific tethered binding sites outnumber the cell-type-specific direct binding sites that nucleate GR binding clusters. Given the extent of amplification of reporter activity that we observe, looping interactions between direct GBSs and tethering TFs may be a major contributor to the transcriptional response to GCs. It is unclear if clusters of interacting GBSs are the result of multiple tethered sites interacting with a single direct GR binding site, or if the observation of clustered GR binding is the result of the aggregated ChIP-seq signal from a population of cells each with binary pairs of distal GR/AP-1 interactions. Consistent with our evidence, separating direct and tethered GBSs minimally improves the ability to predict GC-induced genes (**Appx. 1; Figure 54**). It remains to be determined which specific aspects of the interactions between direct and tethered GBSs will best predict GC-mediated expression responses.

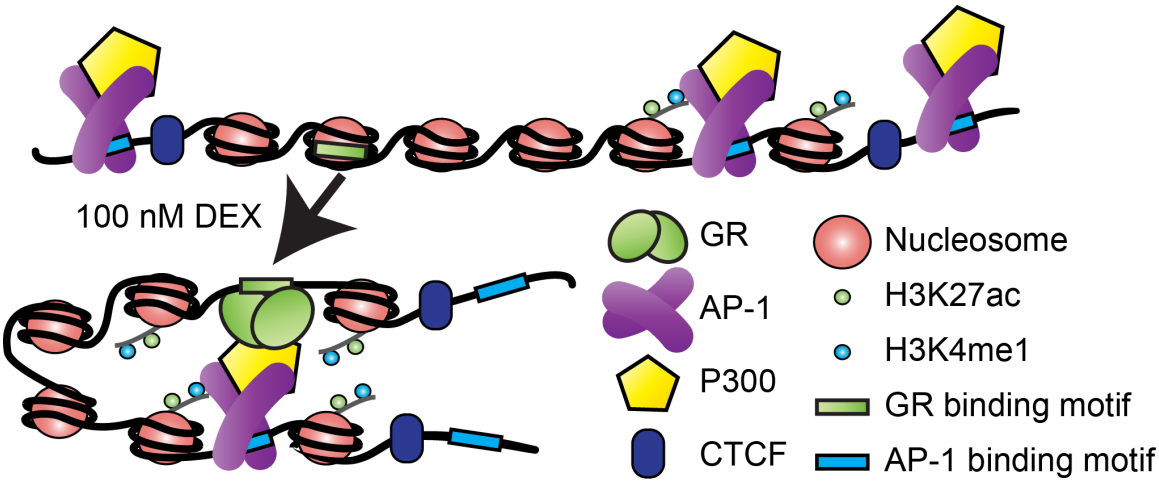

Figure 27: A revised model of GR-mediated enhancer function.

Prior to hormone exposure, AP-1 binds to the genome. AP-1 bound sites are in domains of increased chromatin accessibility that are enriched for H3K27ac, H3K4me1 and P300. After hormone exposure, GR binds directly to the genome in regions of less accessible chromatin. GR interacts with distal AP-1 bound sites within the same CTCF-defined topological domain. The epigenome of the direct GBSs is remodeled, becoming enriched for H3K27ac, H3K4me1 and increased in chromatin accessibility. Direct GBSs confer GC-inducible enhancer function, while AP-1 bound interacting sites modulate the expression output of direct GBSs.

While the results of this study demonstrate an enhancer-cluster model of GC-mediated gene induction, they do not preclude the occurrence of additional regulatory mechanisms. The specific configuration of TFs at direct and tethered sites, interactions with other TFs, covalent modification such as GR and AP-1 phosphorylation, DNA methylation, heterochromatin formation, and the amount of GR in a cell are all likely to contribute to the transcriptional response to GCs. Thus, additional mechanisms may contribute to the lack of repression observed in our results, and to the cell-type specificity of some direct GBSs. For example, while including GR and AP-1 sites on the

same plasmid amplified an activating DEX response in our experiments, others have demonstrated that tethering can also result in repression of gene expression (Luecke and Yamamoto, 2005). Further clarifying the determinants and effects of interactions between direct and tethered sites will therefore likely be informative in understanding the complexities of GC-mediated gene regulation.

Finally, while the focus of this study is on GR and ER binding sites, many other TFs have similar distributions of genomic binding and binding mechanisms. It is therefore likely that our model applies to numerous additional eukaryotic TFs, and may broadly contribute to the coordination of TF binding observed across the human genome.

2.4 Experimental Procedures

ChIP-reporter Input Library Construction

ChIP-seq libraries were adapted to STARR-seq by amplifying with 8 cycles of PCR (New England Biolabs Q5, GC buffer) using primers P1F and P1R to enable cloning into the STARR-seq human screening reporter backbone as previously described (Arnold et al., 2013).

After cloning, the reaction was purified using 1.5X Agentcourt Ampure XP beads (SPRI beads; Beckman Coulter) and eluted in 10 μl water. Purified constructs were split into four 2 μl aliquots and electroporated into 60 μl of MegaX DH10B competent cells (Life Technologies) per aliquot. Cultures were recovered in 3 ml SOC medium for 1 h

and then grown in suspension in 400 ml of Luria Broth medium for 14 h. Library plasmids were purified using the Promega Pure Yield maxiprep system.

ChIP-reporter output library construction

Total RNA was prepared using the Qiagen RNeasy kit. Poly-A RNA was isolated from 70 μg of total RNA by double selecting with Dynabead Oligo-$dT_{(25)}$ beads (Life Technologies). RNA was then treated with turboDNAse (4 U) for 30 min at 37 °C (Invitrogen). DNase treated poly-A RNA was purified using the RNeasy Mini kit. Plasmid specific cDNA was synthesized using Superscript III (Life Technologies) incubated for 2.5 h at 55 °C and inactivated at 70 °C for 15 min. Following synthesis, cDNA was treated with RNaseA (Sigma) at 37 °C for 30 min. cDNA was purified using SPRI beads and then amplified using a two stage PCR as described previously (Arnold et al., 2013).

Quantification of regulatory activity in ChIP-reporter assayed sites

The ChIP-reporter output libraries were sequenced using paired-end 25 bp reads on an Illumina MiSeq.

Cell Culture

A549 cells were grown at 37 °C and 5% CO2 in F-12K medium with 10% fetal bovine serum and 1% penicillin-streptomycin. Cells were treated by adding 0.02% by volume of 5 mM DEX directly to the cell culture media. As a vehicle control, cells were treated in parallel with 0.02% by volume EtOH.

Chromatin Immunoprecipitation (ChIP) and ChIP-seq

Chromatin immunoprecipitation was performed as previously described (Reddy et al., 2009) using 2 x 10^7 A549 cells per replicate. Cells were sonicated using a Bioruptor XL (Diagenode) on the high setting until the resulting chromatin was fragmented to a median fragment size of ~250 nt as assayed by agarose gel electrophoresis. ChIP was performed using the 5 μg rabbit polyconal GR antibody (Santa Cruz Biotechnology sc-1003), and 200 μl of magnetic sheep anti-rabbit beads (Life Technologies M-280). After reversal of formaldehyde crosslinks at 65 °C overnight, DNA was purified using MinElute DNA purification columns (Qiagen). Illumina sequencing libraries were then generated using the Apollo 324 liquid handling platform according to manufacturer's specifications (Wafergen).

GR ChIP-seq analysis

GR ChIP-seq was performed in biological duplicates after 3 h treatment with either 100 nM DEX or 0.02% by volume EtOH. The libraries were sequenced on an

Illumina MiSeq using single-end 50 bp reads. The sequencing reads were aligned to the hg19 version of the reference genome using Bowtie (Langmead and Salzberg, 2012), and binding sites were called in the DEX-treated samples relative to the EtOH-treated samples using the MACS peak calling software (Feng et al., 2012). Peaks were then split into subpeaks using PeakSplitter (Salmon-Divon et al., 2010), and set to a fixed size of 500 bp by extending 250 bp to either side of the sub-peak summit. To combine peaks between replicates in an inclusive manner, subpeaks between replicates were merged using mergeBed from BEDTools (Quinlan, 2014), requiring a 250 bp overlap to combine peaks. The resulting peaks were again fixed to a size of 500 bp by extending 250 bp upstream and downstream from the peak midpoint.

External ChIP-seq data

ChIP-seq data for histone modifications and TF binding in A549 cells was obtained from the ENCODE Data Coordination Center (ENCODE, 2012), the cistrome database (Liu et al., 2011) and the Gene Expression Omnibus (**Table S8**).

Cell transfection and library harvesting

ChIP STARR-seq input libraries were combined in equimolar pools and transfected into T-75 (BAC STARR-seq) or T-150 flasks (ChIP STARR-seq) of A549 cells with Fugene HD (Promega) at a 4:1 ratio of Fugene:DNA. Three replicate transfections

were performed per experimental condition. Cells were treated for 3 h with either 100 nM DEX or 0.02% by volume EtOH. Cells were rinsed with PBS pH 7.4 and incubated for 3 min at 37 °C with DNAse I (5 mg DNAse I in 1 ml of buffer containing 10 mM Tris-HCl pH 7.5, 150 mM NaCl and 1 mM MgCl in DEPC treated water diluted to a total volume of 24 ml in PBS). Cells were rinsed again with PBS and then dissociated with Trypsin-EDTA 0.25% (Life Technologies). Trypsin was neutralized with A549 tissue culture medium and cells were pelleted via centrifugation. Cell pellets were rinsed once with PBS and then lysed in 2 ml of RLT buffer (Qiagen) with 2-mercaptoethanol (Sigma).

STARR-seq Input Library Construction

Six BACs (RP11-806F7, RP11-435L21, RP11-139K17, RP11-788A16, CTD-2340K24, RP11-769H22) that contain previously identified DEX-responsive genes (Reddy et al., 2009) were amplified in E. coli and the resulting DNA was prepared using the NucleoBond BAC 100 kit (Macherey-Nagel). PCR was used to validate that the prepared DNA was from the expected genomic regions. The DNA from each BAC was pooled in equimolar ratios for tagmentation (Illumina) (Adey et al., 2010; Gertz et al., 2012). The tagmentation reactions were performed in six reactions, each with 50 ng of pooled BAC DNA. Three reactions used 1 µl of transposase, and three used 5 µl of transposase. The reactions were incubated at 55 °C for 5 min and then moved to ice. Reactions were neutralized using 30 µl QG buffer (Qiagen) and purified using SPRI beads at a 1.125X

SPRI:reaction ratio. Purified reactions were eluted in 22 μl of nuclease free water (Sigma). The resulting tagged fragments were then PCR amplified to add 15 bp of sequence matching the STARR-seq backbone as for CHIMERA. Specifically, the PCR reaction was primers with primers SSV-nxt-F and SSV-nxt-R, and the DNA was amplified with Phusion High fidelity polymerase with GC buffer (New England Biolabs). The reactions were incubated at 72 °C for 3 min; 98 °C for 30 s; followed by 10 cycles of (98 °C for 10 s, 63 °C for 30s and 72 °C for 3 min). The resulting products were purified using SPRI beads at a 1.8X SPRI:reaction ratio.

The STARR-seq screening vector was digested overnight with SalI and AgeI and linearized backbone was purified with the Wizard SV Gel and PCR Clean-Up kit (Promega). 100 ng of backbone and 17 ng pooled insert were cloned in three 10 μl Infusion HD reactions (Clontech). Infusion reactions were then pooled and electroporated into MegaX DH10B electrocompetent cells at a ratio of 4 μl reaction to 20 μl of competent cells for a total of six electroporations. Transformations were recovered for 1 h in SOC medium while shaking (225 rpm, 37 °C) and then grown for 14 h in 250 ml of Luria Broth while shaking (225 rpm, 37 °C). The resulting STARR-seq input libraries were then purified using the Promega Pure Yield Maxiprep kit.

To assess fragment diversity in the STARR-seq input libraries, the fragments inserted into each was sequenced on an Illumina MiSeq. 10 ng of each input library was PCR amplified using indexed Nextera primers and Phusion DNA polymerase in GC

buffer (New England Biolabs). The following thermal cycling protocol was used: 98 °C for 30 s, followed by 10 cycles of (98 °C for 10 s, 65 °C for 30 s, 72 °C for 2 min), with a final extension at 72 °C for 7 min. PCR products were purified using SPRI beads (1.8X SPRI:DNA ratio) and sequenced on an Illumina MiSeq Instrument using 25 bp paired end reads.

Estimating library diversity in reporter libraries

The number of unique fragments in the ChIP STARR-seq input library was estimated by sequencing on an Illumina MiSeq instrument and estimating the point of saturation. Plasmids were amplified using standard full-length Illumina sequencing adapters and sequenced using 25 bp paired end sequencing. Colony PCR was performed on individual clones of the vector plated from the initial transformation to confirm cloning efficiency. ChIP STARR-seq insert sequence data was aligned to the hg19 version of the human genome using Bowtie (Langmead and Salzberg, 2012), and converted to fragments using SAM tools and an in-house Perl script. Insert diversity was analyzed using a custom perl script, and the number of fragments per library was estimated by B_{max} in a model that accounts for saturation and non-specific sequencing errors (Y = B_{max} X/(K_d active X) active NS X active Background). The model was fit using the Prism software package.

STARR-seq output libraries construction

RNA from transfected A549 cells was prepared as for ChIP STARR-seq output libraries. Reverse transcription was performed in 50 μl reactions using SuperScript III scaled up from the manufacturer's protocol. Reactions were primed using Oligo(dT)$_{12-18}$ primers (Life Technologies) and incubated at 42 °C for 1 h, 50 °C for 90 min and then inactivated at 70 °C for 15 min. cDNA samples were purified using SPRI beads (1.8X SPRI:DNA ratio). Purified cDNA reactions were then PCR amplified using Phusion polymerase with GC buffer under the following conditions: 98 °C for 30 s, 25 cycles of (98 °C for 10 s, 65 °C for 30 s, 72 °C for 2 min) with a final extension at 72 °C for 7 min.

Quantification of regulatory activity in ChIP-reporter assayed sites

Reads were aligned to the hg19 build of the human genome using Bowtie (Langmead and Salzberg, 2012), and the number of aligned reads per GR binding site was determined. The DEX-responsive regulatory activity of each site was evaluated using DESeq2 where read counts at each site were normalized by the total number of aligned reads in each library, and statistically significant changes in regulatory activity after DEX treatment were evaluated using a negative binomial model of the normalized read counts (Love et al., 2014b).

Nuclear localization of plasmids

A549 cells were transfected using Fugene HD in 10 cm dishes as per manufacturer's protocol. Six transfections were performed. After 45 h, 3 plates were treated with 100 nM DEX and 3 plates were treated with 0.02% EtOH for vehicle control. Cells were harvested 3 h after treatment. Purified nuclei were harvested using Sigma's Nuclei EZ Prep Kit according to manufacturer's instructions. Purified nuclei were suspended in Qiagen Plasmid DNA mini-prep buffer P1 rather than nuclei storage buffer and processed according to standard bacterial plasmid mini-prep protocol. Sequencing libraries were generated via PCR using the same protocol used to assay input libraries. Plasmid diversity sequencing was performed using a single lane on an Illumina MiSeq with 25 bp paired end reads.

Luciferase reporter assay validations

GR binding sites were selected from regions of the genome near DEX-responsive genes containing GR binding sites assayed by both ChIP STARR-seq and BAC STARR-seq. Elements were selected to be ~400 bp in length. The elements were PCR amplified from pooled BACs (see "STARR-seq Input Library Construction") using Q5 polymerase and the following cycling conditions: 98 °C for 30 s, 35 rounds of (98 °C for 10 s, 58 °C for 30 s, 72°C for 2 min), and a final extension at 72 °C for 7 min. Primers were designed to anneal to the target region and add 15 nucleotides that match either the HindIII or EcoRV side of the multiple cloning site of the pGL4.24 minimal promoter firefly

luciferase expression vector (Promega) to enable Infusion cloning. Appropriate PCR band size was determined by electrophoresis and amplicons were purified using SPRI beads (1.5X SPRI:reaction ratio) prior to cloning.

pGL4.24 was linearized at the multiple cloning site by incubating overnight with HindIII and EcoRV. Test elements were cloned using Infusion HD (CloneTech) and transformed into Stellar competent cells (CloneTech). Colonies were picked, grown overnight and DNA was purified using PureYield Plasmid Miniprep kits (Promega), and screened for the expected insert with Sanger sequencing.

A549 cells were plated in 96-well plates (10,000 cells/well) and transfected using Fugene HD two days after plating at a 4:1 Fugene:DNA ratio. 100 ng of each test construct and 10 ng pRL-TK *Renilla* luciferase normalization vector (Promega) were transfected per well with 12 replicates. After 24 h, six wells of cells per vector were treated with medium containing either 100 nm DEX or 0.02% EtOH control. After 4.5 h — 3 h for response and 1.5 h for protein folding (Tyedmers, Brunke et al. 1996) — cells were harvested and assayed using the Dual-glo luciferase assay system (Promega) and plates were read using a Victor3 1420 Multilabel Counter (PerkinElmer).

The ratio of firefly luciferase activity to *Renilla* luciferase activity was determined for each well. The six replicates of each condition were averaged and then induction was determined by calculating the $\log_2$(average *Renilla*-normalized luciferase activity in the DEX condition / average *Renilla*-normalized luciferase activity in the EtOH condition).

AP-1 GRE combinatory experiments

For GRE/AP-1 spacing experiments, gene blocks that encode the *TSC22D3* DEX-induced enhancer (**Appx. 1; Figure 51C**) followed by 250 bp of non-GC-inducible DNA (as assayed by STARR-seq; **Appx. 1; Figure 53B**) and a canonical JUND binding element (Mathelier et al., 2014) (MA0489.1) were synthesized (Integrated DNA Technologies). The spacing of the JUN binding motif was varied in 20 bp increments (**Appx. 1; Figure 53A**). DNA fragments were analyzed using the JASPAR database to confirm that no de novo JUN or GR binding motifs were generated while designing the constructs.

For GRE/AP-1 combination experiments, DEX responsive enhancers were synthesized in combination with JUN consensus binding motifs (**Appx. 1; Figure 51A-C, 52A and 52B**) (Mathelier et al., 2014). GC-non-responsive DNA was used as intervening sequence between sites (**Appx. 1; Figure 53B**). Motifs were mutated by altering the top scoring three nucleotides of the position weight matrix of each motif.

Twenty bp of DNA from the 3' and 5' cloning sites were added during gene block synthesis. The resulting fragments were cloned directly into the EcoRV site of the Pgl4.24 backbone (Promega) using Gibson Assembly Master Mix (NEB) according to manufacturer's instructions. GC-inducible Luciferase assays were performed as above, using 4ng pRL-sv40 per well as internal control and assayed using the GloMax multiwall luminometer (Promega). In each experiment 12 replicates were treated with

100 nM DEX and 12 replicates were treated with 0.02% EtOH. Data from these experiments are displayed in **Appx. 1; Figure 52A** and **Appx. 1; Figure 53B**.

GRE addition experiments

Two non-DEX-induced GBSs (~450 nt) were identified from the TSC22D3 and NFKBIA loci (Table S3). Sites were selected based on increased chromatin accessibility in control condition. The highest scoring match to the AP-1 motif was identified (Mathelier et al., 2014) (MA0489.1) in each site and constructs were designed and cloned (as described above; AP-1 GRE combinatory experiments) that contained t a minimal GRE identified from the PER1 locus cloned 20 nt 5′, 3′ or both 5′ and 3′ of the non-induced GBSs. A congruent set of vectors containing the mutated forms of the GRE (as described; AP-1 GRE combinatory experiments) were generated in each cloning position.

Dual luciferase assays (as described above; AP-1 GRE combinatory experiments) were used to determine the effects of adding GREs to non-DEX-induced elements.

DEX washout reporter assay experiments

The upstream *PER1* DEX-induced enhancer was synthesized (Integrated DNA technologies) and cloned into STARR-seq as described above (STARR-seq Input Library Construction). The resulting vectors were transfected into A549 cells and treated with 100 nM DEX, or .02% EtOH vehicle control for 30 minutes. Six replicates of each

condition were treated for 3 h, and 6 replicates of DEX treated cells were washed twice with medium containing EtOH vehicle and treated for an additional 3 h.

RNA was harvested (Qiagen RNeasy miniprep kit), with the addition of ERCC external control RNA cocktail added to each RNA preparation at the cell lysis step. RT-qPCR was performed using primers designed to span the splice junction of the STARR-seq transcript or ERCC external control primers (Devonshire et al., 2010) and the RNA-2-CT cDNA synthesis and SYBR green qPCR kit as per manufacturer instructions. qPCR data was analyzed using the delta-delta-cT method.

RNA-seq

Total RNA was harvested from three replicates of A549 cells treated with DEX or EtOH as previously described in main methods using the Qiagen RNease Mini kit including the on-column DNase digestion. Poly-A RNA was isolated from 1 μg of total RNA as described above, and Illumina RNA-seq libraries were prepared using the Apollo 324 library prep station according to the manufacturer's protocol (Wafergen).

DNA Sequences

The sequences of DNA primers and gene blocks used in this study are located in Appx. 1; Table S3.

Estimation of Regulatory Activity in STARR-seq Assayed Sites

BAC STARR-seq output libraries were sequenced on an Illumina MiSeq Instrument using paired-end 25 bp reads. Reads were aligned to the hg19 reference genome sequence for the target BACs using Bowtie (Langmead and Salzberg 2012), paired end reads were converted to corresponding fragments, and the number of reads per fragment was counted for each replicate. The number of reads in each experiment was then normalized to the median of the number of reads per fragment divided by the geometric mean read count of that fragment. To estimate the significance of differences between EtOH and DEX treatment, a Wilcoxon signed-rank test was performed using a sliding 1 bp window across the target BACs. For each test, the normalized read counts for the fragments overlapping the window were compared between DEX and EtOH treatment conditions, and a p value was reported.

RNA-seq Expression Analysis

RNA-seq reads were sequenced on an Illumina HiSeq using 50 bp single end reads. The resulting reads were aligned to version 19 of the GENCODE reference set of transcripts using Bowtie (Harrow et al., 2006; Langmead and Salzberg, 2012). Differential expression was called using DESeq2 (Love et al. 2014) at a false discovery rate (FDR) < 5% (Benjamini and Hochberg, 1995).

Estimating Proximity Between GR Binding Sites and the TSSs

Transcriptional start site coordinates used in ChIP STARR-seq—gene proximity analyses were obtained from version 19 of the GENCODE annotation for genes that were differentially expressed after treatment for 3 h with 100 nM DEX (FDR < 5%) (Harrow et al., 2006). The ChIP STARR-seq-TSS distances used for empirical cumulative density functions were found using the closestBed function from BEDTools (Quinlan, 2014). GC–induced ChIP STARR-seq sites were defined as having an FDR < 5% and non-GC–induced ChIP STARR-seq sites were defined as having an FDR > 30%. Differing sizes of GC–induced ChIP STARR-seq and non-GC-induced ChIP STARR-seq groups could confound the distribution of distances of the closest ChIP STARR-seq element to a given TSS. To account for these sample size effects, elements from the non-GC-induced ChIP STARR-seq group were randomly sampled and distance calculations were performed 100 times to match the sample size of the GC-induced ChIP STARR-seq group. The mean and standard deviation of these sampled calculations are shown.

Regression Model to Predict Expression Response from ChIP-seq Signal

The association between GR occupancy and proximal gene expression in response to DEX treatment was examined as follows. First, ChIP-seq signal (log fold-change in ChIP-seq in response to DEX) was summed across sites within a maximum distance D from the TSS of each DEX-responsive gene, where values of D = 50 kb, D = 100 kb, and D = 200 kb were examined; ChIP-seq signal was summed separately for

reporter-positive and reporter-negative GR binding sites. Correlations between summed ChIP-seq signal and gene expression were computed via Spearman's rho and plotted separately for each value of D, for reporter-positive and reporter-negative sites (Appx. 1; Figure S6K).

DNA Binding Motif Identification and Scoring

Match scores and p values for motif matches in ChIP STARR-seq tested elements were found using the Motif Alignment and Search Tool (MAST) from the MEME Suite (Bailey and Gribskov, 1998). The top scoring motif occurrence was computed for each ChIP STARR-seq element and used for subsequent correlations and ROC analyses.

Mixture Model and Sequence Analysis to Distinguish Sites With and Without GREs

The set of all tested GBSs was assumed to consist in a mixture of sequences with and without GRE motifs. We fit a two-component Gaussian mixture model to the negative $\log_{10}$ p values for the best GRE motif match for all tested GBSs. The mean match scores corresponding to GRE motif and no GRE motif from the model that was fit with all GBSs were then fixed for reporter positive and reporter negative GBSs. Variances and mixture weights were estimated separately for each class of GBSs. We used the mixture weights to estimate of the proportion of each class of sequences that contained a GRE.

To identify motifs enriched in responsive and non-responsive GR binding sites with GREs, we constructed a matched set of 1,040 sites in each class that do or do not have a GRE according to our mixture model. We then performed a *de novo* motif search using MEME to find twelve enriched motifs in each set under a zero-or-one motif per sequence (ZOOPS) model. We then used Tomtom to compare the identified motifs to known TF binding motifs.

Receiver Operating Characteristic (ROC) Curve Analysis

ROC analysis was performed in R using the pROC software package (Robin et al., 2011). Putative regulatory elements were classified as enhancers if they had an FDR < 0.05 in ChIP-reporter and increased in reporter gene expression. Motif p values for tested motifs were assigned to each of the tested GR sub-peak as determined by MAST and those p values were evaluated as predictors of enhancer status. Predictor baseline was established by repeating ROC analysis following motif prediction in randomly shuffled sequences. Shuffling was performed while preserving dinucleotide frequencies, using software uShuffle (Jiang et al., 2008) version Feb21/2008.

