

Regulatory variation and human disease

Matthew Thomas Maurano

A dissertation

submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

John A. Stamatoyannopoulos, Chair

Stephen J. Tapscott

Joshua M. Akey

Program Authorized to Offer Degree:

Genome Sciences

© copyright 2013
Matthew Thomas Maurano

University of Washington

Abstract

Regulatory variation and human disease
Matthew T. Maurano

Chair of the Supervisory Committee:
Associate Professor John A. Stamatoyannopoulos
Departments of Genome Sciences and Medicine

Non-coding regulatory regions are strongly implicated in human disease via genetic studies. However, it is currently not possible to interpret reliably and systematically the functional consequences of genetic variation within any given transcription factor recognition sequence. To lay the groundwork for the assessment of regulatory variation in human disease, I comprehensively analyzed heritable genome-wide binding patterns of a major sequence-specific regulator (CTCF) in relation to genetic variability in binding site sequences across a three-generation pedigree as well as 19 diverse human cell types. We identified hundreds of genetic variants with reproducible quantitative effects on CTCF occupancy (both positive and negative). While these effects paralleled protein-DNA recognition energetics when averaged, they were extensively buffered by striking local context dependencies. Examining variation across multiple cell types, we observed highly reproducible yet surprisingly plastic genomic binding landscapes, indicative of strong cell-selective regulation of CTCF occupancy. Comparison with massively parallel bisulfite sequencing data indicates that 41% of variable CTCF binding is linked to differential DNA methylation, concentrated at two critical positions within the CTCF recognition sequence. These results establish the feasibility of studying the regulatory architecture of human disease.

I then apply the framework developed in the CTCF model system to the interpretation of genome-wide association studies (GWAS), which have identified many non-coding variants associated with common diseases and traits. We show that these variants are concentrated in regulatory DNA marked by DNase I hypersensitive sites (DHSs). 88% of such DHSs are active during fetal development, and are enriched for gestational exposure-related phenotypes. We identify distant gene targets

for hundreds of DHSs that may explain phenotype associations. Disease-associated variants systematically perturb transcription factor recognition sequences, frequently alter allelic chromatin states, and form regulatory networks. We also demonstrate tissue-selective enrichment of more weakly disease-associated variants within DHSs, and the *de novo* identification of pathogenic cell types for Crohn's disease, multiple sclerosis, and an electrocardiogram trait, without prior knowledge of physiological mechanisms. This dissertation establishes a framework for the study of regulatory variation, suggests pervasive involvement of regulatory DNA variation in common human disease, and provides pathogenic insights into diverse disorders.

Table of Contents

List of Figures	iii
List of Tables	v
1 Acknowledgements	1
2 Introduction.....	2
2.1 Summary	2
2.2 Background.....	3
2.3 Recognizing regulatory DNA	5
2.4 Regulatory involvement in human disease	9
2.5 Functional assessment of variation.....	9
2.6 Connecting regulatory elements to their target genes.....	15
2.7 Conclusion.....	17
2.8 Tables	19
3 Genomic Discovery Of Highly Potent Chromatin Insulators For Human Gene Therapy.....	22
3.1 Introduction	22
3.2 Results	22
3.3 Methods	24
3.4 Figures and Tables.....	26
3.5 Notes	32
4 Widespread Site-dependent Buffering Of Human Regulatory Polymorphism.....	33
4.1 Abstract.....	33
4.2 Author Summary	34
4.3 Introduction	35
4.4 Results	36
4.5 Discussion.....	40
4.6 Figures and Tables.....	45
4.7 Materials and Methods.....	58
4.8 Notes	65
5 Widespread Plasticity In CTCF Occupancy Linked To DNA Methylation.....	66
5.1 Abstract.....	66
5.2 Introduction	67
5.3 Results	68
5.4 Discussion.....	72
5.5 Figures and Tables.....	75
5.6 Methods	87
5.7 Notes	91
6 Chromatin Accessibility And DNA Methylation Patterns.....	92
6.1 Results	92
6.2 Figures	94
6.3 Methods	98
6.4 Notes	99
7 DNA Methylation Modulates Transcription Factor Occupancy At Sites Of High Intrinsic Cell Type Variability	101
7.1 Abstract.....	101
7.2 Introduction	102
7.3 Results	103
7.4 Discussion.....	110
7.5 Figures and Tables.....	113
7.6 Materials and Methods.....	133
7.7 Notes	137

8 Systematic Localization Of Common Disease-Associated Variation In Regulatory DNA	138
8.1 Abstract	138
8.2 Results	139
8.3 Discussion.....	147
8.4 Figures and Tables.....	148
8.5 Materials and Methods.....	173
8.6 Notes	184
9 Conclusion	185
9.1 Summary	185
9.2 Discussion And Future Directions	186
References	189
10 Appendix: Supplementary Tables for Section 8	210

List of Figures

Figure 3-1 DNase I footprinting can resolve different modes of binding.....	26
Figure 3-2 Spacing of CTCF upstream motif.....	27
Figure 3-3 Strategy for the discovery of high-affinity CTCF classes	28
Figure 3-4 High-affinity CTCF classes	29
Figure 3-5 Design of construct for quantification of enhancer blocking potency	30
Figure 4-1 Effect of genetic variation on transcription factor occupancy.....	45
Figure 4-2 Genome-wide survey of the effect of genetic variation.....	45
Figure 4-3 Aggregate linkage analysis confirms that association analysis identifies most but not all heritable signal.....	47
Figure 4-4 Functional SNPs recapitulate the CTCF binding motif.....	48
Figure 4-5 Comparison of association and allele-specific analyses.....	49
Figure 4-6 Sequence context buffers effect of polymorphism on occupancy.....	50
Figure 4-7 SNPs at the weakest and strongest sites are less likely to affect occupancy.....	51
Figure 4-8 Hierarchical clustering confirms strength-dependent buffering.....	51
Figure 4-9 SNPs affecting occupancy despite stronger motif contexts involve more severe perturbations.....	52
Figure 4-10 SNPs not affecting occupancy show no relationship between the magnitude of the perturbation and the strength of motif match.....	52
Figure 4-11 Sequence context buffers effect of polymorphism on occupancy.....	53
Figure 4-12 Statistical significance of the mutual information between sequence context and SNPs affecting occupancy.....	53
Figure 4-13 Power of existing measures to predict the effect of regulatory polymorphism.....	54
Figure 4-14 Positive predictive value of PWM model to predict the effect of SNPs on occupancy...	54
Figure 4-15 Statistical identification of association with differential occupancy and allele-specific occupancy.....	56
Figure 5-1 Summary of CTCF ChIP-seq experiments for 19 surveyed cell types.....	75
Figure 5-2 CTCF <i>in vivo</i> binding exhibits widespread plasticity.....	76
Figure 5-3 CTCF binding variability genome-wide	76
Figure 5-4 Degree of overlapping binding landscape.....	77
Figure 5-5 Location of variable CTCF sites relative to genes.....	78
Figure 5-6 CTCF <i>in vivo</i> occupancy exhibits marked plasticity.....	78
Figure 5-7 CTCF occupancy distinguishes similar cell types.....	79
Figure 5-8 CTCF occupancy landscape distinguishes similar cell types.....	79
Figure 5-9 Impact of DNA methylation on cell-selective CTCF binding.....	80
Figure 5-10 Statistical association of variable methylation with differential occupancy.....	81
Figure 5-11 Overall methylation variability across constitutive and variable CTCF sites.....	82
Figure 5-12 Global impact of methylation at variable CTCF sites monitored by RRBS.....	83
Figure 5-13 Reduction in occupancy at methylated sites.....	83
Figure 5-14 Overall relationship of CTCF occupancy with methylation at 1,076 sites without a significant association.....	83
Figure 5-15 Sites significantly affected by methylation are enriched for CpGs at two positions.....	84
Figure 5-16 Variable CTCF sites without methylation differences.....	84
Figure 5-17 Cell-selective patterns of methylation associated with occupancy differences.....	85
Figure 6-1 Chromatin accessibility and DNA methylation patterns.....	94
Figure 6-2 Further examples of association between methylation and accessibility.....	95
Figure 6-3 Global characterization of the effect of methylation on chromatin accessibility, surveyed at 34,376 DHSs with RRBS data.....	96
Figure 6-4 DNase I footprints mark sites of <i>in vivo</i> protein occupancy.....	96
Figure 6-5 Relationship between TF transcript levels and overall methylation at cognate recognition sequences of the same TFs.....	97

Figure 7-1 Profiling of methylation-silenced transcription factor binding sites in stably demethylated cells.....	113
Figure 7-2 Genome-wide analysis of CTCF binding in 40 cell types.	115
Figure 7-3 Sequence characteristics of CTCF sites.....	117
Figure 7-4 Lack of CTCF reactivation in HCT116 DNMT1 (A) and DNMT3B (B) -/- cells.....	118
Figure 7-5 Reactivation of predetermined lineage-specific sites.....	120
Figure 7-6 Chromatin dynamics at CTCF sites	122
Figure 7-7 Chromatin characteristics of response to demethylation.	123
Figure 7-8 Chromatin dynamics at demethylated CTCF sites.	124
Figure 7-9 Global reactivation of chromatin structure.	125
Figure 7-10 Reactivation frequency relates to sequence characteristics of the CTCF recognition sequence.	126
Figure 7-11 CpGs at key positions in the protein-DNA binding interface drive reactivation	127
Figure 7-12 Cell type-specific reactivation highlights methylation-independent repression.....	129
Figure 7-13 Transient depletion of DNA methylation using 5-aza-2'-deoxycytidine.	130
Figure 7-14 Recognition of labile methylation-sensitive sites.....	131
Figure 7-15 Genomic characteristics associated with reactivation.....	132
Figure 8-1 Diseases and traits studied by GWAS and distribution of GWAS variants	148
Figure 8-2 Disease-associated variation is concentrated in DNase I hypersensitive sites.	151
Figure 8-3 Overlap of non-coding GWAS SNPs and regulatory DNA marked by DHSs.....	152
Figure 8-4 Enrichment for regulatory DNA increases with strength of association.	154
Figure 8-5 Multiple distinct genomic disease associations repeatedly localize within relevant cell-selective DHSs.....	155
Figure 8-6 Disease-associated variation impinges on fetal regulatory elements.....	156
Figure 8-7 Localization of GWAS SNPs in DHSs of fetal and adult tissue classes.....	156
Figure 8-8 Phenotype-specific enrichment of regulatory GWAS SNPs for fetal DHSs.	156
Figure 8-9 Correlation analysis links platelet count-associated SNP and JAK2 gene.....	157
Figure 8-10 Regulatory GWAS variants are linked to distant target genes.....	158
Figure 8-11 Distance from correlated DHSs harboring GWAS variants to target promoters.....	158
Figure 8-12 GWAS variants in DHSs localize within physiologically relevant TF binding sites.	159
Figure 8-13 Examples of allele-specific DNase I sensitivity in cell types derived from heterozygous individuals for GWAS variants that alter TF recognition motifs within DHSs.....	160
Figure 8-14 Allelic imbalance distribution.	160
Figure 8-15 Diabetes-associated SNPs cluster in Mendelian diabetes network.....	161
Figure 8-16 Autoimmune disease-associated SNPs cluster in IRF9 interaction network	162
Figure 8-17 Autoimmune disease-associated SNPs cluster in IRF9 interaction network (full).....	163
Figure 8-18 Repeated involvement of the OTX1 pathway in neuropsychiatric diseases and traits...	164
Figure 8-19 Cancer-associated SNPs cluster in orphan nuclear receptor ESRRA network.	165
Figure 8-20 Common disease networks.	166
Figure 8-21 Psychiatric disease network.....	166
Figure 8-22 Identification of pathogenic cell types.	167
Figure 8-23 Enrichment of GWAS SNPs for DHSs by disease/trait.	168

List of Tables

Table 2-1 Levels of evidence for function role of a non-coding variant.....	19
Table 2-2 Selected associations of regulatory variation and human phenotypes.....	20
Table 2-3 Data sources and tools.....	21
Table 3-1 Classes selected for experimental testing.....	30
Table 3-2 Results of enhancer blocking assay.....	31
Table 3-3 Control results of enhancer blocking assay.....	31
Table 4-1 Survey of the effect of SNPs at transcription factor binding sites - 36,834 CTCF sites.	46
Table 4-2 Summary of ChIP-seq data mapped to customized genomes with Mosaik.....	55
Table 4-3 Pearson correlation of two replicates per individual of ChIP-seq data.....	55
Table 4-4 Summary of resequencing data mapped using bwa.....	57
Table 5-1 Growth conditions for immortal cell lines and primary cells.....	86
Table 7-1 Growth conditions for new cell types.....	114
Table 7-2 Summary of CTCF ChIP-seq experiments.....	116
Table 7-3 Sites susceptible to reactivation.....	119
Table 7-4 Qualitative chromatin dynamics at CTCF sites.....	121
Table 7-5 Summary of H3K4me3 and H3K27ac ChIP-seq experiments.....	121
Table 7-6 Summary of DNaseI experiments.....	121
Table 7-7 Summary of bisulfite experiments.....	128
Table 8-1 Mapping of DHSs in 349 cell and tissue samples.....	150
Table 8-2 Enrichment of GWAS and control sites for DHSs.....	153
Table 8-3 Target genes of distal DHSs harboring GWAS variants.....	157
Table 8-4 Cell types used in DHS correlation analysis from Thurman et al.....	169
Table 8-5 New cell types used in DHS correlation analysis.....	170
Table 8-6 Search terms for association of GO Biological Processes with Disease/Trait classes	170
Table 8-7 GWAS SNPs at high-information positions in TF recognition sequences.....	171
Table 10-1 Genes correlated with distal DHSs harboring GWAS SNPs (Part I).....	210
Table 10-2 Genes correlated with distal DHSs harboring GWAS SNPs (Part II).....	215

1 Acknowledgements

Credit is the lifeblood of advancement in science, yet only on rare occasion do we have opportunity to thank those whose contributions have been so constant as to escape formal record, those who likely contributed the most to our success.

My work has built on data that has been generated over many years both within the lab, by the ENCODE consortium, and across the scientific community. Thanks in particular to Tony Shafer, Jun Neri, Kristen Lee, Tanya Kutuyavin, Sandra Stehling-Sun, Theresa Canfield, Erika Giste, Morgan Diegel, Daniel Bates, Scott Hansen, Pete Sabo, Raj Kaul, Shinny Vong, for generating what is collectively one of the best data sets genomics has ever seen.

The informatics team (each of whom has turned out to be a virtuoso outside of work as well) has provided friendship, vast experience, and so many little shortcuts that made all the difference in the long run. Thanks to Bob, Rich, Eric, Eric, Alex, Shane, Audra and Richard for putting in as much as you do. Thanks to Brendan, Dale, John and Mike for continuously holding off disaster. Finally, thanks to Steve for countless hours spent on paperwork – I've never seen you without a smile.

The Department of Genome Sciences and greater University of Washington community have provided a wonderful environment. In particular, George Stamatoyannopoulos has been a constant source of positive encouragement and his contributions to the globin field and dogged pursuit of the developmental clock are an inspiration. Pat Navas has a rare touch with trainees – thanks for the time you gave me. Thanks to Christine Queitsch, Debbie Nickerson, and Mary Claire King for their generosity, to the Chromatin Journal Club for their company, and Brian Giebel for always having a solution.

My committee has provided a constant hand on the tiller. Josh has been an invaluable reference on all things statistical, genetic, and farcical. Evan has repeatedly provided generous support on a moment's notice, in addition to his boundless enthusiasm and a connoisseur's taste for the right experiment. David has a bottomless source of knowledge and physiology and virology (and for standing in as GSR!). Elhanan is characteristically generous and insightful with his thoughts. Stephen provides inspiration that even if the answer is far more intricate than we could have imagined, it can be found. Thanks to you all for your continued support.

And of course, the biggest decision you make when joining a lab is whom to sit next to. When he's not doing an impossible ChIP seven times before breakfast, Hao is probably freelancing for the New Yorker art department – thanks for your generosity. Thanks to Sam for his amazing depth of knowledge and generosity and for being just about the nicest guy you'll ever meet. And thanks to Andrew for always having a constructive comment and to Jeff for fearing nothing -- your companionship in lab instigated an incomparable focus and energy.

Thanks to all my friends and family, especially Taunya, Frans, Gaby and Mark. I could never begin to repay my parents Tom and Paula for thousands of unrecognized moments -- their dedication, their values, and their love have made me who I am. Stephen and Megan have shared my journey from (almost) the start and now give me the gift of being the proud older brother as I watch theirs. Michela: you and your patience, your honor, your smile, and your passion are the dearest things I have.

Finally, I'd like to thank John for all the opportunities he gave me while in his lab. Your sense of how to foster innovation has created an environment where anything is possible. The independence you gave me and the clarity of thought you demanded have led me to where I never could have gotten alone.

This work has been generously supported by the genetics training grant, the NIH ENCODE project, and a pre-doctoral fellowship from NIMH.

2 Introduction

2.1 Summary

The study of human genetics has been revolutionized by high-throughput techniques, allowing the implication of thousands of loci in a wide variety of human diseases and traits. In parallel, genome-wide experimental techniques have enabled comprehensive cataloging of gene expression regulatory elements. These techniques have revealed a major global role for the regulation of gene expression in the genetic architecture of human traits and diseases. Here I discuss the functional classification and detailed mechanistic interpretation of non-coding sequence variation, as well as its connection to gene expression regulatory networks. I consider informative and experimental techniques for the study of regulatory elements, and revisit past approaches in the context of recent experimental results. I discuss strategies for assessing the functional consequences of sequence variation in these regions, highlighting the importance of studying gene expression regulation in its native context, as well as emerging techniques for the global connection of regulatory elements to target genes.

2.2 Background

Regulatory variation has long been posited to comprise a major source of human phenotypic variation (King and Wilson, 1975). The sequencing of the mouse genome made clear that protein coding sequence does not account for the majority of observed evolutionary constraint (Mouse Genome Sequencing Consortium et al., 2002). But the availability of genome sequences alone has not been adequate for the definition of regulatory elements, whose characteristic cell-type specificity presently requires study by experimental approaches. I discuss new developments in the study of human variation in non-coding regions, with a particular focus on sequence-specific transcription factors. While the investigation of coding variation has benefited from a clear definition of protein-coding regions and non-synonymous changes (Sunyaev, 2012), it is only recently that a comparable framework exists for regulatory variation. Although this nascent framework promises a new frontier of large-scale functional investigation, the complex regulation of a locus *in vivo* can be challenging to fully recapitulate and manipulation in model systems, leading to a natural hierarchy of experimental evidence (Table 2-1). Herein I detail approaches for screening functional non-coding variation and suggest that to implicate such variation as functionally involved in human disease, it must have a measurable, direct molecular effect (e.g. on transcription factor binding), leading to a transcriptional dysregulation which is plausibly related to disease etiology.

(a) A gene-centric interpretation of non-coding genetic variation

This review focuses on the mechanics of the study of regulatory variants. However, it is worthwhile to keep in mind the complexity of human transcriptional regulation. Given the low frequency of many disease-associated alleles, and the low effect size of many others, from a medical or scientific perspective it may be more productive to focus on common principles rather than the esoterics of each disease-associated variant. Detailed mutational studies of genes like CFTR show a broad cross-section vulnerable to perturbation to alleles of varying prevalence, and disease-associated variants (while often clustering in functional regions), occur throughout the protein-coding region. Similarly, it is likely that scores of regulatory variants will perturb the expression of a given gene. The important work is to identify the gene and the circumstances of its mis-expression, which can provide a general etiological insight applicable no matter the exact genetic lesion.

(b) Molecular consequences of non-coding genetic variation

Non-coding variation may have a variety of immediate molecular consequences. Perhaps the simplest conceptually is a direct impact of sequence variation on a regulatory element controlling the expression of a gene transcript. These regulatory elements are composed of clusters of short (6 to 10 bp) recognition sequences whose combinatorial recruitment of sequence-specific transcription factors underlies the complex programs of cell-type specific gene expression crucial for development and physiology. Cognate regulatory factors perform manifold functions, acting proximally to promoters as activators or repressors of transcription, or distally as enhancers, insulators and silencers (Maston et al., 2006). Thus, sequence variation at these elements could affect the dynamics, developmental timing, location, or level of expression of a particular gene (Wray et al., 2003). A functional role for genetic variation in these regulatory elements is supported by the observation that much gene expression varies among individuals in a heritable manner (Brem et al., 2002; Cookson et al., 2009; Monks et al., 2004; Morley et al., 2004). Further, detailed studies at specific loci have provided a bestiary of examples where human traits and diseases result from genetic variation in a transcription factor binding site that affects expression of a particular gene (Table 2-2). Although most disease associations would be expected to ultimately require consequent changes in the level of the protein product, most work has measured steady-state gene expression levels (Melzer et al., 2008).

Sequence variation can also influence the content of the mRNA transcript itself, including the use of alternative sites for transcription start, splicing and polyadenylation, which can now be monitored using RNA-seq. Post-transcriptional effects may not necessarily be visible when studying the sequence or expression level of the mRNA, including modulation of RNA-binding proteins, ribosome pausing, ribosome entry sites, or mRNA stability. High-throughput techniques are becoming available to monitor some of these phenomena genome-wide (Ingolia et al., 2009; Licatalosi et al., 2008), but I will not consider them further here. I also will not consider effects such as alterations in spacing between functional binding sites, susceptibility to methylation, or co-factor specificity, though such effects may indirectly manifest in studies of transcription factor binding alone (Leung et al., 2004; Meijnsing et al., 2009; Stadler et al., 2011) (also see Chapters 5 and 6). The gene product subject to regulation may in turn lead to downstream and compensatory regulatory consequenc-

es. An enrichment of GWAS SNPs in long non-coding RNAs has been observed (Kumar et al., 2013), but there exists no framework for interpreting their regulatory action.

2.3 Recognizing regulatory DNA

Sequence-based approaches

Sequence-specific transcription factors bind variations of a short, degenerate consensus sequences. Their binding sites are spread throughout promoter, intronic, exonic and intergenic regions, making them difficult to identify on a large scale. A factor's binding preferences can be experimentally determined and are quantitatively represented most simply by a position weight matrix (PWM) (Stormo, 2000; Wasserman and Sandelin, 2004), though more complex models are available (Badis et al., 2009; Zhao and Stormo, 2011). Such PWMs are available for several hundred of the ~1,400 human transcription factors from the TRANSFAC, JASPAR and UniPROBE databases (Matys et al., 2006; Newburger and Bulyk, 2009; Portales-Casamar et al., 2010) or by a recent large-scale study (Jolma et al., 2013). A sequenced genome can then be scanned informatically for occurrences of the binding motif, which are scored according to the likelihood that their resemblance exceeds random sequence variation (Wasserman and Sandelin, 2004).

However, sequence-based identification of active recognition sites for transcription factors has been disappointing when applied to metazoan genomes. In contrast to bacterial transcription factors which rely on specific interaction with a modest number of binding sites throughout the genome, the average eukaryotic transcription factor does not distinguish binding sites by recognition sequence alone, but rather relies on cooperative binding with additional factors (Wunderlich and Mirny, 2009). Attempts to map regulatory sites by recognizing clusters of binding sites or co-expressed genes have been hampered by the poor quantitative understanding of cooperativity within the chromatin environment and the lack of a gold standard for measuring success. Thus, in practice, the global binding landscape of a particular factor can not be derived from single metazoan genome sequence.

The sequencing of genomes from throughout the phylogenetic tree led to success in high-resolution quantification of selective constraint using models of neutral sequence evolution (Gumucio et al., 1996; Kheradpour

et al., 2007; Xie et al., 2007). But the low information content of recognition sequences, potential for functional redundancy, and positional flexibility inherent in distal interactions enables high turnover, wherein regulatory elements are lost as their functions are duplicated by newly-created ones (Dermitzakis and Clark, 2002). Only about 10 to 22% of experimentally-determined binding sites can be aligned between representative placental mammals (Schmidt et al., 2010). While this rapid divergence favors the use of highly related species, there remains the opposing need to obtain adequate sequence diversity to model the power of purifying selection. The sequencing of 29 mammalian genomes (Lindblad-Toh et al., 2011) represents the zenith of this technique for the study of regulatory elements. An alternative approach instead studies genome sequences from the same species, offering the potential to study species-specific regulatory elements. Here, purifying selection is inferred from a reduction in nucleotide diversity over functional sequence (Kryukov et al., 2007; Vernot et al., 2012). The power of this approach is limited in the short term by the expense of genome sequencing, and ultimately by the strength of purifying selection in human populations (Tennessen et al., 2012). Further, these sequence-based approaches offer little insight into the range of cell types relevant to a regulatory site.

Genetic approaches to gene expression

Expression quantitative trait locus (eQTL) studies, which offer a genetic approach to associate sequence variation with consequent differences in the expression of particular genes, have been the primary experimental genome-wide approach (Brem et al., 2002; Cookson et al., 2009; Monks et al., 2004; Morley et al., 2004; Stamatoyannopoulos, 2004). This approach offers the advantage of studying a sequence variant in its native context *in vivo*, as well as linking it to the gene(s) whose expression it affects. But, these studies have highest statistical power for common sequence variants and ubiquitously expressed genes tolerant of large perturbations to expression levels. Further, genome-wide eQTL studies require samples from several hundreds to thousands of individuals, and thus have typically been limited to convenient tissue types, in particular peripheral blood. eQTL studies that have included other cell types have found great diversity in regulatory architectures, underlining the importance of relevant cell types for the study of gene expression regulation (Dimas et al., 2009; Price et al., 2011). These practical limitations impeding the study of a representative range of cell types

are likely responsible for the proximity of eQTLs to the transcriptional start site, which contrasts with the promoter-distal localization of cell-type specific enhancers (Dimas et al., 2009; Heintzman et al., 2009; Thurman et al., 2012). Finally, although an eQTL interval is presumed to harbor a variant affecting transcription factor binding or transcript splicing or stability, in practice, fine mapping and direct functional follow-up has been limited and the overall sensitivity of the approach for regulatory elements is unclear (Degner et al., 2012; Veyrieras et al., 2008; Zheng et al., 2010; Zhu et al., 2008).

Direct mapping of regulatory elements

Recently, several experimental approaches have been developed to directly map regulatory elements genome-wide. Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) (Farnham, 2009) has been widely employed, including the recent mapping of the binding sites of 119 transcription factors in 72 cell types (ENCODE Project Consortium, 2012). ChIP-seq both surveys binding sites of a given factor as well as quantifies their occupancy. ChIP-seq can also be used to profile a plethora of covalent histone modifications, whose combined patterns can be recognized as distinct chromatin states (Ernst et al., 2011). ChIP-seq can be applied to any cell type for which sufficient material can be obtained. However, a separate experiment must be performed for each factor and cell type of interest, and the method depends entirely upon the availability of a high-quality antibody. Nevertheless, these experimental approaches have enabled for the first time a complete survey of active regulatory DNA. Notably, concordance between sequence-based and experimental approaches has been poor. For example, many regions occupied in heart tissue by the transcriptional co-activator p300 exhibiting weak conservation nevertheless function as enhancers in the heart tissue of transgenic mice embryos (Blow et al., 2010).

Complementary methods more generally survey chromatin accessibility. For over 30 years, DNase I hypersensitivity has been the sine qua non of regulatory elements (Gross and Garrard, 1988). The combinatorial binding of multiple factors at regulatory elements results in a relaxed chromatin state and increased susceptibility to digestion by an endonuclease like DNase I. These DNase I hypersensitive sites (DHSs) encompass the entire functional spectrum of regulatory elements. Other methods can also be adapted to the survey of regulatory elements, including formaldehyde assisted identification of regulatory elements (FAIRE) (Giresi et

al., 2007). As regulatory elements are depleted for methylation (Lister et al., 2009; Stadler et al., 2011) (also see Chapters 5 and 6), there is also the future potential to survey regulatory binding indirectly through methylation profiling. These techniques offer the potential to identify the entire set of regulatory elements active in any cell type of interest for which adequate material is available.

The ENCODE (ENCODE Project Consortium, 2012) and Roadmap Epigenomics (Bernstein et al., 2010a) projects have published in 115 cell and tissue types (both projects are ongoing) and identified 3.9M DHSs (Chapter 8), offering a picture of the global regulatory landscape. The active regulatory DNA in an average cell type covers 2.1% of the genome, and all cell types together cover 42.2%. Regulatory elements are distributed throughout the genome, with only 3% at transcription start sites, and the remainder evenly distributed between intragenic and intergenic regions (Thurman et al., 2012). Notably, while promoter-proximal elements were active in an average of 25 of 125 cell types, promoter-distal elements were highly cell-type specific (Thurman et al., 2012). The extraordinary diversity of these binding landscapes far exceeds the genetic diversity among the cell types studied, placing a clear limit on the utility of purely sequence-based approaches for finding regulatory elements.

Even within regulatory elements, sequence alone is insufficient to accurately identify sites of protein-DNA interaction. Instead, these regions are peppered with consensus recognition sequences for particular factors, dispersed among intervening unconstrained sequence. Further, even sites of direct protein-DNA interaction may permit point sequence changes to proceed in a selectively neutral fashion (Wray et al., 2003) (Chapter 4). Although patterns of constraint clearly reflect protein-DNA interaction genome-wide (Hesselberth et al., 2009; Xie et al., 2007), its predictive power cannot be relied upon at individual sites (Chapter 4).

Additional resolution can instead be gleaned from the experimental techniques originally used to identify regulatory regions. Indeed, although ChIP-seq peaks are several hundred base pairs wide, they tend to be centered around the protein binding site (Kharchenko et al., 2008), offering a low-resolution (~ 20 bp in ideal cases) indication of the exact binding site, which can be further refined at sites of direct binding by recentering on the recognition sequence (Chapter 4). The ChIP-exo technique uses an exonuclease to digest flanking

sequence up to a protein-DNA crosslink, offering nucleotide-level resolution (Rhee and Pugh, 2011). Nucleotide-level resolution can also be obtained by examining DNase I ‘footprints’ - regions within hypersensitive sites protected from digestion as DNase I is occluded by bound protein (Galas and Schmitz, 1978). In deeply-sequenced DNaseI-seq datasets, footprints identify protein occupancy genome-wide (Boyle et al., 2011; Hesselberth et al., 2009; Nepf et al., 2012b). Footprints are enriched for regulatory sequences, which can be searched against databases of known DNA-binding specificity to suggest factors occupying each site *in vivo* (Nepf et al., 2012b). Although the presence of a DNase I footprint at a specific recognition site does not prove the occupancy of a specific factor, its universal nature is well suited for large-scale screens for functional non-coding variation. Thus, experimental techniques are essential to identify precise regulatory sites with a signal-to-noise ratio acceptable for genome-wide scans.

2.4 Regulatory involvement in human disease

Although regulatory regions are often investigated secondarily to protein-coding regions, there is a wide variety of well-studied examples of regulatory involvement in human disease and traits (Table 2-1) (Kleinjan and van Heyningen, 2005; Knight, 2005; Maston et al., 2006). In conjunction with these new technologies for surveying regulatory elements, the breadth of GWAS undertaken by the genetics community (Stranger et al., 2011) has enabled a global census of the role of regulatory polymorphism in human pathophysiology. Recent global analyses in a handful of cell types have identified an enrichment of eQTLs (Nica et al., 2010; Nicolae et al., 2010; Richards et al., 2012; Zhong et al., 2010) or of regulatory states defined by functional genomics data in GWAS SNPs (Ernst et al., 2011; Trynka et al., 2013). DNase I mapping of regulatory elements in 115 cell types has further shown cell-type specific enrichments across multiple trait/disease categories, and extending below the genome-wide significance association cutoff (Chapter 8). Further, these SNPs perturb common regulatory factors networks, setting the stage for a mechanistic investigation of non-coding variation.

2.5 Functional assessment of variation

I now suggest a general strategy, utilizing high-throughput techniques for initial screening paired with detailed mechanistic investigation of promising loci, to definitively establish a non-coding variant’s contribution to a human phenotype (Table 2-1). A key feature of this strategy is a prudent understanding of the complexity of

mammalian gene regulation, particularly its delicate sensitivity to context - that is, the essential components of its native regulation which cannot be omitted from its study. Critical features of regulatory context that may inadvertently be sacrificed to satisfy experimental constraints include expression beyond physiological levels, failure to capture distal regulatory elements, lack of consideration for chromatin state, or absence of appropriate cofactors. To facilitate a widespread vocabulary for discussing the nature of evidence for regulatory involvement in gene expression, I divide experimental techniques into three tiers, comprising of *in vivo* evidence, artificial constructs, and *in vitro* binding assays (Table 2-1).

In vivo study of regulatory variation

A number of methods are available for the faithful study of regulatory polymorphism *in vivo*. Given that a majority of single nucleotide variants directly overlapping a recognition sequence may fail to affect its occupancy (Chapter 4), it is important to confirm that the variant alters occupancy *in vivo* at the element's native context. Experimental techniques such as ChIP-seq and DNaseI-seq can easily establish this in an appropriate model, ideally in patient-derived cells.

Like for all sequencing-based techniques, allelic imbalance in chromatin states can be assessed directly at heterozygous sites to indicate functional differences between haplotypes, with the advantage that both alleles are present in the same nuclear environment (Kadota et al., 2007; Knight et al., 2003; McDaniell et al., 2010; Meissner et al., 2008). However, allele-specific assays are only informative at heterozygous variants and require extremely high sequencing depth. Alternatively, the association between genotype and transcription factor occupancy in *cis* can be established using just tens of samples (Chapter 4), though less-directed scans require substantially more samples (Degner et al., 2012; Zheng et al., 2010).

Using an association-based approach as a gold standard, I have shown that models generated *in vitro* can be combined with experimental localization of *in vivo* binding sites in a cell type of interest (Chapter 4). By examining the difference in information content between the two alleles of a polymorphic recognition sequence active in a relevant cell type, polymorphism overlapping the recognition sequences of the transcription factor CTCF could be classified informatically with moderate sensitivity and specificity, comparing reasonably with

that obtained within coding regions. Thus, the most practical course is a hybrid combination of experimental approaches to identify regulatory elements and informatics approaches to classify variants is functional.