ChIP-exo Library Construction

Human lung adenocarcinoma cells (A549s) were grown to confluence in 15 cm plates (~20 million cells) and treated with 100 nM DEX for 1 h, in triplicate. Cells were processed as described in the ChIP-seq library construction described above, through to

the LiCl washes. Chromatin-bound IgG beads were then washed with 10 mM Tris-HCl (pH 7.5) and resuspended in 30 μl of end repair mix: 1X T4 Ligase Buffer, 0.4 mM dNTPs, 1.8 U T4 DNA Polymerase, 6 U T4 PNK, and 3 U DNA Polymerase I, Large (Klenow) Fragment. Beads were incubated at room temperature for 30 minutes, and then washed with 10 mM Tris-HCl (pH 8). Beads were then resuspended in 30 μl of A-tailing mix: 1X NEB Buffer 2, 0.2 mM dATP, and 3 U Klenow Fragment (3'->5' exo-). Beads were incubated at 37◎ for 30 minutes. Next, beads were washed with 10 mM Tris-HCl (pH 7.5) and resuspended in 30 μl adapter ligation mix: 1X Quick Ligase Buffer, 4 μl Quick Ligase, and 1 pmol of Illumina Adapter A. Beads were incubated at room temperature for 30 minutes. Next, the beads were washed with 10 mM Tris-HCl (pH 9.5), and resuspended in 20 μL of λ-exonuclease mix: 1X λ-exonuclease Buffer and 5 U λ-exonuclease. The beads were incubated at 37◎ for 30 minutes. Next, the beads were washed with 10 mM Tris-HCl (pH 8), and resuspended in 20 μL $RecJ_f$ mix: 1X NEB Buffer 2 and 15 U $RecJ_f$. The beads were incubated at 37◎ for 30 minutes, and then washed with TE Buffer. The beads were resuspended in 150 μL IP Elution Buffer (1% SDS, 0.1 M $NaHCO_3$), and incubated for 1 hr at 65◎, with vortexing every 15 minutes. The beads were centrifuged at 20,000xg for 3 minutes, and the supernatant was transferred to a new tube and incubated at 65◎ overnight. DNA was purified using a MinElute PCR Purification kit (Qiagen), using 11 μl of EB Buffer for the elution. Purified DNA was mixed with 8 μL of phi29 DNA polymerase mix: 1X phi29 Buffer, 10 pmol

primer P2, 375 nM dNTPs, and 4 μg BSA. The mix was heated to 95◎ for 2 minutes, 63◎ for 5 minutes, 30◎ for 2 minutes, and held at 30◎. 10 U of phi29 DNA polymerase was added to the mix, which was then heated at 30◎ for 20 minutes, 65◎ for 10 minutes, and held at 4◎. DNA fragments were purified using AxyPrep Mag PCR Clean-Up beads at a 2:1 bead volume:sample volume ratio. DNA was eluted off beads using 30 μl A-tailing mix, and incubated at 37◎ for 30 minutes. The DNA was then purified using AxyPrep Mag PCR Clean-Up beads as described above. DNA was eluted from beads using 30 μL of adapter ligation mix (with Illumina Adapter B instead of A), and incubated at room temperature for 30 minutes. The DNA was then purified using AxyPrep Mag PCR Clean-Up beads as described above, and eluted in 50 μl of PCR amplification mix: 1X Q5 Reaction Buffer, 200 nM dNTPs, 10 pmol Primer A, 10 pmol Primer B, and 1 U Q5 polymerase. PCR was carried out according to the manufacturer's specifications, using an annealing temperature of 65◎, and an extension time of 30 seconds, for 20 cycles. The PCR products were purified with AxyPrep Mag PCR Clean-Up beads with a 1:1 ratio, and eluted in 30 μL of EB Buffer. DNA concentration was measured on a Qubit Fluorometer (Life Technologies), and fragment size distribution was assessed on a TapeStation 2200 (Agilent Technologies). Libraries were sequenced on an Illumina HiSeq 2000 with 50 base-pair paired end reads.

ChIP-exo Analysis

Read pairs were aligned to the hg19 reference genome using Bowtie2 (Langmead and Salzberg, 2012) with default settings. Fragments mapping to adapter sequence or ENCODE blacklisted regions (ENCODE, 2012) were removed. PCR duplicates were removed using SAMtools (Li et al., 2009), and reads across triplicates were pooled. The position of the 5'-most base pair of the first read in each read pair was determined. This position should be indicative of where λ-exonuclease was stopped by a DNA-protein interaction. These positions were split according to sense or anti-sense strand, and positions mapping to GR ChIP-seq peaks (as described in ChIP-seq analysis methods) were used for further analysis.

DNase-seq

DNase-seq libraries were made from fresh cell cultures of the control cell line and treated cell line, with three replicates for each cell line. Library preparation and analysis was performed as described (Song and Crawford, 2010) with the modification that oligo 1b was synthesized with a 5′ phosphate to increase the efficiency of ligation.

Aggregate profile plot of DNase-seq, histone ChIP-seq in GR binding sites

We merged aligned reads across the three replicates of DNase-seq in EtOH and DEX conditions separately, from which we created a bigwig signal file. We binned signal from selected ENCODE histone modification ChIP-seq bigwig signal files and

from our DNase-seq data in 10 bp bins 500 bp upstream to 500 bp downstream of the center of each GR binding site. To produce the aggregate profile plot, we normalized by total mapped read counts (in the case of DNase-seq data) and the mean depth per base-pair covered (in the case of ENCODE data) and computed the mean and standard error of the mean in each 10 bp bin for reporter positive and reporter negative GR binding sites.

Comparison of epigenetic signal reporter positive and reporter negative GR binding sites

We merged aligned reads across the two replicates of selected ENCODE histone modification ChIP-seq datasets in EtOH and DEX conditions separately. We counted the number of reads from the above merged DNase-seq and ChIP-seq alignment files for which at least half of the read overlapped a 500 bp window centered on the best match to the GR DNA binding motif within each called sub-peak. We then normalized counts for number of mapped reads and tested for differences in GR binding. To do so, we used regression models in which the log of GR ChIP-seq summed read counts along with binary reporter activity predicted log{DEX read counts}, log{EtOH read counts}, or log{DEX read counts /EtOH read counts}. When testing specifically in the core or flanking regions, we defined the core as the -250 bp to +250 bp region centered on the best match to the GR DNA binding motif, and the flanks as the union of the -500 to -250 bp region and the +250 bp to +500 bp region.

Clustering analysis of low DEX dose GR binding sites

We merged all ENCODE ChIP-seq binding peaks for GR from 0.5 nM, 5 nM, and 50 nM DEX treatments and retained only those union peaks active at 50 nM. We then considered the clustering behavior of sets of peaks among the union that overlapped with lower DEX dose peaks (0.5 nM, 5 nM). We tested whether low dose peaks were closer to one another than expected by chance by computing the median distance for all peaks to the nearest neighboring peak. As a background model, we randomly shuffled low dose peaks among all possible peaks in the cross-DEX-dose union 1000 times and computed the median distance from each active peak to its nearest neighboring active peak for each permutation. We tested whether the set of low dose peaks had fewer stretches of contiguous unbound sites than expected by chance. Again, we randomly shuffled low dose peaks among all possible peaks in the cross-DEX-dose union 1000 times and computed the number of stretches of contiguous unbound sites. We also tested whether pairs of adjacent low dose peaks were less likely to have an intervening ENCODE DEX CTCF peak than by chance. Using the same permutation strategy, we counted the number of times a CTCF peak intervened between adjacent permuted low dose peaks. A pair of peaks was defined as adjacent at a lower dose (e.g. 5 nM) if there were no intervening peaks between the two peaks among the cross-DEX-dose union of peaks and both peaks were bound by GR at the lower dose. For all three analysis

questions above, we fit Gaussian distributions to the distribution of the metrics for the permuted data and tested whether the observed data fit the permuted distribution with a Z-test.

Distance analysis of tethering factors and nuclear receptors

We tested whether certain tethering factors (JUND, JUN; FOXA1) were closer to direct motif-encoded nuclear receptor sites (GR; ER) when co-bound by nuclear receptor. We searched for the nuclear receptor motif (JASPAR,(Mathelier et al., 2014); GR: MA0113.2; ER: MA0112.2) among all nuclear receptor genomic binding sites using MAST (Bailey et al., 2009) motif search tool and designated those sites with p value $< 1 \times 10^{-4}$ as direct motif-encoded. We then compared the set of tethering factor sites that intersect with nuclear receptor sites (e.g. JUND+,GR+) and the set of tethering factor sites that do not intersect with nuclear receptor sites (e.g. JUND+,GR-) in distance to nearest direct motif-encoded nuclear receptor site that does not intersect with a tethering factor (e.g. JUND-,GR+ with GR motif).

Motif analysis of JUND sites gained, maintained, or lost upon DEX exposure

We searched the central 200 bp of JUND DEX binding sites and GR binding sites for the JUND motif (MA0491.1) and the GR motif (MA0113.2) from JASPAR (Mathelier et al., 2014) using MAST motif search tool (Bailey et al., 2009). We also shuffled the

dinucleotides of GR binding sites using ushuffle (Jiang et al., 2008) and searched this shuffled set of sequences for the GR motif. For unshuffled sequences, two-component Gaussian mixture models were fit to the -$\log_{10}${motif p value}.

Analysis of overlap in ER and FOXA1 binding sites with established ER interactions

We took the union of all ER ChIA-PET interactions in the MCF7 cell line with q-value < 0.05 from "IHH015F" and "IHM001F" from Fullwood et al. (Fullwood et al., 2009). Using FOXA1 and ER binding peaks obtained from ChIP-seq in the MCF7 cell line, we distinguished those FOXA1 peaks that intersected ER peaks from those FOXA1 peaks that did not intersect ER peaks (Hurtado et al., 2011). For both FOXA1+,ER+ and FOXA1+,ER- peaks, we computed the best scoring p value to the ER motif (MA0112.2; JASPAR) (Mathelier et al., 2014) using MAST motif search tool (Bailey et al., 2009). We further subsetted peaks by the strength of motif and for each subset we computed the percentage of peaks that overlapped with sites that participate in interactions as identified through ER ChIA-PET.

2.5 Supporting online materials

Supplemental Tables

Table S1. Supplemental ChIP-seq data. Tab 1: The final set of all GR ChIP-seq sub-peak calls are provided as a tab-delimited file of genomic coordinates. The coordinates are for the hg19 version of the human reference genome. The sum of the reads per replicate sub-peak is included in the 4th column. Tab 2: Estimates of H3K27ac modification changes with DEX treatment at GR

binding sites. ChIP-seq data for H3K27ac and H2K27me3 in A549 cells treated for 1 h with either 100 nM DEX or 0.02% EtOH was obtained from the ENCODE Project (Supplementary methods). Results of analysis generated using DEseq2 are provided. The columns of the file are: (1) Location of the summit of the GR ChIP-seq peak call, (2) average normalized signal strength (i.e. "baseMean" from DESeq2), (3) $\log_2$(fold change) in signal between DEX and EtOH conditions, (4) the standard error of the $\log_2$(fold change) estimates, (5) the statistic used for testing (i.e. the $\log_2$(fold change) normalized for standard error in the estimate), (6) the statistical significance of a change in regulatory element activity, and (7) the associated false discovery rate. Tab 3: Estimates of H3K27me3 modification changes with DEX treatment at GR binding sites. The columns in S1 Tab 3 are organized as in S1 Tab 2.

Table S2. Estimates of DEX-responsive regulatory element activity from ChIP-reporter STARR-seq assays. The results of the DESeq2 analysis used to identify changes in regulatory activity between DEX and EtOH conditions are provided as a tab-delimited file. The columns are organized as in Table S1 Tab 2.

Table S3. Plasmid DNA and primer sequences used in this study: DNA primer sequences used for: amplifying DNA fragments used in luciferase validation assays, amplification of tagmented DNA for cloning into BAC STARR-seq assays, adapting ChIP-seq libraries to STARR-seq and ChIP-exo experiments. As well as geneblock sequences for GRE/AP-1 combination experiments

Table S4. GC-induced BAC STARR-seq results The results from our STARR-seq analysis of selected BACs is provided as a UCSC Genome Browser BedGraph file. The data column is the $\log_{10}$-transformed P-value generated from our statistical analysis (described in methods).

Table S5. RNA-seq results. RNA-seq was performed after treatment for 3 h with 100 nM DEX or 0.02% EtOH Differential expression between he two conditions was evaluated using DESeq2. Results from that analysis are provided. Columns are as for Supplementary Dataset 1.2 except that the first column is the GENCODE V19 transcript ID instead of genomic coordinates.

Table S6. Match scores for DNA binding motif MAST was used to identify instances of TF DNA binding motifs in the GR binding sites tested. This dataset contains motif information for GR (NR3C1), AP-1 (as represented by FOSL2), FOXA1, NFκB (as represented by RELA), and CREB1 respectively. The columns of the dataset are: (1-3) Genomic coordinates of the GR binding sites, (4) location of the best match to the motif in the binding site, (5) p value of the motif match as determined by MAST, (6) the location of the best motif in the shuffled sequence, and (7) the associated p value.

Table S7. Related to Figure 4 and Figure 5; Changes in JUND binding in A549 cells after 100 nM DEX exposure. Tabs contain sites that are gained, lost and maintained. Columns contain: Chromosome, start, end, p value JUND motif and p value GR motif for each site.

Table S8. Related to Figure 3, Figure 4, Figure 5 and Figure 6; External datasets used in this study.

Raw and aligned sequencing data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE79432 and GSE77869.

3. Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort

The research presented in this chapter was published in the journal Genome Research in the year 2015.

The complete citation for this article is:

Vockley, C.M.*, Guo, C.*, Majoros, W.H.*, Nodzenski, M., Scholtens, D.M., Hayes, M.G., Lowe, W.L., Jr., and Reddy, T.E. (2015). Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. Genome Res *25*, 1206-1214.

* denotes shared first authorship.

3.1 Introduction

There are now several examples of noncoding genetic variants that alter the activity of regulatory elements and contribute substantially to complex traits and human diseases (Corradin et al., 2014; Guo et al., 2014; Maurano et al., 2012; Nicolae et al., 2010; Olansky et al., 1992; Stadhouders et al., 2014). Such examples are likely representative of a larger trend that genetic variations in regulatory elements are a major contributor to complex phenotypes and disease (Gusev et al., 2014; Maurano et al., 2012). Genetic effects on gene regulation are pervasive, as demonstrated by association studies revealing expression quantitative trait loci (eQTL) for the majority of human genes (Battle et al., 2014; Cantor et al., 2010; Stranger and Raj, 2013). Recent studies have further demonstrated that genetic variants associated with DNase I hypersensitivity, a

strong predictor of the presence of a regulatory element, explain a substantial proportion of eQTLs (Degner et al., 2012), and individuals who are heterozygous in those elements likely have heritable allele-specific open chromatin and transcription factor binding (Birney et al., 2010; McDaniell et al., 2010; Reddy et al., 2012b). Although there is now much evidence supporting the contributions of regulatory variation to human phenotypes, systematically identifying the specific variants and regulatory elements that contribute to phenotype remains a major challenge.

One of the major reasons that challenge remains is that patterns of recombination across the genome limit the resolution of genetic association studies and prevent the identification of specific causal variants. That limitation motivates the development of complementary empirical approaches to assay the consequences of noncoding genetic variation on regulatory element activity (Feng et al., 2013; Fogarty et al., 2014; Guo et al., 2014; Stadhouders et al., 2014). In a reporter gene expression assay, for example, a gene regulatory element is cloned into a plasmid, where the element can control the expression of a fluorescent or chemiluminescent protein. The plasmid is then transfected or infected into cells, and the activity of the regulatory element is estimated by measuring the expression of the reporter gene. Several examples have now shown that reporter assays are a valuable tool to compare the function of genetically different versions of the same regulatory element and to identify noncoding variants that explain genetic associations with gene expression and phenotypes (Fogarty et al., 2014; Guo et

al., 2014). Recent advances have dramatically increased the throughput of reporter assays by embedding molecular barcodes within the reporter gene that can later be observed with DNA sequencing (Kwasnieski et al., 2012; Melnikov et al., 2012; Patwardhan et al., 2009; White et al., 2013), and the regulatory activity of more than one million unique DNA fragments can now be assayed in a single experiment using such massively parallel reporter assays (Arnold et al., 2013).

Here, we have developed a novel high-throughput approach to efficiently measure the activity of regulatory elements captured from the genomes of a human study population. Previous approaches to identify genetic effects on regulatory element activity have used DNA synthesis and random mutagenesis to generate mutations in select regulatory elements (Melnikov et al., 2012; Patwardhan et al., 2009; White et al., 2013). By instead assaying putative regulatory elements captured from donor genomes, the strategy presented here allows for high-throughput empirical measurement of the effects of regulatory variants specific to a study population. Moreover, because haplotypes are maintained within each regulatory element, empirical measurement of the combined effects of all common, rare, and personal variants within a regulatory element are possible. The result is individual-specific measurements of regulatory element activity across the study population. Because candidate regulatory elements are assayed independently of one another, the approach is an effective strategy to identify causal mutations within large regions of statistical association between genotype and

phenotype. Together, these results demonstrate that population-scale functional reporter assays are a valuable strategy for identifying specific causal genetic variants and haplotypes within genomic loci previously associated with phenotype.

3.2 Results

3.2.1 Population-scale reporter assay approach

We designed an empirical strategy to measure the activity of specific candidate regulatory elements across a population of individuals (**Figure 28**). The strategy is based on the STARR-seq assay (Arnold et al., 2013). Briefly, in STARR-seq, candidate regulatory elements are cloned into the 3′ untranslated region (UTR) of a reporter gene. The resulting plasmid pool is then transfected into host cells, where the cloned elements can regulate expression of the reporter gene in which they are embedded. High-throughput sequencing of the 3′ UTR of the expressed reporter gene mRNA can then be used to estimate the regulatory activity of each element.

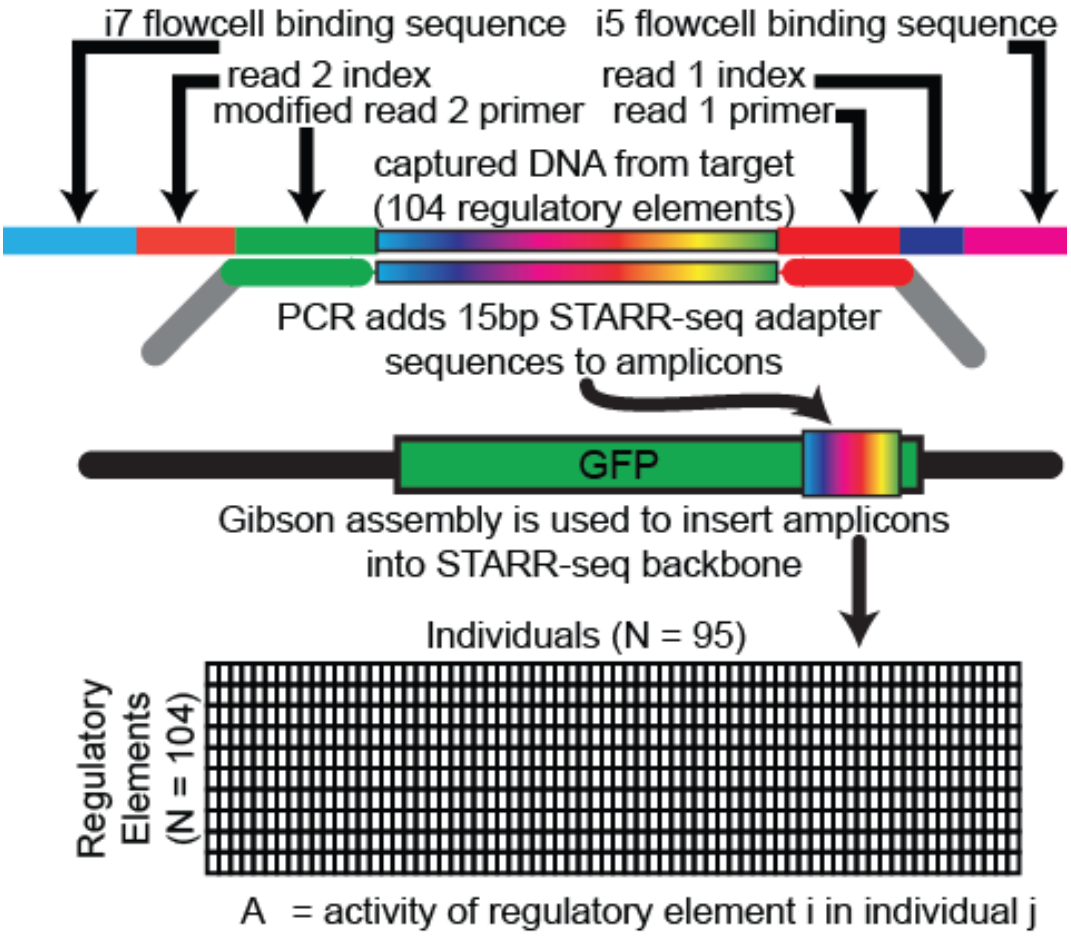

Figure 28: Population STARR-seq assay schematic.

To leverage the STARR-seq approach to measure the activity of candidate regulatory elements across a population of individuals, we first generate a targeted sequencing library of regulatory elements from donor genomes using multiplex PCR. In a subsequent PCR reaction, we then modify the resulting fragment libraries such that the sequence of the terminal 15 bp at each end of each fragment matches the ends of the cloning site in the STARR-seq backbone. We then clone the capture regulatory elements

into the STARR-seq backbone using a homology-based cloning strategy and expand the resulting input library in Escherichia coli. To assay the activity of each captured fragment, we transfect the input library into a human liver carcinoma cell line, HepG2, and use 250-bp paired-end sequencing to observe the abundance of each allele of each element in the input pool of transfected DNA and in the expressed reporter gene mRNA. Using an allele-specific analysis strategy, we then estimate the effect of each allele on regulatory element activity.

3.2.2 Targeted sequencing of candidate regulatory elements from a GWAS population

As demonstration of the aforementioned approach, we focused on candidate regulatory elements from a 250-kb region on Chromosome 3 (3q25) that we previously found to be associated with measures of adiposity at birth (Urbanek et al., 2013). We selected the regions to assay based on evidence from the ENCODE Project Consortium (2012) that suggests potential regulatory activity. Specifically, we aggregated open chromatin data from 40 different cell types relevant to metabolism, which yielded an initial set of 128 open chromatin sites. We further prioritized those sites by selecting DNase I hypersensitive sites (DHSs) that were present in at least two or more cell lines, resulting in a total 104 DHSs (**Figure 29; Supplemental Data 1**). We designed 174 PCR amplicons to amplify from the 104 candidate regulatory elements (Supplemental Data 2). The amplicons had an average length of 409 bp. We then used multiplex PCR to

amplify those elements from 95 individuals at the extremes of adiposity in the genetic association cohort (Urbanek et al., 2013).

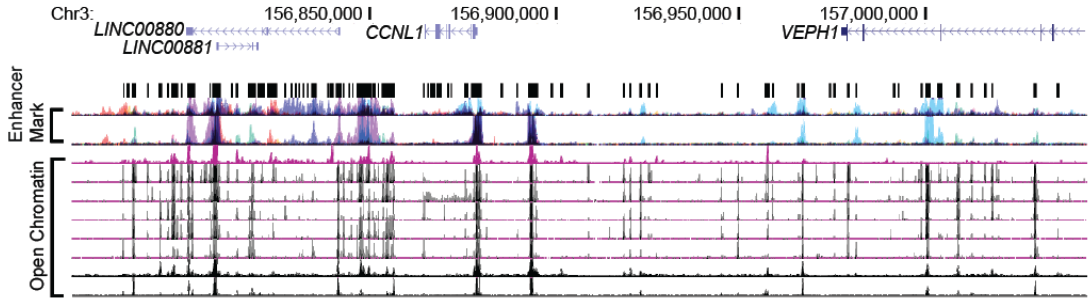

Figure 29: Candidate regulatory sites.

Candidate regulatory sites were sequenced in 95 members of the Hyperglycemia and Adverse Pregnancy Outcome (HAPO) Study (Urbanek et al. 2013) patient cohort using custom amplicon sequencing. The targeted regions overlap open chromatin (DHSs) in multiple cell types as described in Methods.

To quantify the genetic variation in the captured elements, we sequenced the regions using paired-end 250-bp sequencing. That read length was sufficient to observe the entire sequence of each amplicon. Sequencing was completed to a median depth of 1500× (**Appx. 2; Figure 55**), resulting in the identification of 321 genetic variants in the captured elements (**Supplemental Data 3**). Twenty-three percent of the variants identified were specific to the study population as determined by their absence from dbSNP and the 1000 Genomes Project Consortium database (Sherry et al., 2001; Thousand Genomes Project Consortium et al., 2012). The ratio of transitions to transversions was similar between the captured variants and those found in the 1000 Genomes Project (Supplemental Table 1), suggesting that the novel variants were

unlikely due to systematic sequencing errors. We identified a substantially greater fraction of rare and personal variants in our targeted sequencing, likely due to increased sequencing depth that supported more highly powered variant calling (**Appx. 2; Figure 56**). The preponderance of study-specific variants emphasizes the importance of assaying regulatory elements captured from the genomes of the study population rather than from a separate cohort.

3.2.3 Quantifying the effects of noncoding variation in a GWAS population

To quantify the activity of the captured candidate regulatory elements, we cloned the captured amplicons into the 3′ UTR of the STARR-seq reporter gene (Arnold et al., 2013) to generate an input plasmid library. The input library covered 99% of the targeted sequence and included both alleles of 88% of the variants observed in targeted sequencing of the region at a median coverage of approximately 2200× (Supplemental Table 2; **Appx. 2; Figure 57**). We then performed seven independent transfections of the input library into HepG2 cells and used targeted high-throughput sequencing of the expressed reporter gene transcripts to measure the allele-specific regulatory activity for each amplicon. The sequencing generated a median coverage of the target amplicons of approximately 13,000× (**Appx. 2; Figure 58**) and assayed both alleles of 283 of 321 SNPs detected in the input library. Of the assayed SNPs, 83 (29%) were rare, defined as a minor allele frequency <1%. We observed a similar fraction of rare SNPs in the input library (32%), suggesting that there was minimal bias against rare variants in the assays.

There was strong correlation between the allele ratios in each pair of output libraries (Spearman's ϱ between 0.90 and 0.97) (**Figure 31A**), demonstrating reproducibility of the assay. There was also strong correlation between the allele ratios in the input plasmid pool versus the allele ratios in each of the output libraries (Spearman's ϱ between 0.80 and 0.88) (**Appx. 2; Figure 58**), demonstrating that variants had small effects on regulatory activity overall. Cloning the captured candidate regulatory elements into the STARR-seq backbone did not introduce biases in the allele frequency in the assay as demonstrated by a strong correlation between the allele ratios in the plasmid DNA library and the allele ratios in the sequencing of the initial multiplex PCR products ($r2 = 0.94$, two-sided $P < 0.0001$) (**Figure 30**). We therefore concluded that the resulting assay libraries were representative of the genetic diversity in the population. When the allele frequencies of the input plasmid DNA library were compared to the allele frequencies of the variants called in the 95 individuals, we observed enrichment of rare minor alleles in the input plasmid DNA library (**Appx. 2; Figure 59**). Because that bias was specific to the comparison with called variants and was not observed when comparing to the raw sequencing reads, the bias was likely due to underestimation of rare allele frequencies by conservative calling of rare variants (Thousand Genomes Project Consortium et al., 2012).

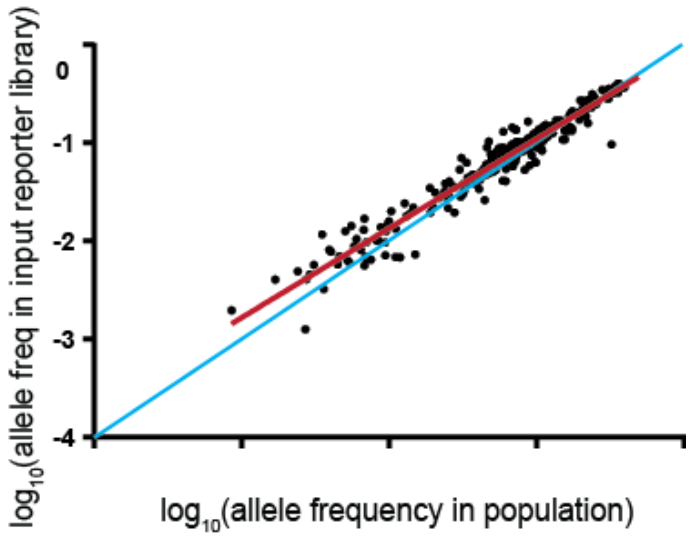

Figure 30: Population STARR-seq reporter libraries are representative of population diversity.