Once a variant has been shown to affect *in vivo* binding, the investigation turns to whether the differential binding affects the expression of a gene. If a sufficient quantity of human tissue samples is available, gene expression differences can be tested for association with natural variation at the native locus. Use of previously published eQTL studies can provide evidence from other populations. Rather than monitoring total expression levels, allele-specific transcription can be tested at heterozygous variants to suggest a functional effect (Pastinen, 2010). However, the steady-state transcript level of a gene is a complex response to multiple present and historical inputs, subject to regulatory feedback (Alon, 2006) and stochastic noise (Eldar and Elowitz, 2010). In this light, clear knowledge of a differentially bound regulatory element can provide an impetus to continue investigation despite the absence of an uncomplicated effect on gene expression.

More meticulous phenotypic confirmation can take place in model systems, starting with focused *in vivo* validation of perturbations to an endophenotype such as gene expression or regulatory factor binding. Given the difficulty of accessing large quantities of appropriate human tissue and the complications of manipulating full-length loci, genome-editing techniques present a valuable opportunity for flexible *in situ* study, especially when combined with *in vitro* differentiation of stem cells (Soldner et al., 2011). These techniques ultimately offer access to a supply of otherwise inaccessible tissues, while ensuring a genetic background permissive for the effect of the variant.

Large constructs such as bacterial artificial chromosomes (BAC) and yeast artificial chromosomes (YAC) offer the potential for transgenic mice using essentially complete loci. This technology offers the opportunity for faithful study of complex regulatory phenotypes such as β -globin switching (Peterson et al., 1995) or of loci not otherwise present in mice (Frazer et al., 1995). Constructs differing by a single nucleotide can be generated to enable the precise attribution of phenotypic consequences to that single variant, such as for the attribution of the Greek $\Lambda\gamma$ form of hereditary persistence of fetal hemoglobin to a single variant in the core promoter (Table 2-2) (Peterson et al., 1995).

A simpler system is a transgenic mouse enhancer assay in which the developmental tissue-specificity of a putative enhancer sequence can be recognized by whole-mount microscopy during embryonic day 11.5 (Pennacchio et al., 2006). Although enhancers are tested outside their native locus with a minimal promoter, this technique allows for moderate-throughput study of non-coding sequence, and precise localization of its tissue specificity. This technique has been applied to 1760 sequence elements (accessed August 2012), 893 of which have demonstrated enhancer activity (Visel et al., 2007). Transgenic mice harboring the disease-associated allele can then be used for phenotypic studies. As human-mouse sequence homology in distal regulatory regions is low, transgenic models incorporating intact human regulatory regions may prove more faithful to the human phenotype (Peterson et al., 1995). Other model organisms can offer advantages of throughput if appropriate for the disease or trait. For example, a recent study of 15 genes harboring novel associations with platelet count and volume used gene silencing in *D. rerio* and *D. melanogaster* to implicate 11 of these as novel hematopoietic regulators (Gieger et al., 2011). To expand mechanistic insight beyond the foothold offered by a variant in a recognition sequence of a single transcription factor, other regulatory targets of that same transcription factor, or its own upstream regulators can be studied for phenotypic impact (Chapter 8).

Enhancer-promoter constructs

These *in vivo* studies can be complemented by more detailed investigation of short enhancer-promoter constructs in an appropriate cell type. Fortunately, catalogs of functional genomics data can provide insight into the set of cell types in which a regulatory element is active (Chapter 8). Classical reporter assays can be used to test the ability of a given haplotype to promote transcription of a reporter gene in cell culture or *in vitro*. These techniques are fundamentally suited to the study of single, short (<1 kb) regulatory regions with a single, defined promoter element. To ensure relevant results, the chosen enhancer must be carefully matched to the promoter element and host cell type. Although enabling easy manipulation, these constructs' small size limits their ability to dissect the complexity of regulation at *in vivo* mammalian loci, including long-distance regulatory interactions between multiple loci.

Once prepared, the construct can be conveniently transfected into a cell line transiently (i.e. as a non-replicating plasmid). Although useful for gross assessment of transcriptional capacity, variable copy number

and the lack of faithful chromatinization compromises the physiological relevance of these transient assays (Smith and Hager, 1997). A stable transfection protocol including a selectable marker results in chromosomal integration and subjects the construct to a more realistic chromatin environment, though this in turn may subject it to position effects and long-term silencing.

Recently, many groups have employed high-throughput techniques to extend reporter gene assays to the study of the transcriptional capacity of tens of thousands of variants of the same site simultaneously *in vivo*. Programmable microarrays can be employed to synthesize regulatory elements, which are then transcribed in a cell-free system, and quantified using high-throughput sequencing (Patwardhan et al., 2009). A related *in vivo* method used randomly-generated oligonucleotides injected as plasmids into rat tail veins, with expression surveyed in the liver (Patwardhan et al., 2012). This method has also been applied to generate models of the transcriptional code governing combinations of regulatory elements (Kinney et al., 2010; Melnikov et al., 2012; Sharon et al., 2012). The most insightful of these was performed in yeast, and quantified the effect of systematic changes in the location, number, orientation, affinity and organization of transcription factor binding sites (Sharon et al., 2012). As a whole, these studies have confirmed the rules of gene expression worked out over the past several decades.

These high-throughput reporter assays promise to become a workhorse in the study of general gene expression regulatory principles, especially in the studying integrated effect of multiple regulatory sites on transcription. However, they share a number of shortcomings for the dissection of specific sites of interest or the contribution of specific sites to complex endogenous loci. First, these techniques depend either on random pools of oligonucleotides, offering little control over the uniformity and characteristics of the variant sites tested, or on synthesis by programmable microarray, the technology for which is expensive and restricted in use. Second, the expense of oligonucleotide synthesis favors the application of these techniques to a limited number of easily-accessible cell types, neglecting the effects of the cell-type specific regulatory milieu. Third and more fundamentally, most of the random sequences tested will be extremely rare in human populations and thus uninformative. Thus these techniques are more adapted to the exhaustive study of a few sites in detail or the

determination of global sensitivities of specific factors to sequence perturbation rather than to sensitive global screening of functional polymorphism.

In vitro study of protein-DNA interaction

Well-established *in vitro* approaches for the determination of protein-DNA specificity can be used to confirm the exact binding sites, their cognate factor, and their sensitivity to perturbation by polymorphism. The electrophoretic mobility shift assay (EMSA) remains the gold standard for the establishment of a specific interaction between a DNA fragment and a protein of interest *in vitro* (Hellman and Fried, 2007), and can be used to identify and localize a difference in affinities between haplotypes. More high-throughput techniques such as protein-binding microarrays (PBM) (Bulyk et al., 2001) can be used to derive the affinity for all possible perturbations of recognition sequences of a particular transcription factor.

Non-coding variation at the 1p13 cholesterol locus

A recent study provides an excellent example of the application of these techniques to the functional characterization of a non-coding GWAS SNP. Musunuru et al. investigated an association for low-density lipoprotein cholesterol (LDL-C) and myocardial infarction (MI), localized to a non-coding region between two genes of unknown function on chromosome 1p13 (Musunuru et al., 2010). By examining eQTLs from human liver (n=960), subcutaneous fat (n=433), and omental fat (n=520), they found a liver-specific association with the expression of three genes, the strongest of which was with SORT1. The authors performed fine mapping and identified two predominant haplotypes composed of 6 SNPs in strong LD. To assess which of these SNPs might be responsible for the observed eQTL, the authors placed each of the haplotypes into a construct harboring an SV40 promoter driving a luciferase reporter. They then transiently transfected these constructs into Hep3B human hepatocellular carcinoma cells. Importantly, the authors used the construct to test individual SNPs in an isogenic background, and showed that one SNP alone (the most highly associated) was responsible for the difference in luciferase expression between haplotypes. Based on the presence of a predicted binding site for C/EBP α , the authors confirmed allele-specific binding *in vitro* using an electrophoretic mobility shift assay (EMSA). The authors were able to confirm reduced expression *in vivo* upon expression of a domi-

nant negative C/EBP protein. Further, in NIH 3T3 mouse fibroblast cells exogenous expression of C/EBP α was required for a difference in expression between the alleles.

To link this difference in transcriptional capacity to the putative target gene SORT1, the authors tested the expression of SORT1 with and without exogenous C/EBP α in two different human embryonic stem cell lines. Upon differentiation into endoderm, only the cell line homozygous for the minor allele showed increased expression. The authors then studied Sort1 expression *in vivo* using knockin mice transfected with a liver-specific adeno-associated virus serotype 8 (AAV8) vector containing murine Sort1 and knockdown mice tail-vein injected with an siRNA. They found Sort1 was associated with decreased hepatocyte secretion of very low-density lipoprotein (VLDL), and lower plasma cholesterol, LDL-C, and VLDL levels. Thus the authors hypothesize that the minor allele confers C/EBP α binding, thus increasing SORT1 expression in liver, which reduces LDL-C and VLDL levels, thereby explaining the reduced risk for myocardial infarction.

2.6 Connecting regulatory elements to their target genes

A major hurdle to a regulatory genomics outlook remains the poor understanding of the sphere of influence of a distal regulatory element across the linear genome sequence. Although no current approach is comprehensive, this question is beginning to be addressed through chromosome conformation capture methods for the genome-wide mapping of physical proximity. A variety of related techniques all involve an intermolecular ligation of formaldehyde-fixed samples to circularize DNA fragments based on *in vivo* proximity. As the ligated fragments include sequence from both the adjacent loci, co-localization can be detected by PCR, hybridization, or sequencing techniques. However, any region of the genome might localize nearby any other region, creating an N^2 problem which requires a step to reduce library complexity. 3C (Dekker et al., 2002), 4C (Simonis et al., 2006), 5C (Dostie et al., 2006), and Hi-C (Lieberman-Aiden et al., 2009) techniques differ in their approach to selecting specific loci for study: 3C identifies co-localizations between two specified loci by the use of specific PCR primers, while 4C uses PCR primers for just one locus to allow study all interactions with that locus. 5C is a multiplexed ligation-mediated technique allowing the study of interactions among arbitrary loci, limited only by the expense of synthesizing a primer pool targeting the loci of interest. Hi-C instead sequences all ligation products, offering genome-wide scale but sacrificing resolution by pooling fragments

from megabase-size windows (Lieberman-Aiden et al., 2009). One promising means to improve its resolution involves the repurposing of genomic capture technologies to target interactions of interest for sequencing. A recent variation, ChIA-PET, uses immunoprecipitation to select for co-localization involving a particular trans factor (Fullwood et al., 2009). This technique is technically extremely challenging, but data is available from a small set of human and mouse cell lines for the factors estrogen receptor alpha (Fullwood et al., 2009), CTCF (Handoko et al., 2011) and RNA polymerase II (Li et al., 2012).

These studies have lent exciting insight into global patterns of connectivity between genes and regulatory elements. First, they highlight the capacity of variants in the core promoter to affect expression of the gene downstream. eQTL approaches found just 5% of associated SNPs were more than 20 kb upstream (Veyrieras et al., 2008) of the transcription start site. Chromatin conformation studies confirm that interactions are concentrated around the transcription start site, but suggest that the highest density of long-range interactions lies much further, 100-200 kb upstream of the transcription start site (Sanyal et al., 2012), confirming suggestions that eQTL studies are underpowered for the identification of cell-type specific promoter-distal elements involved in developmental regulation and inducible gene expression (Dimas et al., 2009; Heintzman et al., 2009; Thurman et al., 2012). Further, these results further complicate the definition of a gene as simply a linear sequence element. In a recent study using ChIA-PET for RNA polymerase II, of all colocalizations rooted at promoters genome-wide, only 5% were with the gene body of that particular promoter (Li et al., 2012). The majority were with extragenic regulatory elements or with promoters of other genes. Similarly, enhancers interact with the promoters of multiple genes, while promoters interact with a number of enhancers (Sanyal et al., 2012) highlighting the complexity of mammalian transcriptional regulation.

These approaches have already been applied to the study of specific disease associations and target loci. For example, 3C was used to demonstrate co-localization of a TCF7L2 binding site harboring a SNP associated with colorectal cancer with the promoter of the gene encoding the proto-oncogene MYC, 335 kb away (Pomerantz et al., 2009). Notably, although the expression of MYC did not correlate with the genotype of the SNP, transgenic mice lacking the putative enhancer element are resistant to tumorigenesis induced by an APC^{min}

mutation (Sur et al., 2012). However, due to the difficulty of these techniques, their application has been limited to a small number of associations.

Given the technical requirements for applying C-based assays on a large scale, a simpler strategy has been proposed to establish enhancer-gene interactions, based on the hypothesis that binding sites of the genome regulator CTCF define domains confining the influence of distal regulatory elements (Martin et al., 2011). However, recent work has shown unexpected degree of binding variability for CTCF among different cell types (Chapter 5), undermining the notion that its binding sites serve as invariant domain boundaries. Further, 5C data shows that 80% of TSS-distal element interactions cross an intervening CTCF site apparently unhindered (Sanyal et al., 2012). Finally, Hi-C experiments have shown that although the boundaries of chromatin interaction domains are enriched in CTCF sites, only a small subset (~15%) of CTCF sites appear to serve this purpose (Dixon et al., 2012). Thus, partitioning the linear genome by CTCF sites to establish the domain of action of a supposed enhancer is likely counterproductive.

Another approach to identify promoter-enhancer interactions postulates that the activity of a distal regulator should correlate across multiple cell types with the activity of its target promoter. This technique has been applied genome-wide to existing maps of chromatin state across a wide variety of cell types (Ernst et al., 2011; Thurman et al., 2012). The results of these analyses broadly correspond to colocalizations supported by C-based methods (Thurman et al., 2012). Given the availability of accessibility data in an expanding range of cell types, the optimization and extension of these techniques is an exciting possibility.

2.7 Conclusion

The importance of regulatory elements to human phenotypes has long been clear (Wray et al., 2003), but only recent experimental techniques have rendered detailed surveys possible. The field has been transformed by the wholesale localization of disease- and trait-associated polymorphism in non-coding regions, contemporaneous with the genome-wide extension of classical and new techniques for the study of regulatory elements. A parsimonious global estimate defines a regulatory compartment roughly twice the size of the exome (Neph et al., 2012b), which is highly enriched for regulatory elements involved in human diseases and traits. The moderate size of this compartment allows for the possibility of focused regulatory association studies.

Knowledge of functional regions can also complement fine mapping efforts and expedite the search to isolate causal polymorphism responsible for a genetic association signal.

Given these technological advances, the relevance of regulatory elements to human genetics will likely only grow. Promoter regions represent the low hanging fruit: much sequence data is already available and identification of the target gene is less of an issue. In studying distal regulatory variation, the principal difficulty is no longer its identification or even the recognition of functional variation. Rather, the biggest present challenge is the association of distal regulatory elements with target genes. These regulatory elements are highly cell-type specific and can be linked to defined cognitive factors, offering immediate context for a genetic association. This challenge is being rapidly addressed by new techniques. As studies of long-distance regulatory interactions have shown, the concept of the gene as a linear sequence element risks overlooking the web of regulatory interactions among elements scattered where expedient throughout neighboring genes.

2.8 Tables

Table 2-1 Levels of evidence for function role of a non-coding variant.

Methods for studying polymorphism and protein-DNA interaction, divided into three classes based on their laboriousness and fidelity to regulatory action at the native locus.

Level 1: Evidence of in vivo regulatory function from in situ models

- 1a In situ / whole locus model of strongly genetically implicated variant that precisely recapitulates the phenotype at the organismal level
- 1b In situ genome modification (e.g., genome editing / knock-in/out)
- 1c Whole-locus transgenic lines (e.g., YAC, BAC; single copy)
- 1d In situ measurement of gain/loss of regulatory protein binding directly coupled to in vivo gene product phenotype
- 1e In situ gain/loss of regulatory protein not coupled to gene product

Level 2: Evidence of transcriptional capacity in artificial/condensed construct models

- 2a Standard transgenic animal
- 2b Stable transfection (integrated into genome)
- 2c Transient transfection (ex-genomic)

Level 3: Non-cellular assays of protein-DNA interaction

- 3a Gel mobility shift
- 3b Fluorescence polarization / Surface plasmon resonance
- 3c Protein binding microarrays / SELEX

Level 4: Predicted binding site

- 4a Evolutionary conservation
- 4b Sequence only

Table 2-2 Selected associations of regulatory variation and human phenotypes.

Selected associations of human phenotypes, including mechanistic investigation to localize the regulatory site, and identified its cognate factor and target gene. * see Table 2-1. T2D, Type 2 diabetes.

Phenotype	Transcription factor	Target gene	Location relative to gene	Reference	Level of Evidence
Hereditary persistence of fetal hemoglobin	CCAAT box	HBG1	Promoter (-117) of HBG1	(Berry et al., 1992; Collins et al., 1985; Gelinas et al., 1985; Peterson et al., 1995)	1a
t(14:18) follicular lymphoma	4x DHS	BCL2	DHSs targeted 3' of BCL2	(Xiang et al., 2011)	1a
Colorectal and other cancers	TCF4	MYC	Intergenic (335 kb away)	(Huppi et al., 2012; Pomerantz et al., 2009; Sur et al., 2012; Tuupanen et al., 2009)	1b
T2D/glucose tolerance		TCF7L2	unlocalized	(Gaulton et al., 2010; Savic et al., 2012; 2011)	1c
Intestinal expression of cystic fibrosis gene	HNF1	CFTR	Intron 1 of CFTR	(Ott et al., 2009; Rowntree et al., 2001)	1c
α thalassemia	GATA1	Novel transcript near HBA	Upstream of HBA (~10 kb)	(De Gobbi et al., 2006)	1d
Preaxial polydactyly	GABP α , ETV4, ETV5	ETS1, SHH	Intron of LMBR1 (1MB away)	(Lettice et al., 2003; 2002; 2012)	1e
Hirschsprung disease	(MCS9.7)	RET	intron(?)	(Grice et al., 2005)	2a
LDL-C & myocardial infarction	C/EBP α	SORT1	3'UTR of CELSR2	(Musunuru et al., 2010)	2b
Coronary artery disease & type 2 diabetes	STAT1	CDKN2B, CDKN2BAS	Latter half of CDKN2BAS	(Harismendy et al., 2011; Visel et al., 2010)	2c
Hyperlipidemia	OCT1	LPL	Promoter of LPL	(Yang et al., 1995)	2c
Rheumatoid arthritis	Unknown	CCR6	Intron of CCR6	(Kochi et al., 2010)	3a
Childhood asthma	C/EBP β	ORMDL3	Intron of GSDML	(Moffatt et al., 2007)	4a

Table 2-3 Data sources and tools.

Shown are selected public data sources of the functional genomics data discussed in this article, as well as tools for making use of it.

Name	URL	Reference	Description
Public data sources			
UCSC genome browser	genome.ucsc.edu/	(ENCODE Project Consortium, 2011)	The primary tool for visualizing functional genomics data. All data tracks from the ENCODE consortium, including RNA-seq, ChIP-seq, DNaseI-seq and methylation data are available from within this browser under the “Regulation” track group. Select data can be conveniently exported using the table browser, or full data sets can be downloaded for local processing.
NIH Roadmap Epigenomics Mapping Consortium	roadmapepigenomics.org	(Bernstein et al., 2010a)	UCSC browser mirror hosting further data in a number of primary tissue and cell types.
GEO ArrayExpress	www.ncbi.nlm.nih.gov/geo www.ebi.ac.uk/arrayexpress/		Originally developed for archiving gene expression microarrays. They now contain large amounts of publicly available functional genomics data such as ChIP-seq.
NHGRI GWAS catalog	www.genome.gov/gwastudies/	(Hindorf et al., 2009)	Compilations of SNPs associated with hundreds of human diseases and traits
Transfac Jaspar UniPROBE Jolma et al. 2013	www.gene-regulation.com/pub/databases.html/ jaspar.cgb.ki.se/ the_brain.bwh.harvard.edu/uniprobe/	(Jolma et al., 2013; Matys et al., 2006; Newburger and Bulyk, 2009; Portales-Casamar et al., 2010)	Databases of experimentally-determined binding specificities
UW de novo motifs	www.uwencode.org/	(Neph et al., 2012b)	Binding specificities inferred de novo from DNaseI footprints
ScanDB Genotype-Tissue Expression (GTEx)	www.scandb.org www.ncbi.nlm.nih.gov/gtex/GTEx2/gtex.cgi		Compilation of published eQTL datasets.
PhyloP	genome.ucsc.edu/	(Pollard et al., 2010)	Per-nucleotide vertebrate conservation scores available from the UCSC genome browser
Tools			
Galaxy	galaxy.psu.edu/	(Goecks et al., 2010)	User-friendly web interface for exploring genomics data sets. Can import data from a number of formats, including directly from the UCSC genome browser.
BEDOPS	code.google.com/p/bedops/	(Neph et al., 2012a)	A suite of command-line UNIX utilities for working with functional data
MEME suite	memes.sdsc.edu/meme/intro.html	(Bailey et al., 2009)	Utilities for working with sequence-specific binding motifs, including searching across the genome (fimo), finding related motifs (tomtom), and de novo discovery (MEME)
HaploReg	www.broadinstitute.org/mammals/haploreg/haploreg.php	(Ward and Kellis, 2012)	Integrated analysis tools for exploring functional annotations at sites of human disease and trait associations. Includes LD, phylogenetic conservation, histone marks and DNaseI-seq data from the ENCODE project, predicted recognition sites, and eQTL associations.

3 Genomic Discovery Of Highly Potent Chromatin Insulators For Human Gene Therapy

3.1 Introduction

Insulator proteins such as CTCF are capable of restricting the effect of an enhancer to a limited domain. In addition to their physiological role (Phillips and Corces, 2009), they are expected to be useful in stably-integrating gene therapy vectors for limiting the off-target effects on the expression of adjacent genes (Emery et al., 2000). The well-studied chicken HS4 insulator element has been tested extensively in this role, however its length (~1.2 kb) renders it cumbersome for use in some viral systems (Emery, 2011). Thus there is a need for additional insulator sequences, offering additional potency and the flexibility to be matched to a particular application. This requires detailed study of the *in vivo* binding properties of known insulator proteins, in particular CTCF.

3.2 Results

Previous genomewide searches for conserved, overrepresented motifs have identified a number of known and unknown regulatory elements (Xie et al., 2007). Figure 3-1 confirms this result by illustrating the nucleotide-level correspondence between protected residues and conservation data, a general feature of DNase I footprinting. Thus conservation can be used to recognize functional regulatory motifs in aggregate. The resolution of DNase I footprinting data reveals a novel motif upstream of the canonical CTCF motif, potentially representing the involvement of additional zinc fingers (Figure 3-1). In addition, this motif occurs at two preferential spacings, 1 nt apart (Figure 3-2). This result indicates the potential for DNase I footprinting to understand the detailed protein-DNA interaction patterns of important regulatory factors. Given CTCF's capability to engage chromatin autonomously, this interaction must be specified in large part by sequence context. A way of identifying the highest-affinity consensus sequences would offer great utility for use in gene therapy viral vectors.

To address this, we designed candidate insulator sequences based on genome-wide *in vivo* occupancy profiling of CTCF in K562 cells (Figure 3-3). Although the majority of CTCF binding sites are distinguished by a conserved sequence motif (Figure 3-3A) (Kim et al., 2007), most instances of this sequence genome-wide are not

bound by CTCF. In fact, we identified 22,579,591 instances of CTCF motifs genome-wide ($P < 10^{-2}$), only 205,227 (0.9%) of which overlapped peaks. To identify sequence features associated with high affinity for CTCF, we grouped these potential recognition sequences genome-wide into classes of exact sequence identity. We focused on the 14-bp sequence at the high-information core of the CTCF motif, a GC-rich region which has been shown sufficient to engage zinc fingers 4-8 (Renda et al., 2007). We found 82,234 classes having >50 instances of the same 14-mer throughout the genome (mean=105 instances). To assess the inherent affinity of each class, we measured the proportion of its instances that were occupied *in vivo*. Occupancy at these classes varied greatly, from 100% (all instances of the 14-mer were bound *in vivo*) to 0% (Figure 3-3D). Strikingly, 11,927 of the 45,696 CTCF binding sites genome-wide (26%) included at least one instance of the top 500 classes.

To study whether the high-affinity CTCF sites identified through this method could potentially function as insulators, we selected representative genomic sequence from instances of high- and low-occupancy classes (Table 3-1). Although CTCF binding is required for the enhancer blocker function of known vertebrate insulators (Bell and Felsenfeld, 2000), other functional sites in the surrounding sequence may also contribute to insulator functionality (Dickson et al., 2010). We therefore identified the genomic sequence surrounding 26 instances of the 6 highest-occupancy classes. As a control, we selected 3 representatives for each of 3 lower-occupancy classes. All candidate loci were bound by CTCF *in vivo* and overlapped a DNase I hypersensitive site.

We found that the candidate sequences exhibited potent enhancer blocking activity, in some cases stronger than the canonical Chicken HS4 element (Table 3-2). For example, the A1 element produced 2.03% as many colonies (relative to the zeo internal control), indicating strong activity. The lower-occupancy classes demonstrated a substantially reduced activity, confirming that this analysis is capable of identifying a graduated range of insulator strength (Table 3-3). These elements are expected to provide a starting point for the design and optimization of gene therapy vectors.

3.3 Methods

Identification of upstream CTCF motif.

The upstream CTCF motif was identified by using MEME (Bailey et al., 2009) on sequences in the location observed upstream footprint at DHSs with strong CTCF ChIP-seq signal.

Identification of high-affinity CTCF sites.

Occurrences of the CTCF recognition sequence were identified throughout the reference human genome (hg18) using the program FIMO (Grant et al., 2011), to scan four published (Kim et al., 2007; Xie et al., 2007), and three curated position weight matrices for CTCF at a significance of $P < 10^{-4}$. Resultant sequences were aligned around the motif location to collate occurrences of classes of unique 14 bp sequences at the core of the CTCF motif. Occupancy of the 14 bp classes with >50 instances was assessed using published CTCF ChIP-seq data in K562 cells (Chapter 5). Data was processed as previously described using ELAND to map tags from 4 biological replicates to the reference genome. For each class, the proportion of instances occupied was computed by considering overlap with peaks identified using SPP (1% FDR) (Kharchenko et al., 2008). Data processing was performed using the BEDOPS suite of software tools (Neph et al., 2012a), awk (Aho et al., 1988) and the R package. Suitable candidate sites for assessing each class were selected manually additionally taking into consideration previously published Digital DNaseI data for K562 (Thurman et al., 2012). Candidates selected for both high- and low-affinity classes were occupied by CTCF in $\geq 16/19$ cell types (Chapter 5).

Colony assay for enhancer-blocking insulators.

Candidate insulator and control sequences were cloned into the *neo* reporter plasmid pJC5-4/P4P2K. This construct is based on the *neo* reporter plasmid pJC5-4 that was initially used to characterize the cHS4 chromatin insulator (Chung et al., 1993). It contains an expression cassette for the drug-resistance reporter gene neomycin phosphotransferase (*neo*) that is transcribed from an $\Lambda\gamma$ - globin gene promoter linked to the DHS 2 (HS2) enhancer from the mouse β -globin gene locus. Candidate elements were amplified by PCR and inserted both upstream of the transcription cassette and between the transcription cassette and the enhancer, using the

Multisite Gateway clonase system (Invitrogen Corp., Carlsbad, CA). Inserts included the insulator candidates listed, the 1.2 kb cHS4 element as a positive control, and a 308 bp fragment from the bacterial drug resistance gene *zoo* as a negative spacer control. Following linearization, plasmid constructs were transfected into K562 cells using the Amaxa Biosystems Nucleofector II electroporator and Nucleofector V kit (Lonza Group Ltd., Basel, Switzerland) following the manufacturer's directions at a dose of 2 µg plasmid per 10⁶ cells. Following electroporation, the cells were resuspended and cultured for 2 more days in 5 mL D8 media (Dulbecco's modified Eagles medium (DMEM) supplemented with 8% fetal bovine serum (FBS)). The cells were then collected, counted, resuspended at a dose of 5 x10⁵ cells in 5 mL DMEM, 20% FBS, 0.8% low-melting agarose, and 1 mg/mL of the neomycin drug analog G418 (active component), and plated in duplicate in 60 mm tissue culture dishes. These semisolid cultures were incubated another 2- 3 weeks and then scored for colony formation.

3.4 Figures and Tables

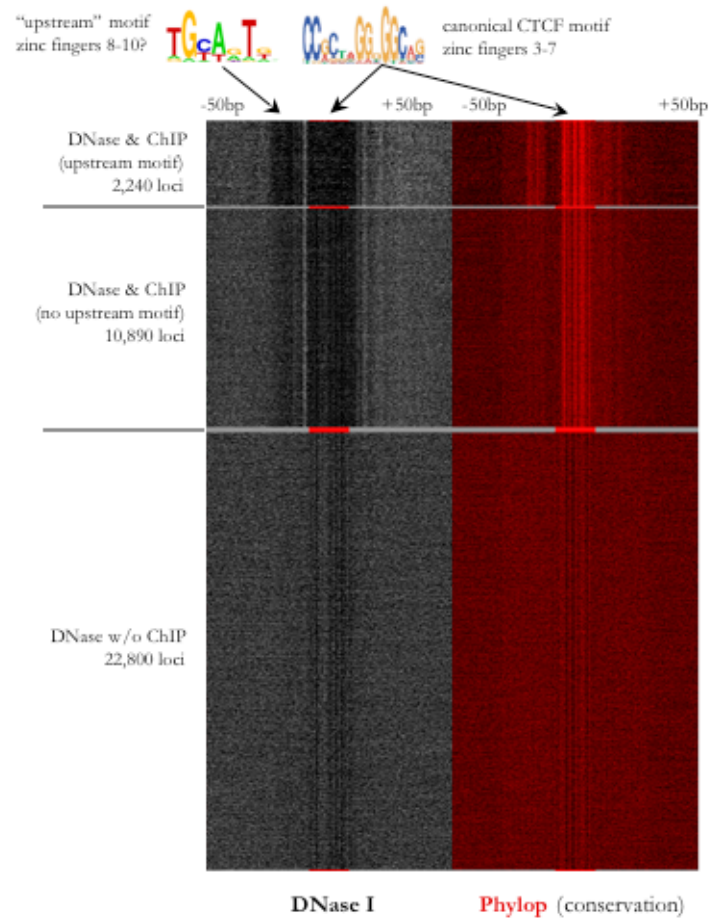


Figure 3-1 DNase I footprinting can resolve different modes of binding.

Plot showing DNase I sensitivity ± 50 bp of CTCF motifs in ChIP-seq peaks (lighter colors indicate higher cleavage). CTCF binds to a second motif upstream (additional dark area) of the canonical one. The region of this upstream motif is also conserved (phyloP, right, measures nucleotide-level conservation. Red indicates higher conservation).

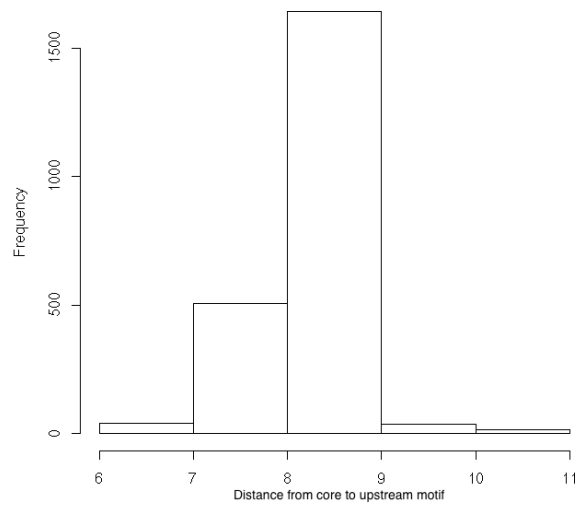


Figure 3-2 Spacing of CTCF upstream motif.

X-axis indicates the distance between the upstream motif and core CTCF motif. Note the two peaks 1 nt apart indicate a flexible spacing in the core 11-ZF DNA-binding domain.

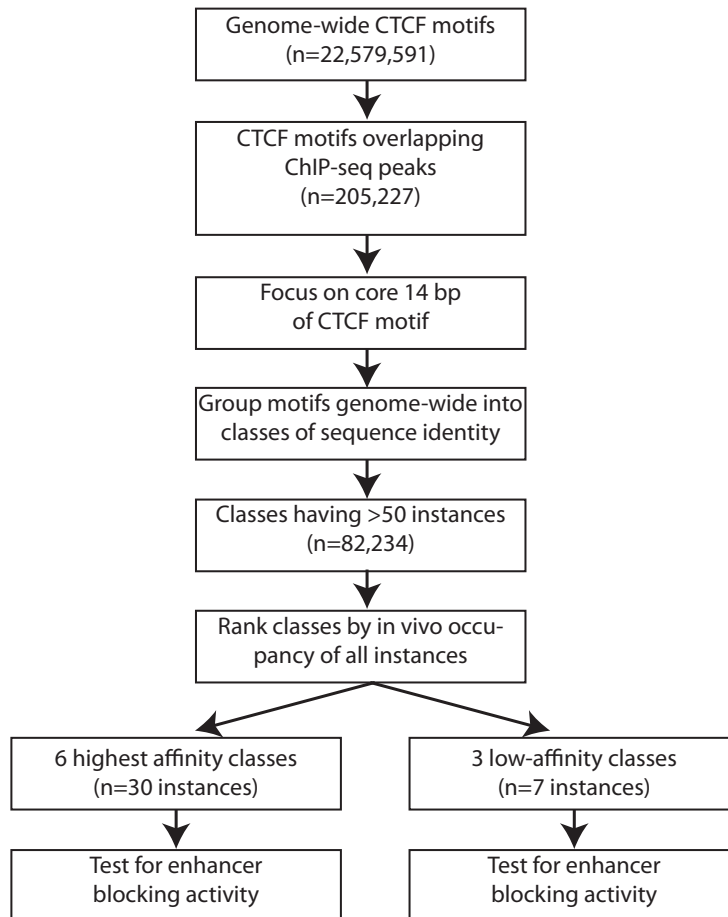


Figure 3-3 Strategy for the discovery of high-affinity CTCF classes

Test construct:

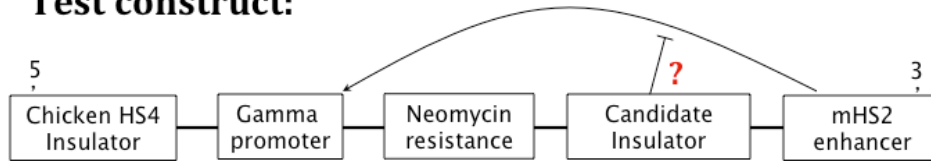


Figure 3-5 Design of construct for quantification of enhancer blocking potency

Construct design follows previously published work (Chung et al., 1993). Potent insulators prevent the interaction of the HS2 enhancer with the neomycin resistance gene required for survival.