Plotted is a comparison of the allele frequency of each SNP in the cohort DNA to the allele frequency of each SNP in the resulting reporter library. Allele frequencies of the cohort DNA used are shown on the x-axis, and the allele frequencies in the resulting reporter library are on the y-axis. The allele frequencies are highly correlated, as evaluated by a Pearson correlation ($r^2 = 0.94$, $P < 1 \times 10^{-5}$). The one-to-one line is shown in blue. The least squares fit is shown in red.

To identify individual variants that have a statistically significant effect on regulatory activity after taking into account differences in read depth, we pooled reads from the replicate output libraries and compared relative variant abundance to the input library using Fisher's exact test. We identified 27 common and nine rare regulatory variants with a false discovery rate (FDR) < 5%. The identified variants had fold changes in regulatory activity ranging from 0.25 to 3.96 (Supplemental Data 4), consistent with previous observations using saturation mutagenesis of enhancers (**Figure 31B**) (Patwardhan et al., 2012). To empirically validate that the results were not due to the

candidate regulatory elements' location in the 3′ UTR of the reporter gene, we used a standard luciferase reporter assay in which the candidate regulatory element is located upstream of the promoter. In all cases, the allele with greater regulatory activity in the STARR-seq assay also had increased luciferase expression (**Figure 31C**). That positive validation indicates that the observed effects were not specific to the location of the candidate regulatory element relative to the reporter gene.

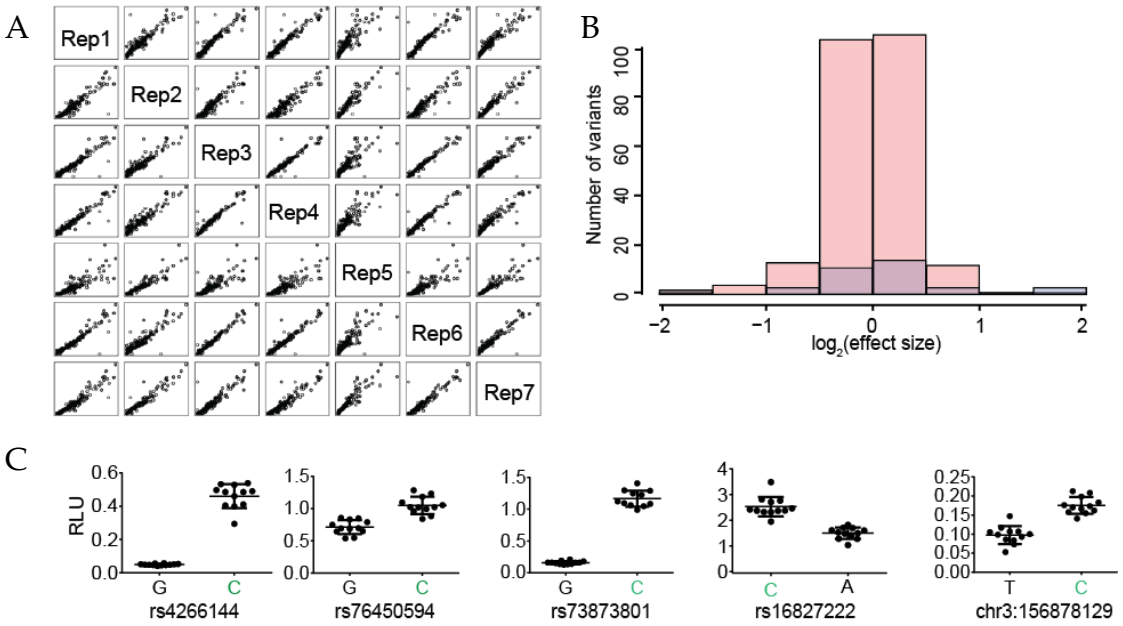

Figure 31: Identifying regulatory variants.

(A) Population STARR-seq is highly reproducible. Rep1–7 are biological replicates generated from independent transfections. The x- and y- axes represent element activity (output RNA reads/input DNA reads). In each case, Spearman's ϱ > 0.90. (B) Log_2(effect sizes) for non-significant (pink) and significant (FDR < 0.05, blue) variants. The effect sizes are small and range between 0.25 and 3.96 fold-change. (C) Firefly luciferase assay validations for population STARR-seq. In all cases, the higher expressing allele in our high-throughput reporter assay, shown in green, also had higher luciferase expression.

3.2.4 Regulatory variants are enriched in active enhancers

We next evaluated whether regulatory variants were enriched in the most active enhancers or could instead be due to noise in low-activity or silent candidate regulatory elements. We defined an enhancer activity score as the proportion of the total reads contributed by a fragment in the targeted RNA-seq output library divided by the proportion of the total reads contributed by that fragment in the input DNA plasmid library. The fragments that contained regulatory variants had higher-ranking enhancer activity scores than those that lacked regulatory variants (U-test, $P < 10{-}4$) (**Figure 32**; Supplemental Data 5,6), consistent with regulatory variants being located in the most active candidate regulatory elements. We also asked whether there was evidence that rare alleles were more likely to have a stronger effect on regulatory activity, and we did not find a statistically significant association between effect size and allele frequency (Spearman $\varrho = -0.18$, $P = 0.28$) (**Appx. 2; Figure 61**).

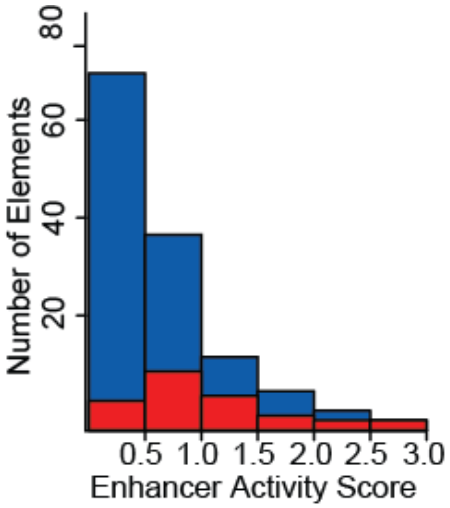

Figure 32: Enhancer activity scores for regulatory variants.

Distribution of enhancer activity scores for fragments containing regulatory variants (red) and fragments containing non-regulatory variants (blue)

3.2.5 Effects of haplotypes on regulatory element activity

For 98 of the amplicons, there was more than one polymorphic site (**Figure 33**), allowing us to ask whether multiple variants act independently to alter regulatory element activity at the haplotype level. To investigate that possibility, we generated phased haplotype sequences based on the targeted sequencing data and used sequence alignment to assign sequencing reads from the expressed reporter library to each haplotype (Supplemental Data 7). That analysis allowed us to estimate the relative expression of each of the more than 450 distinct haplotypes assayed and revealed 24 haplotypes across 16 amplicons that significantly altered regulatory element activity (adjusted $P < 0.05$, Fisher's exact test) (**Supplemental Data 8**). We then evaluated the

extent to which the independent contributions of the estimated effects of each SNP in a haplotype predicted the observed activity of the entire haplotype (**Appx. 2; Figure 62**). The correlation between the effects predicted by individual SNPs and the effects of the haplotype ($r = 0.54$, $P = 0.007$) supports an overall consistency between SNP effects and their combination into haplotype effects. However, there was substantial residual variation that may be due to either experimental noise or synergistic effects between variants within haplotypes. Measuring haplotype-scale effects in larger populations will also be important to establish the distribution of natural functional variation in regulatory elements and may provide insights into the role of gene regulation in a wide variety of biological processes.

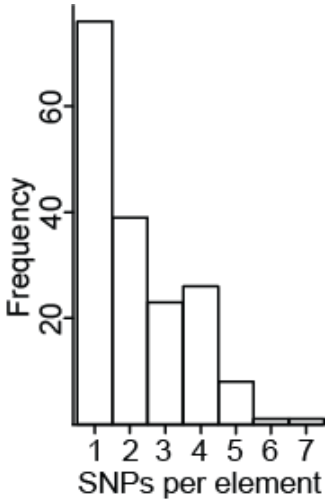

Figure 33: Histogram of number of SNPs per assayed element.

3.2.6 Fine mapping genetic associations with phenotypes

One of the major goals of functionally evaluating regulatory variants is to determine genetic effects on regulatory element activity that may explain genetic associations with phenotypes. To demonstrate that our strategy can support such fine mapping, we investigated a set of SNPs associated with the expression of a long noncoding RNA *LINC00881* in the region. Specifically, the Geuvadis project (Lappalainen et al., 2013) identified a cluster of nine eQTLs associated with the expression of *LINC00881* in lymphoblastoid cell lines (LCLs) (**Appx. 2; Figure 63**). The variants associated with *LINC00881* span ~12 kb of the genome. The statistical significance of the association with *LINC00881* was similar across all nine variants, likely due to high linkage disequilibrium across the region (Figure 34). Four of the nine eQTLs were also assayed in the 95 individuals with our population scale reporter assays. Only one variant, rs73170828, located 242 bp upstream of the annotated *LINC00881* transcription start site, significantly altered reporter gene expression (FDR = 0.02). In the eQTL analysis and in our population scale reporter assays, the reference allele of rs73170828 was associated with increased gene expression and increased regulatory activity, respectively (**Figure 35A**). Together, these results suggest that the promoter-proximal variant rs73170828 is a causal variant that regulates the transcription of *LINC00881* and explains the association of the other eQTLs in the region.

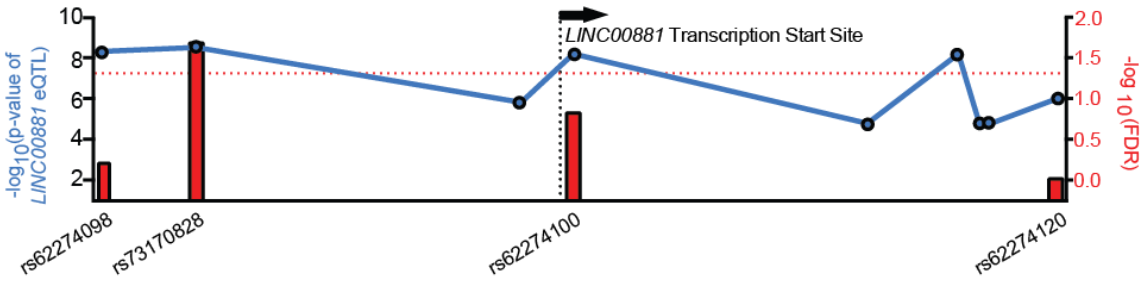

Figure 34: Manhattan plot of eQTLs for the long non-coding RNA *LINC00881*.

Blue dots indicate -log$_{10}$(p value) of *LINC00881* eQTL from the GEUVADIS database (left Y-axis); red bars indicate -log10(FDR) for variants that alter regulatory activity in the population STARR-seq assay (right Y-axis). Red dotted line indicates a FDR = 1.0.

As independent support of the regulatory function of rs73170828, we searched for evidence of allele-specific histone 3 lysine 27 acetylation (H3K27ac), a histone modification associated with active gene regulation (Creyghton et al., 2010). In ChIP-seq experiments performed on LCLs derived from five individuals heterozygous for rs73170828 (Kilpinen et al., 2013), there was substantially higher H3K27ac on the reference allele across the LCLs (P = 0.058, paired Wilcoxon test). Furthermore, there was an overall significant increase in the number of reads aligning to the reference allele when compared to a null model in which the same proportion of reads align to each allele (binomial P = 0.004). Those results are concordant with increased regulatory activity of the reference allele in our reporter assays and increased *LINC00881* expression. The second closest assayed variant, rs62274098, did not have significant allele-specific H3K27ac (binomial P = 0.92), suggesting again that rs73170828 and not neighboring variants mechanistically contributes to the expression of *LINC00881* (Figure

35B). Together, these results show that our novel approach for quantifying the effects of noncoding variation on gene regulation within cohorts reveals likely causal variants that contribute to genotype-phenotype associations.

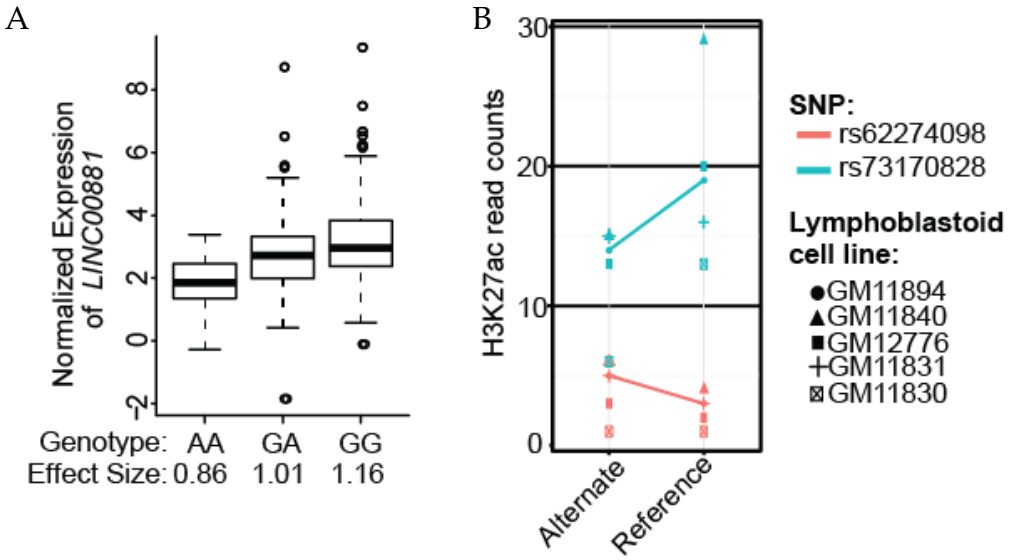

Figure 35: *LINC00881* expression and H3K27 acetylation state as a function of genetic variation

(A) Association between normalized expression of long noncoding gene *LINC00881* in LCLs as measured by the GEUVADIS project (Y-axis) and the measured effect size in population STARR-seq assay (X-axis) for SNP rs73170828 (r2 = 0.07, p = 7.6 x 10^{-9}). (B) Allele-specific H3K27ac analysis of variants rs62274098 and rs73170828, both eQTLs proximal to and 5' of *LINC00881*; read counts (Y-axis) differed substantially between alleles for rs73170828 (Wilcoxon p=0.058, binomial p=0.004) but not for rs62274098 (Wilcoxon p=0.9; binomial p=0.92).

3.2.7 Identifying candidate mechanisms of regulatory element activity

Quantifying genetic effects on regulatory element activity can also give insight into the underlying mechanisms controlling gene expression. As an example, one of the

most significant regulatory variants in our study, the common SNP rs4266144 (minor allele frequency = 0.40), had a 1.34-fold effect on the activity of the regulatory element in which it is located. The variant overlaps a binding site for the transcription factor TEAD4 in the HepG2 cell line that we used in this study (ENCODE, 2012). The C allele more closely matches the TEAD4 consensus motif and also had increased regulatory activity (**Figure 31A, left-most plot; Figure 36**). The higher-activity C allele is also human-specific, whereas the ancestral G allele is conserved across nonhuman members of the Hominidae clade; and it is possible that recent evolution has altered the regulatory activity of that site by changing the TEAD4 recognition sequence (Blanchette et al., 2004). Although only a case study, this example highlights the possibility that combining the identification of regulatory variants with existing maps of transcription factor binding can reveal regulatory factors contributing to regulatory element activity. A systematic evaluation of that possibility will require expanding the catalog of functional noncoding genetic variants in larger populations.

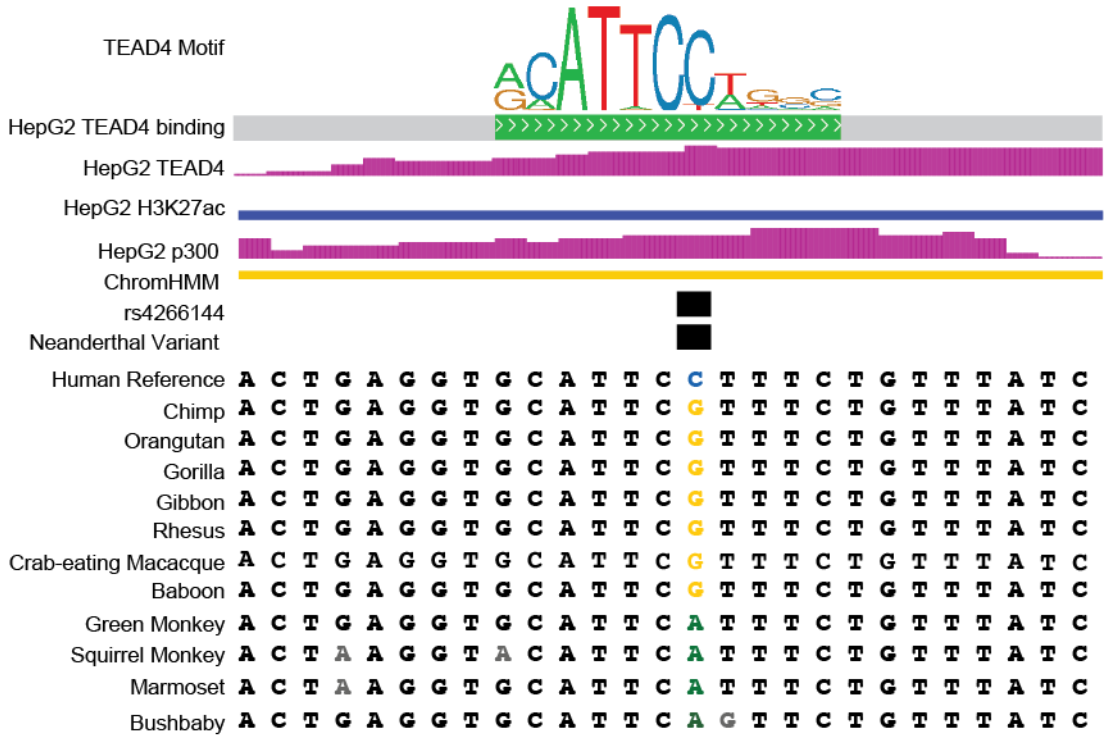

Figure 36: Regulatory variant disruption of a TEAD4 binding site.

SNP rs4266144 resides within a TEAD4 ChIP-seq binding site as assayed in HepG2 cells. The C>G variant is located in a largely invariant region of the TEAD4 canonical consensus binding motif. The binding site is located within a region that is enriched for H3K27ac and P300 occupancy. Concordantly, ChromHMM segmentation analysis scores the locus as a putative weak enhancer (Ernst and Kellis, 2012). Multispecies conservation analysis suggests that this motif resides within a region that is conserved between the great apes.

3.3 Discussion

In this work, we developed a novel high-throughput empirical approach to measure the regulatory effects of noncoding human genetic variation directly from the DNA of individuals from a population-based study cohort. The ability to assay directly from cohort DNA samples is an important distinction from previous high-throughput

reporter assays because it allows investigation of variants and haplotypes that are not present in existing databases of human genetic variation. As rare variants are typically not observed frequently enough to support a statistical association, rare-variant burden tests instead collapse or aggregate variants and correlate the overall burden of those variants with phenotypes (Li and Leal, 2008; Zawistowski et al., 2010). Although burden testing within the coding regions of the genome can leverage predicted effects on the resulting protein (Choi et al., 2012; Hu et al., 2013), modeling regulatory element activity based on sequence alone remains a major challenge. Measuring regulatory activity directly from cohort DNA provides a possible empirical solution that allows the regulatory machinery of the cell to determine the cumulative effects of all regulatory variation in the element tested and allows for inference about the activity of that regulatory element that would not be possible otherwise.

The ability to associate empirically measured regulatory function and phenotype is especially needed in light of recent studies suggesting that coordination of regulatory effects between alleles may explain how weak effects of individual noncoding variants contribute to overall phenotypes (Corradin et al., 2014; Guo et al., 2014; Stadhouders et al., 2014). As we have shown, assaying regulatory elements outside the context of genetic linkage enables identification of individual regulatory elements that contribute to observed associations with gene expression. Importantly, however, genetic linkage is maintained within each individual regulatory element tested. That feature allows for

measuring the effects of regulatory element haplotypes on element activity without the confounding effects of a nearby regulatory element. For those reasons, the approach described here has the ability to both resolve independent effects in multiple regulatory elements while also maintaining local epistatic interactions between variants within an individual element.

For any complex disease, multiple types of cells are likely relevant to an observed phenotype. Additionally, the causal regulatory elements may only be active under certain environmental conditions, or an interaction with the environment may amplify the effect. Transient reporter assays have been shown to recapitulate cell-type- and environment-specific gene regulation (Gisselbrecht et al., 2013; Pennacchio et al., 2006; Shlyueva et al., 2014). Because the input plasmid libraries generated in this study are a renewable resource that can be readily expanded in E. coli, the same captured regulatory elements can be assayed in numerous cell models and environmental contexts. Doing so may have particular benefit for identifying the specific cells or environments that are more relevant to a given genetic association signal.

There are both advantages and disadvantages intrinsic to the architecture of the STARR-seq assay platform. Among the advantages is the potential to characterize dual functioning enhancer-promoters (Arnold et al., 2013). We detected regulatory variants within TSS-proximal regions of two of the three genes located within our test locus, suggesting that the elements that contain these variants serve as dual function enhancer-

promoters. The approach is limited by the observation that enhancers often have promoter-specific activity in transient transfection assays, indicating that alternative promoters may be required in some cases (Zabidi et al., 2015). Addressing those shortcomings will further increase the ability to assign regulatory causes to genetic associations.

Taken together, the approach demonstrated here enables measurement of the functional variation in regulatory activity across human populations and provides a novel and general path forward to identify disease-related perturbations in regulatory mechanisms after the completion of a genome-wide association study.

3.4 Methods

TruSeq Custom Amplicon Sequencing

We defined a target region as the region containing all variants in linkage disequilibrium (LD) ($D' > 0.05$) with the lead SNP previously reported to be associated with fetal adiposity (Urbanek et al. 2013). All annotated exons and all sites with evidence of putative enhancer activity as determined by the presence of DNase I hypersensitive sites (DHSs) in two or more cell lines studied by the ENCODE Project Consortium (2012) were selected for capture (Supplemental Data 1). Captured sites included 10 bp of flanking DNA to ensure that the entire putative regulatory site was included in the study. Lists of annotated DHSs from the ENCODE Project Consortium were downloaded as BED files from http://genome.ucsc.edu/ENCODE/downloads, and the

union of overlapping DHSs was obtained using the "merge" command in BEDTools (Quinlan and Hall 2010). TruSeq custom amplicon probes targeting the regions as well as the exons of CCNL1, *LINC00880*, *LINC00881*, and the five exons of VEPH1 residing within the LD block were designed using the Illumina Design Studio. The probes were designed to not overlap any known SNPs and capture an additional 25 bp flanking each DNase I hypersensitive site. The final design consisted of 174 amplicons with lengths ranging from 398 to 450 bp (mean length of 409 bp and a median length of 402 bp) capturing a total of ~60 kb of DNA (**Supplemental Data 2**). We designed the amplicons to be <450 bp to ensure that paired-end 250-bp sequencing would cover the entire length of the fragment. Library construction was conducted via the standard protocol provided by Illumina using 250 ng genomic DNA per reaction. The libraries were pooled and sequenced using paired-end 250-bp reads on an Illumina MiSeq instrument.

Variant Calling and Phasing

Sequencing reads were demultiplexed and aligned to the target regions using the standard Illumina Custom Amplicon Workflow protocol. Reads were first aligned to the downstream locus-specific and upstream locus-specific oligonucleotide primers used to amplify the targeted regions. Then, the alignment was performed using a banded Smith-Waterman alignment. Variant calling was performed using tools from the Genome Analysis Toolkit (GATK) version 3.2-2, according to GATK Best Practices

recommendations (DePristo et al., 2011; McKenna et al., 2010; Van der Auwera et al., 2013). According to September 2014 guidelines for small targeted experiments, this workflow included using HaplotypeCaller to call variants in target regions individually per subject, followed by joint genotyping using GenotypeGVCFs to produce a multisample VCF. Default settings were used for both tools. After variant calling, the following annotations and thresholds were used to remove low confidence SNPs, based on GATK recommendations for hard filtering: QD < 2.0; MQ < 40.0; FS > 60.0; MQRankSum < −12.5; ReadPosRankSum < −8.0; QUAL < 100.0. Similarly, the following filters were applied to remove low confidence indels: QD < 2.0; FS > 200.0; ReadPosRankSum < −20.0; InbreedingCoeff < −0.8; QUAL < 100.0. After hard-filtering, haplotypes were estimated with SHAPEIT2 software (Delaneau et al. 2012, 2013b; O'Connell et al. 2014) using the "Read Aware Phasing" algorithm (Delaneau et al. 2013a). According to SHAPEIT2 documentation, linkage disequilibrium patterns necessary for haplotype inference can be adequately captured using MCMC sampling in studies with at least 100 subjects; therefore, reference panels were not incorporated, and default algorithm parameters were used.

Reporter Input Library Construction

PCR amplicons from Illumina custom capture libraries from 95 individuals were pooled in equal volume. The resulting pools were then PCR amplified to add 15 bp of

sequence matching the STARR-seq backbone using primers TS2SSF and TS2SSpatientR using Q5 polymerase with GC buffer (New England Biolabs) using the following cycling conditions: for 15 sec at 98°C and cycles of 10 sec at 98°C, 30 sec at 63°C, and 3 min at 72°C. The resulting products were purified using Solid Phase Reverse Immobilization (SPRI) beads at a 1.8× SPRI:reaction ratio.

The STARR-seq screening vector was digested overnight with SalI and AgeI, and linearized backbone was purified with the Wizard SV Gel and PCR Clean-Up kit (Promega). One hundred nanograms backbone and 23 ng pooled insert were cloned in two 20 μL Gibson assembly reactions. The reactions were purified using SPRI beads and eluted in 5 μL ddH20 and then transformed into Stellar chemically competent cells according to the manufacturer's protocol. Transformations were recovered for 1 h in SOC medium while shaking (225 rpm, 37°C) and then grown for 14 h in 250 mL of Luria Broth while shaking (225 rpm, 37°C). The resulting reporter input libraries were then purified using the Promega Pure Yield Maxiprep kit.

To assess variant diversity in the population STARR-seq input libraries, the fragments inserted into each were sequenced on an Illumina MiSeq. Ten nanograms of each input library were PCR amplified using indexed custom sequencing primers and Q5 polymerase in GC buffer (New England Biolabs). The following thermal cycling protocol was used: 30 sec at 98°C followed by 10 cycles of 10 sec at 98°C, 30 sec at 65°C, and 2 min at 72°C, with a final extension for 7 min at 72°C. The reporter input pool PCR

product was purified using SPRI beads (1.8× SPRI:DNA ratio) and sequenced on an Illumina MiSeq Instrument using 250-bp paired-end reads. Primer sequences are available in Supplemental Table 4.

Reporter Output Library Construction

Population STARR-seq input libraries were combined in equimolar pools and transfected into T-150 flasks of HepG2 cells with Fugene (Promega) at a 5.5:1 ratio of Fugene:DNA. Eight replicate transfections were performed. Forty-eight hours after transfection, RNA was harvested as described next.

Cells were rinsed with PBS pH 7.4 and incubated for 3 min at 37°C with DNase I (5 mg DNase I in 1 mL buffer containing 10 mM Tris-HCl pH 7.5, 150 mM NaCl, and 1 mM MgCl in DEPC-treated water diluted to a total volume of 24 mL in PBS). Cells were rinsed again with PBS and then dissociated with Trypsin-EDTA 0.25% (Life Technologies). Trypsin was neutralized with HepG2 tissue culture medium, and cells were pelleted via centrifugation. Cell pellets were rinsed once with PBS and then lysed in 2 mL of RLT buffer (Qiagen) with 2-mercaptoethanol (Sigma).

Total RNA was prepared using the Qiagen RNeasy Midi kit including the on-column DNase I digestion step. Poly-A RNA was isolated from 70 μg total RNA by double selecting with Dynabead Oligo-dT25 beads (Life Technologies). The RNA was then treated with turboDNAse (4 U) for 30 min at 37°C (Life Technologies). DNase

treated poly-A RNA was purified using the RNeasy Mini kit. cDNA was synthesized using the STARR-seq gene-specific primer using SuperScript III (Life Technologies). Reaction volumes were scaled to 50 μL. Reactions were incubated for 2.5 h at 55°C and inactivated by incubating for 15 min at 70°C. Following synthesis, cDNA was treated with RNase A (Sigma) for 30 min at 37°C. cDNA was purified with SPRI beads at a 1.5:1 bead:cDNA ratio (by volume).

The cDNA was then amplified using a two-stage PCR with a protocol similar to the published STARR-seq protocol (Arnold et al., 2013). The cDNA sample from each replicate was used as input into first-round reporter-specific PCR reactions using primers "reporter specific primer1" and "reporter specific primer2," and Q5 high-fidelity polymerase (New England Biolabs) with GC buffer (denaturing for 45 sec at 98°C, amplification with 15 cycles of 15 sec at 98°C, 30 sec at 65°C, and 70 sec at 72°C; final extension for 7 min at 72°C). Samples were then purified using SPRI beads at 1.5× ratio of bead:PCR product and eluted in 15 μL nuclease-free water. The resulting products were used as template for a second round of PCR, which used a standard Illumina TruSeq indexing primer on the p5 end of the library and custom indexing primers (Supplemental Table 3) to barcode the samples for multiplexing prior to sequencing (Illumina). Final sequencing libraries were purified with SPRI beads at a 1.5× SPRI:PCR reaction ratio.

Identifying Regulatory Variants in Population STARR-seq

Haplotype sequences were imputed using the phased VCF file by inserting phased variants into reference sequences from the hg19 genome assembly. Sequencing reads were aligned to these haplotypes using Bowtie 2 (Langmead and Salzberg, 2012) with strict match parameters (mismatch, gap open, and gap extend penalties all set to 100) to ensure exact matching to individual haplotypes. Read counts at each SNP were tallied using SAMtools mpileup (Li et al., 2009). Replicates were pooled to increase statistical power. SNPs having fewer than two reads of either input DNA or pooled RNA were discarded from further analysis. Fisher's exact test was used to detect significant differences in minor allele frequency between input DNA and output RNA; a pseudocount of 1 was added to each table entry in Fisher's exact test. Two-tailed P values were adjusted to control the false discovery rate (FDR) to <5% via procedure p.adjust() in the standard R package "stats" (R Core Team 2015), which implements the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Of 283 SNPs tested, 36 were found significant at an FDR-adjusted level of 0.05. SNP effect sizes for each allele were computed as the ratio of normalized read counts between variants: (RNA0/DNA0)/(RNA1/DNA1) for DNA and pooled RNA read counts for alleles 0 and 1. Haplotype effect sizes were computed as normalized ratios for each haplotype versus all pooled haplotypes at a locus.

$$(RNA_{haplotype} / DNA_{haplotype}) / (RNA_{pooled} / DNA_{pooled})$$

Significance was assessed via Fisher's exact test as above.

Luciferase Validation Assays

Selected regions were amplified from the genomic DNA from individuals who were heterozygous for regulatory variants identified via the population STARR-seq assay. Primer sequences are available in Supplemental Table 4. The amplified regions were then cloned into a modified pGL4.13 luciferase expression vector containing a Supercore1 promoter as described (Arnold et al. 2013). The construct was then transformed into TOP-10 competent cells (Life Technologies) and plated onto LB agar plates with ampicillin and incubated overnight at 37°C. In order to capture both haplotypes from subjects who were heterozygous in those regions, multiple colonies were selected and grown individually in LB media overnight. Plasmids were extracted using the PureYield Plasmid Miniprep System (Promega). Constructs were sequenced using Sanger sequencing, and variants were confirmed in dbSNP31. HepG2 cells were plated into white flat-bottom 96-well plates at a density of 25,000 cells/well. After 48 h, 100 ng of plasmid/well (1:10 Renilla:firefly luciferase ratio) was transfected with Fugene HD (Promega) at a 5.5:1 Fugene:DNA ratio. Twelve biological replicates for each construct were transfected. After 24 h, firefly luciferase and Renilla luciferase signal were quantified using the Dual-glo Luciferase Assay (Promega) using a Victor3 1420 plate reader (PerkinElmer). Normalized luciferase signal was calculated by dividing the

firefly luciferase signal by the Renilla luciferase signal. Statistical significance between the normalized luciferase signals for each allele was determined using a Student's t-test.

GEUVADIS eQTL Analysis

Expression-QTLs and gene expression measurements were obtained from the Geuvadis project (Lappalainen et al., 2013). The expression measurements used in this manuscript were from 462 measurements that passed Geuvadis quality control and that had been PEER-factor normalized (Stegle et al., 2010) and transformed to a standard normal distribution (Lappalainen et al., 2013). Associations between quantile-normalized gene expression levels and genotype were calculated in R via the lm() function.

Allele-specific H3K27ac Analysis

Allele-specific analysis of H3K27ac ChIP-seq reads was completed by using Bowtie (Langmead and Salzberg, 2012) to read to both possible alleles of and flanking regions for rs73170828 and rs62274098. Reads were required to align with no mismatches (Bowtie parameter "-v 0"), and any reads that aligned equally well to both possible alleles were discarded (Bowtie parameter "-m 1"). The approach follows a previously published method that was shown to eliminate alignment biases toward the reference allele (Reddy et al. 2012). To test for allele-specific H3K27ac, the number of

unique reads aligning to each allele was tabulated, and the statistical tests described were performed using R.

Data visualization

Visualization for Figure 1B and rs4266144 case study analysis in Figure 3 was completed on the UCSC Genome Browser using the GRCh37/h19 release of the human genome (Kent et al., 2002).

3.5 Supporting online materials

Supplemental Data 1: Amplicon Probe Coordinates
Supplemental Data 2: DHS coordinates
Supplemental Data 3: Phased VCF file for 95 individuals
Supplemental Data 4: Variant effect sizes determined using population STARR-seq
Supplemental Data 5: Enhancer Activity Scores for Significant Functional Variants
Supplemental Data 6: Enhancer Activity Scores for Non-significant Variants
Supplemental Data 7: Haplotype Sequences
Supplemental Data 8: Haplotype Effects

Raw and aligned sequencing data from the input and output STARR-seq libraries have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE68331.

4. Conclusion and Future Outlook

Understanding how a single genome shared by all somatic cells in the human body can encode the full compliment of information required for cellular specialization and response to environmental stimuli remains a premier challenge in molecular biology. The earliest gene regulation studies focused on the interrogation of individual CRMs often in steady state conditions. These efforts contributed a great deal to the field's understanding of the mechanisms that underlie gene regulation. The development of genome-scale experimental approaches provided an unbiased way to learn the generality of regulatory mechanisms that had previously been discovered in anecdotal studies. The coupling of large-scale gene regulation studies with computational approaches to test sophisticated hypotheses vastly expanded our knowledge of the basic principals of gene regulation. However, these efforts have been limited by the fact that the majority of such studies have been performed on cells at steady state on samples from individual donors.

Genomics studies of nuclear receptors, such as the GR have provided important insights into the causal relationships between transcription factor binding and gene regulation. This is because nuclear receptors allow for ligand dependent induction. However, many such studies have been limited by the assumption that GR binding, as observed by ChIP-seq, can be equated to receptor-induced enhancer activation (Biddie et al., 2011; John et al., 2011). Such studies treated all GR binding sites as equal contributors

to gene regulation. By quantifying the GC-induced enhancer function of each GR binding site, we learned that GR binding sites fall into two main classes. As described in chapter 2, the majority of GR binding sites are not capable of GC-induced enhancer function. These sites instead likely act as steady state enhancers that modulate the activity of GC-induced GR binding sites many kilobases away via protein-protein interactions.

In previous studies, the prevalence of non-GC-induced sites drove statistical relationships between GR binding sites and other genomic features. It is of particular significance that we observed GR to act as a pioneer factor capable of chromatin remodeling at GC-induced enhancers. Previously proposed models suggested that the GR is capable of binding to highly degenerate matches to the GRE motif as long as a pioneer factor binding site is present. This model was supported by a study that observed that GR binding sites that contain highly degenerate instances of the GRE tend to be bound by AP-1, and tend to be in regions of increased chromatin accessibility prior to hormone induction (John et al., 2011). Evidence from our study suggests that many degenerate GRE motifs previously observed at such sites were likely the result of false positive matches to the GRE motif. Meanwhile, our data suggests that sites where AP-1 was previously proposed to act as a pioneer factor likely reflect long-range interactions between AP-1 and the GR. We observed that a similar relationship exists between the ER and AP-1 and between ER and FOXA1. FOXA1 is the archetypal pioneer factor, thus, the

observation that FOXA1 may not be acting as a pioneer factor to enable ER binding suggests the possibility that other classical pioneer factors may serve unexpected roles in distal gene regulation.

The results of our study on the GR have revealed unexpected mechanisms that challenge accepted paradigms of GR-mediated gene regulation. The observation that GR binding sites that contain GC-induced enhancers can bind in regions of decreased chromatin accessibility and then alter the compliment of histone modifications at those sites suggests that at least for the GR, TF binding can be initiating event that potentiates histone modifications. This contrasts accepted paradigms in which epigenetic state poises a locus for function. Our enhancer-cluster model provides a mechanistic explanation for the observed clustering of transcription factor binding sites in the genome and the coordinated epigenetic changes that follow cluster formation.

While our model explains many previously unexplained observations about GR-mediated gene activation, it also opens several avenues for future studies:

1) Given that GR is capable of remodeling the epigenome, it remains unclear how cell-type specific direct GR binding is specified. It is likely that expression level of the GR, post-translational modifications of the GR, GR dimer composition, the regulation of proteins that sequester GR in the cytoplasm and the amount of circulating corticosteroids in the body all contribute to observed differences in direct GR binding.

2) It is unclear if distal chromatin interactions rely on dynamic chromatin loop formation, or rely on the architecture of pre-existing chromatin loops. Our model proposes that distal chromatin looping is fundamentally important for coordinated function of GR binding sites. Studies of multiple cell types at steady state suggest that at the level of topologically associated domains, many chromatin interactions are relatively invariant between cell-types (Dixon et al., 2012). In agreement with this observation, a genome-scale study chromatin looping in cells treated with TNF-alpha found that many TNF-alpha responsive enhancers are already engaged in long range interactions with the promoters of TNF-alpha induced genes (Jin et al., 2013). However, studies using alternative technologies have suggested that stimulation of cells with steroid hormones can induce *de novo* chromatin loop formation (Fullwood et al., 2009; Joseph et al., 2010; Kuznetsova et al., 2015). Thus, it is likely that role of dynamic chromatin looping in hormone-responsive gene expression will become more clear as assay methods used to measure chromatin loop formation improve.

3) Genetic loss and gain of function experiments will further refine our distial interaction model. While our model was developed considering evidence from multiple sources, it will be necessary to further validate these results using

genetic loss and gain of function experiments. According to our model, deletion of a direct GR binding site that nucleates a GR interaction cluster should result in the loss of ChIP-seq signal at adjacent non-GRE motif driven sites. Accordingly, we hypothesize that transplanting GREs into sites that are enriched for AP-1 binding sites but not GR binding would result in the appearance of *de novo* tethered GBSs near the introduced direct GBSs.

4) It is unclear how GC exposure leads to gene repression. The observation that direct GR binding increases reporter gene activity, but does not repress reporter gene activity suggests that GC-mediated increases in gene expression may be mechanistically uncoupled from GC-mediated decreases in gene expression. Under this hypothesis, gene repression in response to GCs might be the result of secondary events propagated by genes that are primary targets of GC-mediated gene activation. This model is supported by the fact that GBSs are farther away from GC-repressed genes than would be expected by chance and by the observation that in time course experiments, GC-mediated activation generally precedes GC-mediated repression (Reddy et al., 2009).

5) Given the role of AP-1 family members as the long-range co-activators of GC-induced genes it is plausible that AP-1 acts generally as a co-activating protein

and does not serve as a canonical pioneer factor. Testing this hypothesis will likely require the synthesis of computational and experimental strategies to identify the instances of non-motif driven (tethered) TF binding events that cluster around motif-driven TF binding events (direct) and then investigating the prevalence of AP-1 binding motifs within tethered sites. Molecular studies including ChIP reporter assays and genetic deletions of cluster nucleating sites will be required.

Evidence for an enhancer-cluster mechanism of gene regulation is particularly relevant in the context of the multiple enhancer variant hypothesis described in the introduction to this dissertation. Given the likelihood that regulatory perturbations contribute to complex disease, our model provides an important framework for understanding how small decreases in activity at multiple sites could interact to decrease the expression of a target gene.

Devising a high throughput method for functional characterization of non coding DNA elements in regions associated with disease in GWAS is of fundamental importance for linking functional genomics with medical genetics. Though only recently published, the use of high-throughput reporters to characterize regulatory variants has implemented in multiple studies that will further resolve the role of regulatory variation on gene expression. Ultimately, the intersection of human genetics and functional

genomics studies will result in a rich biological context for understanding the deleterious effects of regulatory variants. Likewise, understanding which regulatory variants have the most profound impacts on gene expression will inform functional studies of basic genome biology.

Coupling the approaches described in this dissertation to study how regulatory variants alter drug-induced enhancer function will be of particular significance to the pharmacogenomics community. Our data demonstrate that the STARR-seq approach is capable of identifying drug-responsive regulatory elements at high resolution in the absence TF-specific antibodies. Similar screens of pharmacologically active compounds will enable the discovery of the primary site of action of some classes of therapeutic agents. When coupled with antibody-based approaches (ChIP) the method is capable of assaying drug-induced gene regulation across all sites bound by a given factor and identifying the small subset of those sites that encode drug responsive enhancers. In the future, performing similar studies using input DNA from many individuals will yield insights into patient-specific gene regulation and enable the identification of patient-specific drug induced regulatory element activity. The results of such future studies may guide the development of precision diagnostic and therapeutic regimes.

Appendix A

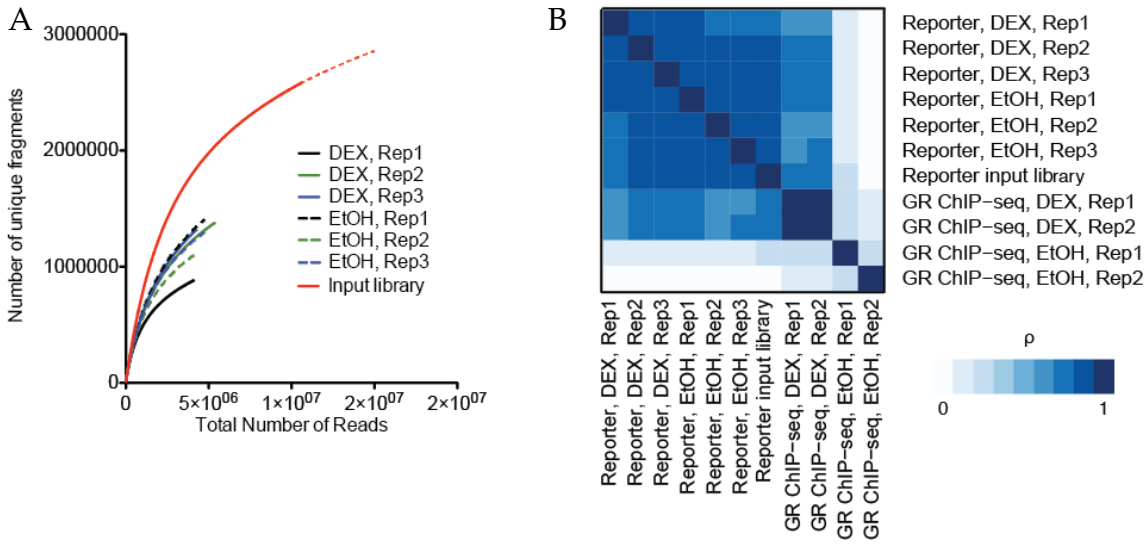

Figure 37: Fragment diversity and composition of GR ChIP-reporter libraries.

(A) Fragment diversity as a function of read depth. The number of unique fragments from sequencing reads plotted as a function of the total number of reads. Number of fragments per library was estimated by Bmax in a model that accounts for saturation and non-specific sequencing errors (Y = Bmax•X/(Kd active X) active NS•X active Background). The diversity of fragments in the input library is indicated in red, where the dotted line is the fit model. The estimated model parameters and standard errors were: Bmax = 2.9 x 10^6 ± 1868, Kd = 2.9 x 10^6 ± 3329, NS = 0.029 ± 1.1 x 10^{-4}, Background = -7910 ± 318.5 **(B)** Correlation between GR ChIP-seq and ChIP-reporter libraries across replicates. After alignment, the number of fragments aligning to each of the called GR binding sites was calculated. The libraries were then compared by correlating the per-binding-site fragment counts between libraries using a Spearman correlation. Plotted is a heat map of the correlations.

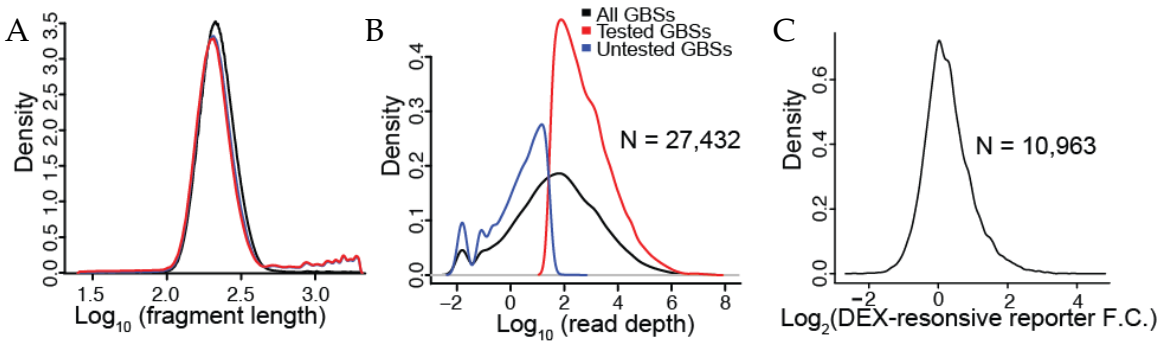

Figure 38: Fragment size, GBS coverage and DEX-induced activity of GR ChIP-reporters

(A) Distribution of fragment sizes as determined by paired-end sequencing in the input GR ChIP-seq library (black), the ChIP-reporter input library (blue), and the ChIP-reporter output library (red). The GR ChIP-seq and ChIP-reporter output libraries were generated from A549 cells after treatment for 3 h with 100 nM DEX. **(B)** Distribution of read depth across GR binding sites. The distribution of log-transformed read depth averaged across all ChIP-reporter output libraries is shown for all GR binding sites (black), the sites tested for response with DESeq2 (red), and the sites excluded from testing by DESeq2 (blue). **(C)** Distribution of DEX-response effect sizes in ChIP-reporter assays. The fold-change in ChIP-reporter activity between DEX and EtOH treatments was calculated for every tested GR binding site after normalizing the read depth for the total number of aligned reads. Plotted is the distribution of those values after performing a $\log_2$ transformation. The mean and median of the distribution was 0.20 and 0.28, respectively, and a heavy right tail (e.g. 1st quartile = -0.16, 3rd quartile = 0.64) indicates an overall prevalence of GR binding sites with increased activity after DEX treatment. **(F)** Negative binomial model of the mean-variance relationship using. To account for over dispersion in sequencing read counts, DESeq2 was used to fit a negative binomial model to estimate the relationship between mean sequencing depth at each GR binding site, and the dispersion in that site. The red line indicates the fit, and blue indicated the final dispersion estimates used in testing.

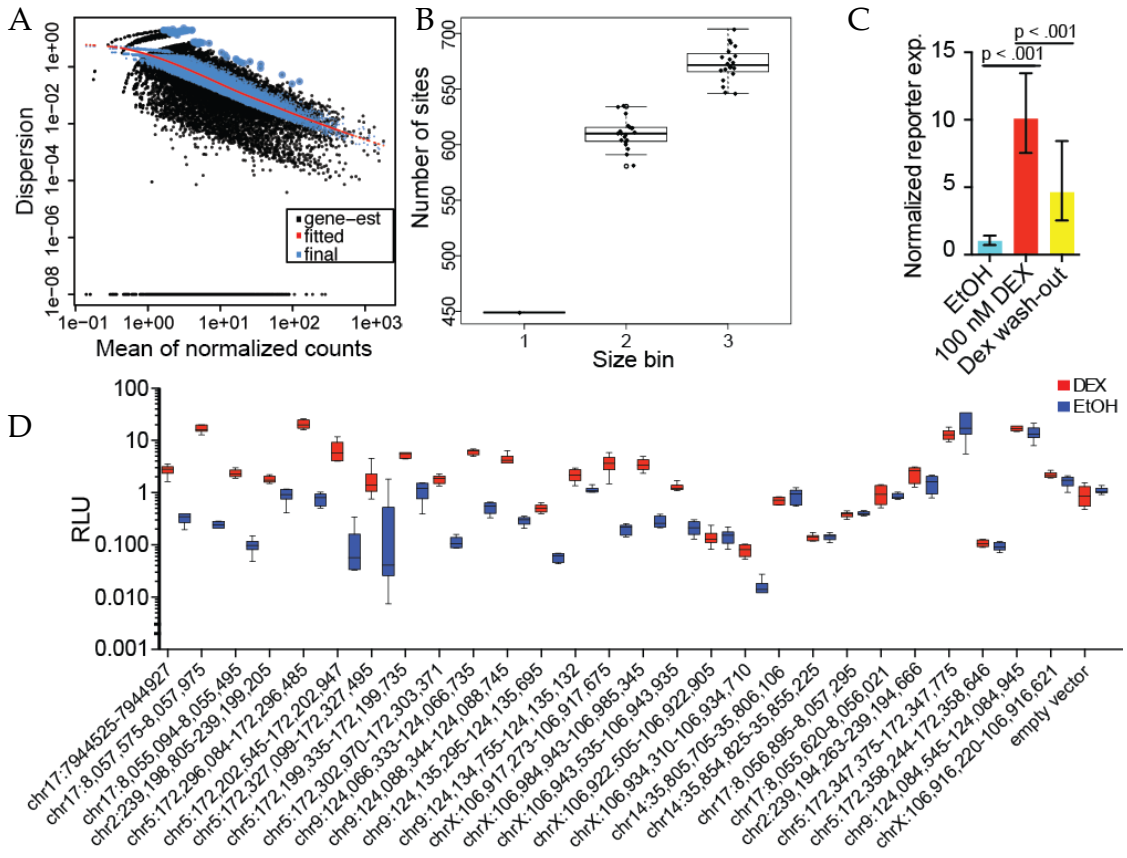

Figure 39:Negative binomial model of sequencing depth and variance, DEX-induced activity as a function of fragment size, assay control and validation data.

(A) Negative binomial model of the mean-variance relationship. To account for over dispersion in sequencing read counts, DESeq2 was used to fit a negative binomial model to estimate the relationship between mean sequencing depth at each GR binding site, and the dispersion in that site. The red line indicates the fit, and blue indicated the final dispersion estimates used in testing. **(B)** ChIP-reporter expression as a function of fragment size. All assayed DNA fragments were placed into three equal bins by fragment size. Sites were mapped to GR binding sites and bins were individually subsampled to normalize for number of fragments in each bin. DEX-inducible activity is plotted for each bin. **(C)** qPCR of STARR-seq transcript levels from a vector containing a DEX-responsive enhancer from the *PER1* locus. Cells were treated with medium containing 0.02% EtOH (3.5 h), 100 nM DEX (3.5 h) or with 100 nM DEX (0.5 h) and then washed with medium containing 0.02% EtOH and treated for an additional 3 h. Error bars indicate S.D. **(D)** Box plots of data from dual luciferase assays used to validate ChIP STARR-seq reporter experiments. Error bars indicate S.D.

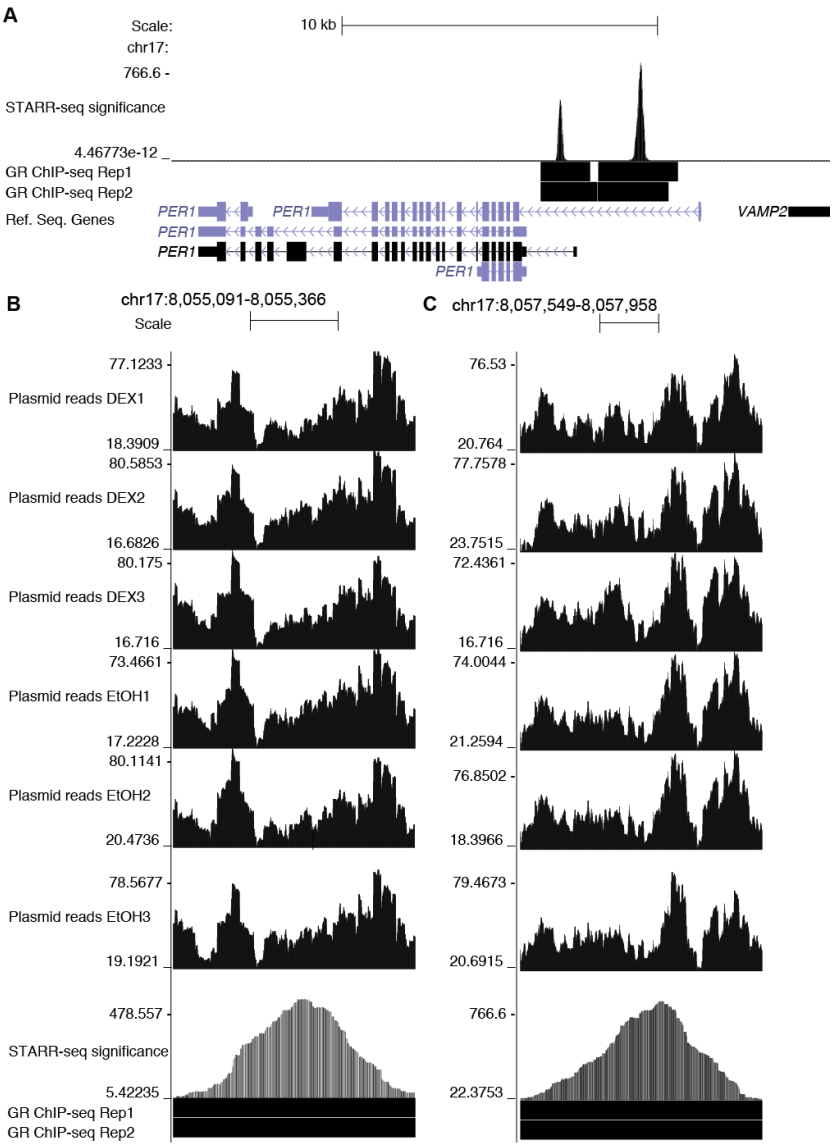

Figure 40: Nuclear localization of reporter plasmids is not biased by GCs.