High-affinity classes		
Rank	Core sequence	Occupancy (ChIP-seq)
1	CACCAGGTGGCGCT	100.00%
2	CCACCAGGGGAGC	100.00%
3	TCAGTAGAGGGCGC	100.00%
4	CCACTAGGGGCGAG	98.80%
5	CAGCAGAGGGCGCT	98.80%
6	CAGTAGAGGGCGCT	98.60%
Low-affinity classes		
932	CCCTCTCCTGGGCA	9.80%
943	GCAGCAGAGAGCAA	9.70%
950	CCCTCTGCTGACTG	9.60%

Table 3-1 Classes selected for experimental testing

Higher rank indicated higher predicted strength based on ChIP-seq occupancy. Core sequence is the identical 14-bp sequence shared by all class members.

Class	chrom	chromStart	chromEnd	Colony Mean	Colony SD
A1	chr1	76229933	76230115	2.03%	0.17%
A2	chr19	46342254	46342440	3.13%	0.79%
A3	chr5	90557806	90557925	8.29%	0.73%
A4	chr7	39526144	39526307	3.03%	1.01%
A5	chr9	123122258	123122485	8.66%	1.28%
B1	chr1	57070274	57070473	3.98%	0.28%
B2	chr14	76320982	76321216	5.50%	0.79%
B3	chr12	106525547	106525690	9.27%	1.89%
B4	chr20	61411429	61411552	18.08%	2.35%
B5	chr22	33876597	33876817	14.10%	3.27%
C1	chr1	30038520	30038763	6.18%	1.64%
C3	chr2	16570154	16570340	5.75%	0.90%
C5	chrX	149602730	149602938	14.66%	2.29%
D1	chr1	161084397	161084593	4.48%	2.14%
D3	chr12	121910863	121911035	20.02%	5.56%
D4	chr17	16207878	16208078	36.55%	4.14%
D5	chr17	24468364	24468528	25.65%	7.79%
E1	chr1	176768124	176768329	4.57%	0.85%
E2	chr13	20397008	20397240	4.57%	0.51%
E3	chr14	68665824	68666051	13.96%	3.25%
E4	chr5	64116126	64116362	6.17%	1.30%
E5	chr5	170702087	170702300	9.85%	1.80%
F1	chr12	55856009	55856202	3.64%	0.73%
F2	chr12	102550590	102550826	3.38%	0.28%
F3	chr12	57793909	57794148	8.12%	1.95%
F4	chr7	95383976	95384189	15.11%	1.08%
F5	chr8	71163374	71163613	40.48%	4.26%

Table 3-2 Results of enhancer blocking assay

Testing of 27 candidate instances belonging to 6 high-affinity CTCF classes for enhancer blocker activity. Strong insulation is indicated by low Colony Mean. Chrom, chromStart and chromEnd indicate the designed genomic location of each class instance (hg18).

Class	chrom	chromStart	chromEnd	Colony Mean	Colony SD
G1	chr1	149974115	149974543	99.89%	5.93%
G2	chr1	204477048	204477285	85.39%	10.93%
G4	chr10	103022102	103022247	103.86%	11.25%
G5	chr14	59501614	59502000	70.37%	3.66%
G6	chr15	71987695	71988039	68.85%	3.93%
G8	chr16	6071437	6071557	38.12%	11.22%
G9	chr16	65242354	65242777	75.19%	11.49%
NTRL-1	chr1	37348429	37348443	82.91%	3.12%

Table 3-3 Control results of enhancer blocking assay

G1-G9 are low-affinity classes. NTRL-1 random human genomic sequence selected for lack of any DNase I or ChIP-seq signal, and is expected to have no activity.

3.5 Notes

This work was performed in collaboration with Mingdong Liu (who performed the insulator assays), Hao Wang, Chao-Zhong Song, Patrick A. Navas, David W. Emery, John A. Stamatoyannopoulos, and George Stamatoyannopoulos.

4 Widespread Site-dependent Buffering Of Human Regulatory Polymorphism

4.1 Abstract

The average individual is expected to harbor thousands of variants within non-coding genomic regions involved in gene regulation. However, it is currently not possible to interpret reliably the functional consequences of genetic variation within any given transcription factor recognition sequence. To address this, we comprehensively analyzed heritable genome-wide binding patterns of a major sequence-specific regulator (CTCF) in relation to genetic variability in binding site sequences across a multi-generational pedigree. We localized and quantified CTCF occupancy by ChIP-seq in 12 related and unrelated individuals spanning three generations, followed by comprehensive targeted resequencing of the entire CTCF-binding landscape across all individuals. We identified hundreds of variants with reproducible quantitative effects on CTCF occupancy (both positive and negative). While these effects paralleled protein-DNA recognition energetics when averaged, they were extensively buffered by striking local context dependencies. In the significant majority of cases buffering was complete, resulting in silent variants spanning every position within the DNA recognition interface irrespective of level of binding energy or evolutionary constraint. The prevalence of complex partial or complete buffering effects severely constrained the ability to predict reliably the impact of variation within any given binding site instance. Surprisingly, 40% of variants that increased CTCF occupancy occurred at positions of human-chimp divergence, challenging the expectation that the vast majority of functional regulatory variants should be deleterious. Our results suggest that, even in the presence of ‘perfect’ genetic information afforded by resequencing and parallel studies in multiple related individuals, genomic site-specific prediction of the consequences of individual variation in regulatory DNA will require systematic coupling with empirical functional genomic measurements.

4.2 Author Summary

A comprehensive understanding of the contribution of individual genome sequences to disease and quantitative traits will require the general ability to predict consequences of genetic variation in non-protein-coding regions, particularly those involved in gene regulation. Here we tested the power to predict such consequences when presented with ‘complete’ information encompassing the genomic DNA binding site patterns of a well-studied regulatory protein across multiple related individuals, coupled with all individual genome sequences at the binding positions. We find that while there is reasonable ability to predict the average effects of variation within the consensus recognition sequence of a transcriptional regulator, it is not possible to determine reliably the consequences of variation at any given genomic instance. This suggests that the interpretation of individual genome sequences will require comprehensive complementation with functional genomic studies.

4.3 Introduction

A growing number of studies associate variation within regulatory DNA and risk of human disease (Harismendy et al., 2011; Hindorff et al., 2009; Musunuru et al., 2010). Variation in regulatory DNA may result in modulation of recognition by sequence-specific transcription factors (TFs), resulting in altered gene expression (Collins et al., 1985; Gelinis et al., 1985; Rockman and Wray, 2002). That the vast majority of variants emerging from human resequencing studies lie in non-coding regions creates an urgent need for determining the consequences of variation within regulatory DNA.

Functionally significant variation within the genomic recognition sequences for certain TFs appears to be correlated in aggregate with nucleotide-level evolutionary conservation and/or position-specific information content (Mirny and Gelfand, 2002; Moses et al., 2003; Nouredine et al., 2009; Veprintsev and Fersht, 2008). Although surveys have identified sites of allele-specific occupancy of TFs and RNA Polymerase II or allele-specific chromatin states (Kadota et al., 2007; Kasowski et al., 2010; Knight et al., 2003; McDaniell et al., 2010; Meissner et al., 2008), these studies have not established the distinguishing characteristics of regulatory sequence variation with an experimentally-observed effect on occupancy. As such, it is currently not possible to interpret reliably the functional consequences of variation within any given TF recognition sequence.

To address this, we apply a novel experimental design to identify comprehensively patterns of genetic variation with heritable effects on the occupancy of the major genomic regulator CTCF (Phillips and Corces, 2009). Unlike most sequence-specific regulators which rely on cooperative interactions with other factors to bind DNA, CTCF is able to access target DNA within chromatin in a relatively autonomous fashion through its rich binding interface. By combining quantitative genome-wide occupancy analysis by ChIP-seq in a multi-generational pedigree with comprehensive resequencing of the binding site landscape across all individuals, we achieve complete knowledge of variation in both sequence and occupancy, thus creating a benchmark for assessing the characteristics of functional and heritable regulatory sequence variation.

4.4 Results

Components of heritable transcription factor occupancy

We mapped binding sites for CTCF by ChIP-seq in B-lymphoblastoid cells derived from 12 members of a three-generation pedigree (Figure 4-1A, B). We identified a total of 51,686 binding sites across all individuals at a false discovery rate (FDR) of 1%. To comprehensively identify genetic variation with potential functional consequences for CTCF binding, we performed targeted resequencing by array capture focused on the 134 bp interval surrounding 46,568 CTCF sites (total ~6 Mbp) in all family members assayed by ChIP-seq. 7,394 of the 35,709 surveyed binding sites (or 21%) overlapped one or more SNPs, some of which had clear associations with occupancy in the direction predicted by the CTCF motif (Figure 4-1C, D). We did not consider other variation such as copy number variants or small indels. In order to minimize reference mapping bias for the ChIP-seq data, we remapped tags to personalized genomes including discovered SNPs (Heap et al., 2009). Additionally, to avoid artifacts resulting from uncertain mapping of 36 bp reads to the genome, we simulated all reads including discovered SNPs from ± 147 bp centered on the ChIP-seq peak and excluded sites with too many ambiguously mapped tags.

We integrated the genetic and functional data sets to survey genome-wide heritable variation in transcription factor occupancy. We reasoned that the strongest signal of heritable variation would be from segregating variants overlapping the binding site. Thus we performed a linear regression of the ChIP-seq density on the SNP genotype in *cis* (Figure 4-1D). Of 5,828 polymorphic sites, this analysis identified 325 (5.6%) sites with a significant association of SNP genotype with occupancy at a false discovery rate (FDR) of 1% (Figure 4-2, Table 4-1). We tested whether several confounding factors might be responsible for our results, however sites at which SNPs were significantly associated with changes in occupancy were similar to polymorphic sites without changes in occupancy in terms of GC content (median 53.7% vs. 53.0%, Mann-Whitney $P < 0.044$) and ChIP-seq input signal (3.61 vs. 3.61, Mann-Whitney $P < 0.84$), and distance to the nearest RefSeq TSS (33 kb vs. 36 kb, $P < 0.96$). Significant sites were only slightly weaker in terms of ChIP-seq density (2.73 vs. 2.93, $P < 2.7 \times 10^{-3}$), and DNase I signal (5.47 vs. 6.92, $P < 7.8 \times 10^{-8}$). Thus we conclude that the SNP genotype is associated with differences in occupancy at 325 of 5,828 sites tested.

We used a hypothesis-driven linkage analysis to assess the heritability of the remaining unexplained differential occupancy. First, we identified 1,376 sites of differential occupancy. Of these, 200 (15%) were already associated to an underlying SNP (FDR 1%), 65 (4.7%) had allele-specific occupancy (FDR 0.1%), and 197 (14%) were on chromosome X. To test for heritable inheritance of occupancy not explained by these factors, we performed Haseman-Elston sib-pair linkage analysis in aggregate at sites differentially occupied among the 6 grandchildren (Figure 4-3A-B, Section 4.6). The 47 binding sites already significantly associated with SNPs had a regression slope of -2.86 (Figure 4-3C), confirming substantial heritability ($P < 8.9 \times 10^{-7}$, permutation). The remaining 50 sites without significant associations had a regression slope of -1.01 (Figure 4-3D), indicating a lower but still significant level of heritable variation ($P < 3.2 \times 10^{-5}$). These results suggest that SNPs directly overlapping the cognate recognition sequence explain most but not all of the heritable variation in this pedigree. Remaining variation in occupancy might be heritable due to sequence variants not considered in our SNP-based analysis or heritable epigenetic variation such as methylation. However, it is quantitatively less significant than the heritability attributable to the direct effect of SNPs on occupancy (Figure 4-3E).

Using motifs to align regulatory polymorphism from multiple sites

Understanding the effect of DNA sequence variation on transcription factor occupancy is critical to a mechanistic interpretation of non-coding variation. To interpret the association results in the context of the CTCF motif, we scanned the center of the ChIP-seq peak with the known position weight matrix (PWM) (Kim et al., 2007), which measures the contribution of each nucleotide in the binding site to the energy of the protein-DNA interaction (Stormo, 2000; Wasserman and Sandelin, 2004). 888 binding sites did not contain a motif match (fimo p-value $< 10^{-2}$) and 1,040 binding sites overlapped multiple SNPs within ± 180 bp of the ChIP-seq peak. Excluding these sites, we analyzed the 4,428 binding sites with a single SNP and a single motif match. These SNPs were distributed throughout the resequenced region surrounding each CTCF motif (Figure 4-4A).

In contrast, we expected that SNPs associated with occupancy differences would be concentrated in the 44 bp region of protein-DNA contact (Filippova et al., 1996). Indeed, despite a slight reduction (1.08-fold) in local sequence diversity, 85% of the SNPs that affected occupancy were within this region (Figure 4-4B). The

allele observed to have higher occupancy matched the energetically more favorable one for 83% of these SNPs. Associated SNPs outside the region of contact had less significant q-values (median q-value of 1.3×10^{-3} outside the versus 1.3×10^{-5} inside), consistent with these SNPs being false positives or sites with ambiguity in the true location(s) of protein-DNA interaction. Alternatively, some of these SNPs might affect CTCF occupancy indirectly by perturbing an adjacent co-factor binding site.

We compared our results to an allele-specific occupancy test performed at heterozygous sites (Figure 4-5). Despite a weaker enrichment of significant sites within the core motif and a less substantial concordance of the higher occupancy allele with the energetically more favorable nucleotide, this allele-specific analysis broadly corresponded to the results of the association analysis. Thus, we interpret the results of our analysis as indicating that we have correctly identified the motifs at most binding sites, and that the significant SNPs directly affect occupancy through modulating the protein-DNA interaction at these sites.

Context-dependent effect of identical changes

Although differences in occupancy were largely associated with SNPs at positions strongly affecting overall binding energy (Figure 4-4B), only 13% of SNPs at the interface of protein-DNA interaction affected occupancy (Table 4-1). Thus although functional SNPs are highly concentrated in the region of protein-DNA contact, the majority of SNPs, even in this region, do not measurably affect occupancy. Since our data set includes multiple sites with the same two alleles at the same position measured relative to the binding motif, we investigated the proportion of sites at which a given change was found to affect occupancy (Figure 4-6). Like in Figure 4-4, the most disruptive changes were observed at positions of high information content in the motif. However, even over the 14 bp core motif, changes affected occupancy at a median of only 36% of the sites where they were observed. We found no changes that uniformly affected occupancy without regard to context. Instead, we observed a strong, progressive depletion in the proportion of changes affecting occupancy at the strongest sites (Figure 4-7), and a smaller depletion at the weakest sites. Indeed, simply clustering the ChIP-seq intensities identified three major groups, of low, medium and high occupancy, which were also distinguished by varying proportions of significant SNPs (Figure 4-8). This result places an upper limit on the accuracy of methods that predict the effects of non-coding SNPs without consideration of their context.

Strength-dependent buffering could be explained by a model where changes in occupancy are observed only when a SNP causes the affinity of a site to cross a threshold for binding. In this case, the strongest and weakest sites will only be affected by the greatest genetic perturbation, while smaller perturbations would affect binding only at sites of intermediate strength. This would create the impression of epistasis between all positions in the cognate recognition site as any affinity-affecting change could potentially buffer another (Berg et al., 2004; Mustonen et al., 2008). We thus compared the inherent affinity of the site with the magnitude of the perturbation caused by each SNP. We divided SNPs affecting occupancy into bins based on the strength of their match to the canonical motif. We found that SNPs at sites matching the CTCF motif more strongly in turn exhibited higher log-odds differences (Figure 4-9). We observed no such trend at SNPs not associated with occupancy differences (Figure 4-10). These results are consistent with stronger motifs being buffered against all but the largest perturbations. Although a linear regression identifies a significant effect ($P < 0.006$), an r^2 of 0.04 indicates that the strength of the motif match alone can not explain the breadth of buffering observed.

Buffering might also be a consequence of the non-additive effect on binding energy of individual positions in the cognate binding site, as has been observed *in vitro* (Badis et al., 2009). The relevance of non-additive interactions for identifying binding sites has been questioned (Benos et al., 2002; Zhao and Stormo, 2011), but the implications for understanding the function of specific variants *in vivo* have remained unclear. To explore the power of our data set to discover epistatic interactions, we measured the mutual information between the sequence context per-base in the core motif and whether a SNP at each location affects occupancy (Figure 4-11A). This analysis identifies two positions in the consensus sequence that significantly buffer the effect of a SNP at another position. First, of the 24 SNPs observed at position 1 in the motif, 13/13 that affected occupancy had an adenine at position 5, compared to only 5/11 for those that did not affect occupancy (Figure 4-11B, above). Interestingly, the second significant buffering interaction is between position 7 and SNPs at the adjacent position 8 (Figure 4-11B, below), suggesting local compensation for the adjacent SNP. These results indicate that higher-order models may be necessary to fully model the effect of polymorphism on pro-

tein-DNA interaction, and are consistent with a model where local factors determine whether polymorphism affects occupancy.

Power to predict functional polymorphism in non-coding regions

Resequencing and association studies are producing large amounts of data on polymorphism in non-coding regions, yet as we have illustrated, their functional classification is difficult. To investigate the power of existing metrics to predict functional polymorphism in non-coding regions, we used as a reference set the 1,368 sites with SNPs within the 44 bp vicinity of a recognizable CTCF motif. We first assessed the predictive power of evolutionary constraint, which has been used successfully to discover regulatory motifs , to highlight functional positions within motifs , and to predict the effect of coding variants (Adzhubei et al., 2010; Cooper et al., 2010; Ng and Henikoff, 2001). Conservation is a particularly attractive operational metric in genome scans, as it can be applied in an unbiased fashion without directly measuring protein-DNA interaction or modeling context effects. Indeed, CTCF binding sites are clearly marked by increased conservation (Xie et al., 2007). Thus, we tested the sensitivity and specificity of per-nucleotide conservation (phyloP 44-way vertebrate alignment, from UCSC browser) to correctly identify the 186 significant SNPs in our reference set. However, despite being applied only across experimentally determined binding sites, conservation had little predictive power on this data set, with an AUC of 0.57 (Figure 4-13).

Then we measured the improvement from evaluating potentially functional SNPs within the context of protein-DNA binding energetics. Applying such predictor showed a marked improvement over conservation, with an AUC of 0.75 (Figure 4-13), although positive predictive power was greatest for the most severe perturbations (Figure 4-14). Nevertheless, these results illustrate the power to be gained from considering non-coding polymorphism within the context of functional genomics data on transcription factor occupancy.

4.5 Discussion

In contrast to the vast diversity of protein function, the elements that regulate gene expression recruit from a shared repertoire of transcription factors, offering the potential for a common regulatory sequence code. The torrent of variants emerging from human resequencing studies – the vast majority of which lie in non-coding

regions – coupled with the growing number of common, disease-associated non-coding variants (Harismendy et al., 2011; Hindorff et al., 2009; Musunuru et al., 2010) has created an urgent need for determining the consequences of variation within regulatory DNA. However, the proportion of variants within regulatory DNA that have reproducible functional consequences on regulatory factor binding is currently unknown, and our ability to predict such outcomes from known rules of protein-DNA interaction is uncertain.

We have described a novel, hypothesis-driven genetic method employing targeted capture and genome-wide *in vivo* occupancy profiling to investigate directly the consequences of heritable variation in regulatory sequence. Our results show that individual transcription factor binding sites are surprisingly robust to genetic variation, even at evolutionarily constrained positions. While previous studies have observed differences in transcription factor occupancy among individuals using occupancy profiling alone (Kasowski et al., 2010), genome-wide linkage scans (Zheng et al., 2010), or allele-specific occupancy approaches (McDaniell et al., 2010), this work is the first systematic analysis of patterns of functional alteration in TF recognition sequences. This study further advances the characterization of heritable variation in TF binding by using highly accurate sequence information throughout a three-generation pedigree.

Our study has revealed a large degree of context dependence for changes to the CTCF recognition sequence. Indeed, even over the core 14 bp motif, only 36% of SNPs affected occupancy (Figure 4-6). Our estimate of the percentage of SNPs that affect occupancy in this 14 bp region ranges between 24% to 42% at FDRs 0.1% and 5%, respectively, indicating that the magnitude of this effect cannot be explained by the choice of significance cutoff. We have suggested that buffering is partly mediated by the strength of the binding site, as well as the sequence context at the local CTCF recognition sequence. In addition, buffering might be facilitated by a feedback process that maintains a constant CTCF occupancy despite alterations to the site's inherent affinity. However, while 21% of the SNPs in the region of protein-DNA contact that were significant in our association analysis also exhibited allele-specific occupancy in heterozygous samples, only 3.7% of the non-significant SNPs did, indicating that buffering is not likely to be the consequence of a feedback process. Alterations in DNA methylation might also mask the effect of otherwise significant genetic changes. However, only 30% of polymorphic CTCF sites contain a CpG at positions 1 or 11 of their recognition sequences. Fur-

thermore, the prevalence of CpGs at these positions is the same at sites where a SNP does and does not affect occupancy, limiting the potential scope of methylation to fully explain the observed buffering. As this study was performed on transformed B-lymphoblastoid cells, it is worth noting that the specific CTCF sites that are buffered may not be extrapolated to primary cell types. However, assuming that EBV transformation does not invoke novel cellular mechanisms to regulate protein-DNA interaction, our primary conclusion stands that TF occupancy is strongly modulated by site-dependent effects.

Our results establish a low level of mutational load directly affecting transcription factor occupancy in the 4 founder genomes. Although variants were found at 21% of surveyed CTCF sites, only 0.9% of binding sites exhibited a difference in occupancy due to polymorphism (Figure 4-2). Previous studies have identified varying levels of positive and negative selection in transcription factor recognition sequences by estimating changes in binding energy (He et al., 2011; Kim et al., 2009; Moses, 2009). However, our results indicate that 87% of polymorphism observed in the region of protein-DNA contact does not affect binding (Table 4-1). This is a higher proportion of silent variation than predicted by binding energy models (Figure 4-14), providing evidence that the scope of sequence change consistent with neutral evolution may be larger than previously thought.

Interestingly, we observed that the allele with higher occupancy was the derived allele in 40% of the cases (assuming the chimpanzee allele is ancestral). This indicates that approximately 40% of the functional substitutions in the human lineage increased occupancy, which is surprisingly high given that most mutations might be expected to reduce binding energy.

Previous work studying the power of comparative genomics has predicted a steep increase in the number of sequenced genomes required to obtain nucleotide resolution, particularly in the absence of perfect conservation (Eddy, 2005). While genome-wide phylogenetic footprinting approaches have highlighted substantial conservation of transcription factor sequence specificities (Kellis et al., 2003; Peterson et al., 1993; Wilson et al., 2008; Xie et al., 2007), functional studies of diverged species have uncovered low conservation in occupancy at orthologous sites (Borneman et al., 2007; Dermitzakis and Clark, 2002; Schmidt et al., 2010). Any

phylogenetic approach is thus a compromise between statistical power gained by sampling more diverged species and the ability to recognize similar functional elements by sequence similarity. The optimum evolutionary distances to sample may be different for assessing functional non-coding elements than for more conserved coding sequence (Blow et al., 2010; Wray et al., 2003). This tradeoff suggests a potential motivation for broad resequencing of natural populations, though even this approach faces the fundamental limitation of ineffective purifying selection in primates and humans (Coventry et al., 2010).

Gene-based studies have successfully identified causal variants using current methods for prediction of functional non-synonymous protein variants (Ng et al., 2010). Coding mutations in the CTCF gene affecting its DNA binding specificity have been identified in cancer samples (Filippova et al., 2002), but lesions in its binding sites are harder to interpret. The link between cognate recognition sequence and cellular consequence is complicated by potential influences from the cell-type specific chromatin landscape (John et al., 2011), maintenance of regulatory function despite sequence rearrangement (Weirauch and Hughes, 2010), altered association with protein complexes (Leung et al., 2004; Meijsing et al., 2009), and the lack of binding specificity and occupancy data for common transcription factors. In spite of these caveats, our results indicate that a motif-based classifier of variation in experimentally-identified CTCF binding sites predicts functional variation with a 59% true positive rate and a false positive rate of 20% (Figure 4-13). In comparison, current methods for prediction of non-synonymous protein variants achieve a 73% to 92% true positive rate at the same false positive rate (Adzhubei et al., 2010) – not dramatically greater considering the comparatively greater depth of variant databases used to assess coding variation. Given the encouraging performance of a straightforward functional genomics approach, that the majority of variants presently associated with human physiology and pathology lie in non-coding regions (Hindorff et al., 2009) should be grounds for optimism.

In summary, our results indicate the existence of widespread recognition site-dependent buffering of polymorphism within regulatory DNA regions. A major implication of our work is that the potential for accurately predicting the consequences of variation affecting regulatory factor recognition sequences is severely limited by complex context dependencies, necessitating empirical assessment using functional genomic approaches. The feasibility of approaches such as the one we describe here has recently dramatically increased

owing to coupled advances in next-generation sequencing technology and molecular biology, and continuation of this trend should in the near future enable further systematic investigations into the effect of polymorphism on protein-DNA interaction on a routine basis.

4.6 Figures and Tables

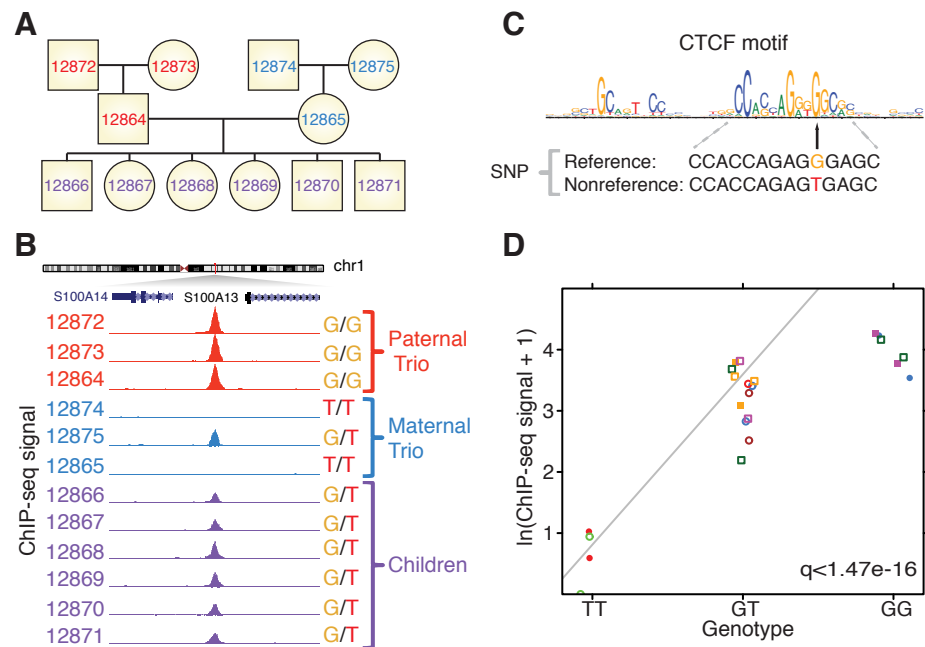


Figure 4-1 Effect of genetic variation on transcription factor occupancy.
(A) We performed ChIP-seq for the transcription factor CTCF followed by targeted resequencing of its complete occupancy landscape in 12 members of CEPH pedigree 1459 (CEU).
(B) Three qualitative levels of occupancy correspond to three genotypes of a SNP located at the binding site, with G/G homozygotes having the highest occupancy (region shown: chr1:151,853,500-151,859,700 [hg18]).
(C) The SNP shown in (B) disrupts a critical position in the CTCF consensus sequence (note that G better matches the consensus recognition sequence).
(D) Regression of ChIP-seq signal on genotype at the site in (B) quantifies the effect of SNPs on occupancy. We applied this strategy genome-wide to identify sites where SNPs are associated with differences in occupancy. At this site, Akaike information criterion favored a dominant effect model (GT and GG coded identically) over an additive model.

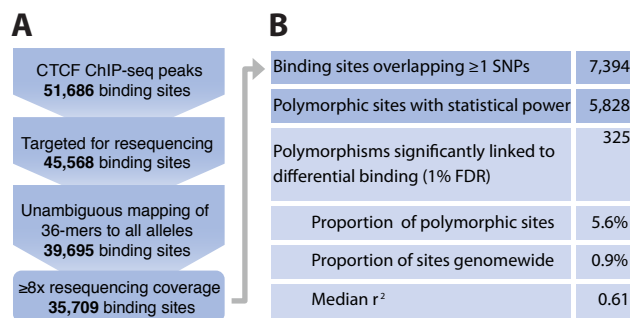


Figure 4-2 Genome-wide survey of the effect of genetic variation.
(A) Filtering strategy for testable CTCF binding sites. A number of binding sites were excluded from the analysis due to microarray probe design constraints, poor mappability, differing mappability between two alleles, or insufficient resequencing coverage.
(B) Summary of the prevalence of SNPs that affect CTCF occupancy at an FDR of 1%. Some sites overlapping SNPs were excluded for having insufficient data points per genotype to perform a robust regression. The model explained a substantial amount of the variance at significant sites (median r^2 of 0.61).

	<i>Polymorphism location known relative to motif match</i> within DNase I footprint (44bp)			<i>flanking the DNase I footprint</i> (2x45bp)			<i>All polymorphisms</i> Entire resequenced region (134bp)		
	# of sites	% polymorphic sites	% sites genome-wide	# of sites	% polymorphic sites	% sites genome-wide	# of sites	% polymorphic sites	% sites genome-wide
Regression of binding signal on genotype									
CTCF binding sites overlapping polymorphisms with respect to reference	1,814			4,071			7,394		
Polymorphic sites with sufficient data-points for regression	1,368			3,060			5,828		
Polymorphisms significantly linked to differential binding ($q < 0.05$)	263	19.2%	0.7%	110	3.6%	0.3%	542	9.3%	1.5%
Polymorphisms significantly linked to differential binding ($q < 0.01$)	186	13.6%	0.5%	32	1.0%	0.1%	325	5.6%	0.9%
Polymorphisms significantly linked to differential binding ($q < 0.001$)	147	10.7%	0.4%	15	0.5%	0.0%	221	3.8%	0.6%
Allele-specific binding									
Sites overlapping heterozygous polymorphisms	1,983			3,537			6,363		
also with sufficient and equal mappability for both alleles	1,962			3,489			6,250		
also with 13x read depth	906			1,309			2,458		
significantly allele-specific sites ($q < 0.01$)	139	15.3%	0.4%	81	6.2%	0.2%	260	10.6%	0.7%
significantly allele-specific sites ($q < 0.005$)	99	10.9%	0.3%	51	3.9%	0.1%	181	7.4%	0.5%
significantly allele-specific sites ($q < 0.001$)	36	4.0%	0.1%	17	1.3%	0.0%	64	2.6%	0.2%

Table 4-1 Survey of the effect of SNPs at transcription factor binding sites - 36,834 CTCF sites.

SNPs are broken down by location relative to motif in the first two columns, either within the region of protein-DNA contact or outside it. The third column summarizes SNPs regardless of location relative to the motif; sites with multiple overlapping SNPs are counted once.

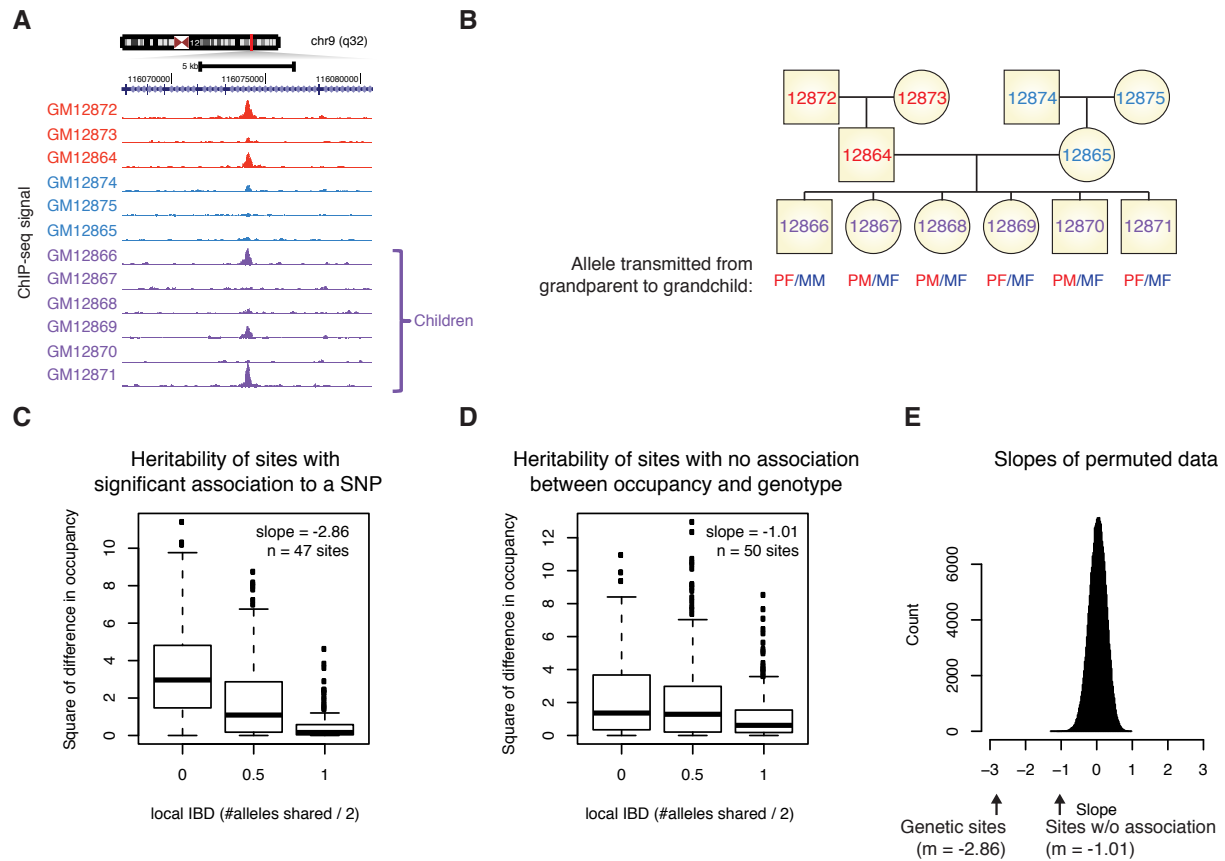


Figure 4-3 Aggregate linkage analysis confirms that association analysis identifies most but not all heritable signal.

(A) Example site showing differential occupancy among the 6 grandchildren

(B) The alleles transmitted to each grandchild illustrate identity by descent (IBD). PF, paternal grandfather; PM, paternal grandmother; MF, maternal grandfather; MM, maternal grandmother. Note that the grandchildren who share alleles from the same grandparent have more similar occupancy; compare the ChIP-seq signal in (A) for the grandchildren who inherited an allele from the paternal grandmother (PM) for (GM12867, GM12868 and GM12870) to the signal for those who inherited an allele from the paternal grandfather (PF; GM12866, GM12869 and GM12871).

(C, D) Aggregate Haseman-Elston linkage analysis sites exhibiting differential occupancy in the 6 grandchildren, analyzing occupancy at (C) a positive control of 47 autosomal sites where already associated with a SNP or (D) the remaining 50 autosomal sites not significantly associated with a SNP. A negative slope indicates that sib pairs with more similar inheritance at the binding site also exhibit more similar occupancy at that site, thus implying that the variation in occupancy is heritable.

(E) Permutation analysis to quantify significance of slopes from (C, D). Both sets of sites (arrows) show significant heritability compared to permuted data, and the signal is weaker for sites not directly implicated in the association analysis (right arrow).

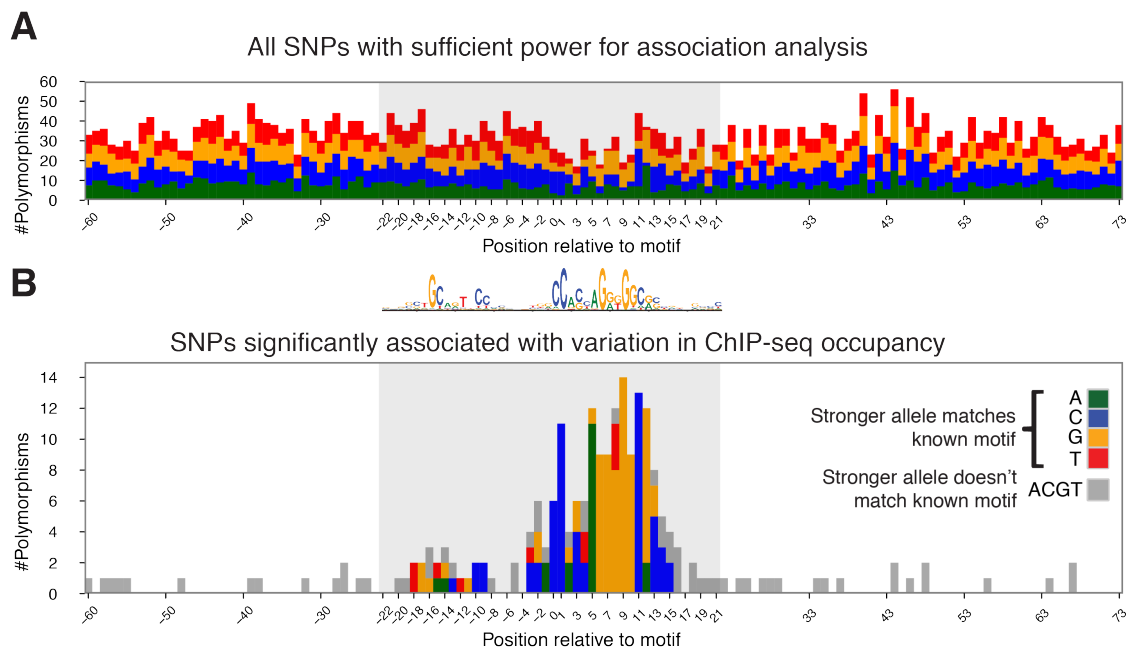


Figure 4-4 Functional SNPs recapitulate the CTCF binding motif.

(A) 4,428 SNPs identified by resequencing at as many sites. Y-axis indicates the number of SNPs identified at a given position (x-axis) relative to the aligned and strand-oriented CTCF motif (below). Bar color indicates alleles of SNPs. Gray shading indicates the 44-bp extent of protein-DNA interaction. Note that SNPs are uniformly distributed throughout the entire window, except for a slight reduction in diversity corresponding to the high-information content positions of the motif.

(B) Of the SNPs in (A), 218 are significantly associated with ChIP-seq occupancy (FDR 1%). Color indicates SNPs for which the higher-occupancy allele (according to association analysis) also had a higher log-odds score in the known motif. Gray indicates SNPs that affected occupancy, but the higher-occupancy allele had a lower score in the motif. See Figure 4-5C for full color. Note that these SNPs are concentrated in the region of protein-DNA contact, and 84% match the allele predicted by the canonical motif (above).

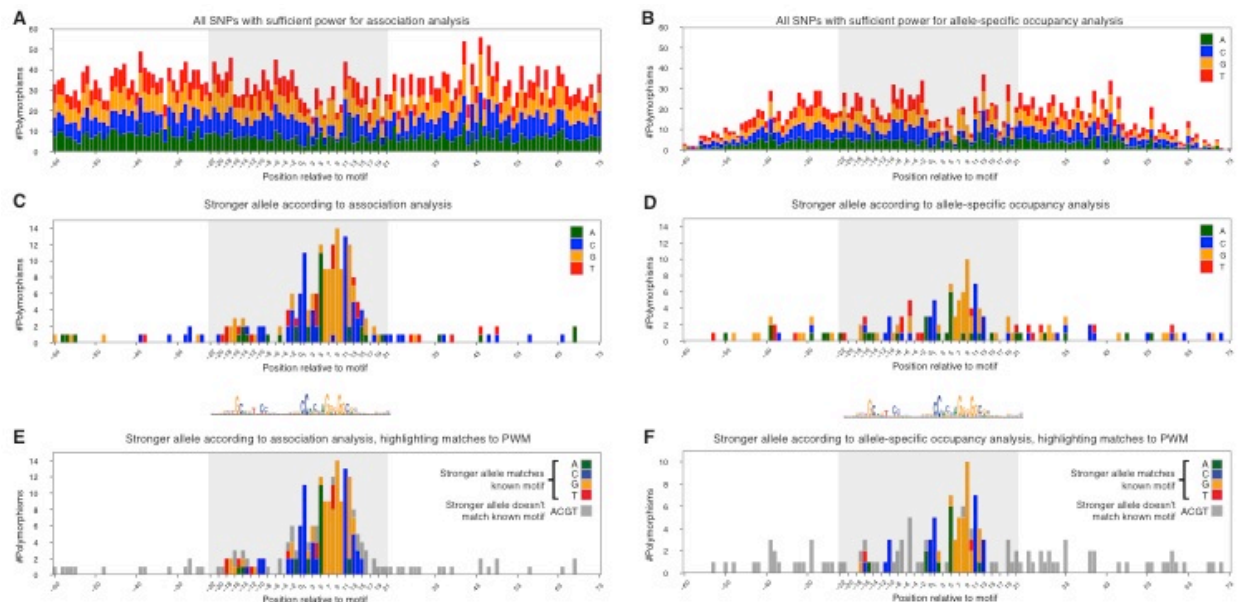


Figure 4-5 Comparison of association and allele-specific analyses.

(A-B) Y-axis indicates the number of SNPs tested for (A) association with occupancy and (B) allele-specific occupancy at a given position (x-axis) relative to the known CTCF motif (below). Color indicates alleles of the SNPs at each position. Gray shading indicates the extent of protein-DNA contact. Although SNPs were found distributed uniformly throughout the resequencing window, note that in (B) fewer SNPs are testable overall and at the flanks given the need for heterozygous SNPs with high ChIP-seq coverage, despite several additional replicates. We observed a slight reduction in SNPs directly over the core motif, corresponding to the most evolutionarily conserved region.

(C-D) SNPs significantly associated with differences in ChIP-seq occupancy (C) and sites demonstrating allele-specific occupancy (D). Color indicates higher-occupancy allele. Both analyses demonstrate an enrichment of significant SNPs over the region of protein-DNA contact, though to a greater extent in (C).

(E-F) Same as (C-D), but color indicates SNPs for which the higher-occupancy allele (according to association analysis) also has a higher log-odds score in the canonical motif and gray indicates SNPs that affect occupancy, but the higher-occupancy allele has a lower score in the motif. In comparison to the allele specific occupancy (F), the regression analysis (E) identifies more sites (325 vs. 181), has a higher concentration of significant SNPs within the region of protein-DNA contact (85% vs. 66%), and has a higher proportion of SNPs matching the expected binding energetics (84% vs. 68%). (A, E) are reproduced from Figure 4-4.

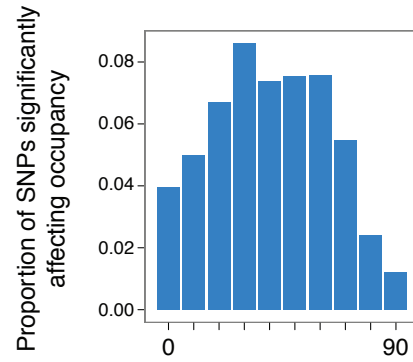


Figure 4-7 SNPs at the weakest and strongest sites are less likely to affect occupancy.

X-axis, decile of ChIP-seq signal for the heterozygote genotype according to the regression model; each decile represents 583 sites. Y-axis, proportion of sites in at which SNPs are associated with differential occupancy.

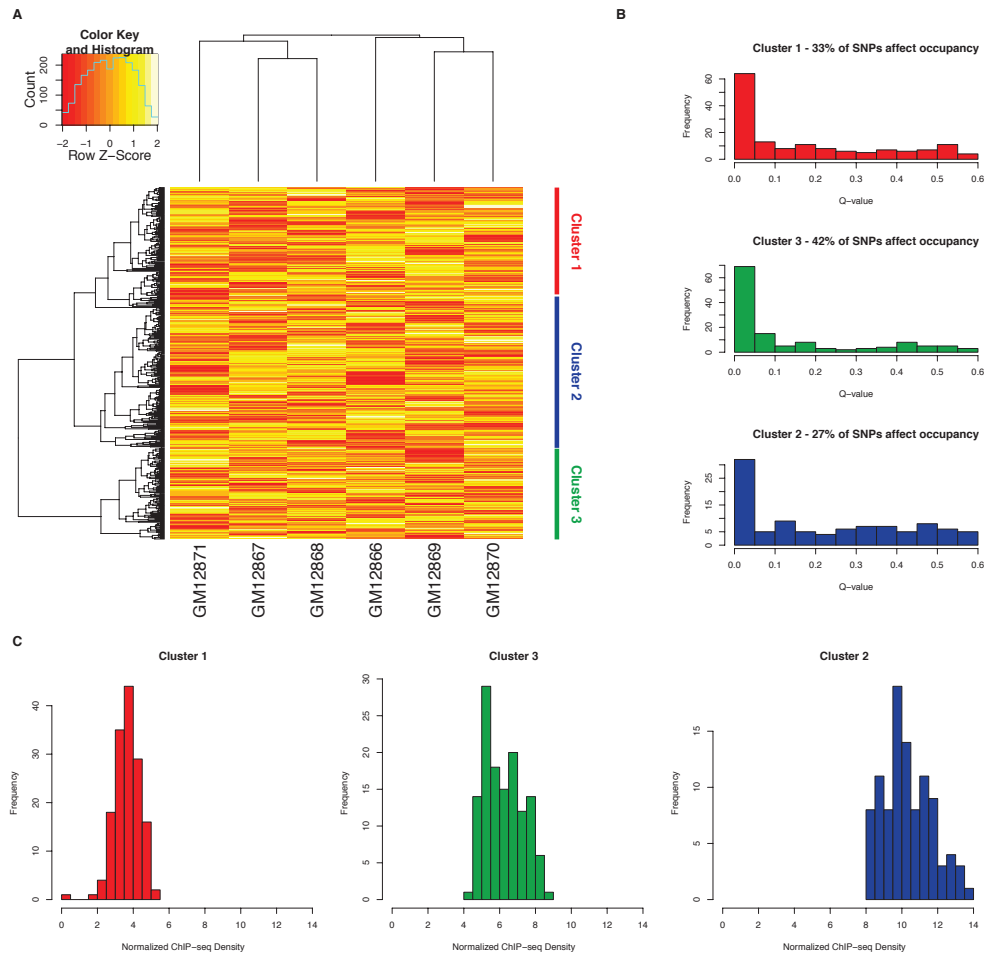


Figure 4-8 Hierarchical clustering confirms strength-dependent buffering.

(A) Normalized ChIP-seq density for 379 CTCF sites (Y-axis) with polymorphism in the core region of protein-DNA contact (positions 0-13) in the six grandchildren (X-axis). Hierarchical clustering resolves three clusters (labeled at right). (B) However, the percentage of polymorphisms that significantly affect binding is higher in cluster 3 (42%) than in cluster 1 and 2 (33% and 27%).

(C) The sites comprising the three clusters are distinguished by their overall ChIP-seq intensity, with cluster 1 being the weakest sites, cluster 3 being intermediate, and cluster 2 being the strongest sites; compare to Figure 4-7.

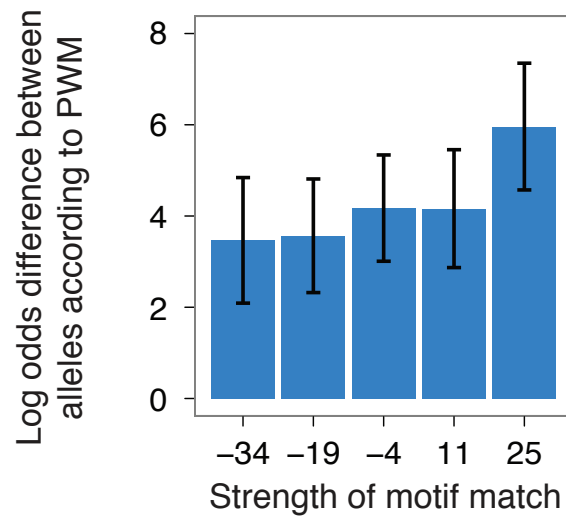


Figure 4-9 SNPs affecting occupancy despite stronger motif contexts involve more severe perturbations. X-axis, log-odds score of motif match, stronger matches at the right, label represents lower limit of bin. Y-axis, magnitude of perturbation, represented by the difference in log-odds scores between the two alleles. Error bars indicate standard deviation. Compare to Figure 4-10, wherein SNPs not affecting occupancy show no such trend.

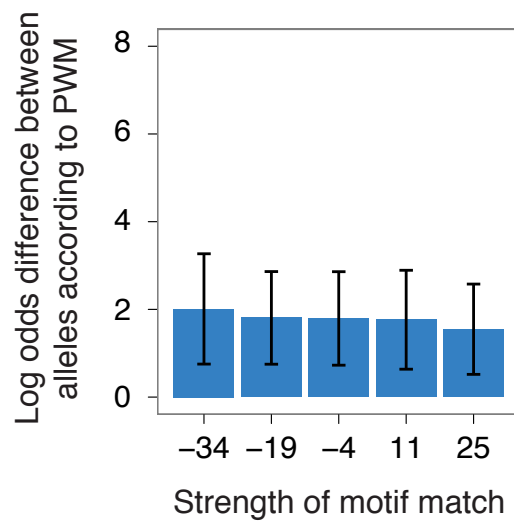


Figure 4-10 SNPs not affecting occupancy show no relationship between the magnitude of the perturbation and the strength of motif match. Y-axis, magnitude of perturbation, represented by the difference in log-odds scores between the two alleles. X-axis, log-odds score of motif match, stronger matches to the right, label represents lower limit of bin.

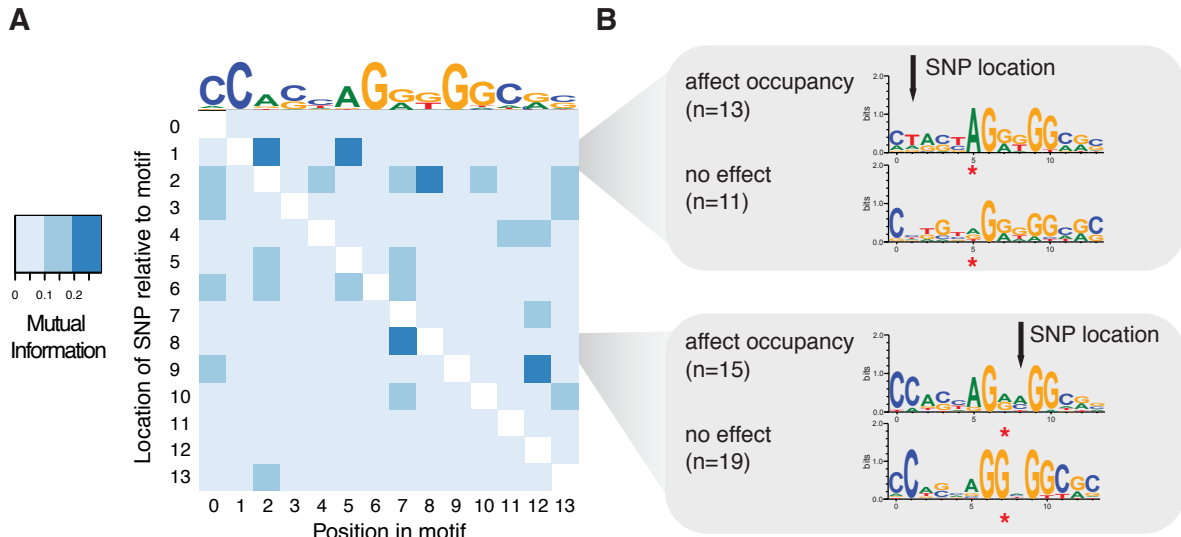


Figure 4-11 Sequence context buffers effect of polymorphism on occupancy

(A) Each cell measures the mutual information between the base pair at positions in the core motif (x-axis) and whether a SNP at another position in the motif (y-axis) affects occupancy (FDR 5%).

(B) Sequence context at sites with SNPs (arrows) at position 1 (above), 6 (below), divided by whether the SNP affected occupancy. Red stars highlight significant sequence differences ($q < 0.05$, see Section 4.6) between buffered and unbuffered sites at positions with elevated mutual information along the x-axis in (A).

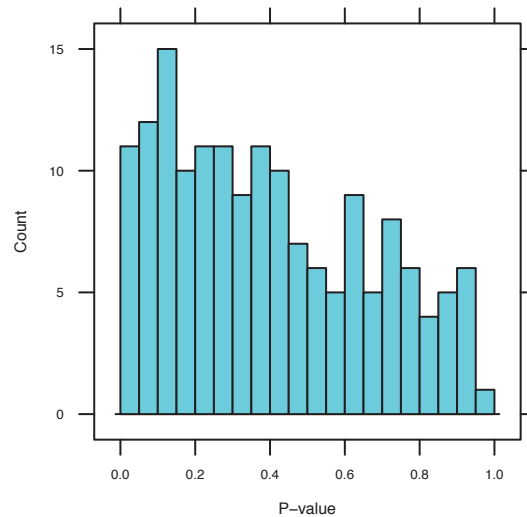


Figure 4-12 Statistical significance of the mutual information between sequence context and SNPs affecting occupancy.

P-values estimated by bootstrap; two interactions with $P < 0.0025$ were considered significant ($q < 0.05$, see Section 4.6).

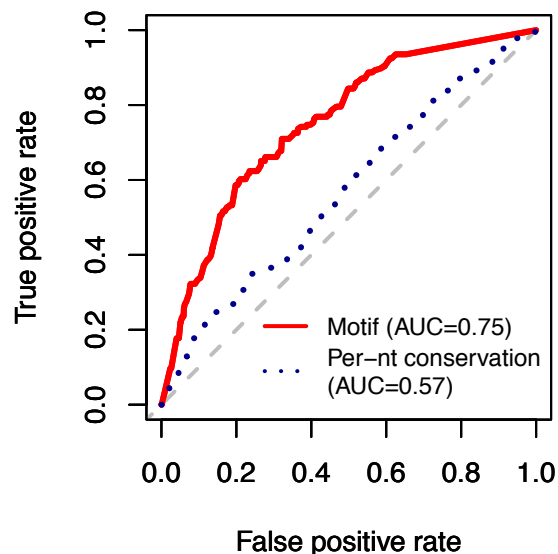


Figure 4-13 Power of existing measures to predict the effect of regulatory polymorphism.

ROC curve evaluating the power of two measures on the 1,368 SNPs in this study found within the region of protein-DNA contact, 186 of which significantly affect occupancy. Dotted blue line represents predictions by ranking SNPs in decreasing order of inferred purifying selection (phyloP per-nucleotide conservation score) at the location of the SNP. Solid red line represents predictions by ranking SNPs based on the difference in log-odds scores between alleles. Area under the curve (AUC) summarizes overall predictive power. Gray line indicates a random predictor and has an AUC of 50%. A perfect predictor would be plotted as a right angle, ranking all functional SNPs ahead of all nonfunctional SNPs, and would have an AUC of 1.0. While per-nt conservation performs little better than chance, consideration of binding energetics substantially improves performance.

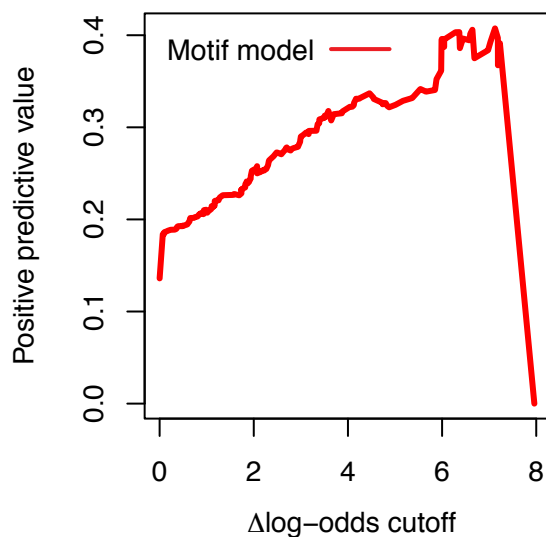


Figure 4-14 Positive predictive value of PWM model to predict the effect of SNPs on occupancy.

X-axis measures stringency of cutoff, represented by the difference in log-odds scores between alleles. At a log-odds difference cutoff of 7.1, 41% of predictions represent true positives.

Cell Line	Sample ID	Total Tags	Uniquely Mapped Tags	Duplicate Tags	Tags used in analysis	Enrichment	Replicate
GM12864	DS-10988	14898735	8914168	622904	8291264	41%	a
GM12864	DS-11499	19392595	11686775	4132298	7554477	39%	b
GM12864	DS-14431	21244507	13987484	3329143	10658341	35%	
GM12865	DS-10991	12514305	7606216	530324	7075892	43%	a
GM12865	DS-11502	16654378	10011914	4179342	5832572	46%	b
GM12865	DS-13756	18616806	11925797	317093	11608704	22%	
GM12866	DS-10994	13262868	7926482	591349	7335133	39%	a
GM12866	DS-11802	14417888	6519505	2672176	3847329	45%	b
GM12867	DS-10995	14859058	8605228	437957	8167271	35%	a
GM12867	DS-11803	16897125	8757167	1682855	7074312	26%	b
GM12868	DS-10996	15066941	8583490	534302	8049188	39%	a
GM12868	DS-11804	16401944	8985881	2411297	6574584	29%	b
GM12869	DS-10997	15260735	8707430	494744	8212686	37%	a
GM12869	DS-11898	16088244	6685107	2061858	4623249	37%	b
GM12870	DS-10998	14357100	7971940	360788	7611152	34%	a
GM12870	DS-10729	6761526	1616470	41625	1574845	39%	b*
GM12871	DS-10999	15467271	10863349	3461453	7401896	33%	a
GM12871	DS-11504	16862546	10500095	3412438	7087657	46%	b
GM12872	DS-10989	14450511	8574405	502377	8072028	38%	a
GM12872	DS-11500	16971110	10714955	2582979	8131976	30%	b
GM12872	DS-14432	25225362	16139763	3215611	12924152	37%	
GM12873	DS-10990	14411318	8864140	885293	7978847	47%	a
GM12873	DS-11501	17128653	10828323	4441784	6386539	45%	b
GM12873	DS-14433	25262030	16777109	3912497	12864612	46%	
GM12874	DS-10992	15065919	8737685	526321	8211364	39%	a
GM12874	DS-11503	14800871	9120343	4039850	5080493	51%	b
GM12875	DS-10993	15182690	9115764	583900	8531864	40%	a
GM12875	DS-11899	14649260	6863906	1721071	5142835	37%	b

Table 4-2 Summary of ChIP-seq data mapped to customized genomes with Mosaik.

Tags used in analysis indicates the uniquely aligned tags remaining after removing duplicates. Enrichment indicates the proportion of tags mapping to a ChIP-seq peak, representing the enrichment over background and indicating the quality of the data. Biological replicate structure: Replicates a and b were used in the regression analysis. 4 further available replicates (for GM12864, GM12865, GM12872, and GM12873) were used to add power for the allele-specific occupancy analysis, but were not used elsewhere. * Replicate “b” of GM12870 was only used for the linkage analysis.

Sample	R (pearson)
GM12872	0.92
GM12873	0.93
GM12874	0.89
GM12875	0.9
GM12864	0.93
GM12865	0.92
GM12866	0.87
GM12867	0.89
GM12868	0.89
GM12869	0.88
GM12870	0.86
GM12871	0.92

Table 4-3 Pearson correlation of two replicates per individual of ChIP-seq data.

Correlation performed after normalization at signal at ChIP-seq binding peaks.

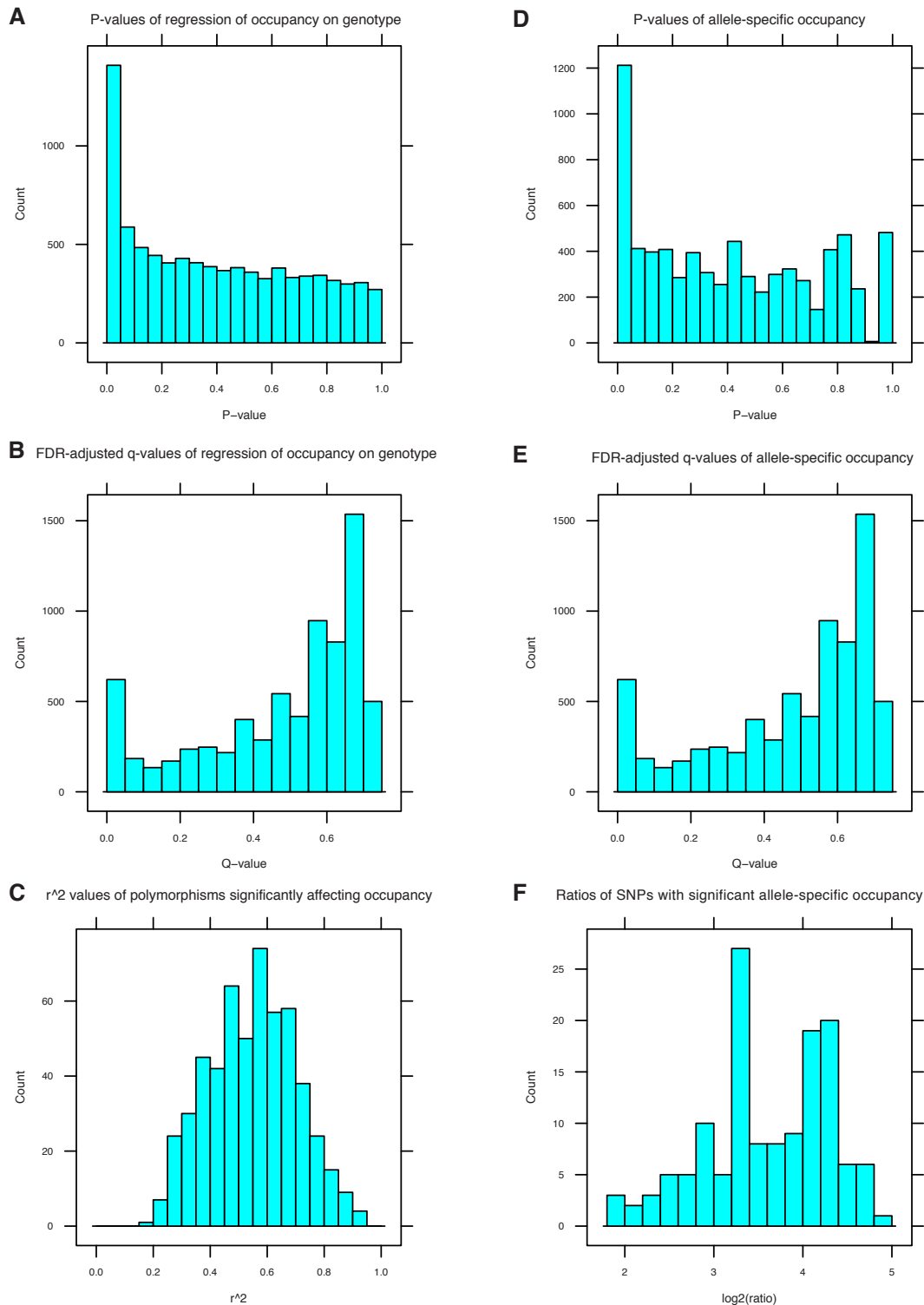


Figure 4-15 Statistical identification of association with differential occupancy and allele-specific occupancy. (A-C) Association of SNPs in *azs* with ChIP-seq occupancy (A) Histogram of p-values for all tested binding sites (B) Histogram of FDR-adjusted q-values (C) Effect size of SNPs associated with differences in occupancy (FDR 1%), measured by r^2 of a linear regression (D-F) Allele-specific occupancy at heterozygous sites (D) Histogram of p-values for all tested binding sites (E) Histogram of FDR-adjusted q-values (F) Effect size of SNPs associated with allele-specific occupancy (FDR 0.005%), measured by the log of the ratio of the counts of reads mapping to the higher and lower alleles.

Cell Line	Sample ID	Total Tags	Uniquely Mapped Tags	Total Lanes	Duplicate Tags	Tags used in analysis	On-target percent	#Sites >=8x coverage	%Sites >=8x coverage
GM12864	DS-14519	93593928	88891393	3	47554426	41336967	100.00%	44151	86.30%
GM12865	DS-14307	69725559	67620616	3	26777287	40843329	100.00%	48129	94.10%
GM12866	DS-14308	102813128	98946807	3	54891039	44055768	100.00%	47950	93.70%
GM12867	DS-14130	53037431	50150666	3	14725478	35425188	100.00%	44648	87.30%
GM12868	DS-14309	74444333	72481070	3	28030175	44450895	100.00%	48328	94.50%
GM12869	DS-14520	95718487	91869195	3	50106159	41763036	100.00%	45919	89.80%
GM12870	DS-14521	92855619	88439230	3	45266342	43172888	100.00%	43348	84.70%
GM12871	DS-13962	69404161	67211079	4	34841881	32369198	100.00%	45420	88.80%
GM12872	DS-14310	94181984	91445563	3	53219474	38226089	100.00%	49314	96.40%
GM12873	DS-14522	101158691	97419508	3	49979729	47439779	100.00%	44225	86.40%
GM12874	DS-14523	105488449	102515934	3	51152704	51363230	100.00%	46489	90.90%
GM12875	DS-14524	91621562	88163680	3	44837718	43325962	100.00%	44268	86.50%

Table 4-4 Summary of resequencing data mapped using bwa.

Tags used in analysis indicates the uniquely aligned tags remaining after removing duplicates. On-target percent indicates the fraction of tags mapping back to within ± 100 bp of a resequencing target.

4.7 Materials and Methods

Cell Culture

The B-lymphoblastoid cell lines from CEPH pedigree 1459 were obtained from Coriell and cultured in RPMI1640 medium (Cellgro), supplemented with 15% fetal bovine serum (FBS, Hyclone), 2 mM L-Glutamine, and 25 IU/mL penicillin and 25 µg/mL streptomycin (Cellgro).

ChIP-seq

B-lymphoblasts (5 million cells) were crosslinked with 1% formaldehyde (Sigma), lysed in lysis buffer (50 mM Tris-HCl pH8.0, 10 mM EDTA pH8.0, 1% SDS), and sheared by Bioruptor (Diagenode). The supernatant was further diluted 10-fold with dilution buffer (50 mM Tris-HCl pH8.0, 166 mM NaCl, 1.1% Triton X-100, 0.11% sodium deoxycholate). For each immunoprecipitation, 100 µL Dynabeads (M-280, sheep anti-rabbit IgG, Invitrogen) were incubated with 20 µL CTCF antibody (#2899, Cell Signaling) for at least 6 hours at 4 °C. The antibody-conjugated beads were then incubated overnight with sheared chromatin. The complexes were washed with IP wash buffer I (50mM Tris-HCl pH8.0, 0.15 M NaCl, 1mM EDTA pH8.0, 0.1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate), high salt buffer (50mM Tris-HCl pH8.0, 0.5 M NaCl, 1mM EDTA pH8.0, 0.1% SDS, 1% Triton X-100, 0.1% sodium deoxycholate), and TE buffer (10 mM Tris-HCl pH8.0, 0.1 mM EDTA pH8.0). Crosslinking was then reversed in elution buffer (10 mM Tris-HCl pH8.0, 0.3 M NaCl, 5 mM EDTA pH8.0, 0.5% SDS) at 65 °C overnight. The DNA was separated from the beads and treated with Proteinase K (Fermentas) and purified by phenol-chloroform extraction and ethanol precipitation.