(A) The DEX-inducible enhancer landscape of the *PER1* locus. Y-axis is $-\log_{10}$(p value) from Wilcoxon rank sum test. **(B, C)** High-resolution view of chr17:8,055,091-8,055,366 and chr17:8,057,549-8,057,958 enhancer loci showing read-depth normalized mapped plasmid reads from purified nuclei isolated from cells transfected with GC-inducible STARR-seq libraries and treated with 100 nM DEX or 0.02% EtOH for 3 h (top six tracks), STARR-seq significance and ChIP peak location. No loci of ~1 Mb assayed had significant differences in nuclear-localized plasmid content between DEX and EtOH conditions.

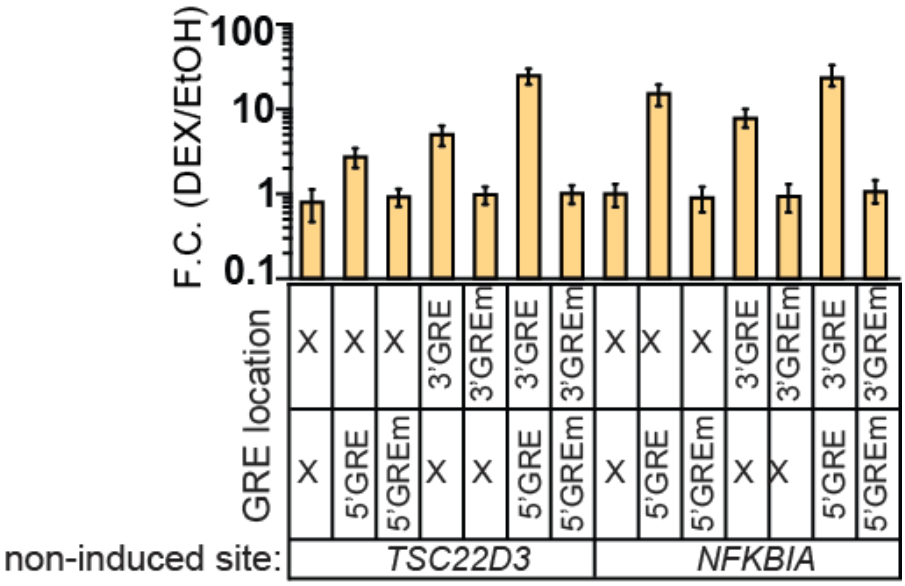

Figure 41: Addition of GREs increases DEX-induced reporter gene expression from sites bound by the GR but not induced in ChIP-reporter assays.

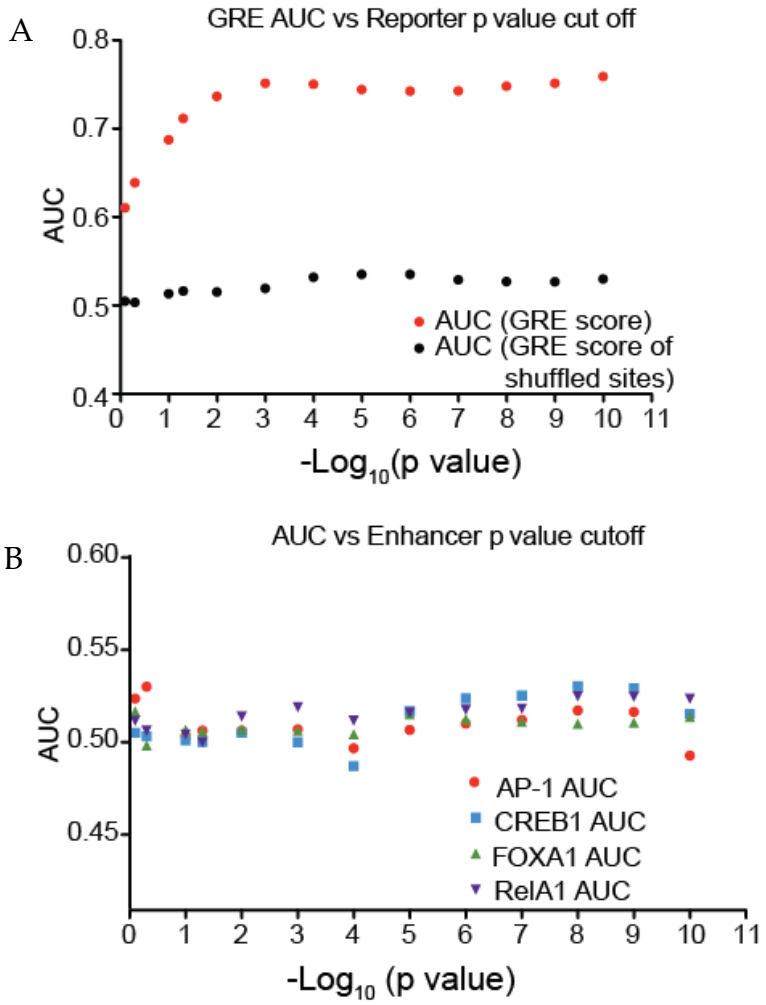

Figure 42: The presence of GRE motifs but not co-factor motifs predicts DEX-induced reporter activity.

(A) GRE prediction of DEX-induced ChIP-reporter activity as a function of FDR threshold. The GRE was used to predict positive DEX-responsiveness (i.e. increased activity after DEX treatment) in ChIP-reporter assays across a range of ChIP-reporter FDR thresholds. For each FDR threshold (x-axis), the AUC for the resulting ROC curve is shown. Red points are for the experimental data, and black are for dinucleotide-shuffled versions of the GBS sequences. **(B)** Prediction of DEX-induced ChIP-reporter activity using DNA binding motifs for co-binding TFs. Responsiveness in ChIP-reporter assays was predicted using DNA motifs for TFs known to bind near or interact with the GR. Data were analyzed and plotted as above. None of the AUCs were significantly better than our null model generated using dinucleotide-shuffled sequences of the GR binding sites.

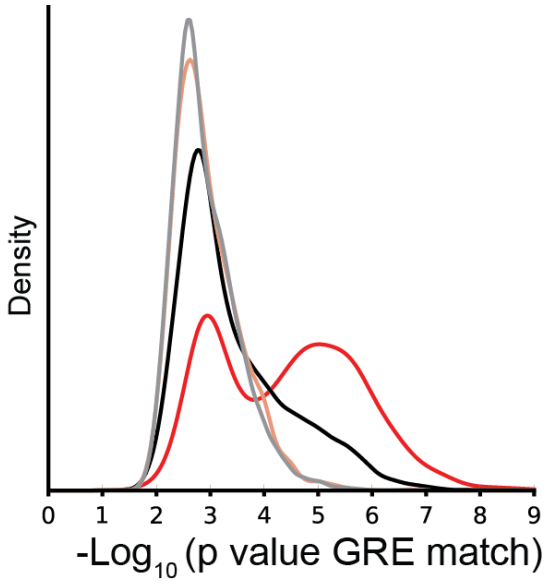

Figure 43: GR motif strength vs. reporter density.

GR motif strength vs. reporter density (red = DEX-induced sites, pink = dinucleotide shuffled sequences from red, black = non-DEX-induced sites, gray= dinucleotide-shuffled sequences from black).

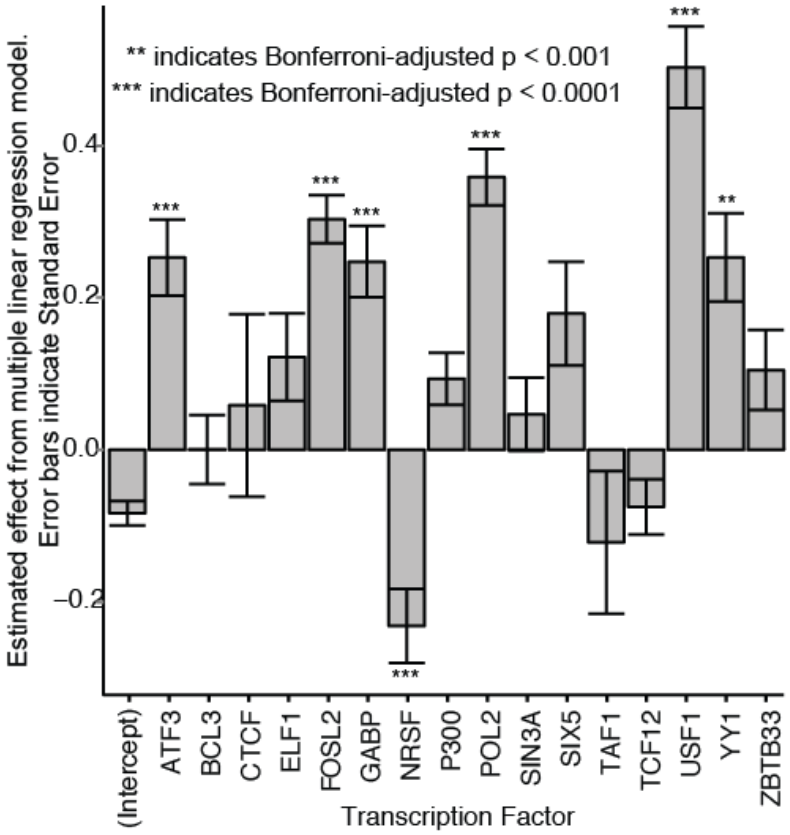

Figure 44: Additive linear regression model of activity in ChIP-reporter assays

Additive linear regression model of activity in ChIP-reporter assays predicted by overlap with TF binding sites. TF binding sites in A549 cells after treatment for 1 h with 0.02% EtOH — similar to our vehicle-control treatment — were obtained from the ENCODE project (**Table S8**). The number of called binding sites that overlapped each GR binding site from our ChIP-seq analysis was calculated. An additive linear regression model was then used to predict the estimated $\log_2$(fold change) in ChIP-reporter activity between the vehicle control library and the input ChIP-reporter plasmid library. All $\log_2$(fold change) estimates were normalized by the standard error of the estimate. Each bar indicates the corresponding regression coefficients, and the error bars are the standard error of the estimate. Stars show statistical significance, as indicated.

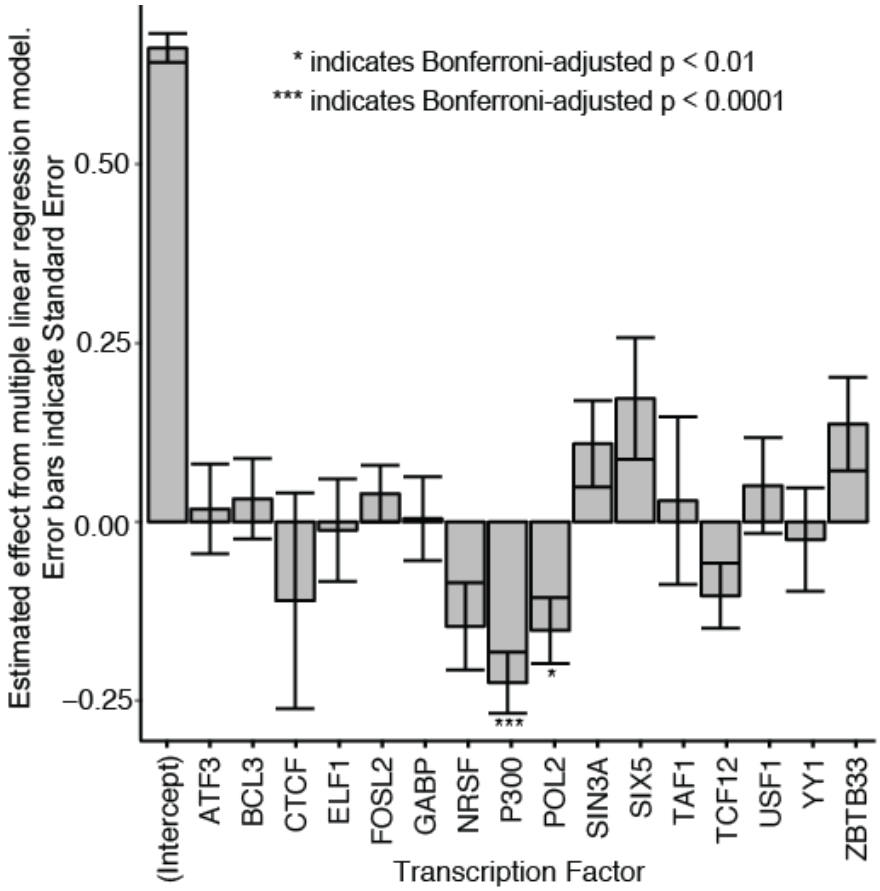

Figure 45: Additive linear regression model of DEX-responsive activity in ChIP-reporter assays.

Additive linear regression model of DEX-responsive activity in ChIP-reporter assays predicted by overlap with TF binding sites. The data and analysis are the same as above, but predicting the $\log_2$(fold change) between DEX and EtOH treatments rather than between EtOH treatment and the plasmid input library.

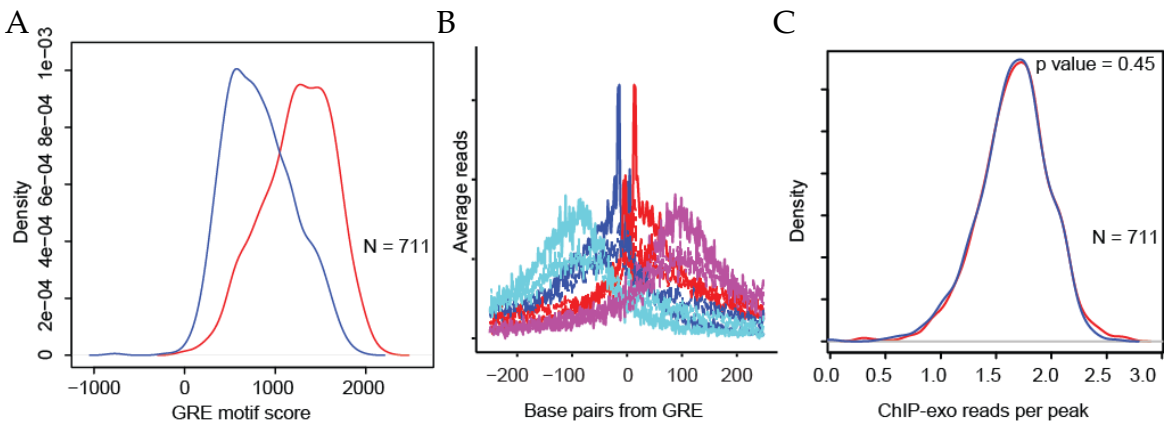

Figure 46: GR ChIP-exo supporting evidence.

(A) Distribution of GRE motif scores among binding footprints at GR sites quantified by ChIP-exo. Motifs from elements that make up the most significant quartile of reporter expression are plotted in red and the least significant are plotted in blue. **(B)** Overlay of ChIP-exo reads for DEX-induced (blue, red) and non-DEX-induced sites (cyan, magenta). **(C)** Distribution of read depth from ChIP-exo reads per peak from DEX-induced and equal number of non-dex-induced sites matched for read depth. Significance calculated with Wilcoxon rank sum test.

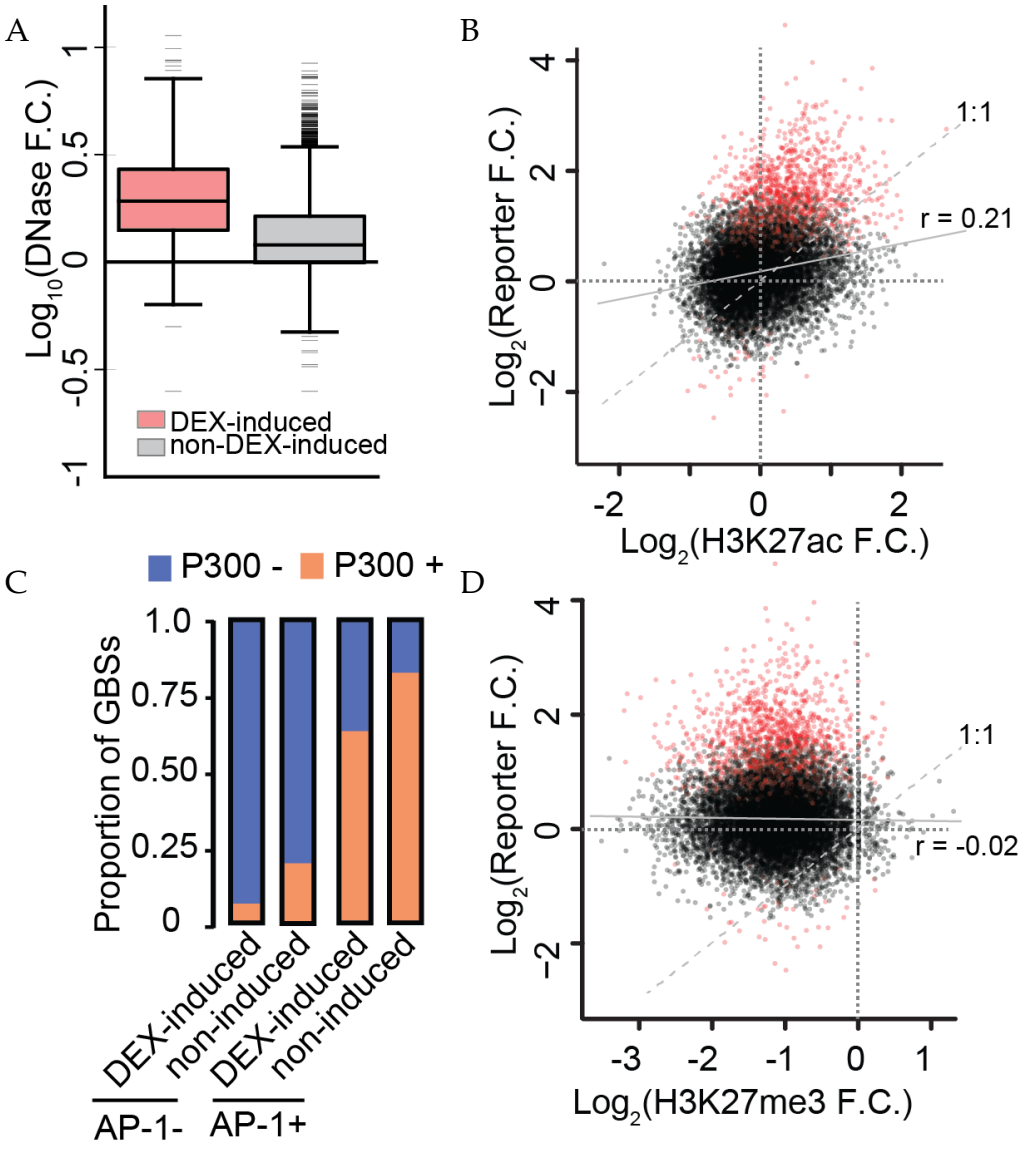

Figure 47: Epigenetic remodeling at ChIP-reporter assayed sites.

(A) Change in DNase-seq signal at sites that are DEX-induced and non-DEX-induced in ChIP-reporter assays. **(B)** Change in in H3K27ac at the reporter tested GR binding sites in A549 cells after 1 h DEX treatment relative to ethanol vs reporter fold change. Correlation coefficient factor line in solid gray. **(C)** Distribution of P300 prior to DEX exposure at among DEX-inducible and non-DEX-inducible ChIP-reporter sites at AP-1 bound and AP-1 unbound sites. **(D)** Change in in H3K27me3 at the ChIP-reporter tested GR binding sites in A549 cells after 1 h DEX treatment relative to ethanol vs reporter fold change. Correlation coefficient factor line in solid gray.

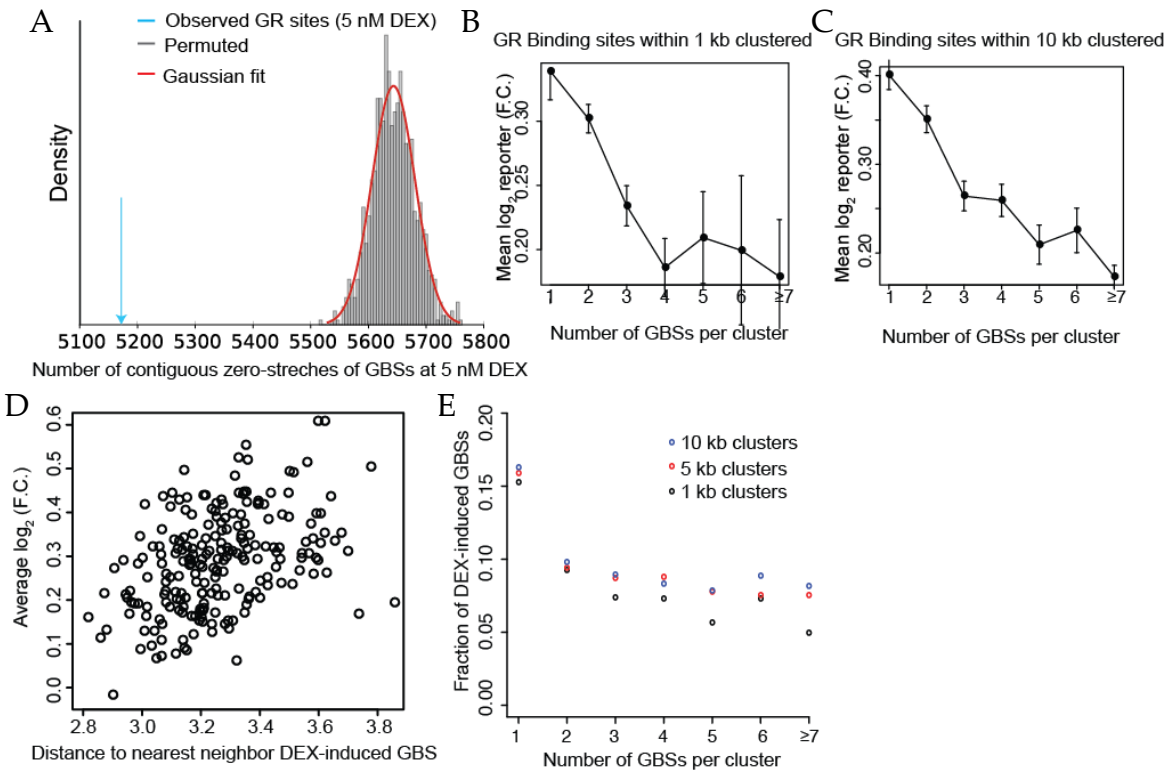

Figure 48: GBSs cluster in the genome.

(A) The number of stretches of unbound GR binding sites at 5 nM DEX (relative to possible binding sites at 50 nM DEX) is shown in blue. Sites were permuted across possible GBS locations at 50 nM DEX 1000 times. The distribution of the stretches of unbound sites in the shuffled background model is shown in red. **(B and C)** Activity as a function of GR cluster isolation thresholds. GR binding sites were assigned to local clusters based on the presence of another GR binding site within **(B)** 1 kb or **(C)** 10 kb. Mean Reporter activity is plotted as a function of number of GR binding events per local cluster. Error bars reflect the standard error of the mean (SEM) **(D)** Average Reporter activity as a function of distance between GR binding sites. GR binding sites were ordered by distance to the nearest adjacent GR binding site. The average Reporter activity was then calculated for non-overlapping windows of GR binding sites. Plotted are the average ChIP-reporter activity and average distance between sites for each window. **(E)** The fraction of GBSs that are DEX-responsive is shown as a function of the cardinality of the cluster for clustering based on a 1 kb (black), 5 kb (red), or 10 kb (blue) window.

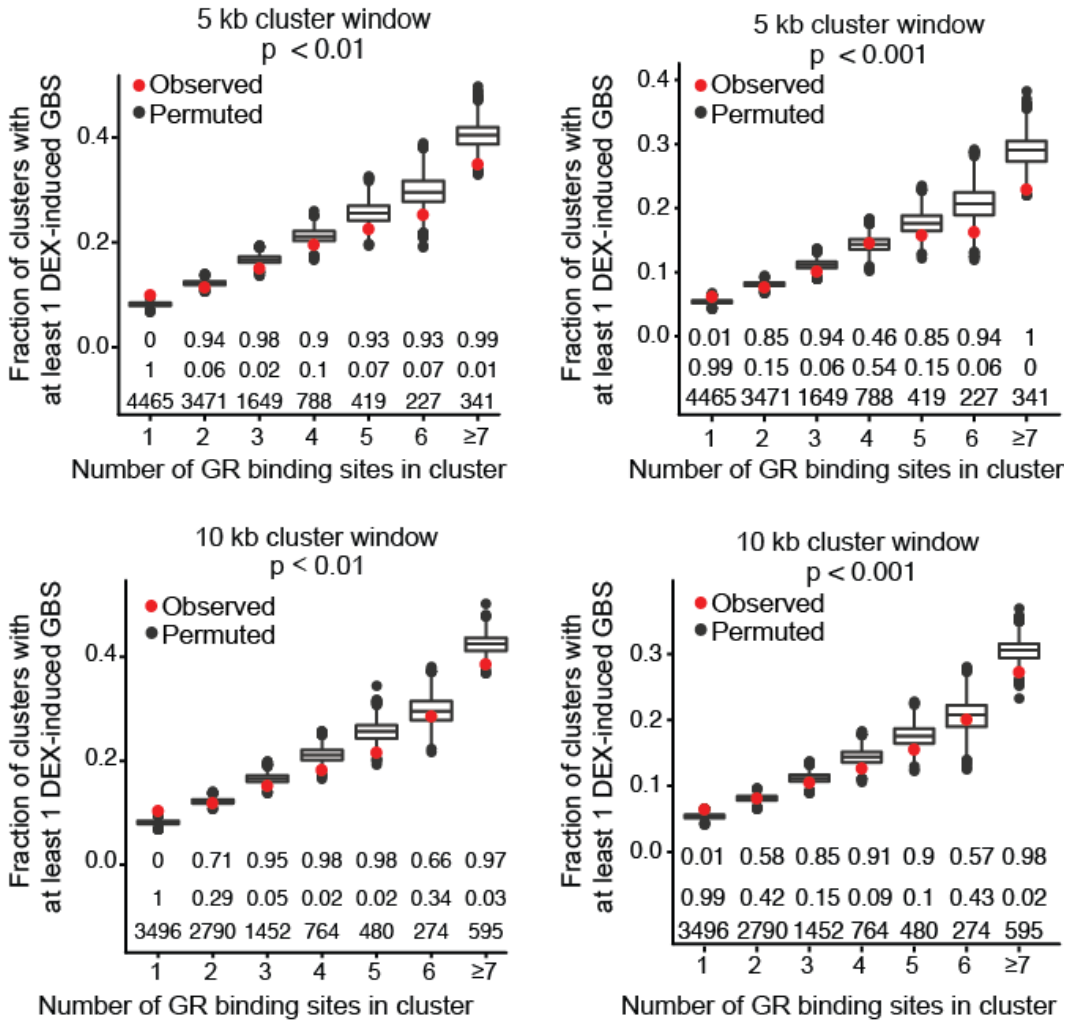

Figure 49: Fraction of GBS clusters with at least one DEX-responsive GBS.