Sequencing libraries were constructed according to Illumina's genomic prep kit protocol as previously described (Hesselberth et al., 2009). Briefly, ChIP DNA was end-repaired using the End-it DNA repair kit (Epicentre). Adenines were added to the 3' ends of the blunt-ended DNA using Taq DNA polymerase (NEB). PE adapter (1:20 dilution, 1 µL for 15-50 ng starting ChIP DNA, Illumina) was ligated to the ends of the A-tailed ChIP DNA with T4 DNA ligase (NEB). 1/3-1/4 of the purified ligation product was PCR amplified for 16 cycles with High-fidelity PCR master mix (NEB) and PE primer 1.0/2.0 (Illumina). Libraries were run

on 2% agarose gels, size-selected, and purified with QIAquick gel extraction kit (Qiagen). The libraries were sequenced to 36 bp on an Illumina Genome Analyzer by the High-throughput Genomics Unit (University of Washington) according to standard protocol. ChIP-seq experiments were performed on 2-3 independently cultured biological replicates per sample (Table 4-2, Table 4-3).

ChIP-seq peak calls

High quality reads were aligned to the reference genome using the Eland aligner. SPP (Kharchenko et al., 2008) was used to call peaks on total tag data from the 12 samples, resulting in 51,686 peaks at a Poisson-derived FDR of 1%. Using the aggregate of all tags was more conservative (in terms of number of peaks) than taking the union of peak calls on individual samples, but less conservative than taking the intersection. The MTC method (“tag.lwcc”) was used to call point binding positions Motif representations used Weblogo (Crooks et al., 2004).

Capture resequencing

We designed an Agilent 244k SureSelect microarray for targeted resequencing on the 51,686 ChIP-seq peaks identified in the 12 samples. We used fimo (<http://meme.sdsc.edu/meme/>) to scan for instances of the core 14 bp of the canonical motif (Kim et al., 2007) with a p-value of 10^{-2} . We adjusted the target locations to center on matches to the nearest CTCF motif if the motif was within 50 bp, and added flanking targets to capture additional nearby motifs. 5 potential probes were tiled at 15 bp spacing to the 120 bp surrounding each target. We adjusted probe binding energy similarly to Ng et al. (Ng et al., 2009), adjusting the spacing of probes by up to 5 bp and adjusting the lengths to between 40-60 bp to reach a predicted T_m between 60-72 °C. We used the Duke Uniqueness 20 bp track (UCSC genome browser) to filter out 5,828 probes with potential for cross-hybridization. We further excluded 145 probes in satellite repeats (RepeatMasker, UCSC genome browser) or with high blast scores to multiple genomic locations. The final design had 242,380 probes targeting 46,652 CTCF sites.

Genomic DNA was extracted from cultured cells, and targeted capture was performed based on Supplementary Protocol 3 of Mamanova et al. 2010 (Mamanova et al., 2010) with modifications. Briefly, 12 µg of ge-

nomic DNA was sheared in a Covaris S2 (Covaris, Inc.) using a duty cycle of 10%, intensity 5, cycle/burst 100, time 600 sec. DNA was end-repaired, A-tailed, and ligated to SE adapters (Illumina), and purified with Agencourt AMPure XP beads (Beckman Coulter). A trial PCR amplification using Phusion HF polymerase (NEB) and SE primers SLXA_FOR_AMP and SLXA_REV_AMP (Ng et al., 2009) (IDT) was performed on a fraction of ligated DNA with an Roche LightCycler 480 to determine the optimal number of cycles. PCR for all ligated DNA was then performed in eight 200 μ L tubes using the following program: 98 °C for 30 s, then a previously determined number of cycles of 98 °C for 10 s, 65 °C for 30 s, 72 °C for 30 s, followed by 72 °C for 5 min. The final DNA library was pooled and concentrated to 5 μ L in a SpeedVac. For hybridization, 10 μ g of DNA library was combined with formamide (Ambion), blocking oligos (SLXA_FOR_AMP, SLXA_REV_AMP, SLXA_REV_AMP_rev and SLXA_FOR_AMP_rev (Ng et al., 2009), Human C₀t-1 DNA (Invitrogen), 2X Hi-RPM Hybridization Buffer (Agilent) and 10x Blocking Buffer (Agilent). Hybridization was performed using the Maui Hybridization System (BioMicro Systems, Inc.) at 55 °C for 48 hours according to “MAUI® Mixer LC on Agilent® 244K CGH Microarrays” protocol. After hybridization, microarrays were washed with Agilent aCGH Wash Buffers 1 and 2, sealed with Secure-Seal (GRACE Bio-Labs) and placed on heat block (VWR) for elution of DNA. DNA was eluted with 3 mL of 95 °C PCR-grade water, concentrated, and amplified with SE primers using the same PCR program as above. The libraries were sequenced to 36 bp on an Illumina Genome Analyzer by the High-throughput Genomics Unit (University of Washington) according to standard protocol.

Reads were aligned to the human genome (hg18) using bwa 0.5.8 (Li and Durbin, 2009) using default settings, allowing up to 2 mismatches (Table 4-4). Some lanes exhibited an excess of mismatches to the reference sequence at the 5' or 3' end, so tags were trimmed by up to 9 bp. Reads with identical 5' ends were presumed to be PCR duplicates and were excluded using Picard v1.22 MarkDuplicates (<http://picard.sourceforge.net/>). SAMtools v0.1.7 (Li et al., 2009) was used to generate a pileup of potential SNPs from uniquely-mapping reads with a mapping quality above 30. SNPs were called as biallelic variants with at least 8x resequencing coverage, a Phred-scaled SNP quality of at least 30, at least 20% of reads matching the allele with lower cov-

erage, and Mendelian segregation according to PLINK v1.07 (Purcell et al., 2007). The chimpanzee allele was identified using axtNet alignment files for PanTro2 from the UCSC Genome Browser.

We performed two validations of the SNP calls from our targeted resequencing. First, we performed Sanger sequencing on PCR products from genomic DNA. We tested 33 SNPs in all 12 samples. 0 of the 388 genotype calls were discordant. Second, we compared genotypes with genotypes available from the HapMap project for 6 of the 12 samples. Release 27 genotypes were obtained from <http://hapmap.ncbi.nlm.nih.gov>, and matched to capture resequencing SNPs by location. 244 of the 27,808 genotype calls in common (0.88%) were discordant.

Measuring occupancy differences among individuals at a common set of peaks

For each individual, a custom human genome was created from hg18 including autosomes, unscaffolded contigs, mitochondrial DNA, a Y chromosome for males, and the Epstein-Barr virus genome (gi|9625578|ref|NC_001345.1). Each genome was personalized to reflect SNPs identified by resequencing, using IUPAC codes to represent heterozygous position. ChIP-seq data was mapped using MosaikAligner v1.1.0021 (<http://code.google.com/p/mosaik-aligner/>) with the options "-bw 13 -act 20 -mhp 100 -m unique -mm 4 -minp 1.0", requiring a unique mapping considering up to 4 mismatches. Reads with more than 2 mismatches were then discarded. Reads with identical 5' ends were presumed to be PCR duplicates and were excluded using Picard v1.22 MarkDuplicates. Smoothed density tracks were generated using bedmap (<http://code.google.com/p/bedops/>) to count the number of tags overlapping a sliding 150 bp window, with a step width of 20 bp. Density tracks were normalized for sequencing depth by a global linear scaling to fix an arbitrary value of 25 as the 50th percentile of bins with more than 15 reads. We identified instances of the canonical motif (fimo p-value < 10^{-2}) within 15 bp of the center of the resequencing target, keeping the motif with the best p-value. We measured occupancy by the maximum normalized ChIP-seq tag density over the 14 bp motif.

Regression of occupancy on genotype

We applied a regression method to measure whether a particular biallelic SNP is associated with occupancy, and if so which allele is associated with higher occupancy. We tested only sites with $\geq 8\times$ resequencing coverage in ≥ 6 samples, sufficient mappability, and data for ≥ 4 data points each for ≥ 2 genotypes and ≥ 12 data points overall. We further excluded 242 sites overlapping indels in the CEU population identified by the 1000 Genomes Project release 2010_07 (1000 Genomes Project Consortium, 2010). We used a negative binomial generalized linear model (glm.nb in the R package MASS) to measure the significance of the effect of polymorphism on occupancy using an additive effect model, and including a replicate term to account for batch effects. We used two ChIP-seq replicates per sample, except for GM12870, which had one replicate (Table 4-2).

For sites with more than one SNP, we tested only the SNPs with more data points and those inside the region of protein-DNA contact. If there were still multiple SNPs per window, we chose the SNP with the lowest p-value, though these sites were then excluded from analyses depending on the known position of the SNP relative to the motif. We also tried fitting a dominant effect model where permitted by sample size: we chose between additive and dominant effect models using the Akaike information criterion. We separately fitted a linear model on the same data to estimate the r^2 .

We used the R package qvalue to estimate q-values (Figure 4-15) (Storey and Tibshirani, 2003), which established a cutoff of $P < 9.6 \times 10^{-4}$ as an FDR of 1% (325 sites). Using a Benjamini-Hochberg FDR strategy confirms a similar cutoff for FDR 1% (4.9×10^{-4} , 293 sites). To independently confirm our FDR methodology, we considered the proportion of significant SNPs within the region of protein-DNA contact (Figure 4-4), which ranged from 71%, to 85%, to 91% at FDRs 5%, 1%, 0.1%, respectively (Table 4-1).

Aggregate Haseman-Elston linkage analysis

We used 31,128 SNPs identified in our resequencing as markers to generate a map of identity-by-descent (95th percentile marker spacing of 0.5 cM). We used recombination rate data (1000 Genomes Project Consortium, 2010) to place our SNP coordinates onto a genetic map, choosing SNPs with fewer missing genotypes at du-

plicate map positions. We then used MERLIN (Abecasis et al., 2001) to filter out improbable genotypes and compute IBD at our marker locations (option “-ibd”). We used the nearest marker at each binding site to estimate IBD for the 15 possible sib pairs. IBD values were placed into 3 bins (0, 0.5, and 1.0); values with >0.05 uncertainty were excluded.

We then used the package DESeq (Anders and Huber, 2010) to identify differentially occupied regions, both throughout all 12 samples as well as among just the grandchildren, using two replicates per sample (Table 4-2). We then applied a variance stabilizing transformation, and normalized the occupancy at each site to a mean of 0 and standard deviation of 1. We then averaged the occupancy of the two replicates.

We then performed Haseman-Elston regression at the 97 autosomal sites differentially occupied among the grandchildren, treating separately sites whose differential occupancy was already associated with a SNP or allele-specific binding. We considered all sib pairs and all sites simultaneously. Although regression methods exist with higher power (Feingold, 2002) or that use data from all members of the pedigree (Olson and Wijsman, 1993), we applied the original method of regressing squares of trait differences for sib pairs, reasoning that the robustness of a simple method would have more forgiving assumptions. To account for the non-independence of measuring multiple sib pairs from the same family, we assessed the significance of the regression slope by permuting IBD vectors to random sib pair difference vectors for all differentially occupied sites, thus maintaining any correlation structure in the data.

Allele-specific occupancy analysis

We tested allele-specific occupancy in pooled replicate data for each sample. Given the reliance of allele-specific occupancy tests on high coverage at heterozygous sites, we included data from an additional replicate for 4 samples (Table 4-2). We had sufficient power to test 2,535 heterozygous binding sites with adequate mappability and 13x ChIP-seq read coverage. We used a chi-squared test against a 50:50 null expectation to derive a p-value from the counts of the ChIP-seq tags mapping to the two alleles. We used the R package qvalue to estimate an FDR (Figure 4-15) (Storey and Tibshirani, 2003). In interpreting our FDR threshold, we considered the observed concentration of significant SNPs within the 44 bp region of protein-DNA interac-

tion (Figure 4-5), which increased at more conservative FDRs: 63%, 66%, 68% at FDRs 1%, 0.5%, 0.1%, respectively (Table 4-1). At sites for which multiple samples had testable heterozygous sites with the same alleles, the sample with the most total reads was picked as representative for plotting.

Position weight matrix (PWM) motif models

We used PWM models (Wasserman and Sandelin, 2004) to measure the effect of a polymorphism on information content. CTCF binds in a multivalent fashion (Bowers et al., 2009; Filippova et al., 1996; Ohlsson et al., 2001), wherein three modes of binding are distinguished by the presence and position of an upstream motif. At each site we chose the best-matching of the three motif models. To measure information content, we converted frequencies to log-odds, using a pseudo-frequency of 0.01 (the minimum observed frequency).

Mutual information analysis

We calculated the mutual information between whether a given SNP affected occupancy (FDR 5%) and sequence context at 14 positions in the core CTCF motif using the `mutualInfo` function of the R package `bioDist`. To estimate the significance of the mutual information values, we applied a bootstrap method for each pair of positions tested. We calculated p-values from 2,000 iterations of resampling per pair of positions (Figure 4-12). To account for multiple testing across all positions, we used the R package `qvalue` (Storey and Tibshirani, 2003).

ROC curve

We downloaded phyloP based on a 44-way vertebrate alignment from the UCSC Genome Browser. SNPs were ranked in decreasing order of the phyloP score at the location of the SNP, thus ranking sites with the most indication of purifying selection the likely to be disrupted by a SNP. To measure the predictive power of models of protein-DNA binding energy, we used a position weight matrix to compute the difference in log-odds score between the two alleles of each SNP. Sites with the largest difference between alleles at the location of the SNP were ranked as most likely to be disruptive. Plots were generated using the R package `ROCR` (Sing et al., 2005).

Data availability

ChIP-seq data are viewable in the UCSC Genome Browser (<http://genome.ucsc.edu>, version hg18), and have been deposited in GEO (GSE30263). Resequencing data are available in the SRA (SRP009457), and the capture resequencing array design in GEO (GPL14147).

4.8 Notes

This chapter has been published as **Matthew T. Maurano***, Hao Wang*, Tanya Kutyaev, John A. Stamatoyannopoulos. Widespread Site-dependent Buffering of Human Regulatory Polymorphism. *PLoS Genet.* 8, e1002599 (2012).

** co-first authors*

Acknowledgements

We thank Andrew Stergachis, Shamil Sunyaev, and Jeff Vierstra for critical reading of the manuscript and many helpful suggestions. We thank Eric Haugen, Eric Rynes, and the staff of the University of Washington Genome Sciences High-Throughput Genomics Center for technical assistance; Molly Weaver for help with Illumina library construction; and Richard Sandstrom for help with repository submission. We also thank members of the Stamatoyannopoulos and Akey laboratories for their helpful discussions.

5 Widespread Plasticity In CTCF Occupancy Linked To DNA Methylation

5.1 Abstract

CTCF is a ubiquitously-expressed regulator of fundamental genomic processes including transcription, intra- and inter-chromosomal interactions, and chromatin structure. Because of its critical role in genome function, CTCF binding patterns have long been assumed to be largely invariant across different cellular environments. Here we analyze genome-wide occupancy patterns of CTCF by ChIP-seq in 19 diverse human cell types including normal primary cells and immortal lines. We observed highly reproducible yet surprisingly plastic genomic binding landscapes, indicative of strong cell-selective regulation of CTCF occupancy. Comparison with massively parallel bisulfite sequencing data indicates that 41% of variable CTCF binding is linked to differential DNA methylation, concentrated at two critical positions within the CTCF recognition sequence. Unexpectedly, CTCF binding patterns were markedly different in normal vs. immortal cells, with the latter showing widespread disruption of CTCF binding associated with increased methylation. Strikingly, this disruption is accompanied by up-regulation of CTCF expression, with the result that both normal and immortal cells maintain the same average number of CTCF occupancy sites genome-wide. These results reveal a tight linkage between DNA methylation and the global occupancy patterns of a major sequence-specific regulatory factor.

5.2 Introduction

The polyfunctional regulator CTCF plays a central role in multiple complex genomic processes, including transcription (Baniahmad et al., 1990; Filippova et al., 1996; Vostrov and Quitschke, 1997), imprinting (Bell and Felsenfeld, 2000; Hark et al., 2000), and long-range chromatin interactions and subnuclear localization (Hou et al., 2008; Splinter, 2006; Yusufzai et al., 2004). Cohesin, a major mediator of chromosomal contacts during mitosis (Seitan et al., 2011) is tightly co-localized with CTCF, indicating a key function for CTCF in chromosome pairing (Parelho et al., 2008; Rubio et al., 2008; Wendt et al., 2008). CTCF has also been connected with multiple malignancies, including by the association of mutations in its gene locus (Filippova et al., 1998), its antiproliferative effect (Rasko et al., 2001), and through regulatory interactions with tumor suppressor genes (Butcher et al., 2004; Dávalos-Salas et al., 2011; Soto-Reyes and Recillas-Targa, 2010; Witcher and Emerson, 2009).

CTCF is ubiquitously expressed, and it is widely believed that CTCF binding patterns are largely invariant between cell types (Cuddapah et al., 2008; Heintzman et al., 2009; Kim et al., 2007), though diverse regulatory mechanisms at individual loci have been described (Lai et al., 2010; Lefevre et al., 2008; Sekimata et al., 2009; Shukla et al., 2011; Witcher and Emerson, 2009). In addition, at a small number of loci, variable CTCF occupancy has been linked with DNA methylation *in vivo* (Kanduri et al., 2000; Pant et al., 2003), and *in vitro* studies suggest that methylation may hinder CTCF binding at certain sequence elements (Bell and Felsenfeld, 2000; Filippova et al., 2001; Hark et al., 2000; Renda et al., 2007). However, neither the degree to which CTCF binding patterns vary between different cell types, nor the relationship of such variability with DNA methylation is currently known.

We therefore sought to establish the cellular selectivity of CTCF binding, and to define its relationship with methylation on a global scale. By using genome-wide occupancy profiling and reduced representation bisulfite sequencing, we establish that a majority of CTCF sites are cell-selective, and link 41% of this variable CTCF occupancy to differential DNA methylation. We further observe markedly different CTCF binding patterns distinguishing normal and immortal cells, which are associated with increased methylation and up-regulation

of CTCF expression. These results indicate a global linkage between DNA methylation and the occupancy patterns of an important genome regulator.

5.3 Results

Widespread plasticity of CTCF occupancy patterns

To assess CTCF binding variation genome-wide, we localized and quantified CTCF occupancy by ChIP-seq in 19 diverse cell types, including 7 immortal cell lines and 12 normal cell types. We generated two biological replicates for each cell type. Both replicates were of high enrichment and exhibited high concordance (average correlation of 0.93, Figure 5-1). We found that CTCF binds an average of about 55,000 sites in each tested cell type (Figure 5-1A). In total, we identified 77,811 distinct binding sites across all 19 cell types.

To survey binding variability genome-wide, we conservatively assessed how many cell types demonstrated binding at each site using a dual-threshold strategy to prevent bias towards variable sites (see Section 5.5). 27,662 binding sites were present in all 19 cell types. However, 50,149 binding sites were found to be unbound in at least one cell type. Thus, 64% of CTCF sites are found to vary in at least one cell type, demonstrating the existence of a widespread variability in CTCF occupancy. These variable sites exhibited clear occupancy differences between bound and unbound cell types, including at the well-known *H19/Igf2* imprinted locus (Figure 5-2). Variable binding sites were occupied in an average of 10 of 19 cell lines, implying a high degree of shared regulation between cell types (Figure 5-3). Indeed, between any two cell types, an average of 72% of bound sites were in common (Figure 5-4). Variable sites had a similar genomic localization (Figure 5-5) compared to constitutive sites.

Distinct CTCF binding landscapes in normal vs. immortal cells

To understand whether binding variability follows a similar pattern in related cell types, we performed an unsupervised hierarchical clustering of variable CTCF binding sites (see Section 5.5). We found that the variable CTCF binding landscape distinguished three groups (Figure 5-7A). The first group of immortal cells consists of malignancy-derived and EBV-immortalized cell lines including several carcinomas (colorectal, Caco-2; cervical, HeLa-S3; hepatocellular, HepG2), neuroblastoma (SK-N-SH_RA), and retinoblastoma (WERI-RB-1)

and EBV-transformed lymphoplastoid (GM06990). The remaining two groups consist of normal cell types of limited proliferative potential: the second group consists of three epithelial cell types, including renal cortical (HRE), small airway (SAEC) and esophageal (HEEpiC) mucosal epithelia, and the third group consists of fibroblasts, including abdominal (AG10803), toe skin (AG09309), gum (AG09319), aortic adventitial (AoAF), foreskin (BJ), mammary (HMF), pulmonary artery (HPAF), and pulmonary (HPF) and brain microvascular endothelium (HBMEC). Principal component analysis and bootstrap assessment of the uncertainty in the hierarchical clustering confirmed a separation between the normal cell types and remaining cell lines, although the epithelial line HRE was less clearly distinguished (Figure 5-8). We then sought to identify the specific binding differences characterizing these three groups. We identified 4,146 specific binding sites whose occupancy was significantly different between these groups at a false discovery rate (FDR) of 1% (see Section 5.5) (Figure 5-7). These results suggest that CTCF occupancy exhibits major regulatory differences distinguishing immortal cell lines from normal epithelium, endothelium and fibroblasts.

Variable CTCF occupancy linked to CpG methylation

Pre-existing methylation can antagonize CTCF binding *in vitro* (Bell and Felsenfeld, 2000; Hark et al., 2000; Kanduri et al., 2000). Therefore we asked whether differential methylation was associated with variable sites *in vivo*. To study this, we compared CTCF occupancy and reduced representation bisulfite sequencing (RRBS) data (Figure 5-9A). We studied a subset of CTCF sites in 13 cell types (n=6,707) for which RRBS data was available from the ENCODE project (Varley et al., 2013). We obtained methylation status of 44,048 CpGs dinucleotides in the region centered on these sites (see Section 5.5), with each CpG monitored in an average of 12 out of 13 cell types (Figure 5-10).

First, we assessed overall methylation status at the 6,707 CTCF sites with RRBS data. We found that methylation was substantially more variable at variable CTCF sites than at constitutive ones (Figure 5-11). Only 10% of these sites tested showed intermediate methylation status (between 25% and 75% methylation) (Figure 5-10). Overall, 98% of CTCF sites were unmethylated (defined as <50% methylation) in at least one of the cell types tested, confirming an inverse relationship between methylation and CTCF occupancy. However,

47% of CTCF sites were methylated (>50% methylation) in at least one cell type, suggesting a widespread potential link between methylation and CTCF occupancy.

To quantify the global association of differential methylation status with variable CTCF occupancy, we performed a linear regression analysis at the 6,707 sites for which we had RRBS data (Figure 5-9; see Section 5.5). 4,099 (61%) of these sites exhibited variable CTCF binding in the 13 cell types tested. Of the 4,099 variable sites with RRBS data, 1,677 (41%) showed a significant association (5% FDR) between methylation and occupancy (Figure 5-12). At significant sites, increased methylation was negatively associated with occupancy in 98% of cases. The magnitude of the association between methylation and occupancy was strong – occupancy was on average 87% lower at significant sites in the methylated cell types relative to the unmethylated cell types (Figure 5-13). Further supporting a strong link to methylation, 67% of variable methylation was associated with a concomitant affect on occupancy. The remaining 36% of sites with variable methylation that was not associated with occupancy nevertheless demonstrated an aggregate reduction in occupancy in methylated cell types (Figure 5-14), confirming the overall inverse association of methylation with CTCF occupancy, but suggesting that this relationship may be complicated by additional factors at this subset of sites.

Next we asked whether the inverse relationship between methylation and CTCF occupancy is characterized by regional hypermethylation or if instead methylation is concentrated specifically at the region of protein-DNA interaction. We examined the location of all CpG dinucleotides relative to the CTCF motif at sites with variable methylation. Indeed, sites of differential methylation associated with occupancy differences showed an enrichment of CpG dinucleotides at two positions in the CTCF recognition sequence (Figure 5-15). This finding is consistent with previous reports showing methylation outside the recognition sequence does not affect CTCF binding *in vitro* (Chadwick, 2008; Engel et al., 2004). Within the recognition sequence, methylation at one of these CpGs (position 1) has been shown to inhibit binding of CTCF *in vitro* (Renda et al., 2007). The second (position 11) is the predominant CpG in the motif, which has been shown to have a higher rate of C-T transitions at vertebrate-conserved binding sites (Kim et al., 2007), consistent with germline methylation. Interestingly, constitutively unmethylated CTCF sites also showed an enrichment of CpGs at these two positions compared to differentially methylated sites without an association to occupancy (Figure 5-16). Giv-

en that the latter sites nevertheless exhibit substantial methylation variability, this suggests that the absence of CpGs at these positions may decouple CTCF occupancy from differential methylation at these sites. Overall, 29% of CTCF recognition sequences genome-wide contain a CpG at positions 1 and/or 11, and 52% of recognition sequences contain a CpG anywhere in the sequence. The genome-wide prevalence of "susceptible" CTCF sites suggests a widespread potential for interaction between CTCF and methylation.

Methylated-associated remodeling of CTCF binding in immortal cell lines

Paralleling prior reports of widespread hypermethylation in cancer (Jones and Baylin, 2007; Varley et al., 2013), we observed a bimodal pattern of methylation at CTCF sites distinguishing normal and immortal cell types (Figure 5-17A). At 31% of the sites where differential methylation was associated with CTCF occupancy, methylation was observed throughout the thirteen normal and immortal cell types (average number of methylated cell types, 7.3). In contrast, the remaining 69% of sites were characterized by cell-specific hypermethylation constrained to the 6 immortal lines (average number of methylated cell lines, 2.1; Figure 5-17A, strip at right). Notably, although the neuroblastoma line SK-N-SH_RA clusters with epithelial cell types based purely on CTCF binding (Figure 5-7A), it exhibits the hypermethylation characteristic of the other immortal lines. Surprisingly, the increased methylation in immortal lines does not correspond to a decrease in the total number of bound CTCF sites (Figure 5-17B). Strikingly, we also observed that CTCF transcript levels are significantly higher in the immortal cell lines (Figure 5-17C). This disruption of CTCF binding in immortal cell lines is further distinguished by a unique association between CTCF occupancy and methylation at promoter sites. Of the promoter CTCF sites where methylation was significantly associated with occupancy, 98% (281 of 288) of these sites were characterized by hypermethylation in the immortal lines (Figure 5-17D). These results suggest a widespread methylation-associated remodeling of the CTCF binding landscape in immortal cell lines.

5.4 Discussion

Surprising plasticity of the CTCF occupancy landscape

This study exposes a previously unappreciated degree of plasticity within the binding landscape of the master genomic regulator CTCF. Previous studies in a small number of cell types had uncovered only limited cell-type specificity (Cuddapah et al., 2008; Heintzman et al., 2009; Kim et al., 2007). We further associate differential methylation with 41% of this variable binding at a subset of sites overlapping existing RRBS data in 13 cell types. We specifically linked this variable methylation to the presence of a CpG at two key positions relative to the consensus motif. Finally, we observe the maintenance of a stable total amount of CTCF genomic binding sites in immortal cell lines despite their altered localization associated with increased methylation. Our results show that methylation is indeed a global feature of the regulatory diversity of CTCF, and our approach is readily extensible to the repertoire of vertebrate transcription factors.

Methylation-associated disruption of CTCF binding in immortal lines

Although CTCF binding varied across all 13 cell types, we observed unique patterns of CTCF occupancy specific to the immortal cell lines. Interestingly, CTCF overexpression has previously been associated with resistance to apoptosis in breast cancer cell lines (Docquier et al., 2005) and with DNMT3B overexpression (Butcher et al., 2004). Further, the unique occurrence of hypermethylation-associated abrogation of CTCF occupancy at promoters in immortal lines is notable, given the involvement of CTCF in the methylation-associated silencing of known tumor suppressors and oncogenes (Lai et al., 2010; Soto-Reyes and Recillas-Targa, 2010; Witcher and Emerson, 2009). We found that the immortal cell lines we profiled have the same overall amount of genomic CTCF binding sites despite a redistribution of CTCF occupancy from binding sites subject to hypermethylation. The concomitant up-regulation of CTCF expression may therefore represent a cancer-associated compensatory mechanism. This inverse correlation is compatible with the existence of a stabilizing mechanism acting through increased CTCF expression to maintain a constant level of genomic binding despite increased methylation at its target sites, although further study in an expanded set of cell types will be necessary.

Although DNA methylation is widely invoked as a causal mechanism for transcriptional repression, surprisingly little *in vivo* evidence is available. While experimentally-directed methylation can prevent binding of CTCF and other factors *in vitro* (Renda et al., 2007; Tate and Bird, 1993), the mechanisms establishing methylation patterns *in vivo* remain unknown, and its precise relationship with gene expression remains unclear (Enver et al., 1988; Selker, 1990; Walsh and Bestor, 1999). Likewise, our results do not distinguish whether demethylation facilitates subsequent CTCF binding, or whether bound CTCF maintains an unmethylated domain.

An alternative model has DNA methylation deposited passively in the wake of independent abrogation of transcription factor binding. This model is equally consistent with evidence that transcription factor binding sites appear to be generally depleted for DNA methylation (Lister et al., 2009; Mukhopadhyay et al., 2004; Thurman et al., 2012), and that binding sites recognized by certain sequence-specific factors have been associated with lack of methylation (Dickson et al., 2010; Gebhard et al., 2010; Lienert et al., 2011; Straussman et al., 2009). Indeed, there is evidence that the binding of some transcription factors, including CTCF, is sufficient to effect a local demethylated state (Lin et al., 2000; Matsuo et al., 1998; Stadler et al., 2011). But if *in vivo* methylation were deposited generally at unoccupied binding sites, then how would this process interact with the *in vitro* methylation sensitivity of common transcription factors?

The well-investigated *H19/Igf2* imprinted locus offers an appropriate example: CTCF binding there has been shown necessary to maintain an existing unmethylated state (Pant et al., 2004; Schoenherr et al., 2002). However, CTCF is not the originator of the unmethylated state (Matsuzaki et al., 2010), implying a limited capacity to directly affect methylation. Perhaps methylation instead acts as a cooperative switch to prevent the return of CTCF after a reprogramming event. In this model, rather than guiding binding localization, methylation is a general amplifier of perturbations to transcription factor occupancy.

Other sources of variable CTCF binding

Although we have shown that 41% of overall CTCF occupancy variation is significantly linked to methylation at tested sites, 36% of variable CTCF sites overlap no variable methylation at all. It is unlikely that much of this variability is associated with genetic variability in CTCF recognition sequences (Section 4), though some sites may associate with modified forms of CTCF (Klenova et al., 2001; MacPherson et al., 2009; Yu et al., 2004). One likely possibility is that the constantly-unmethylated variable CTCF sites may represent instances of cooperative regulation that complicate a direct relationship between methylation and CTCF occupancy. Accordingly, CTCF has been known to interact with a number of cofactors which could potentially govern its selectivity at these sites or, alternatively, maintain demethylation in the absence of CTCF binding (Chernukhin et al., 2007; Donohoe et al., 2009; 2007; Liu et al., 2011; Ohlsson et al., 2010; Parelho et al., 2008; Rubio et al., 2008; Wendt et al., 2008). Interestingly, we found that of the 36% of variable sites despite constant methylation, 76% were within 2.5 kb of a RefSeq transcription start site, compared to 38% of the variable sites associated with methylation differences. Recent work has further observed an enrichment of tethered CTCF peaks at promoters (Neph et al., 2012a) suggesting that the remaining variation in CTCF occupancy may derive from complex regulation of co-factors or variation in its specific interaction partners. Given the breadth of CTCF's regulatory functionality, our observation of global binding variation implies a widespread potential role in the translation of epigenetic marks to genome organization at thousands of sites.

5.5 Figures and Tables

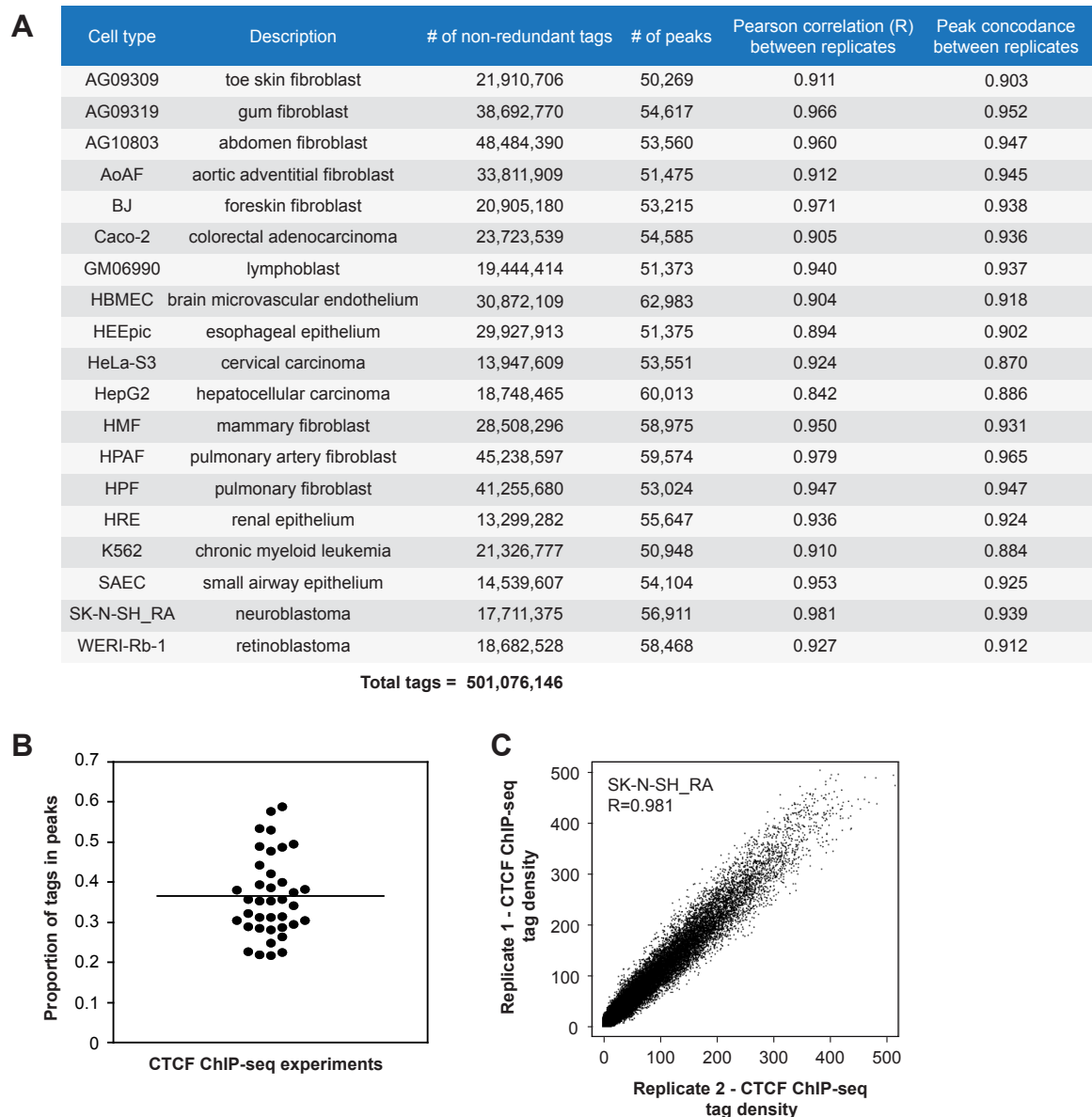


Figure 5-1 Summary of CTCF ChIP-seq experiments for 19 surveyed cell types.

(A) Number of tags represents the total uniquely-mapped reads (excluding PCR duplicates) for both biological replicates of each cell type. Number of peaks refers to the number of CTCF binding sites considered present in a given cell type (see Section 5.5). Pearson correlation coefficient (R) measures concordance between replicate density tracks on chromosome 19 (average=0.93). Peak concordance between replicates refers to the average of the proportion of peaks in one replicate that are found in the second replicate and vice versa.

(B) SPOT scores (Signal Portion of Tags, representing overall enrichment) for 19 ChIP-seq experiments, showing degree of enrichment. Y-axis indicates proportion of all sequenced tags mapping to hotspot regions.

(C) Example of Pearson correlation R for two K562 replicates. X-axis, Y-axis, replicate 1 and 2 tag densities on chr19.

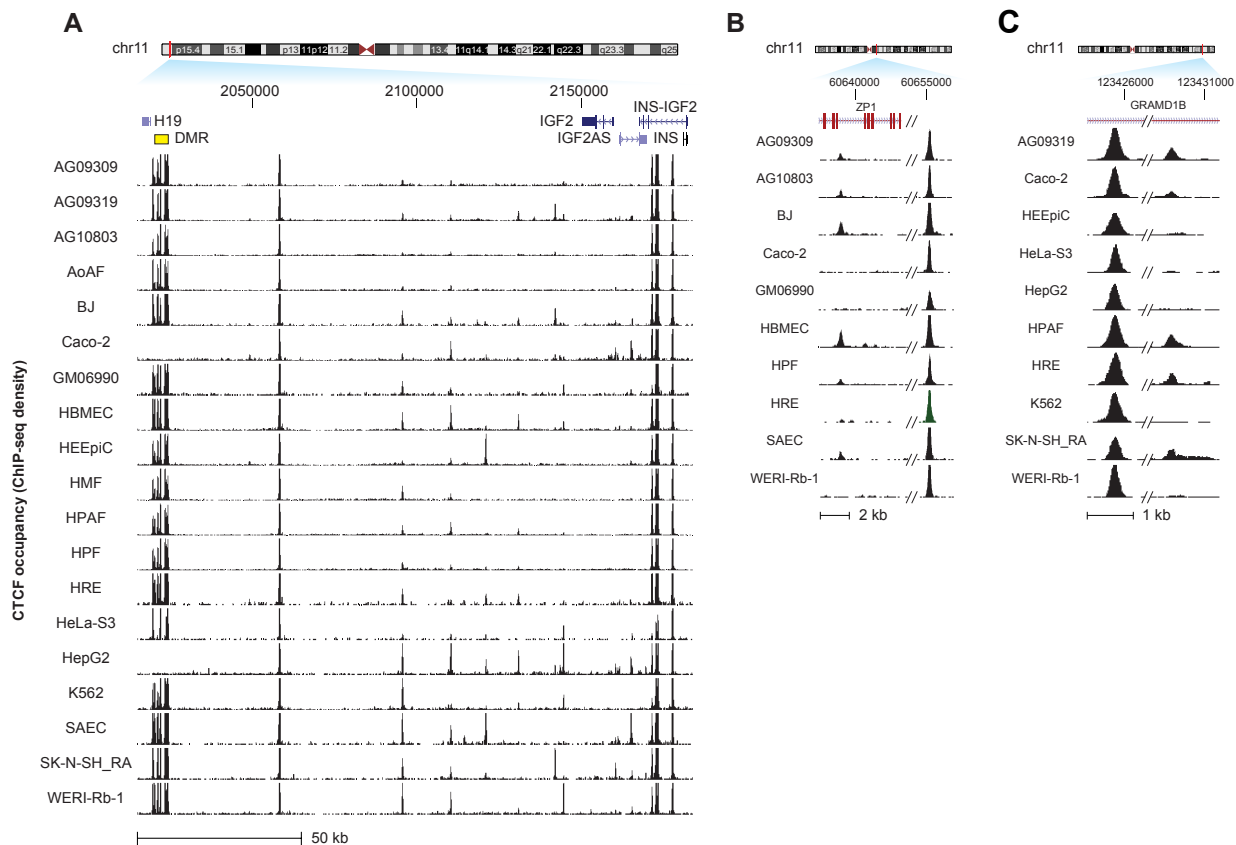


Figure 5-2 CTCF *in vivo* binding exhibits widespread plasticity.

Constitutive and variable CTCF sites

(A) The *H19/Igf2* imprinted locus in multiple human cell types. Note the total silencing in two cell lines of the seven CTCF sites in the differentially methylated region (DMR, yellow box at left), and the complex pattern of cell-selective CTCF binding flanked by constitutive sites. Location (hg19), chr11:2,015,000-2,184,000

(B-C) additional sites

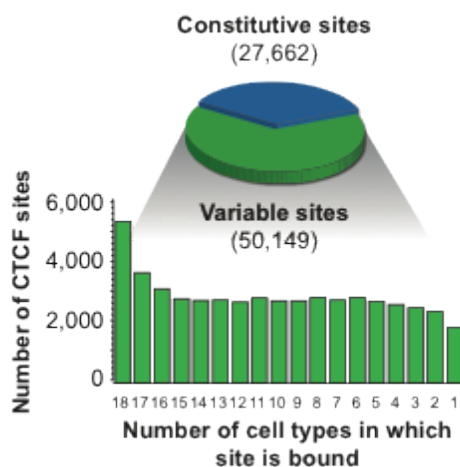


Figure 5-3 CTCF binding variability genome-wide

Genome-wide analysis of CTCF binding in 19 cell types reveals 77,811 distinct binding sites. 27,662 sites are constitutively present in all cell types. 50,149 variable sites exhibiting a wide range of selectivity are present in a subset of 1 to 18 cell types (below).

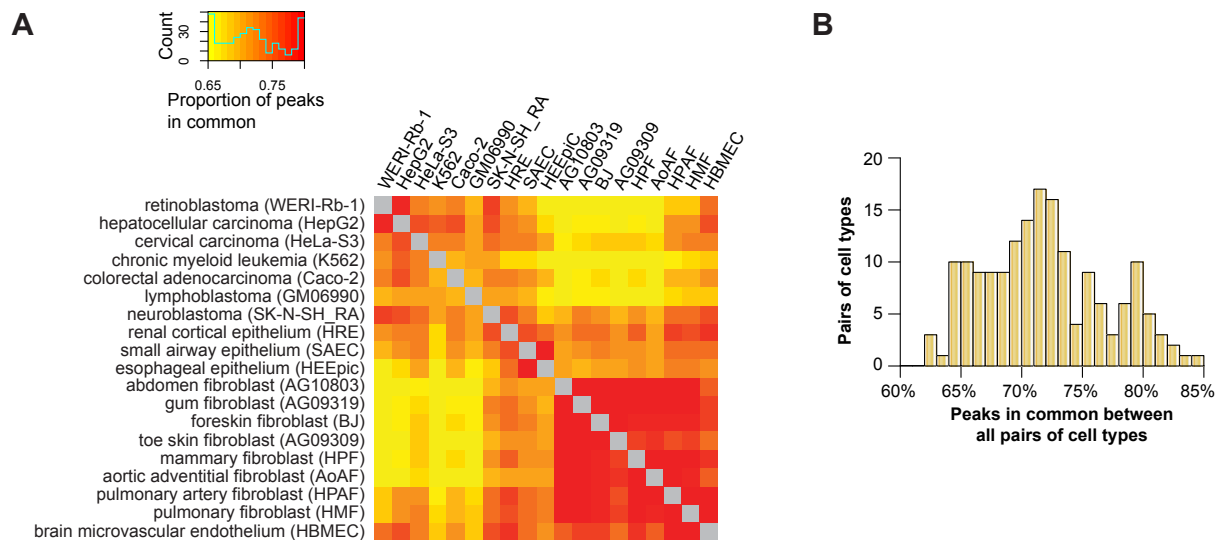


Figure 5-4 Degree of overlapping binding landscape.

(A) Overlap between pairs of cell types, in terms of proportion of peaks in common.

(B) Percentage of binding sites shared between all possible pairs of individual replicate samples.

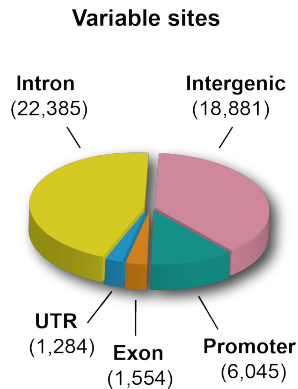


Figure 5-5 Location of variable CTCF sites relative to genes
Genomic distribution of variable sites is similar to constitutive sites (Figure 5-6).

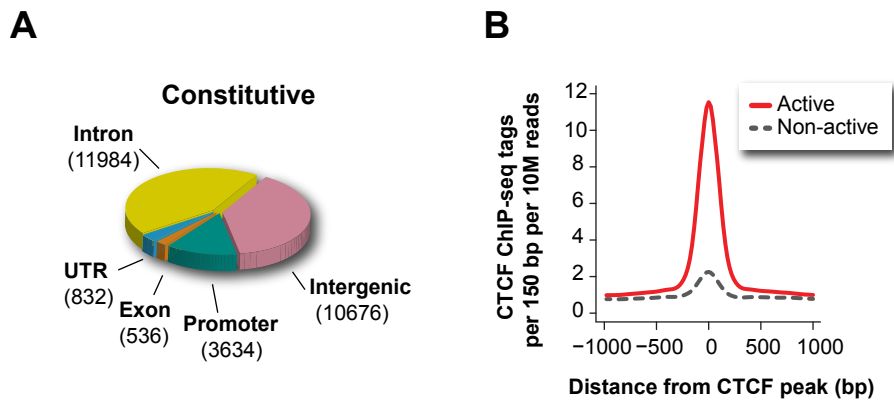


Figure 5-6 CTCF *in vivo* occupancy exhibits marked plasticity.
(A) Genomic distribution of constitutive sites (Gencode V7 annotations). Promoter, 1 kb region upstream of transcription start site.
(B) Aggregate normalized CTCF signal in active (red) and non-active (grey) cell types. Vertical axis, normalized mean ChIP-seq tag density. Horizontal axis, distance from peak center.

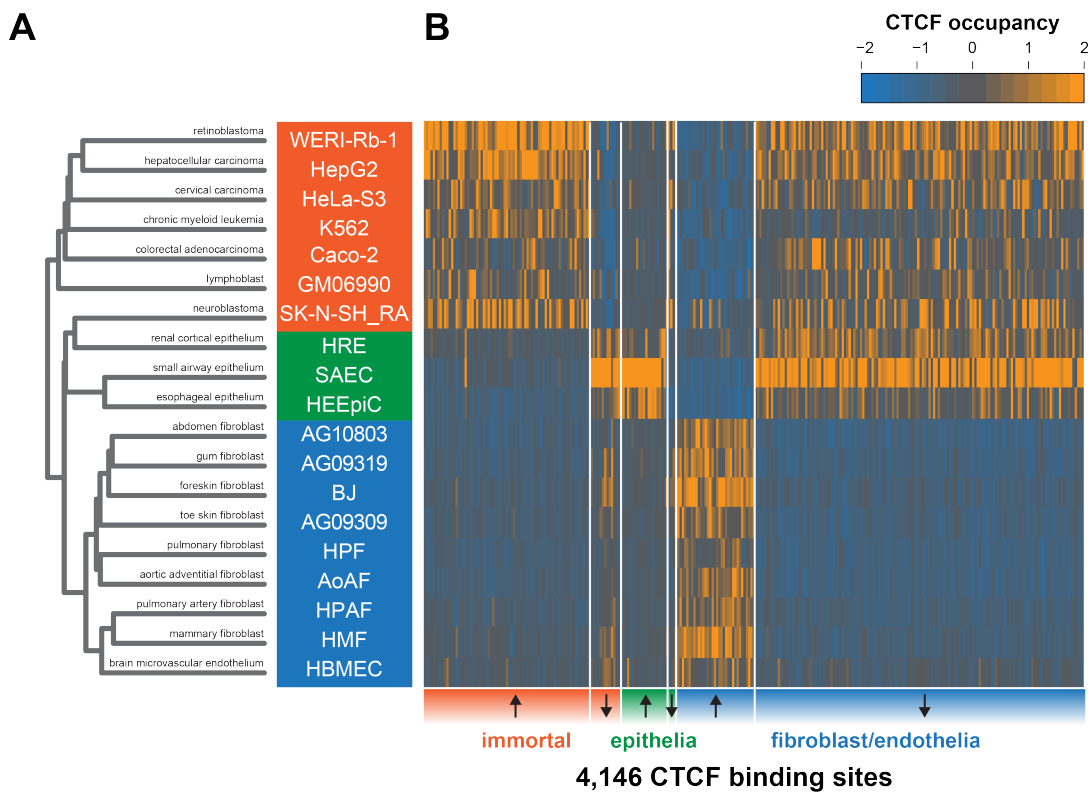


Figure 5-7 CTCF occupancy distinguishes similar cell types.

(A) Unsupervised hierarchical clustering of binding at all CTCF sites.

(B) CTCF occupancy at 4,146 variable binding sites that distinguish immortal cell lines, epithelia, fibroblasts and endothelia (Section 5.5). X-axis, CTCF binding sites in chromosomal order, separated into sites which are up-regulated and down-regulated (arrows) in each of the three groups (immortal, epithelial, fibroblast and endothelial). Color corresponds to z-score of normalized ChIP-seq density.

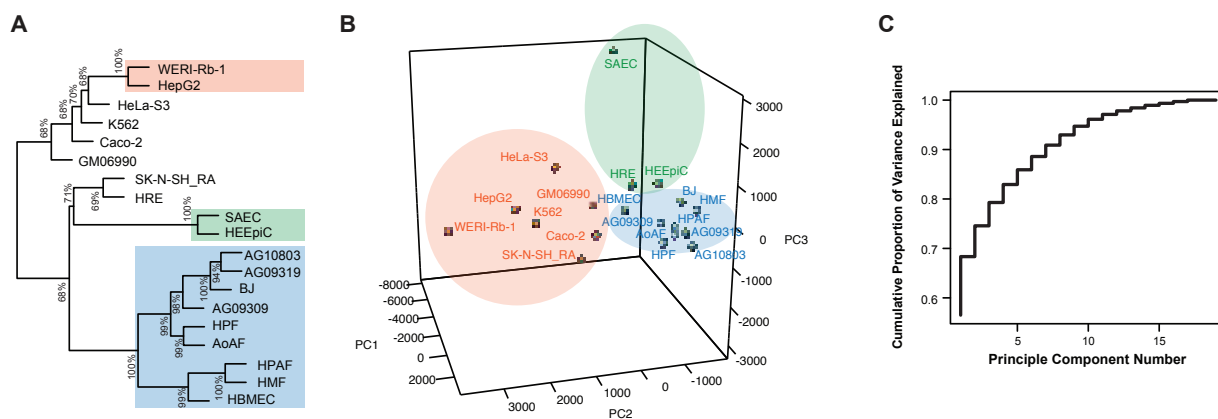


Figure 5-8 CTCF occupancy landscape distinguishes similar cell types.

(A) Bootstrap confidence values of hierarchical clustering analysis of CTCF binding sites. Percentages indicate approximately unbiased (AU) p-values; boxes indicate significant clusters.

(B) Principle components analysis separation of cell types based on CTCF occupancy.

(C) Cumulative proportion of variance explained by each principle component.

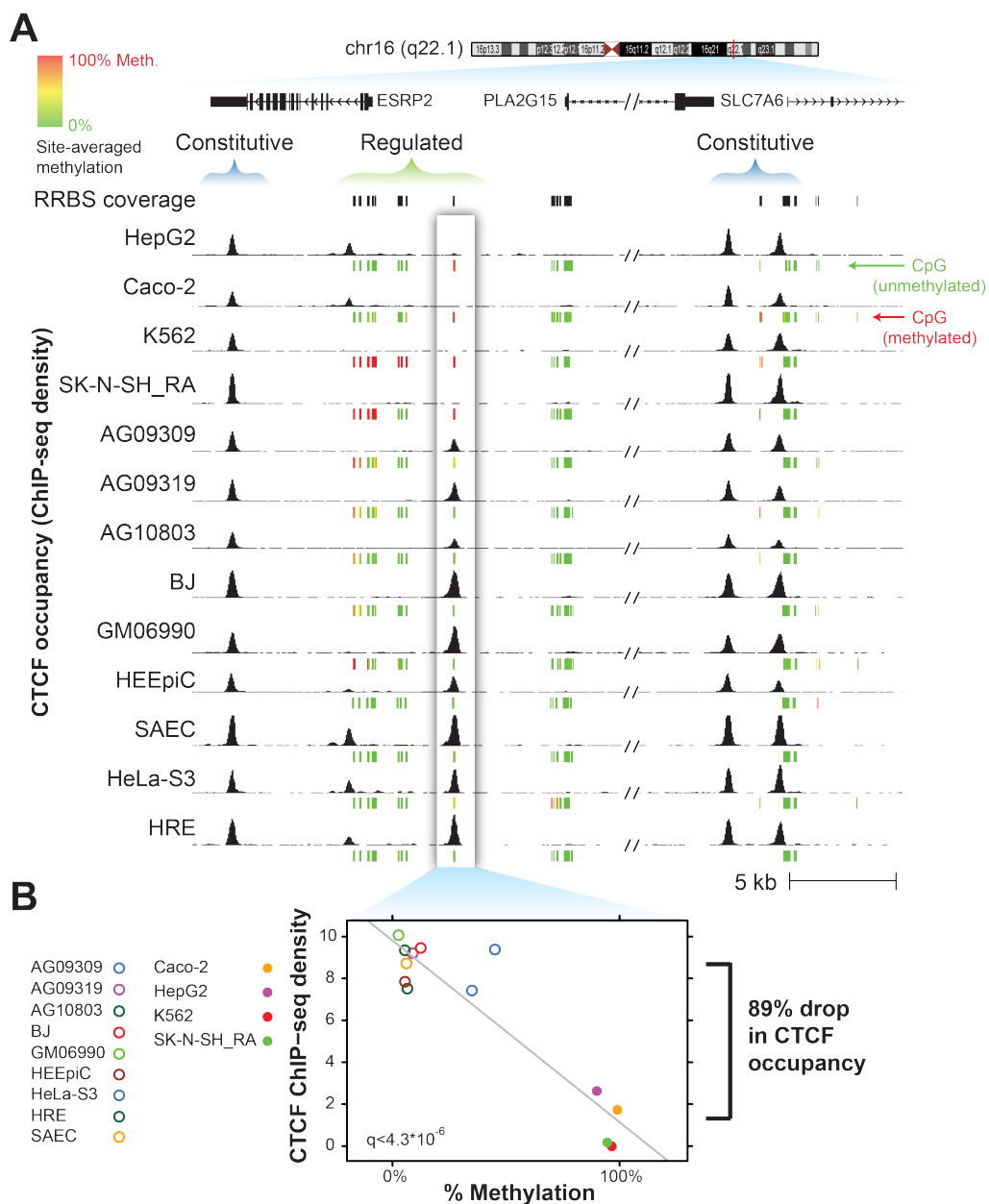


Figure 5-9 Impact of DNA methylation on cell-selective CTCF binding.

(A) Example CTCF binding sites, where occupancy (above) quantitatively increases as local CpG methylation decreases (below). Green, CpG is 0% methylated; yellow, 50%; red, 100%.

(B) Quantitative analysis of methylation at the boxed CTCF binding site in (A).

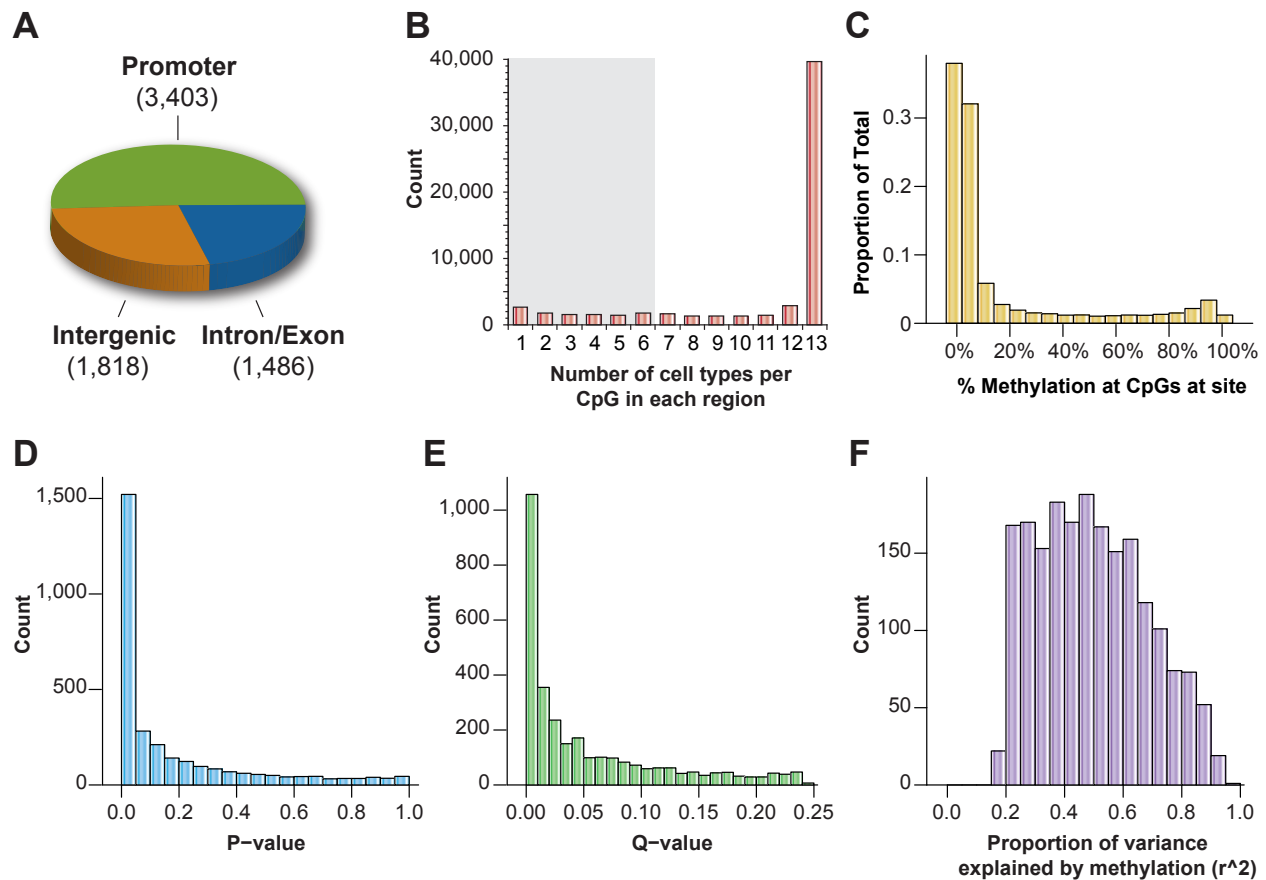


Figure 5-10 Statistical association of variable methylation with differential occupancy.

(A) Location relative to genes of sites surveyed by reduced representation bisulfite sequencing, which enriches for genic and CpG-island regions, including promoters.

(B) Most CpGs were surveyed in all 13 samples. Sites without with at least one surveyed CpG for at least 6 samples (gray shading) were excluded from association analysis.

(C) Methylation levels (X-axis) observed across all CTCF sites in all cell types; most CTCF sites are unmethylated.

(D-E) Association of methylation with ChIP-seq occupancy identifies 905, 1,677 and 2,046 significant sites at FDR levels 1%, 5% and 10%, respectively. Histogram of p-values (D) and FDR-adjusted q-values (E) for all tested binding sites.

(F) Effect size of methylation differences associated with differences in occupancy (FDR 5%), measured by r^2 of a linear regression.

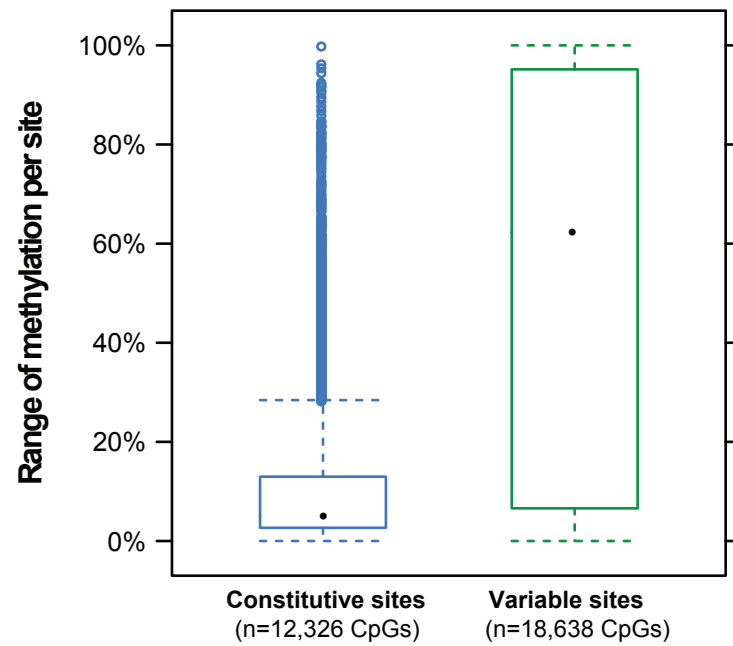


Figure 5-11 Overall methylation variability across constitutive and variable CTCF sites.

Y-axis denotes the range between the most- and least-methylated cell types at each CpG overlapping constitutive and variable CTCF sites.

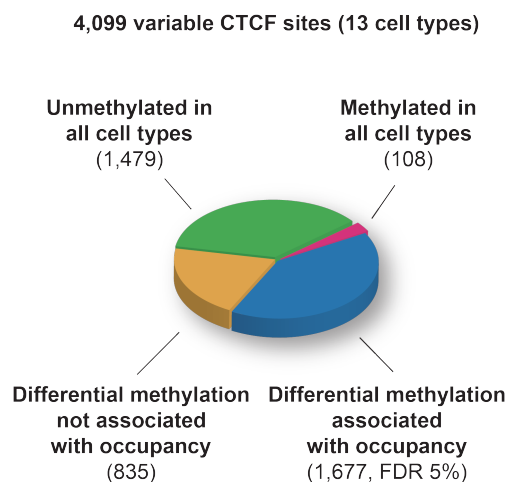


Figure 5-12 Global impact of methylation at variable CTCF sites monitored by RRBS.
65% of sites with cell-type selective patterns of methylation also exhibited differences in occupancy.

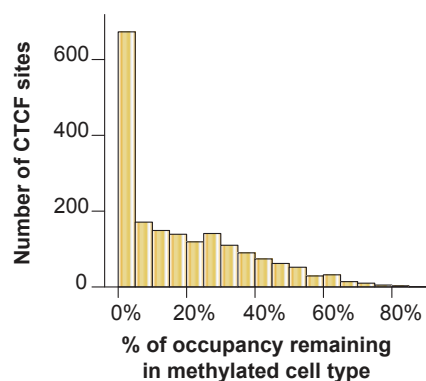


Figure 5-13 Reduction in occupancy at methylated sites

At methylated binding sites, occupancy was reduced on average by 87% when compared to cell lines without methylation at the same site. Shown are sites where increased methylation was associated with decreased occupancy (98% of all significant sites).

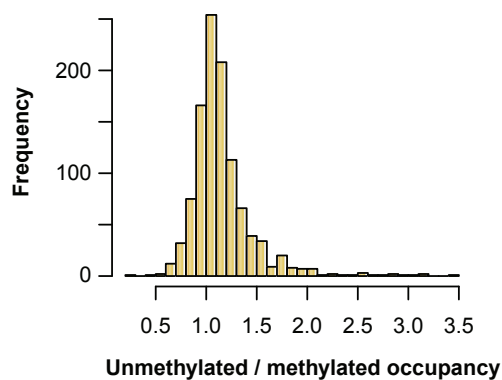


Figure 5-14 Overall relationship of CTCF occupancy with methylation at 1,076 sites without a significant association.

X-axis denotes ratio of average occupancy in unmethylated cell types divided by the average in methylated cell types.

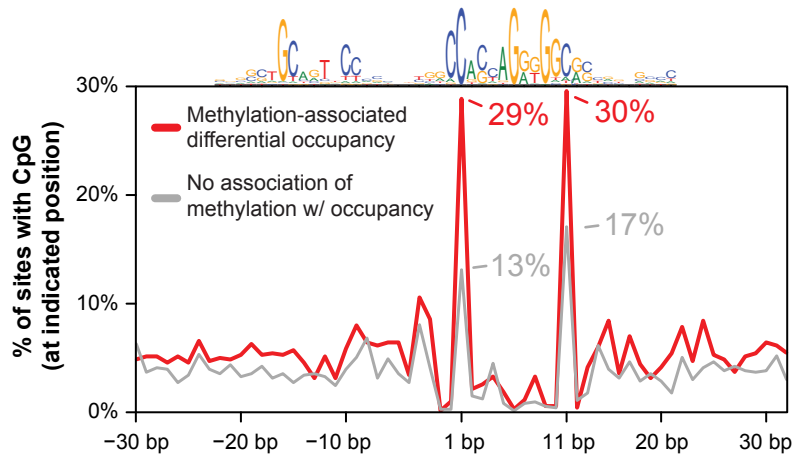


Figure 5-15 Sites significantly affected by methylation are enriched for CpGs at two positions.

Frequency of a CpG (Y-axis) at positions relative to the CTCF motif (X-axis) shown for sites with variable methylation that is associated (red) and is not associated (gray) with occupancy changes. Note that at positions 1 and 11, there is a 2.2- and 1.8-fold enrichment, respectively, for the presence of a CpG at sites where the variable methylation was not associated with occupancy. 29% of CTCF motifs genome-wide contain a CpG at one or both of these positions.

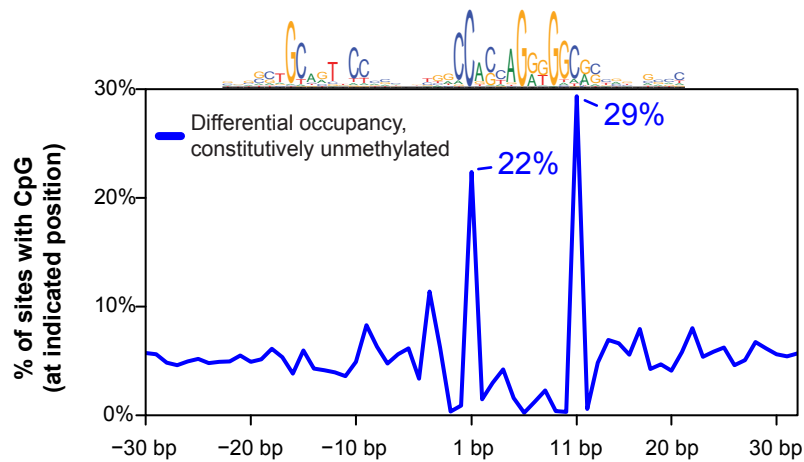


Figure 5-16 Variable CTCF sites without methylation differences.

Frequency of a CpG (Y-axis) at positions relative to the CTCF motif (X-axis) shown for sites with differential binding but no differential methylation (blue). Note that the presence of a CpG at positions 1 and 11 is similar to that at sites where the variable methylation was associated with occupancy (Figure 5-15, red), but sites with variable methylation that is not associated with occupancy (Figure 5-15, gray) are depleted for these CpGs.