The fraction of clusters with at least one DEX-responsive GBS is shown as a function of the cardinality of the cluster. Within each panel, red indicates fraction of clusters with a DEX-responsive GBS observed in our ChIP-reporter assays. Black box-plots indicate the fraction of clusters with a DEX-responsive GBS across 2000 permutations of the cluster assignments. Empirical p values were calculated to assess whether the observed values deviate significantly from the permutations. The numbers below each column indicate (top) the upper-tail empirical p value, (middle) the lower-tail empirical p value, and (bottom) the number of clusters with the indicated cardinality. Each panel reports data from a different clustering window and a different significance threshold for calling a GBS DEX-responsive, as indicated in the title above each panel.

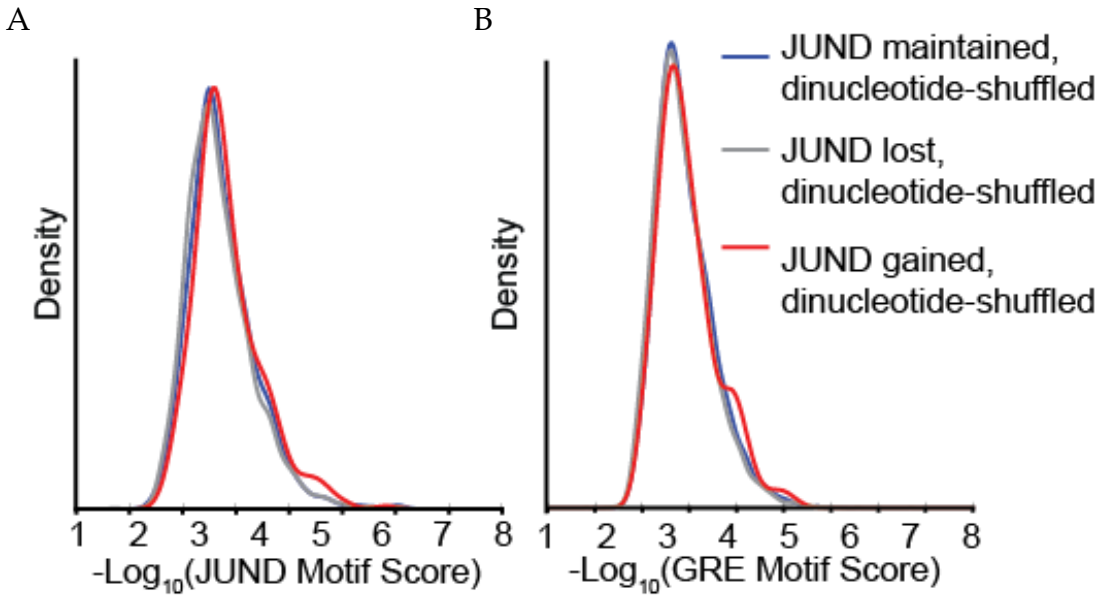

Figure 50: Negative control motif analysis for JUND-GR interaction experiments.

(A) Distribution of AP-1 motif scores calculated after dinucleotide shuffling of sites that gained, maintained, and lost JUND binding at sites that overlap reporter active GR binding sites. **(B)** Distribution of GRE motif scores calculated after dinucleotide shuffling of sites that gained, maintained and lost JUND binding at sites that overlap reporter active GR binding sites.

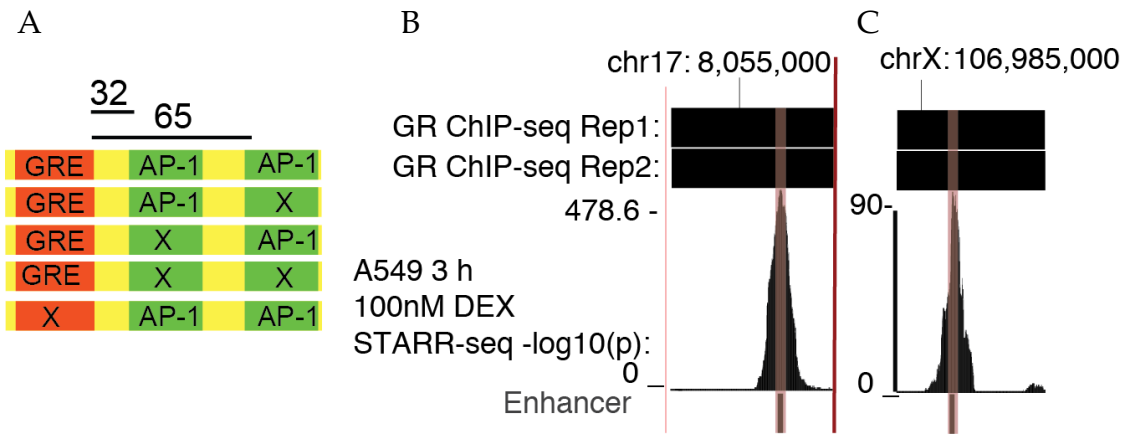

Figure 51: Schematic of GRE/AP-1 epistasis experiments.

(A) Schematic of GRE/AP-1 combinatorial activation vectors used in dual luciferase assay experiments. **(B)** BAC-STARR-seq data showing the distribution of DEX-induced enhancer activity at a GR binding site proximal to the *PER1* gene on Chr. 17. Cloned GC-inducible enhancer highlighted in red. **(C)** BAC-STARR-seq data showing the distribution of GC-inducible enhancer activity at a GR binding site proximal to the *TSC22D3* gene on chromosome X. Cloned DEX-inducible enhancer highlighted in red.

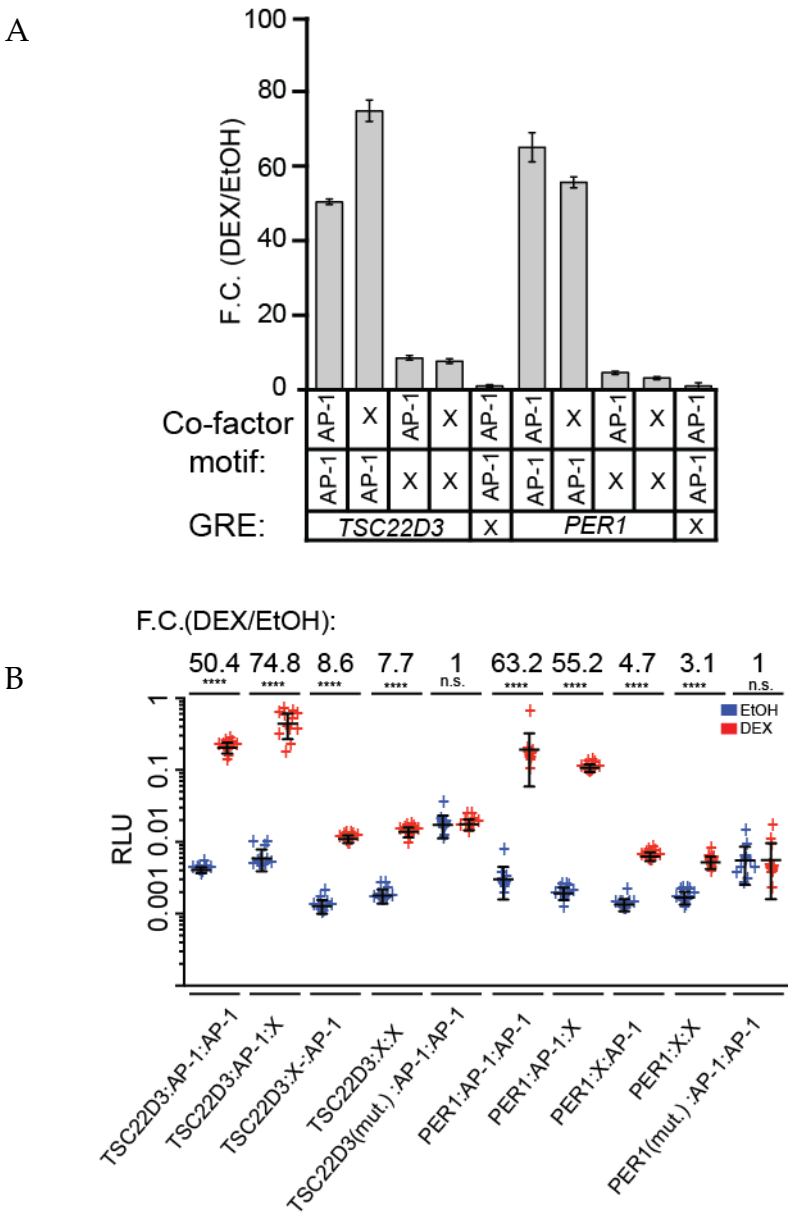

Figure 52: Functional data from GR/AP-1 epistasis experiments.

(A) Dual luciferase assays in A549 cells treated with 100 nM DEX or vehicle control performed using plasmids that contain the DEX-inducible enhancer proximal to the *TSC22D3* or *PER1* gene with combinations two proximal AP-1 binding motifs. **(B)** Dot plots showing the distribution of luciferase activity observed in GRE/AP-1 combinatorial activation experiments displayed in Appx. 1; Figure S6F.

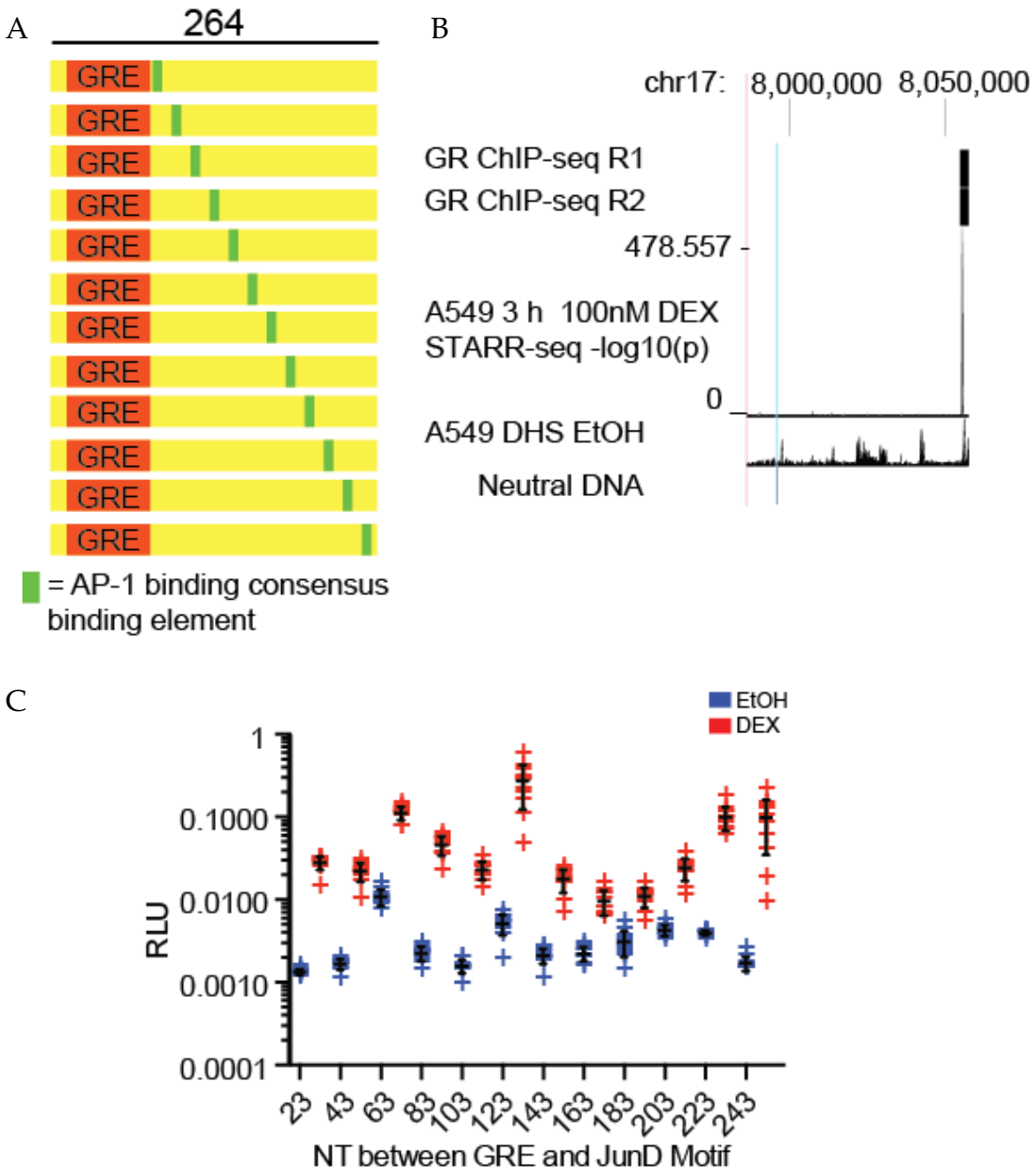

Figure 53: GR/AP-1 motif spacing experiments.

(A) of vectors generated for GRE/AP-1 distal gene activation experiments. **(B)** BAC STARR-seq of neutrally acting DNA used in GRE/AP-1 distal gene activation experiments. **(C)** Dot plots showing the distribution of luciferase activity observed in GRE/AP-1 distal gene activation experiments.

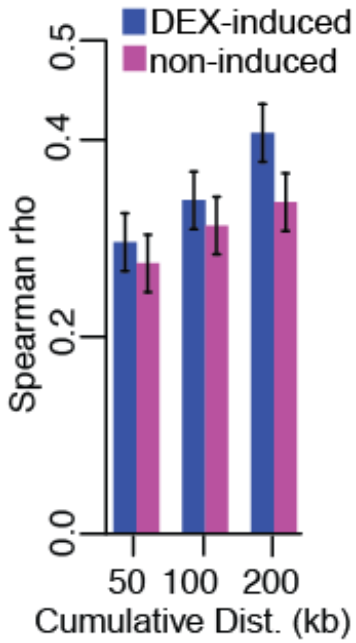

Figure 54: Correlation between reporter activity and endogenous gene regulation.

Correlation (Spearman ϱ +/- 95% CI) between cumulative ChIP-seq signal (log fold-change of DEX response) and gene expression, as a function of distance from TSS, for DEX-induced (blue) and non-DEX-induced (pink) GBSs.

Appendix B

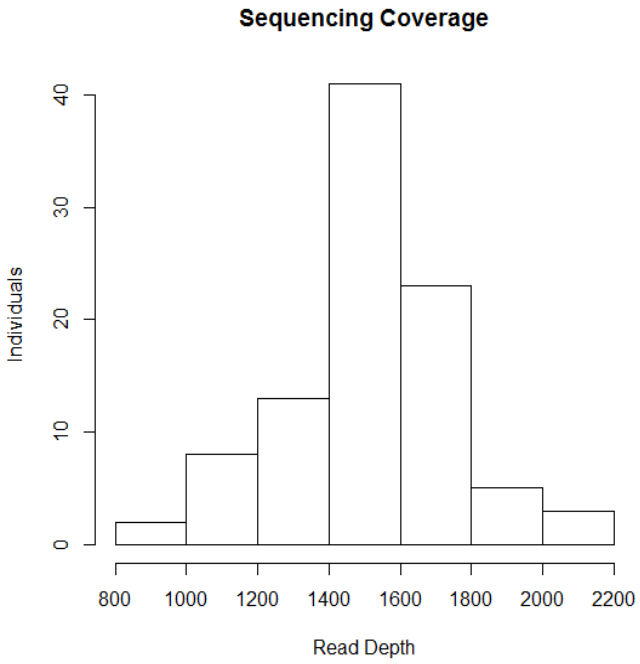

Figure 55: Distribution of TruSeq Custom Amplicon sequencing coverage for 95 individuals.

For each individual, the read depth was determined by calculating the median coverage per amplicon for that specific individual. The median read depth for an individual in the cohort is 1500x.

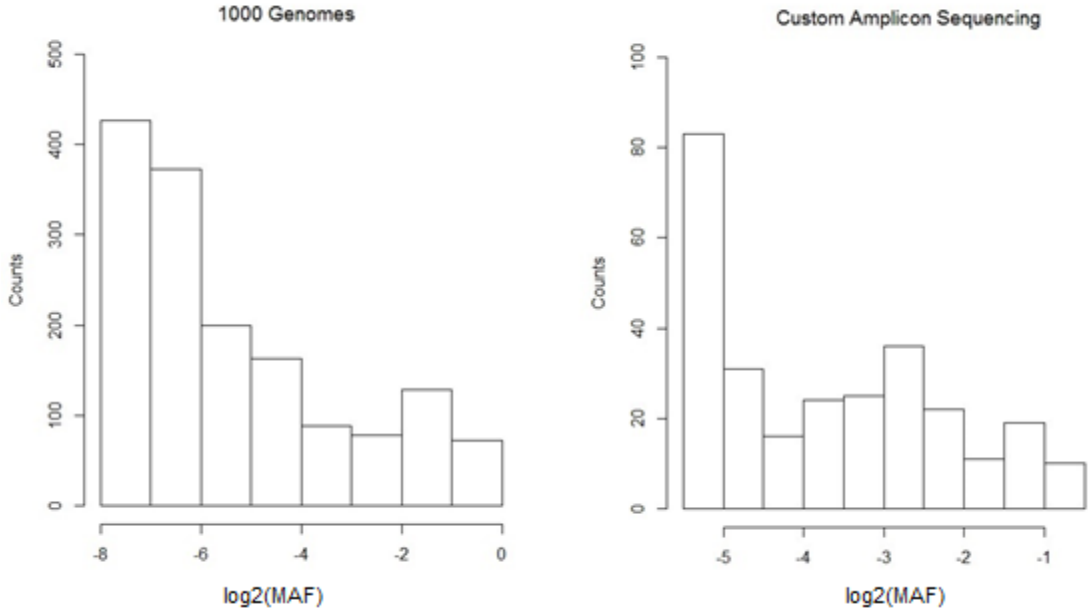

Figure 56: Distribution of allele frequencies.

Distribution of allele frequencies in the 1000 Genomes Project and our TruSeq Custom Amplicon Sequencing for 95 individuals. The high coverage permitted confidently calling a higher number of private variants per individual.

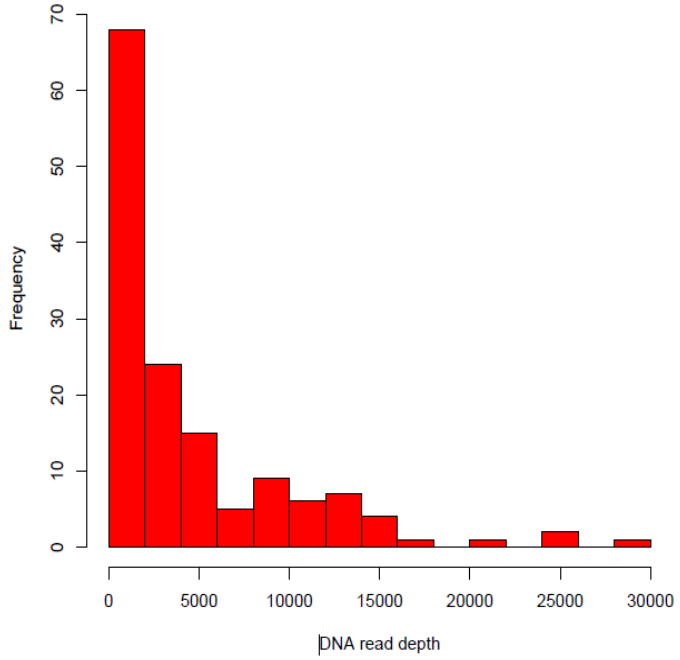

Figure 57: Distribution of median coverage per amplicon in reporter input libraries.

Distribution of median coverage per amplicon of STARR-seq DNA plasmid input library sequencing. The Y-axis represents the number of amplicons and the X-axis represents the median depth per fragment. The median number of times an amplicon was sequenced was 2200 times.

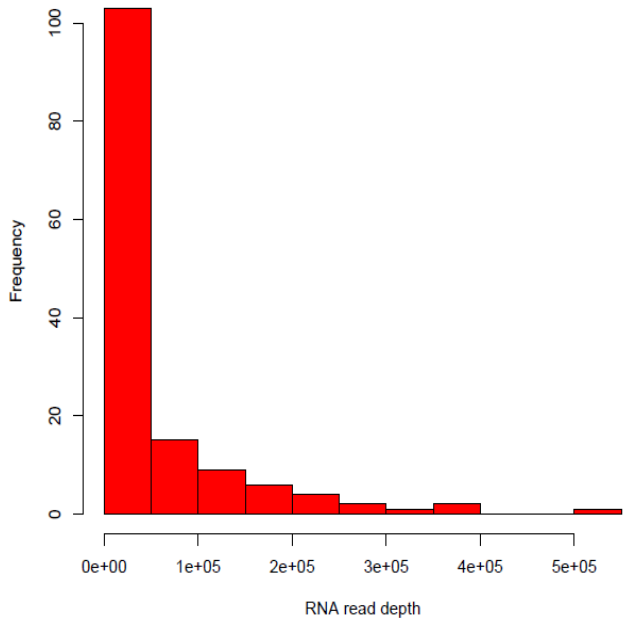

Figure 58: Distribution of reporter RNA-seq coverage.

Distribution of median coverage per amplicon of the RNA-seq output library sequencing. The Y-axis represents the number of amplicons and the X-axis represents the median depth per fragment. The median number of times an amplicon was sequenced was 13,000 times.

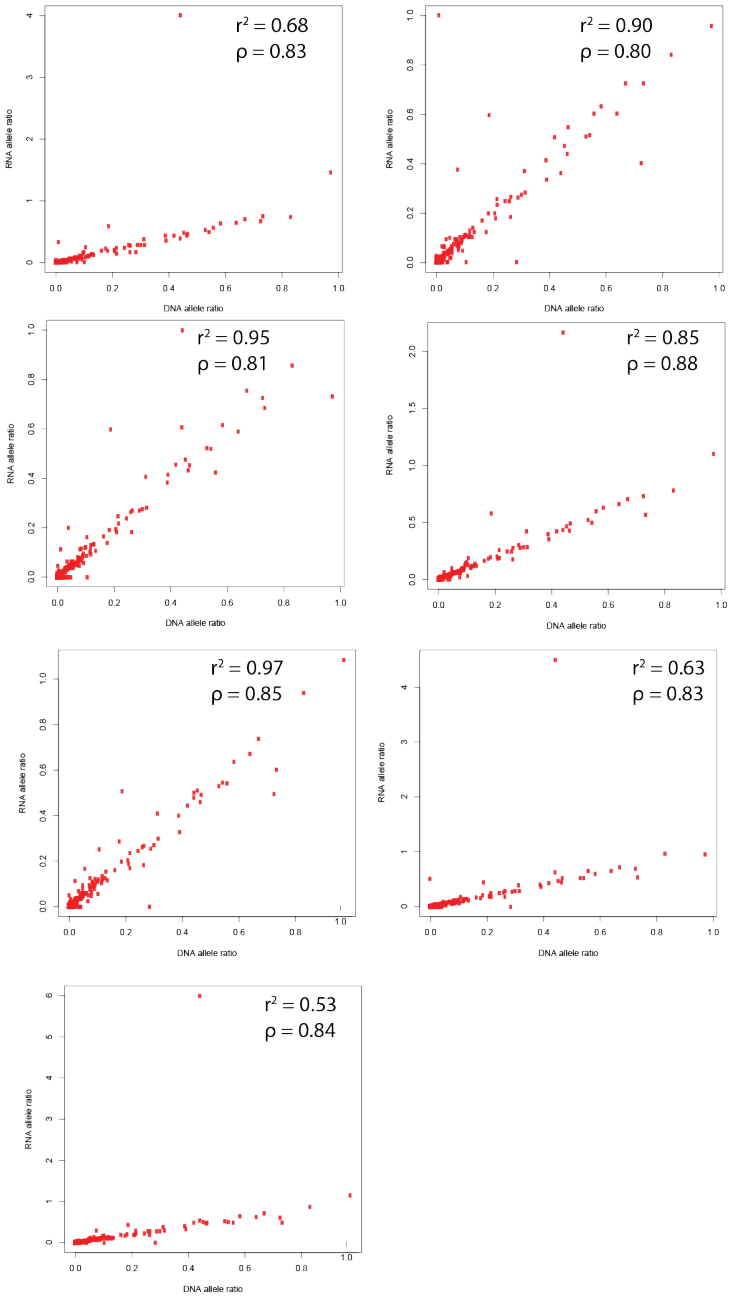

Figure 59: Allele ratios in plasmid vs. RNA-seq reporter libraries.

Allele ratio in the DNA plasmid library (X-axis) versus allele ratio in RNA-seq output libraries (Y-axis) for each replicate. Allele ratio is defined as (number of reads containing Allele_0) / (number of reads containing Allele_1). Allele_0 is the reference allele and Allele_1 is the alternate allele defined by the VCF file. Generally, the alternative allele has lower frequency, although this is not always the case.

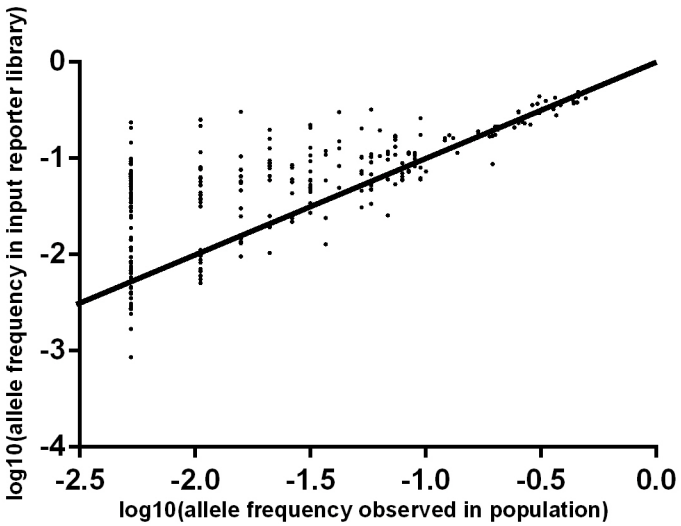

Figure 60: Allele frequency in reporter libraries vs. allele frequency observed in the population.

Plotted is a comparison of the allele frequency of each SNP in the cohort DNA determined by variant calling to the allele frequency of each SNP in the resulting reporter library. Allele frequencies of the cohort DNA used are shown on the X-axis; and the allele frequency in the resulting reporter library are on the Y-axis.

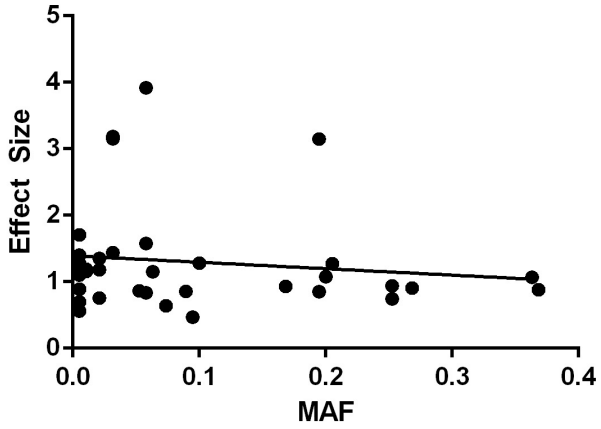

Figure 61: Minor allele frequency vs. variant effect size.

Correlation between minor allele frequency (MAF: X-axis) and variant effect size (Y-axis) for functional variants identified in our population STARR-seq assay (Spearman $\rho = -0.18$, $p = 0.28$).

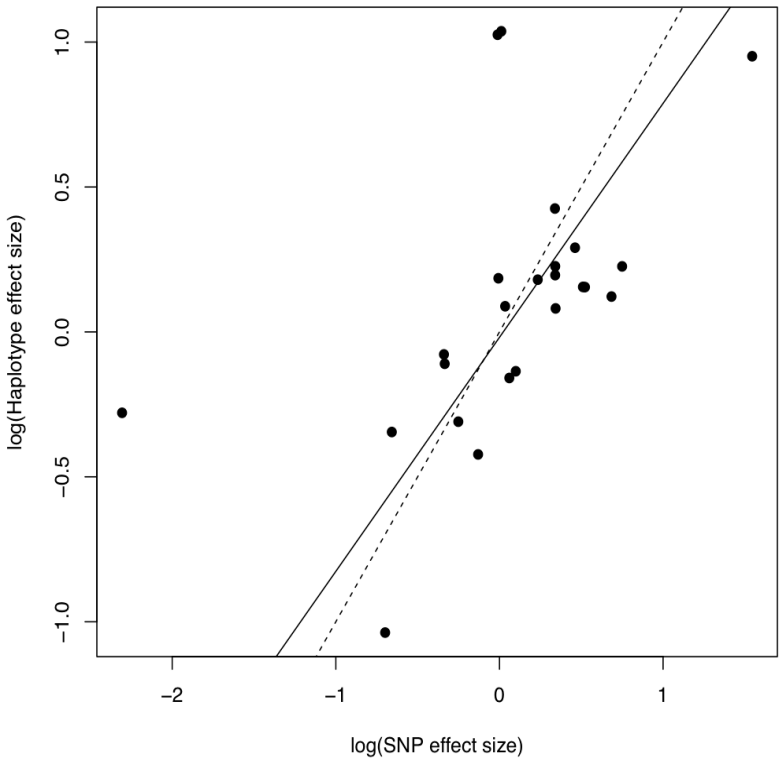

Figure 62: SNP effects vs. haplotype effects.

Correlation between SNP effect sizes (X-axis: log of product of effect sizes of SNPs on haplotype) and log of observed haplotype effects (Y-axis) for putative regulatory haplotypes containing more than one SNP (r = 0.54, p = 0.007). Observed haplotype effect sizes were computed as normalized ratios for each haplotype versus all pooled haplotypes at a locus: ($RNA_{haplotype}/DNA_{haplotype}$)/($RNA_{pooled}/DNA_{pooled}$). Solid line: regression line (slope=0.8, intercept=-0.019); dotted line: 1:1 diagonal.

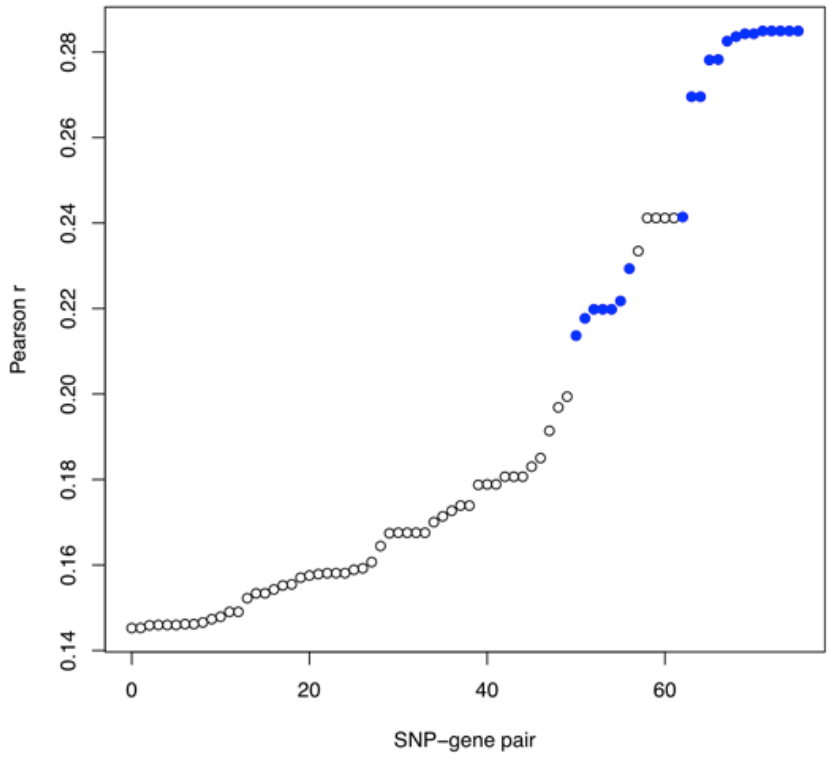

Figure 63: eQTLs associated with the expression of *LINC00881*.

Pearson correlation coefficients for five genes (*LINC00881*:ENSG00000241135.1, *LINC00880*:ENSG00000243629.1, *CCNL1*:ENSG00000163660.7, *TIPARP*:ENSG00000163659.8, *LEKR1*:ENSG00000197980.6) and 67 proximal SNPs. Blue points correspond to eQTLs for *LINC00881* as defined by the GEUVADIS consortium

Table 2: Vockley et al., 2015 Supplemental Table 1: Transition:Transversion ratios in 1000 Genomes and Custom Amplicon Sequencing

	1000 genomes	Custom Amplicon Sequencing
Exons	2.71	2.83
DHS Peaks	2.36	2.31
DHS Peak middle	2.65	2.63
DHS Peak edges	2.25	2.18
Non DHS Intergenic	1.61	1.97
Overall	2.02	2.17

Table 3: Vockley et al., 2015 Supplemental Table 2: Proportion of assayed variation in Population STARR-seq.

	Number of Fragments Sequenced	Number of Variants Called
Custom Amplicon Assay	174	321
Population STARR-seq	173	283
Percent	99.42528736	88.16199377

Table 4: Vockley et al., 2015 Supplemental Table 3: STARR-seq Primers

TS2SSF:	TAGAGCATGCACCGGACACTCTTTCCCTACACGACGCTCTTCCGATCT
TS2SSpatientR:	GGCCGAATTCGTCGATCGCGAGTTAATGCAACGATCGTCGAAATTCGC
PPRead2	TCGCGAGTTAATGCAACGATCGTCGAAATTCGC
PPBCread	GCGAATTTCGACGATCGTTGCATTAACTCGCGA
PPBC1	CAAGCAGAAGACGGCATACGAGATCGTGATTCGCGAGTTAATGCAACGATCGTCGAAATTCG*C
PPBC2	CAAGCAGAAGACGGCATACGAGATACATCGTCGCGAGTTAATGCAACGATCGTCGAAATTCG*C
PPBC3	CAAGCAGAAGACGGCATACGAGATGCCTAATCGCGAGTTAATGCAACGATCGTCGAAATTCG*C
PPBC4	CAAGCAGAAGACGGCATACGAGATTGGTCATCGCGAGTTAATGCAACGATCGTCGAAATTCG*C
PPBC5	CAAGCAGAAGACGGCATACGAGATCACTGTTCGCGAGTTAATGCAACGATCGTCGAAATTCG*C
PPBC6	CAAGCAGAAGACGGCATACGAGATATTGGCTCGCGAGTTAATGCAACGATCGTCGAAATTCG*C
PPBC7	CAAGCAGAAGACGGCATACGAGATGATCTGTCGCGAGTTAATGCAACGATCGTCGAAATTCG*C
PPBC8	CAAGCAGAAGACGGCATACGAGATTCAAGTTCGCGAGTTAATGCAACGATCGTCGAAATTCG*C
PPBC9	CAAGCAGAAGACGGCATACGAGATCTGATCTCGCGAGTTAATGCAACGATCGTCGAAATTCG*C
	*= phosphorothioate bond.

Table 5: Vockley et al., 2015 Supplemental Table 4: Luciferase Validation Primers

chr3 156800768 F	CTGGCCTAACTGGCCGGTACCCCAGCCTGTGTGGATGTTGC
chr3 156806431 F	CTGGCCTAACTGGCCGGTACCGGGGAAGATCAGGGGATGAA
chr3 156812738 F	CTGGCCTAACTGGCCGGTACCAGTTCGTTTTCCGGGGGTGA
chr3 156852592 F	CTGGCCTAACTGGCCGGTACCTGACAGCCCCCTCTAGTGCAG
chr3 156878129 F	CTGGCCTAACTGGCCGGTACCCGGCAGCAGTAGCTGTCGAA
chr3 156898104 F	CTGGCCTAACTGGCCGGTACCTTCTGTAGATGATTGAAATATTTTGGA
chr3 156800768 R	TACCCTAGGGAGATCTCCCTTGTCCCCAGGAAGCTC
chr3 156806431 R	TACCCTAGGGAGATCTCTCTTCCCGCTCGCAGCA
chr3 156812738 R	TACCCTAGGGAGATCTTGAGGGAGCTGTCTTCAGTTCAGA
chr3 156852592 R	TACCCTAGGGAGATCTGCATCAGTTGAGCTGAGGGACA
chr3 156878129 R	TACCCTAGGGAGATCTCCGCCACTTCCCTTGGTACA
chr3 156898104 R	TACCCTAGGGAGATCTCTTCATGGAGAGGTGGAGGA

References

Adey, A., Morrison, H.G., Asan, Xun, X., Kitzman, J.O., Turner, E.H., Stackhouse, B., MacKenzie, A.P., Caruccio, N.C., Zhang, X., and Shendure, J. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol *11*, R119.

Adoue, V., Schiavi, A., Light, N., Almlof, J.C., Lundmark, P., Ge, B., Kwan, T., Caron, M., Ronnblom, L., Wang, C., Chen, S.H., Goodall, A.H., Cambien, F., Deloukas, P., Ouwehand, W.H., Syvanen, A.C., and Pastinen, T. (2014). Allelic expression mapping across cellular lineages to establish impact of non-coding SNPs. Mol Syst Biol *10*, 754.

Allfrey, V.G., Faulkner, R., and Mirsky, A.E. (1964). Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. Proc Natl Acad Sci U S A *51*, 786-794.

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., and Shiroishi, T. (2009). Chromosomal dynamics at the Shh locus: limb bud-specific differential regulation of competence and active transcription. Developmental cell *16*, 47-57.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science *339*, 1074-1077.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res *37*, W202-208.

Bailey, T.L., and Gribskov, M. (1998). Combining evidence using p-values: application to sequence homology searches. Bioinformatics *14*, 48-54.

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

Barski, A., Cuddapah, S., Cui, K., Roh, T.Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823-837.

Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R., Urban, A.E., Montgomery, S.B., Levinson, D.F., and Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res *24*, 14-24.

Belikov, S., Astrand, C., and Wrange, O. (2009). FoxA1 binding directs chromatin structure and the functional response of a glucocorticoid receptor-regulated promoter. Mol Cell Biol *29*, 5413-5425.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological) *57*, 289-300.

Benoist, C., and Chambon, P. (1981). In vivo sequence requirements of the SV40 early promotor region. Nature *290*, 304-310.

Bernstein, E., Duncan, E.M., Masui, O., Gil, J., Heard, E., and Allis, C.D. (2006). Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. Mol Cell Biol *26*, 2560-2569.

Beysen, D., Raes, J., Leroy, B.P., Lucassen, A., Yates, J.R., Clayton-Smith, J., Ilyina, H., Brooks, S.S., Christin-Maitre, S., Fellous, M., Fryns, J.P., Kim, J.R., Lapunzina, P., Lemyre, E., Meire, F., Messiaen, L.M., Oley, C., Splitt, M., Thomson, J., Van de Peer, Y., Veitia, R.A., De Paepe, A., and De Baere, E. (2005). Deletions involving long-range conserved nongenic sequences upstream and downstream of FOXL2 as a novel disease-causing mechanism in blepharophimosis syndrome. American journal of human genetics *77*, 205-218.

Biddie, S.C., John, S., Sabo, P.J., Thurman, R.E., Johnson, T.A., Schiltz, R.L., Miranda, T.B., Sung, M.H., Trump, S., Lightman, S.L., Vinson, C., Stamatoyannopoulos, J.A., and Hager, G.L. (2011). Transcription factor AP1 potentiates chromatin accessibility and glucocorticoid receptor binding. Mol Cell *43*, 145-155.

Birney, E., Lieb, J.D., Furey, T.S., Crawford, G.E., and Iyer, V.R. (2010). Allele-specific and heritable chromatin signatures in humans. Human molecular genetics *19*, R204-209.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., Haussler, D., and Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res *14*, 708-715.

Botta, M., Haider, S., Leung, I.X., Lio, P., and Mozziconacci, J. (2010). Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. Mol Syst Biol *6*, 426.

Brockmann, D., Putzer, B.M., Lipinski, K.S., Schmucker, U., and Esche, H. (1999). A multiprotein complex consisting of the cellular coactivator p300, AP-1/ATF, as well as NF-kappaB is responsible for the activation of the mouse major histocompatibility class I (H-2K(b)) enhancer A. Gene expression *8*, 1-18.

Butz, K., and Hoppe-Seyler, F. (1993). Transcriptional control of human papillomavirus (HPV) oncogene expression: composition of the HPV type 18 upstream regulatory region. Journal of virology *67*, 6476-6486.

Cantor, R.M., Lange, K., and Sinsheimer, J.S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. American journal of human genetics *86*, 6-22.

Chandler, V.L., Maler, B.A., and Yamamoto, K.R. (1983). DNA sequences bound specifically by glucocorticoid receptor in vitro render a heterologous promoter hormone responsive in vivo. Cell *33*, 489-499.

Chen, J., Rozowsky, J., Galeev, T.R., Harmanci, A., Kitchen, R., Bedford, J., Abyzov, A., Kong, Y., Regan, L., and Gerstein, M. (2016). A uniform survey of allele-specific binding and expression over 1000-Genomes-Project individuals. Nature communications *7*, 11101.

Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the functional effect of amino acid substitutions and indels. PloS one *7*, e46688.

Cirillo, L.A., Lin, F.R., Cuesta, I., Friedman, D., Jarnik, M., and Zaret, K.S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. Mol Cell *9*, 279-289.

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.

Collado-Vides, J. (1991). The search for a grammatical theory of gene regulation is formally justified by showing the inadequacy of context-free grammars. Computer applications in the biosciences : CABIOS *7*, 321-326.

Collado-Vides, J. (1992). Grammatical model of the regulation of gene expression. Proc Natl Acad Sci U S A *89*, 9405-9409.

Comings, D.E. (1967). Histones of genetically active and inactive chromatin. The Journal of cell biology *35*, 699-708.

Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal lari, R., Lupien, M., Markowitz, S., and Scacheri, P.C. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res *24*, 1-13.

Crawford, G.E., Holt, I.E., Mullikin, J.C., Tai, D., Blakesley, R., Bouffard, G., Young, A., Masiello, C., Green, E.D., Wolfsberg, T.G., Collins, F.S., and National Institutes Of Health Intramural Sequencing, C. (2004). Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. Proc Natl Acad Sci U S A *101*, 992-997.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A., Boyer, L.A., Young, R.A., and Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A *107*, 21931-21936.

Crisponi, L., Uda, M., Deiana, M., Loi, A., Nagaraja, R., Chiappe, F., Schlessinger, D., Cao, A., and Pilia, G. (2004). FOXL2 inactivation by a translocation 171 kb away: analysis

of 500 kb of chromosome 3 for candidate long-range regulatory sequences. Genomics *83*, 757-764.

Davies, A.F., Mirza, G., Flinter, F., and Ragoussis, J. (1999). An interstitial deletion of 6p24-p25 proximal to the FKHL7 locus and including AP-2alpha that affects anterior eye chamber development. Journal of medical genetics *36*, 708-710.

de Kok, Y.J., Vossenaar, E.R., Cremers, C.W., Dahl, N., Laporte, J., Hu, L.J., Lacombe, D., Fischel-Ghodsian, N., Friedman, R.A., Parnes, L.S., Thorpe, P., Bitner-Glindzicz, M., Pander, H.J., Heilbronner, H., Graveline, J., den Dunnen, J.T., Brunner, H.G., Ropers, H.H., and Cremers, F.P. (1996). Identification of a hot spot for microdeletions in patients with X-linked deafness type 3 (DFN3) 900 kb proximal to the DFN3 gene POU3F4. Human molecular genetics *5*, 1229-1235.

Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., Stephens, M., Gilad, Y., and Pritchard, J.K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature *482*, 390-394.

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

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., and Daly, M.J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet *43*, 491-498.

Devonshire, A.S., Elaswarapu, R., and Foy, C.A. (2010). Evaluation of external RNA controls for the standardisation of gene expression biomarker measurements. BMC genomics *11*, 662.

Diamond, M.I., Miner, J.N., Yoshinaga, S.K., and Yamamoto, K.R. (1990). Transcription factor interactions: selectors of positive or negative regulation from a single DNA element. Science *249*, 1266-1272.

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.

Dominguez, A.A., Lim, W.A., and Qi, L.S. (2016). Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. Nature reviews Molecular cell biology *17*, 5-15.

Driscoll, M.C., Dobkin, C.S., and Alter, B.P. (1989). Gamma delta beta-thalassemia due to a de novo mutation deleting the 5' beta-globin gene activation-region hypersensitive sites. Proc Natl Acad Sci U S A *86*, 7470-7474.

Emison, E.S., McCallion, A.S., Kashuk, C.S., Bush, R.T., Grice, E., Lin, S., Portnoy, M.E., Cutler, D.J., Green, E.D., and Chakravarti, A. (2005). A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk. Nature *434*, 857-863.

Enattah, N.S., Sahi, T., Savilahti, E., Terwilliger, J.D., Peltonen, L., and Jarvela, I. (2002). Identification of a variant associated with adult-type hypolactasia. Nat Genet *30*, 233-237.

ENCODE (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

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

Fang, J., Dagenais, S.L., Erickson, R.P., Arlt, M.F., Glynn, M.W., Gorski, J.L., Seaver, L.H., and Glover, T.W. (2000). Mutations in FOXC2 (MFH-1), a forkhead family transcription factor, are responsible for the hereditary lymphedema-distichiasis syndrome. American journal of human genetics *67*, 1382-1388.

Fang, X., Xiang, P., Yin, W., Stamatoyannopoulos, G., and Li, Q. (2007). Cooperativeness of the higher chromatin structure of the beta-globin locus revealed by the deletion mutations of DNase I hypersensitive site 3 of the LCR. J Mol Biol *365*, 31-37.

Fantes, J., Redeker, B., Breen, M., Boyle, S., Brown, J., Fletcher, J., Jones, S., Bickmore, W., Fukushima, Y., Mannens, M., and et al. (1995). Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype. Human molecular genetics *4*, 415-422.

Farrall, M. (2004). Quantitative genetic variation: a post-modern view. Human molecular genetics *13 Spec No 1*, R1-7.

Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X.S. (2012). Identifying ChIP-seq enrichment using MACS. Nature protocols *7*, 1728-1740.

Feng, Q., Vickers, K.C., Anderson, M.P., Levin, M.G., Chen, W., Harrison, D.G., and Wilke, R.A. (2013). A common functional promoter variant links CNR1 gene expression to HDL cholesterol level. Nature communications *4*, 1973.

Fernandez, B.A., Siegel-Bartelt, J., Herbrick, J.A., Teshima, I., and Scherer, S.W. (2005). Holoprosencephaly and cleidocranial dysplasia in a patient due to two position-effect mutations: case report and review of the literature. Clinical genetics *68*, 349-359.

Fogarty, M.P., Cannon, M.E., Vadlamudi, S., Gaulton, K.J., and Mohlke, K.L. (2014). Identification of a regulatory variant that binds FOXA1 and FOXA2 at the CDC123/CAMK1D type 2 diabetes GWAS locus. PLoS Genet *10*, e1004633.

Frank, C.L., Liu, F., Wijayatunge, R., Song, L., Biegler, M.T., Yang, M.G., Vockley, C.M., Safi, A., Gersbach, C.A., Crawford, G.E., and West, A.E. (2015). Regulation of chromatin accessibility and Zic binding at enhancers in the developing cerebellum. Nature neuroscience *18*, 647-656.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., and Mirny, L.A. (2016). Formation of Chromosomal Domains by Loop Extrusion. Cell reports *15*, 2038-2049.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., Chew, E.G., Huang, P.Y., Welboren, W.J., Han, Y., Ooi, H.S., Ariyaratne, P.N., Vega, V.B., Luo, Y., Tan, P.Y., Choy, P.Y., Wansa, K.D., Zhao, B., Lim,

K.S., Leow, S.C., Yow, J.S., Joseph, R., Li, H., Desai, K.V., Thomsen, J.S., Lee, Y.K., Karuturi, R.K., Herve, T., Bourque, G., Stunnenberg, H.G., Ruan, X., Cacheux-Rataboul, V., Sung, W.K., Liu, E.T., Wei, C.L., Cheung, E., and Ruan, Y. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58-64.

Gabellini, D., Green, M.R., and Tupler, R. (2002). Inappropriate gene activation in FSHD: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle. Cell *110*, 339-348.

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

Gao, X., Vockley, C.M., Pauli, F., Newberry, K.M., Xue, Y., Randell, S.H., Reddy, T.E., and Hogan, B.L. (2013). Evidence for multiple roles for grainyhead-like 2 in the establishment and maintenance of human mucociliary airway epithelium.[corrected]. Proc Natl Acad Sci U S A *110*, 9356-9361.

Ge, B., Pokholok, D.K., Kwan, T., Grundberg, E., Morcos, L., Verlaan, D.J., Le, J., Koka, V., Lam, K.C., Gagne, V., Dias, J., Hoberman, R., Montpetit, A., Joly, M.M., Harvey, E.J., Sinnett, D., Beaulieu, P., Hamon, R., Graziani, A., Dewar, K., Harmsen, E., Majewski, J., Goring, H.H., Naumova, A.K., Blanchette, M., Gunderson, K.L., and Pastinen, T. (2009). Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. Nat Genet *41*, 1216-1222.

Gertz, J., Savic, D., Varley, K.E., Partridge, E.C., Safi, A., Jain, P., Cooper, G.M., Reddy, T.E., Crawford, G.E., and Myers, R.M. (2013). Distinct properties of cell-type-specific and shared transcription factor binding sites. Mol Cell *52*, 25-36.

Gertz, J., Varley, K.E., Davis, N.S., Baas, B.J., Goryshin, I.Y., Vaidyanathan, R., Kuersten, S., and Myers, R.M. (2012). Transposase mediated construction of RNA-seq libraries. Genome Res *22*, 134-141.

Gisselbrecht, S.S., Barrera, L.A., Porsch, M., Aboukhalil, A., Estep, P.W., 3rd, Vedenko, A., Palagi, A., Kim, Y., Zhu, X., Busser, B.W., Gamble, C.E., Iagovitina, A., Singhania, A., Michelson, A.M., and Bulyk, M.L. (2013). Highly parallel assays of tissue-specific enhancers in whole Drosophila embryos. Nature methods *10*, 774-780.

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.

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

Guo, C., Ludvik, A.E., Arlotto, M.E., Hayes, M.G., Armstrong, L.L., Scholtens, D.M., Brown, C.D., Newgard, C.B., Becker, T.C., Layden, B.T., Lowe, W.L., Jr., and Reddy, T.E. (2014). Coordinated Regulatory Variation Associated with Gestational Hyperglycemia Regulates Expression of a Novel Hexokinase HKDC1. Nature Communications *In Press*.

Guo, C., Ludvik, A.E., Arlotto, M.E., Hayes, M.G., Armstrong, L.L., Scholtens, D.M., Brown, C.D., Newgard, C.B., Becker, T.C., Layden, B.T., Lowe, W.L., and Reddy, T.E. (2015a). Coordinated regulatory variation associated with gestational hyperglycaemia regulates expression of the novel hexokinase HKDC1. Nature communications *6*, 6069.

Guo, Y., Xu, Q., Canzio, D., Shou, J., Li, J., Gorkin, D.U., Jung, I., Wu, H., Zhai, Y., Tang, Y., Lu, Y., Wu, Y., Jia, Z., Li, W., Zhang, M.Q., Ren, B., Krainer, A.R., Maniatis, T., and Wu, Q. (2015b). CRISPR Inversion of CTCF Sites Alters Genome Topology and Enhancer/Promoter Function. Cell *162*, 900-910.

Gusev, A., Lee, S.H., Trynka, G., Finucane, H., Vilhjalmsson, B.J., Xu, H., Zang, C., Ripke, S., Bulik-Sullivan, B., Stahl, E., Schizophrenia Working Group of the Psychiatric Genomics, C., Consortium, S.-S., Kahler, A.K., Hultman, C.M., Purcell, S.M., McCarroll, S.A., Daly, M., Pasaniuc, B., Sullivan, P.F., Neale, B.M., Wray, N.R., Raychaudhuri, S., Price, A.L., Schizophrenia Working Group of the Psychiatric Genomics, C., and Consortium, S.-S. (2014). Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. American journal of human genetics *95*, 535-552.

Hagerty, T., Morgan, W.W., Elango, N., and Strong, R. (2001). Identification of a glucocorticoid-responsive element in the promoter region of the mouse tyrosine hydroxylase gene. J Neurochem *76*, 825-834.

Harju, S., Navas, P.A., Stamatoyannopoulos, G., and Peterson, K.R. (2005). Genome architecture of the human beta-globin locus affects developmental regulation of gene expression. Mol Cell Biol *25*, 8765-8778.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C.K., Chrast, J., Lagarde, J., Gilbert, J.G., Storey, R., Swarbreck, D., Rossier, C., Ucla, C., Hubbard, T., Antonarakis, S.E., and Guigo, R. (2006). GENCODE: producing a reference annotation for ENCODE. Genome Biol *7 Suppl 1*, S4 1-9.

Herrlich, P. (2001). Cross-talk between glucocorticoid receptor and AP-1. Oncogene *20*, 2465-2475.

Hertel, K.J., Lynch, K.W., and Maniatis, T. (1997). Common themes in the function of transcription and splicing enhancers. Curr Opin Cell Biol *9*, 350-357.

Hilton, I.B., D'Ippolito, A.M., Vockley, C.M., Thakore, P.I., Crawford, G.E., Reddy, T.E., and Gersbach, C.A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. Nat Biotechnol *33*, 510-517.