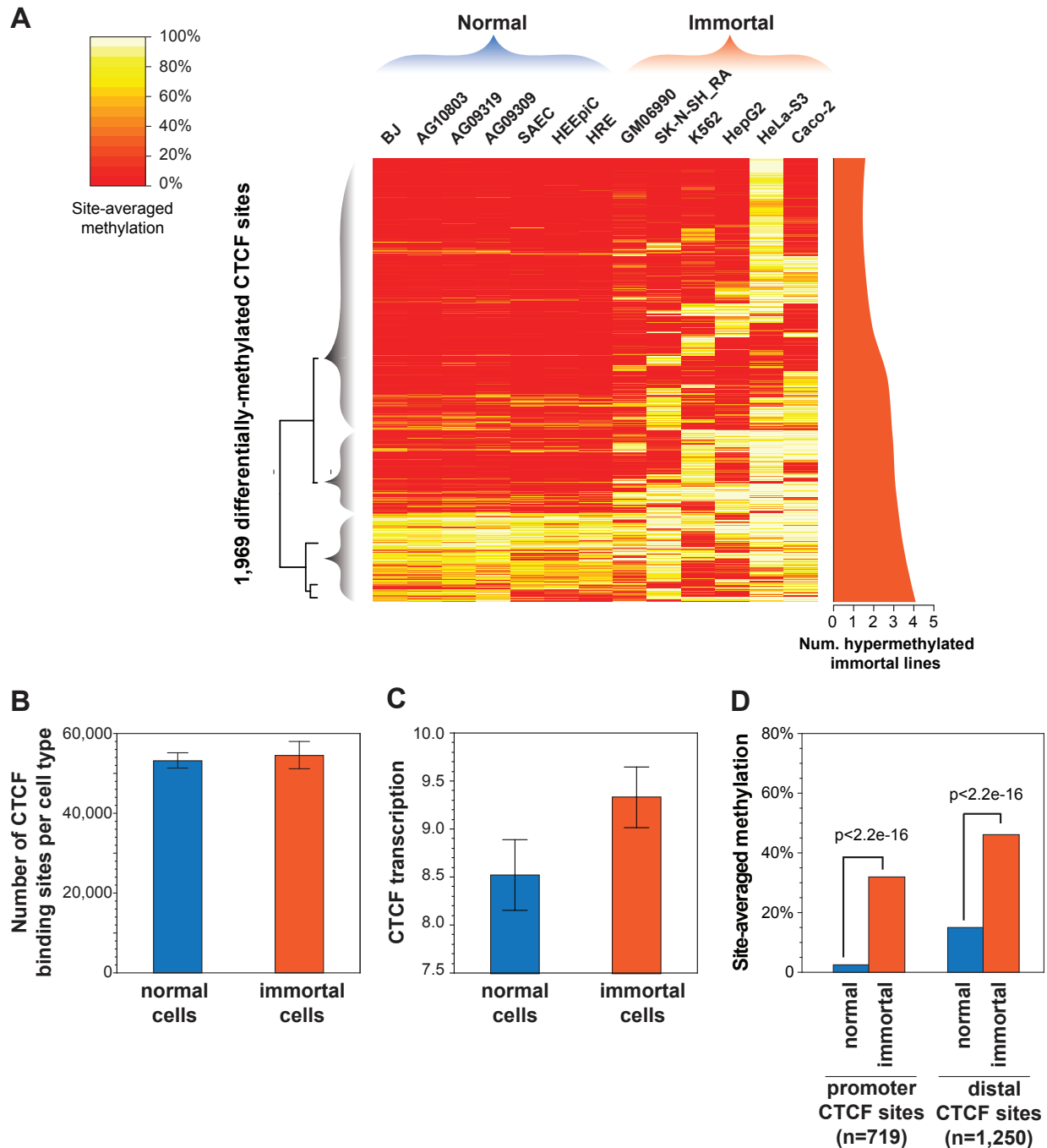


Figure 5-17 Cell-selective patterns of methylation associated with occupancy differences.

(A) Methylation status at 1,969 CTCF sites where differential methylation is significantly associated with occupancy differences. Color corresponds to the percentage of bisulfite sequencing tags at each site overlapping methylated CpG positions. Dendrogram (left) highlights pattern of hypermethylation in immortal cell lines. Right, smoothed plot of number of immortal lines exhibiting hypermethylation at each site.

(B) Immortal lines show no significant difference in number of occupied CTCF sites (Y-axis, mean). Error bars, standard deviation.

(C) immortal lines demonstrate increased CTCF transcript levels (Y-axis, mean). Error bars, standard deviation.

(D) Immortal lines exhibit increased methylation relative to the other cell types, though significant promoter methylation is rarely observed in normal lines. Y-axis, genome-wide median of per-site methylation. P-values, Wilcoxon. Promoter, ± 2.5 kb of RefSeq transcription start site.

Cell	Description	Medium	Supplements		
			%FBS	NEAA ^A	NaPyruvate ^B
AG09309	adult toe fibroblast	MEM ^C	15%	✓	
AG09319	adult gum fibroblast	MEM ^C	15%	✓	
AG10803	adult abdomen fibroblast	MEM ^C	15%	✓	
AoAF	human aortic adventitial fibroblast	Lonza SCBM basal medium+SCGM SingleQuot Kit ^D			
BJ	skin fibroblast	MEM ^C	10%	✓	✓
Caco-2	colorectal adenocarcinoma	MEM ^C	20%	✓	✓
GM06990	lymphoblast	RPMI ^E	15%		
HBMEC	human brain microvascular endothelium	Endothelial cell medium (ECM) ^F			
HEEpiC	human esophageal epithelium	Epithelial cell medium-2 (EpiCM2) ^F			
HMF	human mammary fibroblast	Fibroblast medium (FM) ^F			
HPAF	human pulmonary artery fibroblast	Fibroblast medium (FM) ^F			
HPF	human pulmonary fibroblast	Fibroblast medium (FM) ^F			
HRE	human renal epithelium	REBM Basal Medium+REGM SingleQuot Kit ^D			
HeLa-S3	cervical adenocarcinoma	DMEM ^G	10%		
HepG2	hepatocellular carcinoma	MEM ^C	10%	✓	✓
K562	chronic myeloid leukemia	IMDM ^H	10%		
SAEC	human small airway epithelium	Lonza SABM Basal Medium+SAGM SingleQuot Kit ^D			
SK-N-SH_RA ^I	neuroblastoma	RPMI ^E	10%		✓
WERI-Rb-1	retinoblastoma	RPMI ^E	10%		

Table 5-1 Growth conditions for immortal cell lines and primary cells

All cells are supplemented with 25 IU/L penicillin and 25 mg/L streptomycin (Invitrogen)

Supplements: A: **NEAA**=0.1 mM non-essential amino acids (Invitrogen); B: **NaPyruvate**=1 mM sodium pyruvate (Invitrogen)

Media: C: **MEM**=Eagle's minimum essential medium with Earle's salts and 2 mM L-Glutamine (Cellgro); D: Lonza medium and supplements; E: **RPMI**= RPMI 1640 Medium with 2 mM L-Glutamine (Cellgro); F: ScienCell medium; G: **DMEM**=Dulbecco's Modification of Eagle's Medium with 4.5 g/L Glucose, 110 mg/L Sodium Pyruvate and 4mM L-Glutamine (Cellgro); H: **IMDM**=Iscove's modification of DMEM with L-glutamine and 25 mM HEPES, without α -thioglycerol and β -mercaptoethanol (Cellgro)

I: At 60% confluency, treat with 6 μ M retinoic acid for 48 hours to differentiate

5.6 Methods

Cell culture

Cells were cultured in an appropriate growth medium, with the addition of growth factors and supplements according to suppliers' instruction (Table 5-1). Cell lines were maintained in humidified incubator at 37 °C in the presence of 5% CO₂.

ChIP-seq

Suspension cells were crosslinked with formaldehyde (Sigma) at a final concentration of 1% for 10 min at room temperature. Adherent cells were first detached from the plates by 0.05% Trypsin-EDTA and Trypsin neutralizer solution (Invitrogen), and then crosslinked by 1% formaldehyde. Glycine was added to a final concentration of 0.125 M for 5 min. Cells were rinsed twice with phosphate buffered saline, lysed in lysis buffer (50 mM Tris-HCl pH8.0, 10 mM EDTA, 1% SDS) containing protease inhibitor cocktail (Roche), and sheared by Bioruptor (Diagenode). The chromatin was incubated with Dynabeads (M-280, sheep anti-rabbit IgG, Invitrogen)-conjugated anti-CTCF polyclonal antibody (Cell Signaling #2899).

The CTCF-DNA complexes were washed, eluted, and reverse-crosslinked. The DNA was RNase A-, Proteinase K-treated, and purified by phenol-chloroform-isoamyl alcohol extraction and ethanol precipitation. DNA was end-repaired (End-it DNA End-repair kit, Epicentre), followed by the addition of adenine to the 3' ends (Taq DNA polymerase, NEB), and ligated to an adapter (Illumina). Purified ligation product was PCR amplified and run on a 2% agarose gel. The size-selected libraries were sequenced on an Illumina Genome Analyzer (Illumina) by the High-Throughput Genomics Center (University of Washington) according to a standard protocol.

For each cell type, experiments were conducted on 2 independent biological replicates.

Identification and quantification of CTCF binding sites

We obtained Uniform Element Calls from the ENCODE project for each cell line. Briefly, peaks were called using SPP (Kharchenko et al., 2008). The set of peaks reproducible in both replicates were identified based on an irreproducible discovery rate (IDR) of 0.25% (Li et al., 2011). We then combined peak calls from 19

cell types to generate a master list of all distinct CTCF binding sites. We adjusted the peak locations to center on matches to the nearest CTCF motif ($P < 10^{-5}$, fimo) if the motif was within 50 bp.

To distinguish between variable and constitutive binding sites, for each site we examined the presence of a peak in each of 19 cell types. We used the peak calling program Hotspot (John et al., 2011) to enable a conservative procedure for the identification of variable binding sites. To reduce the misclassification of sites near the peak-calling threshold as variable, we employed separate cutoffs for calling peak presence and absence. First, for each CTCF binding site called above, we additionally required that it overlap a 0.5% FDR hotspot in both replicates of at least one cell line. Then, a binding site was counted as occupied in subsequent cell lines if a looser 1% FDR hotspot was present in one or both replicates for that cell line. Employing this looser criteria for binding in subsequent cell types results in conservative identification of variable sites. We confirmed that binding sites in cell types considered absent were substantially closer to background than sites in cell types considered active (Figure 5-6B).

ChIP-seq data were mapped to the human genome (GRCh37/hg19) using bowtie (Langmead et al., 2009) with the options "bowtie --mm -n 3 -v 3 -k 2 --phred64-quals", allowing up to 3 mismatches. Reads mapping to multiple locations were then excluded, and reads with identical 5' ends and strand were presumed to be PCR duplicates and were excluded. Smoothed density tracks were generated using bedmap (<http://code.google.com/p/bedops/>) to count the number of tags overlapping a sliding 150 bp window, with a step width of 20 bp (Neph et al., 2012a). Density tracks were normalized for sequencing depth by a global linear scaling to 10 million tags. We measured occupancy by the maximum normalized ChIP-seq tag density over the 134 bp region.

Reproducibility of ChIP-seq experiments was tested using Pearson correlation on normalized density tracks of chromosome 19 between each replicate.

Clustering of cell-selective CTCF binding sites

We converted the presence and absence of a given peak to 1 and 0 respectively in 19 cell lines. We then performed hierarchical clustering with the hclust function in R, using the "average" method and Euclidean dis-

tance metric. We cut the dendrogram (Figure 5-7A) into 3 groups, of immortal cell lines, epithelia, and fibroblasts. To assess the significance of these three groups, we used the R package pvclust (Suzuki and Shimodaira, 2006) and principal components analysis (Figure 5-8). We then used the package DEseq (Anders and Huber, 2010) on the tag count at each peak to identify differentially occupied sites between each of these three groups (FDR 1%).

RRBS genome-wide methylation profiling

We downloaded RRBS methylation data for 13 cell lines from the “HAIB Methyl RRBS” track (Varley et al., 2013) of the UCSC Genome Browser. To measure methylation in each cell line, we combined counts for both strands in both replicates and removed data for samples with <8x coverage. We retained only CpGs monitored in at least 6 samples (Figure 5-10B)

We applied a linear regression to measure whether methylation status is associated with occupancy. We normalized CTCF occupancies using the `getVarianceStabilizedData` function of DEseq, then averaged replicate signals. We regressed CTCF occupancy onto the average proportion methylated of all monitored CpGs in a 134 bp region centered around the CTCF peak. We excluded 1,806 sites missing RRBS data and ChIP-seq data for 7 or more cell types, or having too great a difference in the number of CpGs monitored between any two cell types (>6 CpGs monitored). We averaged the methylation level of all CpGs within a 134 bp window to increase sensitivity and reliability. We excluded sites where the number of monitored CpGs differed by more than 4 among any two cell lines. We used the R package qvalue to estimate an FDR (Storey and Tibshirani, 2003).

RNA expression analysis

For each cell lines, total RNA was extracted in 2 replicates from 5×10^6 cells using Ribopure (Ambion) according to manufacturer’s instructions. RNA quality was ascertained using RNA 6000 Nano Chips on a bioanalyzer (Agilent, Santa Clara, CA). Approximately 3 μ g of total RNA for each sample was used for labeling and hybridization (University of Washington Center for Array Technology) to Affymetrix Human Exon 1.0 ST arrays (Affymetrix) using a standard protocol. Exon expression data were analyzed through Affymetrix Ex-

pression Console using gene-level RMA summarization and sketch-quantile normalization method. Measurements from both replicates were then averaged.

5.7 Notes

This chapter has been published as Wang H*, **Maurano MT***, Qu H*, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R, Thurman RE, Kaul R, Myers RM, and Stamatoyannopoulos JA. "Wide-spread Plasticity in CTCF Occupancy Linked to DNA Methylation." **Genome Research** 22: 1680–1688 (2012). * *co-first authors*

Data Access

CTCF ChIP-seq data are available from the GEO repository (<http://www.ncbi.nlm.nih.gov/geo/>) under the accession GSE30263. Affymetrix exon array data are available under the accession GSE19090. RRBS methylation data are under the accession GSE27584. All three sets are available for viewing in the UCSC Genome Browser (<http://genome.ucsc.edu/>).

Acknowledgments

We thank Jeff Vierstra, Andrew Stergachis, and Sam John for critical reading of the manuscript and many helpful suggestions. We also thank Daniel Bates, Morgan Diegel, and Doug Dunn at the University of Washington High-Throughput Genomics Center for technical assistance. This work was supported by National Institutes of Health grants U54HG004592 (J.A.S.) and U54HG004576 (R.M.M.).

Authors' contributions: HW MTM JAS conceived the study. HW TC cultured cells. HW KL produced ChIP-seq data. HW MW generated Illumina libraries. KEV JG FP generated RRBS data under supervision of RMM. MTM RS RET processed data. MTM HQ analyzed data. HW MTM wrote the manuscript, with contributions from JAS.

6 Chromatin Accessibility And DNA Methylation Patterns

6.1 Results

CpG methylation has been closely linked with gene regulation, based chiefly on its association with transcriptional silencing (Siegfried et al., 1999). However, the relationship between DNA methylation and chromatin structure has not been clearly defined. We analyzed ENCODE reduced-representation bisulphite sequencing (RRBS) data, which provide quantitative methylation measurements for several million CpGs (Varley et al., 2013). We focused on 243,037 CpGs falling within DHSs in 19 cell types for which both data types were available from the same sample. We observed two broad classes of sites: those with a strong inverse correlation across cell types between DNA methylation and chromatin accessibility (Figure 6-1 and Figure 6-2), and those with variable chromatin accessibility but constitutive hypomethylation (Figure 6-1, right). To quantify these trends globally, we performed a linear regression analysis between chromatin accessibility and DNA methylation at the 34,376 CpG-containing DHSs (see Section 6.3). Of these sites, 6,987 (20%) showed a significant association (1% FDR) between methylation and accessibility (Figure 6-3A). Increased methylation was almost uniformly negatively associated with chromatin accessibility (>97% of cases). The magnitude of the association between methylation and accessibility was strong, with the latter on average 95% lower in cell types with coinciding methylation versus cell types lacking coinciding methylation (Figure 6-3B). Fully 40% of variable methylation was associated with a concomitant effect on accessibility.

Sites of protein–DNA interactions are specifically sensitive to cytosine methylation (Lister et al., 2009; Tate and Bird, 1993). To examine whether the depletion of methylation was specifically localized to the nucleotides responsible for transcription factor binding genome-wide, we examined genomic DNase I footprinting data and whole-genome bisulphite sequencing methylation data from pulmonary fibroblasts (IMR90). We found that CpG dinucleotides contained within DNase I footprints were significantly less methylated than CpGs in non-footprinted regions of the same DHS (Mann–Whitney U-test; P , 2.2×10^{-16} ; Figure 6-4). Footprints therefore seem to be selectively sheltered from DNA methylation, indicating a widespread connection between regulatory factor occupancy and nucleotide-level patterning of epigenetic modifications.

The role of DNA methylation in causation of gene silencing is presently unclear. Does methylation reduce chromatin accessibility by evicting transcription factors? Or does DNA methylation passively ‘fill in’ the voids left by vacating transcription factors? Transcription factor expression is closely linked with the occupancy of its binding sites (O’Geen et al., 2007). If the former of the two above hypotheses is correct, methylation of individual binding site sequences should be independent of transcription factor gene expression. If the latter, methylation at transcription factor recognition sequences should be inversely correlated with transcription factor abundance (Figure 6-5A).

Comparing transcription factor transcript levels to average methylation at cognate recognition sites within DHSs revealed significant negative correlations between transcription factor expression and binding site methylation for most (70%) transcription factors with a significant association ($P < 0.05$). Representative examples are shown in Figure 6-5B-E. These data argue strongly that methylation patterning paralleling cell-selective chromatin accessibility results from passive deposition after the vacation of transcription factors from regulatory DNA, confirming and extending other recent reports (Stadler et al., 2011).

Interestingly, a small number of factors showed positive correlations between expression and binding site methylation (Figure 6-5E), including MYB and LUN-1 (also known as TOPORS). Both of these transcription factors showed increased transcription and binding site methylation specifically within acute promyelocytic leukaemia (PML) cells (NB4), and both interact with PML bodies, a sub-nuclear structure disrupted in PML cells. The anomalous behavior of these two transcription factors with respect to chromatin structure and DNA methylation may thus be related to a specialized mechanism seen only in pathologically altered cells.

6.2 Figures

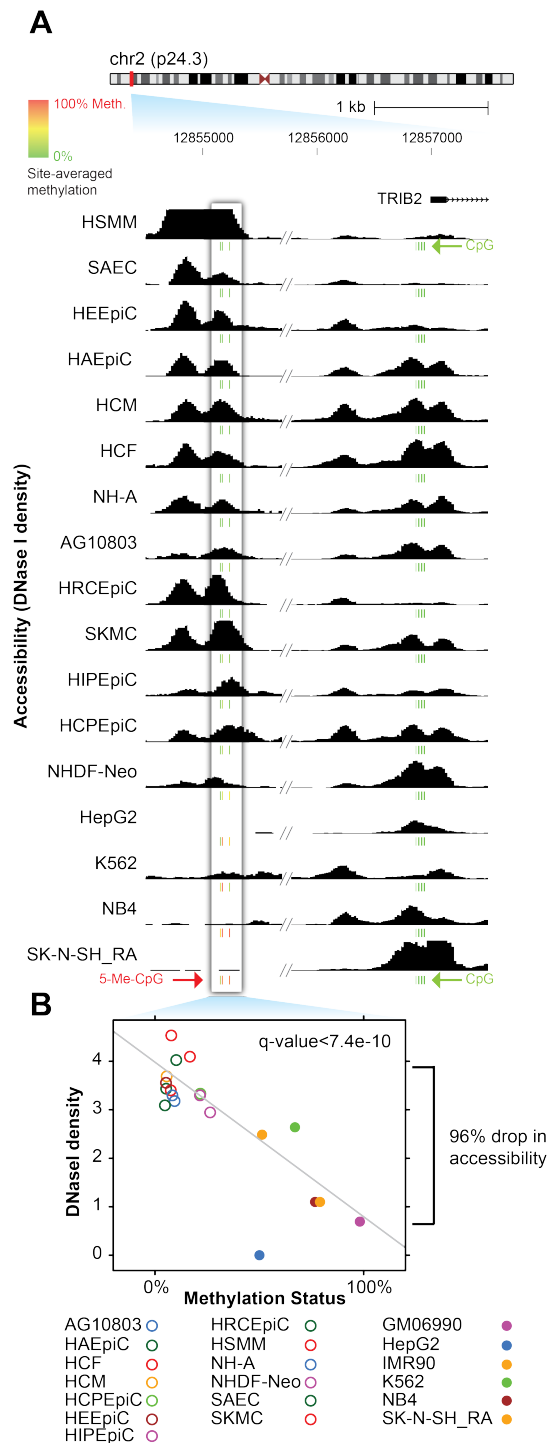


Figure 6-1 Chromatin accessibility and DNA methylation patterns. DNase I sensitivity in 17 of 19 cell types with ENCODE reduced representation bisulphite sequencing data. Inset box: accessibility (y axis) decreases quantitatively as methylation increases. Other DHSs (right) show low correlation between accessibility and methylation. CpG methylation scale: green, 0%; yellow, 50%; red, 100%.

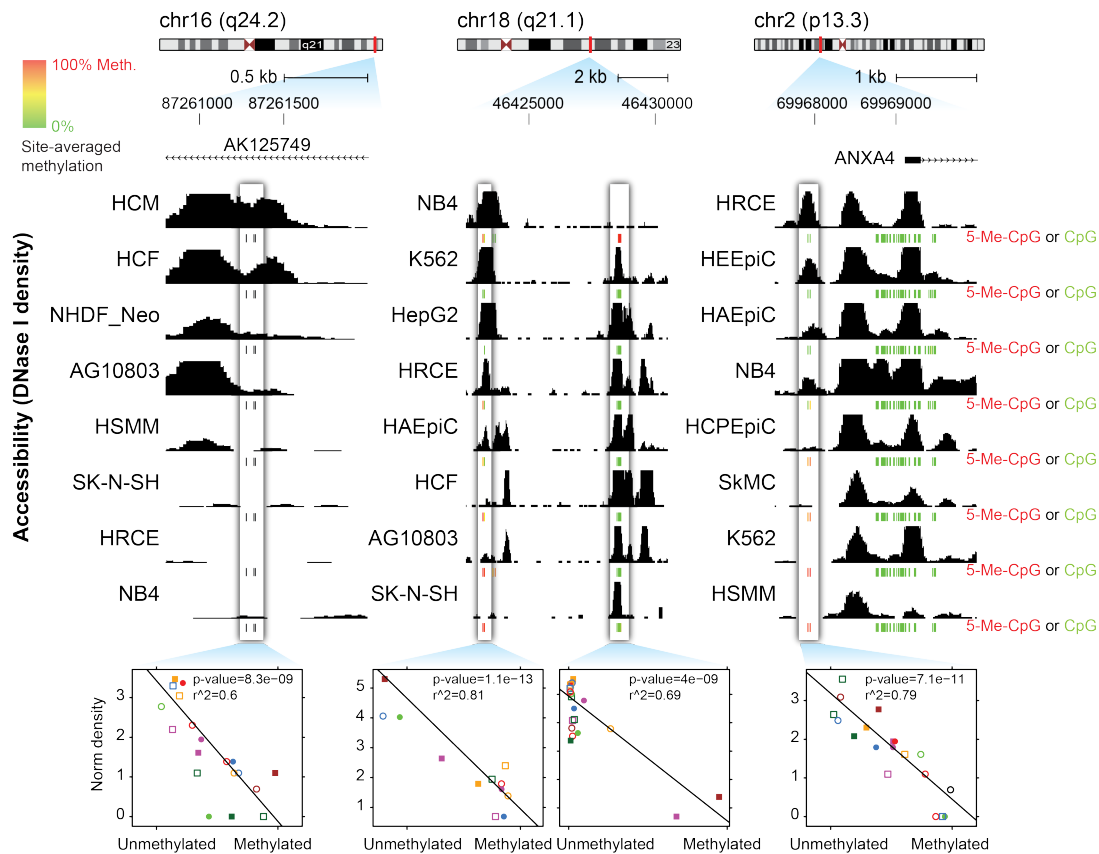


Figure 6-2 Further examples of association between methylation and accessibility.

Data tracks show DNase I sensitivity in selected cell types. Green bars, CpG is 0% methylated; yellow, 50% methylated; red, 100% methylated. Association is quantified in the plots below the tracks. Each point in the graph represents one of 19 cell-types (a superset of the tracks above). X-axis is the percent methylation of the site in that cell-type; y-axis is the normalized DNase I tag density at the site in that cell type. In each example, accessibility (y-axis) quantitatively decreases as methylation increases (left to right).

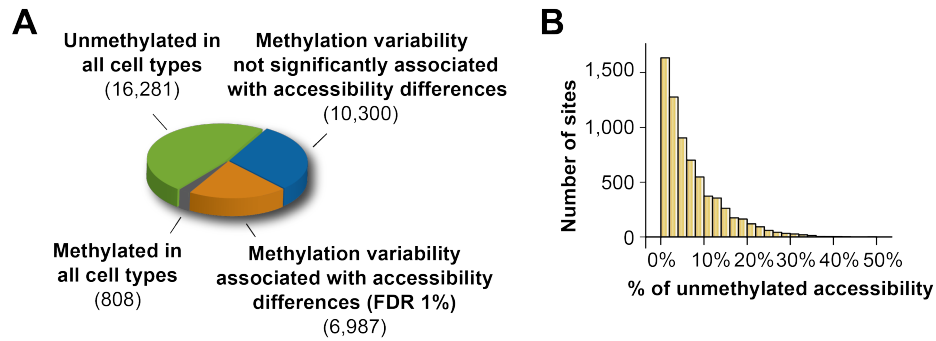


Figure 6-3 Global characterization of the effect of methylation on chromatin accessibility, surveyed at 34,376 DHSs with RRBS data.

(A) 40% of sites with variable methylation across cell-types were associated with differences in chromatin accessibility. (B), In cell lines with methylated DHSs, site accessibility was reduced on average by 95%. Shown are sites where increased methylation was significantly associated with decreased accessibility (= 97% of all sites in the orange slice shown in (A)).

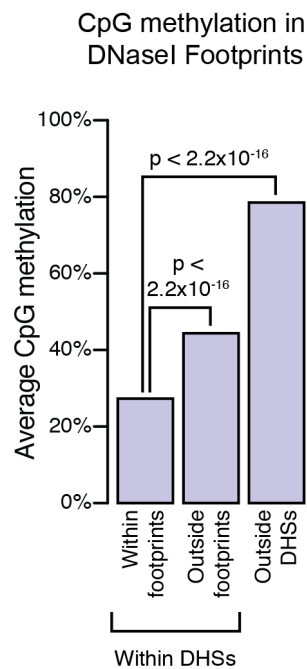


Figure 6-4 DNase I footprints mark sites of in vivo protein occupancy.

The average CpG methylation within IMR90 DNase I footprints, IMR90 DHSs (but not in footprints) and non-hypersensitive genomic regions in IMR90 cells. CpG methylation is significantly depleted in DNase I footprints ($P < 2.2 \times 10^{-16}$, Mann-Whitney U-test).

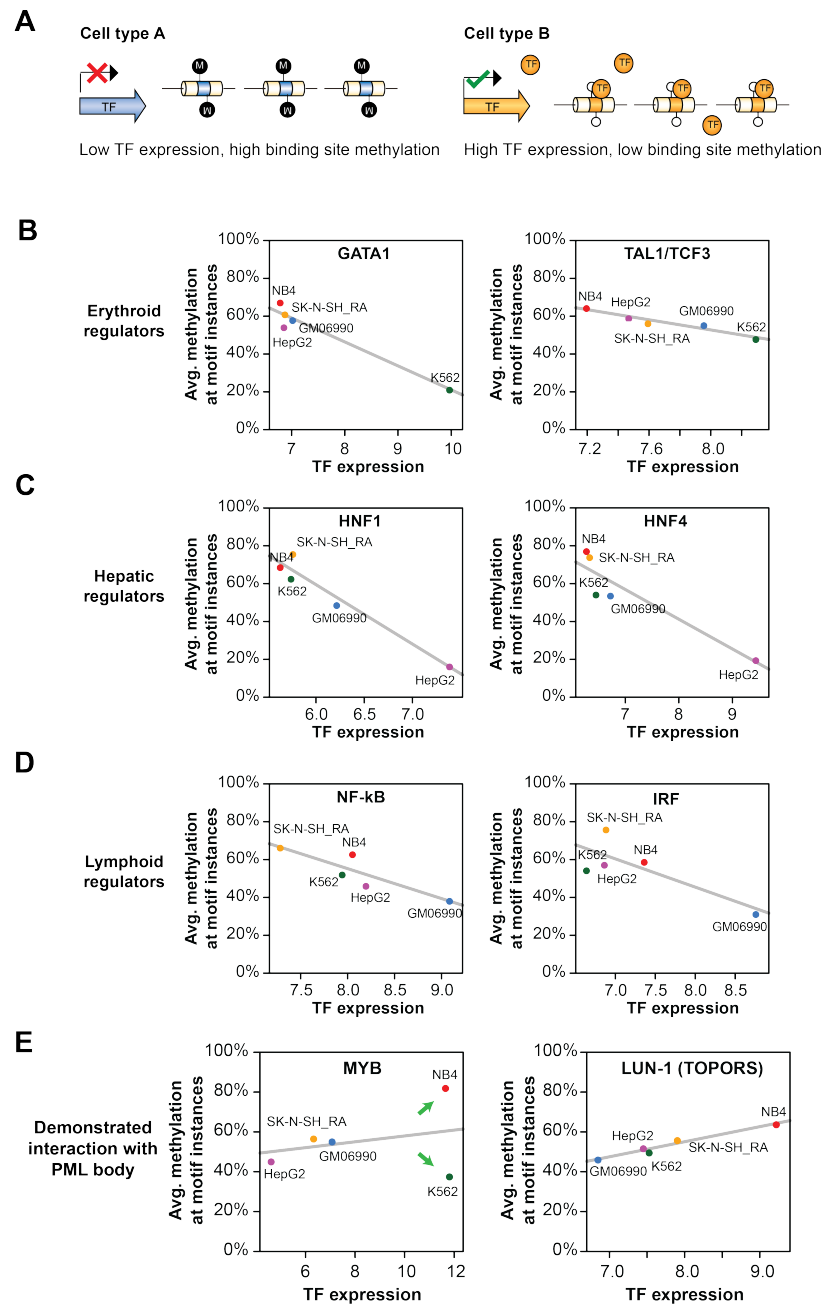


Figure 6-5 Relationship between TF transcript levels and overall methylation at cognate recognition sequences of the same TFs.

(A) Model of transcription factor (TF)-driven methylation patterns in which methylation passively mirrors transcription factor occupancy. (B-E) Relationship between transcription factor transcript levels and overall methylation at cognate recognition sequences of the same transcription factors. Negative correlation suggests that site-specific DNA methylation follows transcription factor vacuation of differentially expressed transcription factors. (B) Erythroid regulators in the erythroleukemia line K562, (C) Hepatic regulators in the liver carcinoma HepG2, and (D) Lymphoid regulators in the B lymphoblast line GM06990. (E) MYB and LUN1 have both been demonstrated to interact with PML bodies, and show increased transcription and binding site methylation in the acute promyelocytic leukemia (APL) line NB4. Although Myb expression is upregulated in both erythroid K562 and the APL line NB4 (green arrows), its putative binding sites exhibit altered methylation only in the APL line NB4.

6.3 Methods

DNase I profiling

DNase I hypersensitivity mapping was performed using protocols developed by University of Washington (John et al., 2011) on a total of 125 cell types. Data sets were sequenced to an average depth of 30 million uniquely mapping sequence tags (27–35 bp) per replicate. For uniformity of analysis, some cell-type data sets that exceeded 40M tag depth were randomly subsampled to a depth of 30 million tags. Sequence reads were mapped using the Bowtie aligner, allowing a maximum of two mismatches. Only reads mapping uniquely to the genome were used in our analyses. Mappings were to male or female versions of hg19/GRCh37, depending on cell type, with random regions omitted. Data were analyzed using a single algorithm (Thurman et al., 2012) to localize DNase I hypersensitive sites. Transcription factor recognition motif occurrences within DHSs were defined with FIMO (Grant et al., 2011) at significance $P < 10^{-5}$ using motif models from the Transfac database.

RNA expression

For each cell line, total RNA was extracted in 2 replicates from 5×10^6 cells using Ribopure (Ambion) according to manufacturer’s instructions. RNA quality was ascertained using RNA 6000 Nano Chips on a bioanalyzer (Agilent, Santa Clara, CA). Approximately 3 μ g of total RNA for each sample was used for labeling and hybridization (University of Washington Center for Array Technology) to Affymetrix Human Exon 1.0 ST arrays (Affymetrix) using a standard protocol. Exon expression data were analyzed through Affymetrix Expression Console using gene-level RMA summarization and sketch-quantile normalization method. Measurements from both replicates were then averaged. Raw data have been deposited in GEO under accession number GSE19090.

RRBS genome-wide methylation profiling

We downloaded RRBS methylation data for 19 cell lines from the “HAIB Methyl RRBS” track (Varley et al., 2013) of the UCSC Genome Browser. To measure methylation in each cell line, we combined counts for

both strands in both replicates and removed CpGs with <8x coverage. We retained only CpGs monitored in at least 6 samples.

We applied a linear regression to measure whether methylation status is associated with accessibility. First, we generated a master list of DHSs found in any of the 19 cell lines. We then regressed accessibility onto the average proportion methylated of all monitored CpGs in a 150 bp region centered around the DNase I peak. We tested only sites with both RRBS data for at least one CpG within the 150 bp window and ChIP-seq data for at least 6 cell lines. We excluded sites where the number of monitored CpGs differed by more than 4 among any two cell lines. We performed a linear regression at each remaining site, and used the R package *qvalue* to estimate a global FDR (Storey and Tibshirani, 2003).

To assess the relationship between expression and TFBS methylation, we determined a set of putative binding sites for transcription factors, based on matches to database motifs inside DHSs where methylation was significantly associated with accessibility. For each transcription factor, we regressed the average methylation at all of these motif instances onto the gene expression in each immortal cell type. We tested only motif models including a CpG.