Hiragami-Hamada, K., Soeroes, S., Nikolov, M., Wilkins, B., Kreuz, S., Chen, C., De La Rosa-Velazquez, I.A., Zenn, H.M., Kost, N., Pohl, W., Chernev, A., Schwarzer, D., Jenuwein, T., Lorincz, M., Zimmermann, B., Walla, P.J., Neumann, H., Baubec, T., Urlaub, H., and Fischle, W. (2016). Dynamic and flexible H3K9me3 bridging via HP1beta dimerization establishes a plastic state of condensed chromatin. Nature communications *7*, 11310.

Hong, S., Cho, Y.W., Yu, L.R., Yu, H., Veenstra, T.D., and Ge, K. (2007). Identification of JmjC domain-containing UTX and JMJD3 as histone H3 lysine 27 demethylases. Proc Natl Acad Sci U S A *104*, 18439-18444.

Hu, H., Huff, C.D., Moore, B., Flygare, S., Reese, M.G., and Yandell, M. (2013). VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. Genetic epidemiology *37*, 622-634.

Hurtado, A., Holmes, K.A., Ross-Innes, C.S., Schmidt, D., and Carroll, J.S. (2011). FOXA1 is a key determinant of estrogen receptor function and endocrine response. Nat Genet *43*, 27-33.

Imhof, A., Yang, X.J., Ogryzko, V.V., Nakatani, Y., Wolffe, A.P., and Ge, H. (1997). Acetylation of general transcription factors by histone acetyltransferases. Current biology : CB *7*, 689-692.

Itzen, F., Greifenberg, A.K., Bosken, C.A., and Geyer, M. (2014). Brd4 activates P-TEFb for RNA polymerase II CTD phosphorylation. Nucleic Acids Res *42*, 7577-7590.

Jamieson, R.V., Perveen, R., Kerr, B., Carette, M., Yardley, J., Heon, E., Wirth, M.G., van Heyningen, V., Donnai, D., Munier, F., and Black, G.C. (2002). Domain disruption and mutation of the bZIP transcription factor, MAF, associated with cataract, ocular anterior segment dysgenesis and coloboma. Human molecular genetics *11*, 33-42.

Janson, L., and Pettersson, U. (1990). Cooperative interactions between transcription factors Sp1 and OTF-1. Proc Natl Acad Sci U S A *87*, 4732-4736.

Jiang, G., Yang, F., van Overveld, P.G., Vedanarayanan, V., van der Maarel, S., and Ehrlich, M. (2003). Testing the position-effect variegation hypothesis for facioscapulohumeral muscular dystrophy by analysis of histone modification and gene expression in subtelomeric 4q. Human molecular genetics *12*, 2909-2921.

Jiang, M., Anderson, J., Gillespie, J., and Mayne, M. (2008). uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. BMC bioinformatics *9*, 192.

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290-294.

John, S., Sabo, P.J., Thurman, R.E., Sung, M.H., Biddie, S.C., Johnson, T.A., Hager, G.L., and Stamatoyannopoulos, J.A. (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. Nat Genet *43*, 264-268.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497-1502.

Joseph, R., Orlov, Y.L., Huss, M., Sun, W., Kong, S.L., Ukil, L., Pan, Y.F., Li, G., Lim, M., Thomsen, J.S., Ruan, Y., Clarke, N.D., Prabhakar, S., Cheung, E., and Liu, E.T. (2010). Integrative model of genomic factors for determining binding site selection by estrogen receptor-alpha. Mol Syst Biol *6*, 456.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., Alston, J., Mikkelsen, T.S., and Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. Genome Res *23*, 800-811.

Kilpinen, H., Waszak, S.M., Gschwind, A.R., Raghav, S.K., Witwicki, R.M., Orioli, A., Migliavacca, E., Wiederkehr, M., Gutierrez-Arcelus, M., Panousis, N.I., Yurovsky, A., Lappalainen, T., Romano-Palumbo, L., Planchon, A., Bielser, D., Bryois, J., Padioleau, I., Udin, G., Thurnheer, S., Hacker, D., Core, L.J., Lis, J.T., Hernandez, N., Reymond, A., Deplancke, B., and Dermitzakis, E.T. (2013). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. Science *342*, 744-747.

Kioussis, D., Vanin, E., deLange, T., Flavell, R.A., and Grosveld, F.G. (1983). Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. Nature *306*, 662-666.

Kleinjan, D.A., and Lettice, L.A. (2008). Long-range gene control and genetic disease. Advances in genetics *61*, 339-388.

Kleinjan, D.A., Seawright, A., Schedl, A., Quinlan, R.A., Danes, S., and van Heyningen, V. (2001). Aniridia-associated translocations, DNase hypersensitivity, sequence comparison and transgenic analysis redefine the functional domain of PAX6. Human molecular genetics *10*, 2049-2059.

Koch, C.M., Andrews, R.M., Flicek, P., Dillon, S.C., Karaoz, U., Clelland, G.K., Wilcox, S., Beare, D.M., Fowler, J.C., Couttet, P., James, K.D., Lefebvre, G.C., Bruce, A.W., Dovey, O.M., Ellis, P.D., Dhami, P., Langford, C.F., Weng, Z., Birney, E., Carter, N.P., Vetrie, D., and Dunham, I. (2007). The landscape of histone modifications across 1% of the human genome in five human cell lines. Genome Res *17*, 691-707.

Kollias, G., Wrighton, N., Hurst, J., and Grosveld, F. (1986). Regulated expression of human A gamma-, beta-, and hybrid gamma beta-globin genes in transgenic mice: manipulation of the developmental expression patterns. Cell *46*, 89-94.

Kulkarni, M.M., and Arnosti, D.N. (2003). Information display by transcriptional enhancers. Development *130*, 6569-6575.

Kuznetsova, T., Wang, S.Y., Rao, N.A., Mandoli, A., Martens, J.H., Rother, N., Aartse, A., Groh, L., Janssen-Megens, E.M., Li, G., Ruan, Y., Logie, C., and Stunnenberg, H.G. (2015). Glucocorticoid receptor and nuclear factor kappa-b affect three-dimensional chromatin organization. Genome Biol *16*, 264.

Kwasnieski, J.C., Mogno, I., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. Proc Natl Acad Sci U S A *109*, 19498-19503.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature methods *9*, 357-359.

Lappalainen, T., Sammeth, M., Friedlander, M.R., t Hoen, P.A., Monlong, J., Rivas, M.A., Gonzalez-Porta, M., Kurbatova, N., Griebel, T., Ferreira, P.G., Barann, M., Wieland, T., Greger, L., van Iterson, M., Almlof, J., Ribeca, P., Pulyakhina, I., Esser, D., Giger, T., Tikhonov, A., Sultan, M., Bertier, G., MacArthur, D.G., Lek, M., Lizano, E., Buermans, H.P., Padioleau, I., Schwarzmayr, T., Karlberg, O., Ongen, H., Kilpinen, H., Beltran, S., Gut, M., Kahlem, K., Amstislavskiy, V., Stegle, O., Pirinen, M., Montgomery, S.B., Donnelly, P., McCarthy, M.I., Flicek, P., Strom, T.M., Geuvadis, C., Lehrach, H., Schreiber, S., Sudbrak, R., Carracedo, A., Antonarakis, S.E., Hasler, R., Syvanen, A.C., van Ommen, G.J., Brazma, A., Meitinger, T., Rosenstiel, P., Guigo, R., Gut, I.G., Estivill, X., and Dermitzakis, E.T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. Nature *501*, 506-511.

Lee, D.H., and Schleif, R.F. (1989). In vivo DNA loops in araCBAD: size limits and helical repeat. Proc Natl Acad Sci U S A *86*, 476-480.

Lee, J.A., Madrid, R.E., Sperle, K., Ritterson, C.M., Hobson, G.M., Garbern, J., Lupski, J.R., and Inoue, K. (2006). Spastic paraplegia type 2 associated with axonal neuropathy and apparent PLP1 position effect. Annals of neurology *59*, 398-403.

Lettice, L.A., Hill, A.E., Devenney, P.S., and Hill, R.E. (2008). Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. Human molecular genetics *17*, 978-985.

Lettice, L.A., Horikoshi, T., Heaney, S.J., van Baren, M.J., van der Linde, H.C., Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B.A., Endo, N., Shibata, M., Suzuki, M., Takahashi, E., Shinka, T., Nakahori, Y., Ayusawa, D., Nakabayashi, K., Scherer, S.W., Heutink, P., Hill, R.E., and Noji, S. (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc Natl Acad Sci U S A *99*, 7548-7553.

Levene, P.A. (1903). On the Chemistry of the Chromatin Substance of the Nerve Cell. The Journal of medical research *10*, 204-211.

Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. American journal of human genetics *83*, 311-321.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Li, X., and Liao, W.S. (1992). Cooperative effects of C/EBP-like and NF kappa B-like binding sites on rat serum amyloid A1 gene expression in liver cells. Nucleic Acids Res *20*, 4765-4772.

Liu, T., Ortiz, J.A., Taing, L., Meyer, C.A., Lee, B., Zhang, Y., Shin, H., Wong, S.S., Ma, J., Lei, Y., Pape, U.J., Poidinger, M., Chen, Y., Yeung, K., Brown, M., Turpaz, Y., and Liu,

X.S. (2011). Cistrome: an integrative platform for transcriptional regulation studies. Genome Biol *12*, R83.

Liu, X., Wang, L., Zhao, K., Thompson, P.R., Hwang, Y., Marmorstein, R., and Cole, P.A. (2008). The structural basis of protein acetylation by the p300/CBP transcriptional coactivator. Nature *451*, 846-850.

Love, M.I., Huber, W., and Anders, S. (2014a). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

Love, M.I., Huber, W., and Anders, S. (2014b). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2.

Luecke, H.F., and Yamamoto, K.R. (2005). The glucocorticoid receptor blocks P-TEFb recruitment by NFkappaB to effect promoter-specific transcriptional repression. Genes Dev *19*, 1116-1127.

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature *389*, 251-260.

Magram, J., Chada, K., and Costantini, F. (1985). Developmental regulation of a cloned adult beta-globin gene in transgenic mice. Nature *315*, 338-340.

Mathelier, A., Zhao, X., Zhang, A.W., Parcy, F., Worsley-Hunt, R., Arenillas, D.J., Buchman, S., Chen, C.Y., Chou, A., Ienasescu, H., Lim, J., Shyr, C., Tan, G., Zhou, M., Lenhard, B., Sandelin, A., and Wasserman, W.W. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. Nucleic Acids Res *42*, D142-147.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A.K., Canfield, T.K., Giste, E., Diegel, M., Bates, D., Hansen, R.S., Neph, S., Sabo, P.J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S.R., Kaul, R., and Stamatoyannopoulos, J.A. (2012). Systematic

localization of common disease-associated variation in regulatory DNA. Science *337*, 1190-1195.

McDaniell, R., Lee, B.K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A., Keefe, D., Collins, F.S., Willard, H.F., Lieb, J.D., Furey, T.S., Crawford, G.E., Iyer, V.R., and Birney, E. (2010). Heritable individual-specific and allele-specific chromatin signatures in humans. Science *328*, 235-239.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*, 1297-1303.

Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr., Kinney, J.B., Kellis, M., Lander, E.S., and Mikkelsen, T.S. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol *30*, 271-277.

Mittal, R., Kumar, K.U., Pater, A., and Pater, M.M. (1994). Differential regulation by c-jun and c-fos protooncogenes of hormone response from composite glucocorticoid response element in human papilloma virus type 16 regulatory region. Molecular endocrinology *8*, 1701-1708.

Montgomery, S.B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R.P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E.T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. Nature *464*, 773-777.

Muncke, N., Wogatzky, B.S., Breuning, M., Sistermans, E.A., Endris, V., Ross, M., Vetrie, D., Catsman-Berrevoets, C.E., and Rappold, G. (2004). Position effect on PLP1 may cause a subset of Pelizaeus-Merzbacher disease symptoms. Journal of medical genetics *41*, e121.

Ni, Y., Hall, A.W., Battenhouse, A., and Iyer, V.R. (2012). Simultaneous SNP identification and assessment of allele-specific bias from ChIP-seq data. BMC genetics *13*, 46.

Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. PLoS Genet *6*, e1000888.

Nielsen, A.L., Ortiz, J.A., You, J., Oulad-Abdelghani, M., Khechumian, R., Gansmuller, A., Chambon, P., and Losson, R. (1999). Interaction with members of the heterochromatin protein 1 (HP1) family and histone deacetylation are differentially involved in transcriptional silencing by members of the TIF1 family. The EMBO journal *18*, 6385-6395.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Bluthgen, N., Dekker, J., and Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381-385.

O'Brien, E.K., Zhang, X., Nishimura, C., Tomblin, J.B., and Murray, J.C. (2003). Association of specific language impairment (SLI) to the region of 7q31. American journal of human genetics *72*, 1536-1543.

Olansky, L., Welling, C., Giddings, S., Adler, S., Bourey, R., Dowse, G., Serjeantson, S., Zimmet, P., and Permutt, M.A. (1992). A variant insulin promoter in non-insulin-dependent diabetes mellitus. J Clin Invest *89*, 1596-1602.

Pai, A.A., Pritchard, J.K., and Gilad, Y. (2015). The genetic and mechanistic basis for variation in gene regulation. PLoS Genet *11*, e1004857.

Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S.I., Cooper, G.M., Ahituv, N., Pennacchio, L.A., and Shendure, J. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. Nat Biotechnol *30*, 265-270.

Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D., and Shendure, J. (2009). High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat Biotechnol *27*, 1173-1175.

Pearce, D., Matsui, W., Miner, J.N., and Yamamoto, K.R. (1998). Glucocorticoid receptor transcriptional activity determined by spacing of receptor and nonreceptor DNA sites. J Biol Chem *273*, 30081-30085.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., Plajzer-Frick, I., Akiyama, J., De Val, S., Afzal, V., Black, B.L., Couronne, O., Eisen, M.B., Visel, A., and Rubin, E.M. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*, 499-502.

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

Quinlan, A.R. (2014). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *47*, 11 12 11-11 12 34.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665-1680.

Rastegar, S., Hess, I., Dickmeis, T., Nicod, J.C., Ertzer, R., Hadzhiev, Y., Thies, W.G., Scherer, G., and Strahle, U. (2008). The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. Developmental biology *318*, 366-377.

Ratman, D., Vanden Berghe, W., Dejager, L., Libert, C., Tavernier, J., Beck, I.M., and De Bosscher, K. (2013). How glucocorticoid receptors modulate the activity of other transcription factors: a scope beyond tethering. Mol Cell Endocrinol *380*, 41-54.

Reddy, T.E., Gertz, J., Crawford, G.E., Garabedian, M.J., and Myers, R.M. (2012a). The hypersensitive glucocorticoid response specifically regulates period 1 and expression of circadian genes. Mol Cell Biol *32*, 3756-3767.

Reddy, T.E., Gertz, J., Pauli, F., Kucera, K.S., Varley, K.E., Newberry, K.M., Marinov, G.K., Mortazavi, A., Williams, B.A., Song, L., Crawford, G.E., Wold, B., Willard, H.F.,

and Myers, R.M. (2012b). Effects of sequence variation on differential allelic transcription factor occupancy and gene expression. Genome Res *22*, 860-869.

Reddy, T.E., Pauli, F., Sprouse, R.O., Neff, N.F., Newberry, K.M., Garabedian, M.J., and Myers, R.M. (2009). Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. Genome Res *19*, 2163-2171.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., and Muller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics *12*, 77.

Rye, M., Saetrom, P., Handstad, T., and Drablos, F. (2011). Clustered ChIP-Seq-defined transcription factor binding sites and histone modifications map distinct classes of regulatory elements. BMC Biol *9*, 80.

Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. Development *132*, 797-803.

Sakai, D.D., Helms, S., Carlstedt-Duke, J., Gustafsson, J.A., Rottman, F.M., and Yamamoto, K.R. (1988). Hormone-mediated repression: a negative glucocorticoid response element from the bovine prolactin gene. Genes Dev *2*, 1144-1154.

Salmon-Divon, M., Dvinge, H., Tammoja, K., and Bertone, P. (2010). PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. BMC bioinformatics *11*, 415.

Sanborn, A.L., Rao, S.S., Huang, S.C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K.P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E.K., Lander, E.S., and Aiden, E.L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc Natl Acad Sci U S A *112*, E6456-6465.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res *32*, D91-94.

Schultz, D.C., Ayyanathan, K., Negorev, D., Maul, G.G., and Rauscher, F.J., 3rd (2002). SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. Genes Dev *16*, 919-932.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458-472.

Sheppard, K.A., Phelps, K.M., Williams, A.J., Thanos, D., Glass, C.K., Rosenfeld, M.G., Gerritsen, M.E., and Collins, T. (1998). Nuclear integration of glucocorticoid receptor and nuclear factor-kappaB signaling by CREB-binding protein and steroid receptor coactivator-1. J Biol Chem *273*, 29291-29294.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res *29*, 308-311.

Shlyueva, D., Stelzer, C., Gerlach, D., Yanez-Cuna, J.O., Rath, M., Boryn, L.M., Arnold, C.D., and Stark, A. (2014). Hormone-responsive enhancer-activity maps reveal predictive motifs, indirect repression, and targeting of closed chromatin. Mol Cell *54*, 180-192.

Slater, E.P., Rabenau, O., Karin, M., Baxter, J.D., and Beato, M. (1985). Glucocorticoid receptor binding and activation of a heterologous promoter by dexamethasone by the first intron of the human growth hormone gene. Mol Cell Biol *5*, 2984-2992.

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.

So, A.Y., Chaivorapol, C., Bolton, E.C., Li, H., and Yamamoto, K.R. (2007). Determinants of cell- and gene-specific transcriptional regulation by the glucocorticoid receptor. PLoS Genet *3*, e94.

Soccio, R.E., Chen, E.R., Rajapurkar, S.R., Safabakhsh, P., Marinis, J.M., Dispirito, J.R., Emmett, M.J., Briggs, E.R., Fang, B., Everett, L.J., Lim, H.W., Won, K.J., Steger, D.J., Wu, Y., Civelek, M., Voight, B.F., and Lazar, M.A. (2015). Genetic Variation Determines PPARgamma Function and Anti-diabetic Drug Response In Vivo. Cell *162*, 33-44.

Somma, M.P., Pisano, C., and Lavia, P. (1991). The housekeeping promoter from the mouse CpG island HTF9 contains multiple protein-binding elements that are functionally redundant. Nucleic Acids Res *19*, 2817-2824.

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

Stadhouders, R., Aktuna, S., Thongjuea, S., Aghajanirefah, A., Pourfarzad, F., van Ijcken, W., Lenhard, B., Rooks, H., Best, S., Menzel, S., Grosveld, F., Thein, S.L., and Soler, E. (2014). HBS1L-MYB intergenic variants modulate fetal hemoglobin via long-range MYB enhancers. J Clin Invest *124*, 1699-1710.

Staller, M.V., Vincent, B.J., Bragdon, M.D., Lydiard-Martin, T., Wunderlich, Z., Estrada, J., and DePace, A.H. (2015). Shadow enhancers enable Hunchback bifunctionality in the Drosophila embryo. Proc Natl Acad Sci U S A *112*, 785-790.

Starick, S.R., Ibn-Salem, J., Jurk, M., Hernandez, C., Love, M.I., Chung, H.R., Vingron, M., Thomas-Chollier, M., and Meijsing, S.H. (2015). ChIP-exo signal associated with DNA-binding motifs provides insight into the genomic binding of the glucocorticoid receptor and cooperating transcription factors. Genome Res *25*, 825-835.

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS computational biology *6*, e1000770.

Stormo, G.D. (1990). Consensus patterns in DNA. Methods in enzymology *183*, 211-221.

Strahl, B.D., and Allis, C.D. (2000). The language of covalent histone modifications. Nature *403*, 41-45.

Stranger, B.E., and Raj, T. (2013). Genetics of human gene expression. Current opinion in genetics & development *23*, 627-634.

Swinstead, E.E., Miranda, T.B., Paakinaho, V., Baek, S., Goldstein, I., Hawkins, M., Karpova, T.S., Ball, D., Mazza, D., Lavis, L.D., Grimm, J.B., Morisaki, T., Grontved, L., Presman, D.M., and Hager, G.L. (2016). Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin Transitions. Cell *165*, 593-605.

Tang, Z., Luo, O.J., Li, X., Zheng, M., Zhu, J.J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S.Z., Penrad-Mobayed, M., Sachs, L.M., Ruan, X., Wei, C.L., Liu, E.T., Wilczynski, G.M., Plewczynski, D., Li, G., and Ruan, Y. (2015). CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. Cell *163*, 1611-1627.

Teurich, S., and Angel, P. (1995). The glucocorticoid receptor synergizes with Jun homodimers to activate AP-1-regulated promoters lacking GR binding sites. Chemical senses *20*, 251-255.

Tewhey, R., Kotliar, D., Park, D.S., Liu, B., Winnicki, S., Reilly, S.K., Andersen, K.G., Mikkelsen, T.S., Lander, E.S., Schaffner, S.F., and Sabeti, P.C. (2016). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. Cell *165*, 1519-1529.

Thakore, P.I., D'Ippolito, A.M., Song, L., Safi, A., Shivakumar, N.K., Kabadi, A.M., Reddy, T.E., Crawford, G.E., and Gersbach, C.A. (2015). Highly specific epigenome editing by CRISPR-Cas9 repressors for silencing of distal regulatory elements. Nature methods *12*, 1143-1149.

Thoma, F., Koller, T., and Klug, A. (1979). Involvement of histone H1 in the organization of the nucleosome and of the salt-dependent superstructures of chromatin. The Journal of cell biology *83*, 403-427.

Thousand Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., Ibrahim, M., Omar, S.A., Lema, G., Nyambo, T.B., Ghori, J., Bumpstead, S., Pritchard, J.K., Wray, G.A., and Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet *39*, 31-40.

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F., and de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. Mol Cell *10*, 1453-1465.

Townes, T.M., Lingrel, J.B., Chen, H.Y., Brinster, R.L., and Palmiter, R.D. (1985). Erythroid-specific expression of human beta-globin genes in transgenic mice. The EMBO journal *4*, 1715-1723.

Trembath, D.G., Semina, E.V., Jones, D.H., Patil, S.R., Qian, Q., Amendt, B.A., Russo, A.F., and Murray, J.C. (2004). Analysis of two translocation breakpoints and identification of a negative regulatory element in patients with Rieger's syndrome. Birth defects research Part A, Clinical and molecular teratology *70*, 82-91.

Tufarelli, C., Stanley, J.A., Garrick, D., Sharpe, J.A., Ayyub, H., Wood, W.G., and Higgs, D.R. (2003). Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. Nat Genet *34*, 157-165.

Ulirsch, J.C., Nandakumar, S.K., Wang, L., Giani, F.C., Zhang, X., Rogov, P., Melnikov, A., McDonel, P., Do, R., Mikkelsen, T.S., and Sankaran, V.G. (2016). Systematic Functional Dissection of Common Genetic Variation Affecting Red Blood Cell Traits. Cell *165*, 1530-1545.

Urbanek, M., Hayes, M.G., Armstrong, L.L., Morrison, J., Lowe, L.P., Badon, S.E., Scheftner, D., Pluzhnikov, A., Levine, D., Laurie, C.C., McHugh, C., Ackerman, C.M., Mirel, D.B., Doheny, K.F., Guo, C., Scholtens, D.M., Dyer, A.R., Metzger, B.E., Reddy, T.E., Cox, N.J., Lowe, W.L., Jr., and Group, H.S.C.R. (2013). The chromosome 3q25 genomic region is associated with measures of adiposity in newborns in a multi-ethnic genome-wide association study. Human molecular genetics *22*, 3583-3596.

Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., and DePristo, M.A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis [et al] *11*, 11 10 11-11 10 33.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nature reviews Genetics *10*, 252-263.

Viprakasit, V., Kidd, A.M., Ayyub, H., Horsley, S., Hughes, J., and Higgs, D.R. (2003). De novo deletion within the telomeric region flanking the human alpha globin locus as a cause of alpha thalassaemia. British journal of haematology *120*, 867-875.

Vockley, C.M., Guo, C., Majoros, W.H., Nodzenski, M., Scholtens, D.M., Hayes, M.G., Lowe, W.L., Jr., and Reddy, T.E. (2015). Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. Genome Res *25*, 1206-1214.

Wakui, K., Gregato, G., Ballif, B.C., Glotzbach, C.D., Bailey, K.A., Kuo, P.L., Sue, W.C., Sheffield, L.J., Irons, M., Gomez, E.G., Hecht, J.T., Potocki, L., and Shaffer, L.G. (2005). Construction of a natural panel of 11p11.2 deletions and further delineation of the critical region involved in Potocki-Shaffer syndrome. European journal of human genetics : EJHG *13*, 528-540.

Wang, J.C., Derynck, M.K., Nonaka, D.F., Khodabakhsh, D.B., Haqq, C., and Yamamoto, K.R. (2004). Chromatin immunoprecipitation (ChIP) scanning identifies primary glucocorticoid receptor target genes. Proc Natl Acad Sci U S A *101*, 15603-15608.

Wang, Y., Zhang, J.J., Dai, W., Lei, K.Y., and Pike, J.W. (1997). Dexamethasone potently enhances phorbol ester-induced IL-1beta gene expression and nuclear factor NF-kappaB activation. J Immunol *159*, 534-537.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res *42*, D1001-1006.

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. Proc Natl Acad Sci U S A *110*, 11952-11957.

Wild, A., Kalff-Suske, M., Vortkamp, A., Bornholdt, D., Konig, R., and Grzeschik, K.H. (1997). Point mutations in human GLI3 cause Greig syndrome. Human molecular genetics *6*, 1979-1984.

Yikrazuul (2009). CREB-1 binding to DNA https://commons.wikimedia.org/wiki/File:CREB.png (Wikipedia).

Zabidi, M.A., Arnold, C.D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature *518*, 556-559.

Zawistowski, M., Gopalakrishnan, S., Ding, J., Li, Y., Grimm, S., and Zollner, S. (2010). Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. American journal of human genetics *87*, 604-617.

Zheng, C., and Hayes, J.J. (2003). Structures and interactions of the core histone tail domains. Biopolymers *68*, 539-546.

Biography

Christopher Michael Vockley was born in Bryn Mawr, Pennsylvania in 1984. He earned a Bachelors of Science in Cell Biology and Molecular Genetics at the University of Maryland in 2006. After graduation Chris participated in three competitive fellowship training programs: The Oak Ridge Institute for Science and Education Fellowship at the NIH/US FDA, the Post-baccalaureate Intramural Research Training Award Fellowship at the NIH's National Human Genome Research Institute, and a fellowship in the Department of Stem Cell and Regenerative Biology at Harvard University.

Chris has authored or co-authored 14 manuscripts that have been cited nearly 600 times in total. These include the primary subjects of this dissertation; *Direct GR binding sites potentiate clusters of TF binding across the human genome* and *Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort*.

Chris is the recipient of the Ruth L. Kirschstein National Research Scholar Award Individual Pre-doctoral Fellowship from the National Institutes of Health funded by the National Heart, Lung, and Blood Institute (NHLBI). He is also a Lung Repair and Regeneration Consortium Young Investigator and the recipient of a competitive technology development grant from the NHLBI LRRC.