CpG methylation calculation within footprints, DHSs and non-DHSs

IMR90 methylation calls (Lister et al., 2009) were filtered to CpGs covered by at least 40 reads. Methylation at each CpG is defined as the count of reads showing methylation (protection from bisulphite conversion) divided by the total read depth. We generated three sets of genomic coordinates with this signal: IMR90 FDR 1% footprints, IMR90 DNase I peaks (subtracting overlapping footprint bases), and locations of CpGs in the GRCh37/hg19 genome reference sequence, removing elements that overlap IMR90 DNase I hotspots. For each contiguous region in these data sets, we took the mean methylation of all overlapping CpGs that passed the 40-read coverage threshold. Regions with no such overlap were ignored. To compute P values, vectors of mean methylation values were compared using a two-sided Mann–Whitney U-test.

6.4 Notes

This chapter has been published in the following two papers:

1. Thurman RE, Rynes E, Humbert R, Vierstra J, **Maurano MT**, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kuttyavin T, Lajoie B, Lee B-K, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, and Stamatoyannopoulos JA. "The Accessible Chromatin Landscape of the Human Genome." **Nature** 489: 75–82 (2012)
2. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, **Maurano MT**, Humbert R, Rynes E, Wang H, Vong S, Lee K, Bates D, Diegel M, Roach V, Dunn D, Neri J, Schafer A, Hansen RS, Kuttyavin T, Giste E, Weaver M, Canfield T, Sabo P, Zhang M, Balasundaram G, Byron R, MacCoss MJ, Akey JM, Bender MA, Groudine M, Kaul R, and Stamatoyannopoulos JA. "An Expansive Human Regulatory Lexicon Encoded in Transcription Factor Footprints." **Nature** 489: 83–90 (2012).

Acknowledgements

Particular thanks to Eric Haugen for informatics support. We thank our ENCODE colleagues for many insights into the data types generated by different centers and for help with coordinated analyses. We thank many colleagues, particularly F. Urnov and S. John, for their helpful critiques of the manuscript and figures. This work was funded by NIH grants HG004592 (J.A.S.).

7 DNA Methylation Modulates Transcription Factor Occupancy At Sites Of High Intrinsic Cell Type Variability

7.1 Abstract

The nuclear genome of every cell encodes millions of unoccupied transcription factor (TF) recognition sequences that harbor methylated cytosines. Although DNA methylation is commonly invoked as a repressive mechanism, the extent to which it actively silences specific TF occupancy sites is unknown. To define the role of DNA methylation in modulating TF binding, we quantified the effect of DNA methyltransferase abrogation on the occupancy patterns of a ubiquitous TF capable of autonomous binding to its target sites in chromatin (CTCF). Here we show that the vast majority of unoccupied, methylated CTCF recognition sequences remain unbound upon depletion of DNA methylation. Rather, methylation-regulated binding is restricted to a small fraction of elements that exhibit high intrinsic variability in CTCF occupancy across cell types. We conclude that DNA methylation is not a major gatekeeper of genomic transcription factor occupancy landscapes, but rather a specialized mechanism for stabilizing epigenetically labile sites.

7.2 Introduction

DNA methylation is required for development and plays a central role in imprinting (Jones, 2012; Li et al., 1992). Cytosine methylation in the context of CpG dinucleotides is also widely invoked as a causal mechanism for transcriptional repression at promoter regions (Jones and Taylor, 1980). However, the global relationship between DNA methylation and the binding of the sequence-specific factors that direct chromatin structure remains unresolved (Deaton and Bird, 2011). DNA methylation itself might directly confer repressive chromatin structure (Collings et al., 2013; Ramirez-Carrozzi et al., 2009), or instead alter recruitment of sequence-specific or methyl-cytosine-specific factors (Baubec et al., 2013; Lewis et al., 1992; Tate and Bird, 1993). Transcription factor binding occupied sites *in vivo* are specifically depleted for DNA methylation (Groudine and Conkin, 1985; Lister et al., 2009; Neph et al., 2012b; Thurman et al., 2012), there is evidence that it is the binding of transcription factors that initiates and maintains a local demethylated state rather than the reverse (Brandeis et al., 1994; Lienert et al., 2011; Lin et al., 2000; Macleod et al., 1994; Matsuo et al., 1998; Stadler et al., 2011; Thurman et al., 2012). Given the recurrence of epigenetic disruption in malignancy and the dynamic pattern of DNA methylation during development, it is critical to know whether these altered methylation patterns themselves invoke transcriptional repression or are instead downstream consequences of other regulatory factors.

Previous work has studied the relationship between DNA methylation and gene expression through broad depletion of genomic methylation. homozygous disruptions of the DNA methyltransferases DNMT1 and DNMT3B stably reduces global methylation is by 83-95% in the human colorectal carcinoma line HCT116 DKO (Akalın et al., 2012; Rhee et al., 2002), resulting in genome-wide changes in gene expression and chromatin structure (Hagemann et al., 2011; Komashko and Farnham, 2010; Pandiyan et al., 2013; Qiu et al., 2010). 5-aza-2'-deoxycytidine (5-aza-CdR), a DNA methyltransferase inhibitor and cytotoxic agent in clinical use for myelodysplastic syndromes and acute myeloid leukemia, transiently reduces global methylation levels and has been associated with reactivation of transcription factor binding or increased gene expression (Rodriguez et al., 2010; Witcher and Emerson, 2009). However, it remains unclear to what extent alterations in chromatin structure and gene expression are directly caused by demethylation of specific sites versus second-

ary effects or cytotoxicity (Bernstein et al., 2010b; Komashko and Farnham, 2010; Stresemann et al., 2008; Tsai et al., 2012; Yan et al., 2012). Furthermore, a mechanistic link between the depletion of methylation and its consequences on chromatin structure and gene expression has remained obscure as previous work has not specifically studied the effects of alterations in methylation on transcription factor binding *in vivo*.

The polyfunctional genome regulator CTCF exhibits a tight anticorrelation with DNA methylation at its binding sites *in vivo* (Phillips and Corces, 2009; Wang et al., 2012), and its binding is abrogated by methylation *in vitro* (Bell and Felsenfeld, 2000; Filippova et al., 2001; Hark et al., 2000; Renda et al., 2007). However, CTCF appears to have some ability to bind methylated recognition sequences *in vivo* (Stadler et al., 2011; Wang et al., 2012), although sodium bisulfite methods do not distinguish 5-hydroxymethylcytosine from 5-methylcytosine. Given the prominent role of methylation ascribed to the regulation of CTCF binding, we sought to investigate the causal relationship between genome-wide methylation and its occupancy.

We find that the CTCF binding landscape remains largely unchanged upon depletion of DNA methylation, implying the presence of alternative repressive mechanisms at most sites *in vivo*. However, we observe that a small minority of sites demonstrate strong, reproducible activation. We find these sites are distinguished by high CpG content localizing to key positions in the CTCF binding interface, and an absence of preexisting regulatory activity. Despite the presence of tens of thousands of equally good sequence matches that remain silent, reactivation only occurs at CTCF sites whose binding varies across other cell types. These results suggest that alterations in methylation result in limited regulatory effect according to sequence and chromatin context.

7.3 Results

Limited reactivation of methylation-dependent sites

We investigated CTCF occupancy in the context of reduced methylation by performing genome-wide profiling with chromatin immunoprecipitation (ChIP-seq) in HCT116 cells and DNMT1 and DNMT3B double knockout (DKO) HCT116 cells (Figure 7-1A) including multiple biological replicates (3 for HCT116 and 2 for DKO). To establish the cell type activity profile of each CTCF site, we additionally surveyed CTCF occu-

pancy in 39 varied cell types, including new data for 20 cell types, each with at least two biological replicates (Table 7-1). We identified a total of 107295 CTCF sites in these 40 cell types, HCT116 DKO and 5-aza-2'-deoxycytidine-treated K562 erythroleukemia cells (5-aza-CdR; described later), representing a 38% increase in the number of known CTCF sites (Figure 7-1B, Figure 7-2, Table 7-2).

To precisely gauge the alteration to CTCF binding in the context of reduced genomic methylation, we categorized sites based on whether occupancy differed significantly (FDR 1%) between HCT116 and DKO and whether or not they were already occupied in HCT116 and (Figure 7-1C and Methods). The majority of sites were 'preexisting' sites already bound in HCT116 and with unchanged CTCF occupancy in DKO (Figure 7-1D), suggesting a minor effect on a global scale. To confirm a lack of reactivation, we surveyed sequence matches genome-wide for the CTCF binding motif not exhibiting CTCF occupancy *in vivo* in any cell type (Figure 7-3 and Section 7.6). These sequences demonstrated a complete lack of reactivation, and exhibited a lack of occupancy similar to a set of random sequences without the CTCF recognition sequence. Thus the CTCF binding landscape remains largely similar despite a drastic reduction in genomic methylation.

However, we did identify a small compartment of 4204 sites that differed significantly in occupancy (FDR 1%), comprising 158 'downregulated' and 809 'upregulated' sites (bound in HCT116, but decreased or increased significantly); and 3237 'reactivated' sites (DKO but not HCT116). Reactivated sites were strongly occupied, showing similar occupancy in DKO to preexisting sites in either cell type (Figure 7-1E). To exclude the possibility that binding differences might be due to indirect effects from the targeting of the DNA methyltransferases rather than their subsequent effect on DNA methylation, we also profiled single knockouts of DNMT1 and DNMT3B. Each single knockout is known to exhibit little difference in methylation (Rhee et al., 2002), and correspondingly, we found that each deletion resulted in effectively no significant changes in occupancy (Figure 7-4).

To verify whether reactivation was linked specifically to the relief of methylation-dependent repression, we examined a subset of CTCF sites for which quantitative profiling of methylation in HCT116 was available using representation bisulfite sequencing (RRBS) (Varley et al., 2013). Confirming prior results (Lister et al.,

2009; Thurman et al., 2012; Wang et al., 2012), methylation was low at preexisting sites and high at a majority of unbound sites (Figure 7-1F, Table 7-3). Reactivation occurs almost exclusively (96%) at methylated sites. Yet only 50% of methylated unbound sites were reactivated, suggesting either highly localized resistance to demethylation or alternate regulatory mechanisms that prevent binding. Similarly, 90% of CTCF motifs without ChIP-seq occupancy were methylated but exhibited virtually no reactivation. Given the low global level of remaining methylation in DKO (Akalin et al., 2012; Rhee et al., 2002), it is unlikely that a lack of reactivation represents localized persistence of methylation at unreactivated sites. These results suggest that while loss of methylation is necessary for CTCF binding at a subset of sites, methylation is not the sole barrier to *in vivo* occupancy at the majority of sites genome-wide.

Reactivation of predetermined cell type-specific CTCF sites

We asked whether reactivated sites demonstrate CTCF binding in other cell types or if demethylation instead activates novel binding sites. Strikingly, we found that fully 91% of peaks found in DKO but not HCT116 were found in at least 1 of 39 other cell types, suggesting reactivation affects a methylation-sensitive subset of sites with an inherent capacity for CTCF binding (Figure 7-5A). However, there remain 44008 sites found in other cell types but not HCT116, suggesting that methylation-sensitive sites are distinguished by specific characteristics. We use these potentially reactivated sites to compute the measurement of the frequency of reactivation at various classes of CTCF sites.

By investigating CTCF binding across 39 other cell types, we found that upregulated and reactivated sites were far more likely than preexisting sites to vary epigenetically across cell types (Figure 7-5B). Methylation sensitivity was highest at sites found in almost all cell types (excluding HCT116) and at sites found exclusively in malignancy-derived or immortalized lines (Figure 7-5C). Methylation-dependent CTCF sites have been shown to regulate several tumor suppressors and oncogenes (Butcher et al., 2004; Dávalos-Salas et al., 2011; Soto-Reyes and Recillas-Targa, 2010; Witcher and Emerson, 2009). A subset of methylation-associated cell-type selective CTCF sites exhibits characteristic hypermethylation in immortal but not normal lines (Wang et al., 2012), especially at CpG islands (Varley et al., 2013), which is strikingly recapitulated in patterns of reactivation (Figure 7-5D). Indeed, the vast majority (93%) of reactivated CTCF sites had a CpG within the 44-bp

region of protein-DNA interaction, compared to only 54% of unreactivated sites. Further, reactivation frequency increased with the quantity of CpGs at the binding sites (Figure 7-5E). Similarly, while 29% of the reactivated sites were in CpG islands, only 10% of unreactivated sites were. Thus, malignancy is associated with the disruption of a specific set of labile, methylation-sensitive CTCF sites.

Chromatin dynamics at reactivated CTCF sites

The observation that demethylation reactivates experimentally determined labile sites rather than other methylated sequence matches implies that unoccupied CTCF sites are clearly distinguished *in vivo* from a separate class of inert sequence matches. To examine whether chromatin accessibility prior to demethylation distinguishes these predetermined sites (John et al., 2011), we profiled HCT116 and DKO using DNaseI-seq and ChIP-seq for trimethylation of histone 3 lysine 4 (H3K4me3) and acetylation of histone 3 lysine 27 (H3K27ac), covalent modifications of histones associated with transcriptional activity and chromatin accessibility (Figure 7-6A, Table 7-4). At the vast majority of CTCF sites, its occupancy is tightly associated with the presence of a DNase I hypersensitive site (DHS) marking increased chromatin accessibility (Figure 7-6B). Only 17% of reactivated sites overlapped DHSs in HCT116, suggesting that the novel recruitment of CTCF is not directed to chromatin accessible prior to demethylation (Figure 7-6B).

In contrast to a lack of accessibility in HCT116 at reactivated CTCF sites, we observed that unreactivated CTCF sites were more frequently DHSs in HCT116 (Figure 7-6B). Further confirming the lack of reactivation at sites of co-binding, the association of increased reactivation with increased number of CpGs was specific to sites without preexisting accessible chromatin (Figure 7-6C). The presence of trimethylation of histone 3 lysine 4 (H3K4me3) was similarly antagonistic to reactivation (Figure 7-7B), and the presence of both DNase I hypersensitivity and H3K4me3 was associated with a lack of methylation specifically at unreactivated sites despite a lack of CTCF occupancy (Figure 7-1F, Figure 7-7C). We found that reactivated sites were depleted for the occupancy of 17 transcription factors (Figure 7-7D) studied by the ENCODE project (ENCODE Project Consortium, 2012). Thus, sensitivity to methylation is reduced by the co-binding of additional regulatory factors permitting a relaxed chromatin state nevertheless inhospitable to CTCF binding.

Given that the removal of DNA methylation might indirectly enable CTCF occupancy through a nonspecific relaxation of chromatin state (Ooi et al., 2007; Thomson et al., 2010), we then asked whether alterations in chromatin structure depended upon the reactivation of CTCF binding. We found that chromatin accessibility measured by DNaseI-seq matched the overall change in CTCF occupancy at reactivated, upregulated, and downregulated sites, but that only a subset of reactivated CTCF sites were accompanied by reactivation of H3K4me3 (Figure 7-8). In fact, H3K4me3 reactivation occurred regardless of whether CTCF was recruited to the site and depended strongly upon CpG content (Figure 7-6D), consistent with the observation that Cfp1 (part of the Set1 methyltransferase complex) recognizes unmethylated CpG islands (Thomson et al., 2010). In contrast, the correlation between DNase I reactivation and CpG content was almost entirely conditional on CTCF reactivation (Figure 7-6E). While only a minority of H3K4me3 reactivation was associated with CTCF reactivation, most increased DNase I accessibility was at sites with CTCF reactivation (Figure 7-6F). Thus, we conclude that while active histone marks such as H3K4me3 are deposited in regions of high CpG content regardless of CTCF occupancy, DNase I accessibility specifically marks regulatory factor occupancy.

The extent to which other mammalian regulatory factors interact with the global methylation landscape remains obscure (Lienert et al., 2011). DNase I accessibility provides a global picture of the regulatory landscape within which to contextualize the contribution of CTCF in connecting DNA methylation to regulatory sites genome-wide. DKO cells showed a generally expanded accessible chromatin landscape, comprising a 61% increase in the number of DHSs (Figure 7-9), but this represents a surprisingly small increase given the even greater reduction in methylation. We found that the vast majority of H3K4me3 (84%) and H3K27ac (88%) peaks in DKO overlapped with DHSs, confirming the global sensitivity of DNase I profiling for regulatory activity. Using a stringent threshold, we found 12279 DHSs with statistically significant differential accessibility (1% FDR), 96% of which were reactivated DHSs not present in HCT116. We found that 70% of reactivated DHSs were present in a survey of 124 cell types (Maurano et al., 2012a), confirming preferential reactivation of silence sites from other lineages. We found that between 48% (1% FDR) and 77% (5% FDR) of reactivated CTCF sites without a DHS in HCT116 were recognizable from chromatin accessibility alone,

suggesting a high sensitivity to reactivation of individual factors. Although CTCF has a large number of binding sites plays a pivotal role at imprinted loci, 90% of reactivated DHSs were not CTCF sites, suggesting a broad pool of methylation-responsive regulatory sites (Figure 7-9D).

CpGs at key positions in the protein-DNA binding interface drive reactivation

Although increased CpG content is strongly associated with CTCF reactivation (Figure 7-6C), it is unclear whether this is due to a generally increased response of high-CpG regions to removal of methylation or a specific disruption of the CTCF protein-DNA interface. Consistent with a direct effect on CTCF binding, we observed a strong overall increase in reactivation at stronger matches to the CTCF consensus sequence, with the strongest demonstrating 27% reactivation (Figure 7-10A). We classified CTCF binding sites by sequence similarity into three known binding modes (Bowers et al., 2009; Filippova et al., 1996; Nakahashi et al., 2013; Ohlsson et al., 2001; Rhee and Pugh, 2011), including engagement of zinc fingers 3-7 (core sites) and additional specific interactions by zinc fingers 8-10 (upstream and extended spacing sites). We found that core sites were reactivated 42% more frequently than upstream sites (Figure 7-10B), supporting speculation that different modes of engagement of its 11 zinc fingers may confer susceptibility to different modes of regulation or functional selectivity (Ohlsson et al., 2010). The markedly increased reactivation frequency for core sites may represent increased sensitivity to CpG methylation due to the lower binding energy from the reduced protein-DNA binding interface. Despite the correlation of sequence features with reactivation, it is important to reiterate that local sequence content alone is insufficient to distinguish *in vivo* occupied sites from other recognition sequences in the genome (Figure 7-3C).

To examine whether the association of CpG content with reactivation might localize to specific positions in the protein-DNA interface (Renda et al., 2007; Wang et al., 2012; Wiench et al., 2011), we used a linear regression model to quantify the contributions of CpG dinucleotides in the recognition sequence. We found that critical CpGs were concentrated in the core binding region, consistent with *in vitro* sensitivity to methylation (Renda et al., 2007), the positions of CpGs at sites of methylation-associated cell-type selective binding (Wang et al., 2012), and sensitivity to single nucleotide variants (Maurano et al., 2012b) (Figure 7-11). Although it is possible that different CTCF binding modes exhibit differential sensitivity to methylation at spe-

cific locations within the recognition sequence, we did not find a significant effect of binding mode on susceptibility at any given position.

Cell type-specific reactivation highlights methylation-independent repression

Given our discovery of a clearly distinguishable subset of methylation-sensitive CTCF sites, we studied whether transient depletion of methylation resulted in similar patterns of reactivation. We performed ChIP-seq on mock-treated and 5-aza-CdR-treated K562 cells, identifying 61658 CTCF sites active in these cells. We observed weaker reactivation with 5-aza-CdR than in DKO cells, including 767 reactivated, 191 upregulated, and 4 downregulated sites, DKO (Figure 7-12A-B). Given the limited extent of reactivation using 5-aza-CdR, we assessed the extent of demethylation with high resolution using a capture sodium bisulfite design including 12,539 CTCF sites (Section 7.6). Consistent with previous reports (Hagemann et al., 2011; Pandiyan et al., 2013; Rhee et al., 2002), these results confirm a more limited and uneven reduction in methylation than the almost completely demethylated DKO cells (Figure 7-13). CTCF reactivation was largely specific to sites of preexisting methylation: 79% of reactivated sites exhibited >25% methylation in K562 cells. However, only 30% of reactivated sites showed >5% methylation change and 38% of unreactivated sites showed no methylation change, suggesting widespread secondary effects and a similarly muted response of CTCF binding to demethylation. Conversely, only 12% of the CTCF sites methylated and unoccupied in K562 exhibited reduced methylation (>5%) upon treatment (Figure 7-13D). This complex relationship is consistent with previous observations that 20% of variable CTCF sites are unmethylated in all cell types (Wang et al., 2012), and that most 5-aza-CdR-induced alterations in gene expression or chromatin structure occur at previously unmethylated sites (Komashko and Farnham, 2010).

Despite the lesser extent of reactivation with 5-aza-CdR, we found substantial overlap of reactivation: 69% of sites reactivated with 5-aza-CdR overlapped sites also reactivated in DKO (Figure 7-11C). Sites reactivated concordantly in both cell types demonstrated CpG-dependent reactivation (Figure 7-11D), and exhibited high preexisting methylation in both K562 and HCT116 cells, further supporting the existence of a predetermined class of methylation-dependent sites. We sought to specifically identify sites reactivated in only one cell type upon depletion of methylation, as a given unoccupied CTCF site might be silenced in a different fashion be-

tween K562 and HCT116. Indeed, sites discordantly reactivated in only one of the two cell types demonstrated significantly less methylation (Figure 7-11E), suggesting the presence of a methylation-independent silencing mechanism in the unreactivated cell type (Figure 7-11F).

Recognition of labile methylation-sensitive sites

Given the importance of methylation at regulatory sites in imprinting and malignancy, we examined the predictive power of genomic characteristics to recognize the minority of sites that were reactivated in comparison to the unaffected CTCF binding landscape (Figure 7-14). The number of CpGs at the critical positions in the recognition sequence (Figure 7-11A) was the single most predictive factor (Figure 7-14A). The number of CpGs in the flanks was almost as predictive, probably due to the generally high CpG density of CTCF sites (Figure 7-14B). The second most predictive factor was lack of chromatin accessibility in HCT116 (Figure 7-14A). As many of the features surveyed are not independent, we used a logistic regression model to assess the combined predictive power of the characteristics we identified (Section 7.6). A model considering just chromatin accessibility and number of CpGs was more predictive (AUC of 0.82) than either factor alone, but almost as predictive as the full model (AUC 0.84), indicating that the remaining factors provide only minor additional discriminative power (Figure 7-14B). Although we observed increased reactivation at exons, the increased G+C content of protein-coding sequence is presumably responsible for a markedly increased reactivation frequency relative to other genomic sequence (Figure 7-15A), which suggests an altered regime of methylation-regulatory factor interaction at these sites. At the subset of sites for which methylation data were available, the degree of methylation alone had strong predictive power (Figure 7-14B). But, this predictive power was largely redundant when DNase I accessibility and number of CpGs was additionally considered, suggesting that the latter features primarily provide indication of methylation status (Figure 7-7C). Thus, reactivated sites can be reliably distinguished based on *in vivo* profiling and a simple model considering CpG content and chromatin accessibility (Figure 7-14C).

7.4 Discussion

We describe a surprisingly muted response of the regulatory landscape to large-scale depletion of methylation, using both transient and stable inhibition of DNA methyltransferases in two different cell types. Although

some sites showed exquisite sensitivity to methylation, demonstrating increased occupancy despite a small changes in methylation, the binding landscape as a whole remains mostly unchanged. This is consistent with the moderate effect on binding of single nucleotide changes in the CTCF recognition sequence (Maurano et al., 2012b), and suggests that its rich binding interface is buffered at most sites *in vivo* to the effect of methylation. Given the global lack of correlation between demethylation and altered binding, these results place a limit on the extent of CTCF-mediated coupling between DNA methylation and genome organization. Thus although DNA methylation globally shows a tight relationship with reduced CTCF binding, this relationship is only causal at a small subset of sites, cautioning against a facile interpretation of alterations of DNA methylation in oncogenesis.

That reactivation is strictly limited to *in vivo* CTCF sites from other cell types suggests that either these silent sites are globally marked in HCT116 or that other sequence matches to the CTCF recognition sequence are silenced independent of methylation. Approximately half of both reactivated and unreactivated sites are accessible to DNase I in human embryonic stem cells, compared to only 11% of sequence matches, suggesting that the marking of bona fide CTCF sites occurs early in development. But it remains unclear how unreactivated sites are distinguished from mere sequence matches. These silent sites are not marked by DNaseI accessibility, and in fact reactivation appears to be less frequent in the presence of cofactors. Thus, although CTCF preferentially recognizes a DNA sequence with higher information content than most other human transcription factors, it appears likely that cooperative binding with as-yet undetermined factors plays an even larger role in the determination of its *in vivo* binding sites.

Despite the secondary role of methylation at most sites, a set of reproducibly reactivated sites is distinguished by solitary CTCF binding and clear sequence characteristics including the presence of CpGs at key positions in the binding interface. The strongest factor distinguishing reactivated sites is the presence of CpG dinucleotides at key positions of the protein-DNA recognition interface. This localization is consistent with our previous findings showing that the depletion of methylation is specifically increased at sites of transcription factor occupancy marked by DNaseI footprints (Neph et al., 2012b). The conferral of methylation-sensitive binding

by the presence of CpG dinucleotides at key positions is reflected in epigenetically lability across cell types and suggests that CTCF's rich binding interface is tuned to permit a range of sensitivity to methylation.

7.5 Figures and Tables

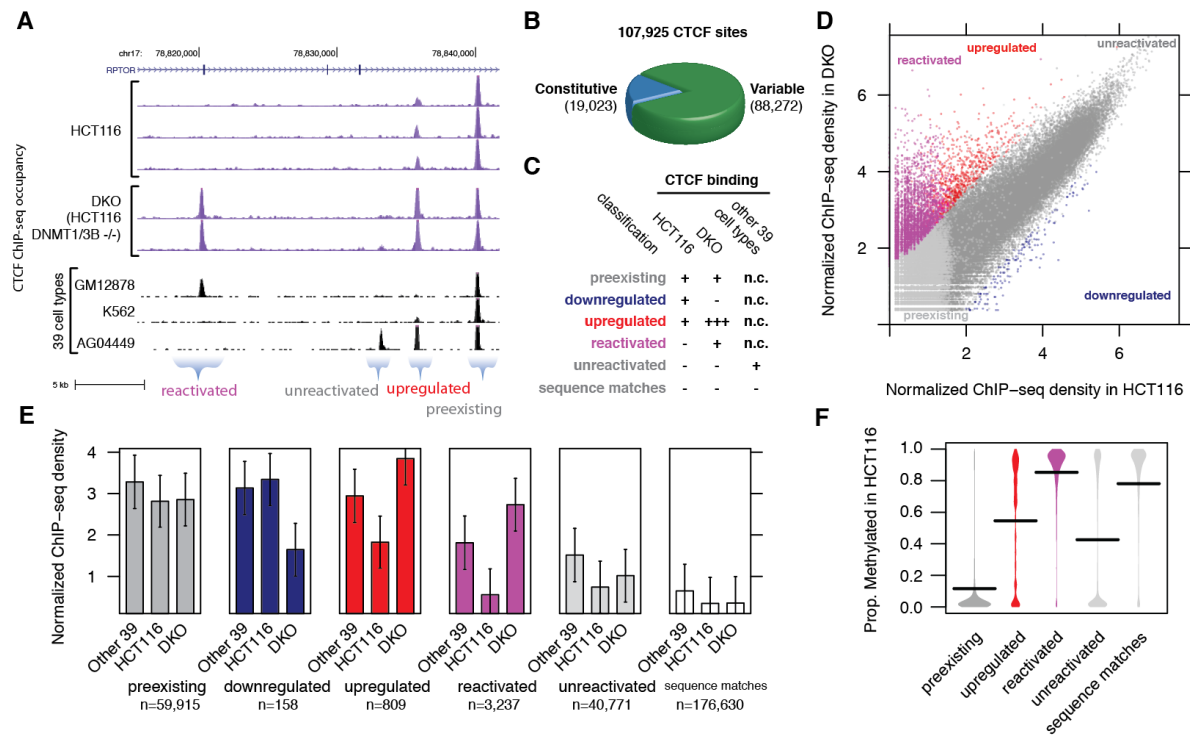


Figure 7-1 Profiling of methylation-silenced transcription factor binding sites in stably demethylated cells.

(A) ChIP-seq profiling of CTCF occupancy in colorectal cancer cells depleted for methylation (HCT116, wild type; DKO, double knockout of DNMT1 and DNMT3b) shows reactivated sites at RPTOR locus present in DKO but not HCT116.

(B) Profiling across 40 cell types comprehensively maps the variable CTCF *in vivo* binding landscape.

(C) Classification scheme for CTCF sites based on ChIP-seq peak presence/absence (+/-) in HCT116; significant difference in occupancy upon demethylation in DKO (+ or +++, increased occupancy at 1% FDR); and peak presence/absence (n.c., not considered) in other 39 cell types.

(D) A minority of sites exhibited increased occupancy in DKO (above diagonal).

(E) Quantitative comparison of mean occupancy at each class of sites in HCT116, DKO or other cell types (measured as 90th percentile occupancy across 39 other cell types per site). Error bars represent standard deviation.

(F) Average methylation across 9208 sites using RRBS profiling in HCT116. High methylation characterizes reactivated sites; unreactivated sites exist in both unmethylated and unmethylated populations. Downregulated sites (n=5 with methylation data) not shown. Horizontal bars represent mean.

Table 7-1 Growth conditions for new cell types.

Cell type	Description	Medium	Supplement				Notes
			%FBS	NEAA ^a	NaPy- ruvate ^b	Pen- strep ^c	
A549	human lung carcinoma epithelial cells	F-12K ^d	10%			1 ×	
AG04449	fetal buttock/thigh dermal fibroblast	MEM ^e	15%	1 ×		1 ×	
AG04450	fetal lung fibroblast	MEM ^e	15%	1 ×		1 ×	
BE2_C	human brain neuroblastoma cells	1:1 mixture of F-12Kd and MEM ^e	10%	1 ×	1 ×	1 ×	
GM12878	human lymphoblastoid cell line	RPMI ^f	15%			1 ×	
HA-c	human astrocytes-cerebellar	AM astrocyte medium (ScienCell, #1801)					(h)
HA-sp	human astrocytes-spinal cord	AM astrocyte medium (ScienCell, #1801)					(h)
HCFaa	human cardiac fibroblasts-adult atrial	FM fibroblast medium (ScienCell, #2301)					(h)
HCM	human cardiac myocytes	CMM cardiac myocyte medium (ScienCell, #6201)					(h)
HCPEpiC	human choroid plexus epithelial cells	EpiCM, epithelial cell medium (ScienCell #4101)					(h)
HCT116	human colorectal carcinoma cells	McCoy's 5A modified medium (ATCC #30-2007)	10%			1 ×	
HCT116 DNMT1	human colorectal carcinoma cells with deletion of DNMT1	McCoy's 5A modified medium (ATCC #30-2007)	10%			1 ×	
HCT116 DNMT3B	human colorectal carcinoma cells with deletion of DNMT3b	McCoy's 5A modified medium (ATCC #30-2007)	10%			1 ×	
HCT116 DKO	human colorectal carcinoma cells with deletion of DNMT1 and DNMT3b	McCoy's 5A modified medium (ATCC #30-2007)	10%			1 ×	
HEK-293	human embryonic kidney epithelial cells	MEM ^e	10%	1 ×	1 ×	1 ×	
HFF	human foreskin fibroblasts	DMEM ^g	10%			1 ×	
HFF-Myc	human foreskin fibroblasts expressing canine cMyc	DMEM ^g	10%			1 ×	
HMEC	human mammary epithelial cells	MEGM BulletKit (Lonza #CC-3150)					
HRPEpiC	human retinal pigment epithelial cells	EpiCM, epithelial cell medium (ScienCell #4101)					(h)
HUVEC	human umbilical vein endothelial cells	EGM-2 BulletKit (Lonza #CC-3162)					
HVMF ^f	human villous mesenchymal fibroblast cells	FM fibroblast medium (ScienCell, #2301)					(h)
K562 aza	human chronic myelogenous leukemia cells treated with 1 μM 5-aza-2'-deoxycytidine in DMSO daily for 3 days	RPMI ^f	10%		1 ×	1 ×	
K562 mock	human chronic myelogenous leukemia cells treated with 1 μM DMSO daily for 3 days	RPMI ^f	10%		1 ×	1 ×	
MCF7	human mammary gland adenocarcinoma	MEM ^e	10%	1 ×		1 ×	bovine insulin ⁱ
NB4	human acute promyelocytic leukemia cells	RPMI ^f	10%			1 ×	
NHDF-Neo	neonatal normal human dermal fibroblasts	FGM-2 BulletKit (Lonza #CC-3132)					
NHEK	normal human epidermal keratinocytes	KGM BulletKit (Lonza #CC-3111)					
NHILF	normal human lung fibroblasts	FGM-2 Bullet Kit (Lonza #CC-3132)					
RPTEC	renal proximal tubule epithelial cells	REGM™BulletKit® (Lonza #CC-3190)					
WI- 38/hTERT /GFP- RAF-ER	human embryonic lung fibroblasts	MEM ^e	10%	1 ×	1 ×	1 ×	

Supplements:

- a: NEAA 0.1 mM non-essential amino acids (Invitrogen, #11140-050)
b: NaPyruvate 1 mM sodium pyruvate (Cellgro, #25-000-CI)
c: Pen-Strep 100 IU penicillin and 100 μg/ml streptomycin (Cellgro, #30-002CI)

Media:

- d: F-12K Kaighn's modification of Ham's F-12 medium (ATCC, #30-2004)
e: MEM MEM with 2mM L-glutamine and Earle's salts (Cellgro, #10-010-CM)
f: RPMI RPMI 1640 with 2 mM L-glutamine (Cellgro, #10-040-CM)
g: DMEM DMEM with 4.5 g/L glucose, L-glutamine, and sodium pyruvate (Cellgro, 10-013-CV)

Notes:

- h: coat flasks with poly-L-lysine 2 μg/cm², ScienCell #0413
i: bovine insulin with 0.01 mg/ml bovine insulin (Sigma #I6634)

Other cells cultured as previously described (Wang 2012).

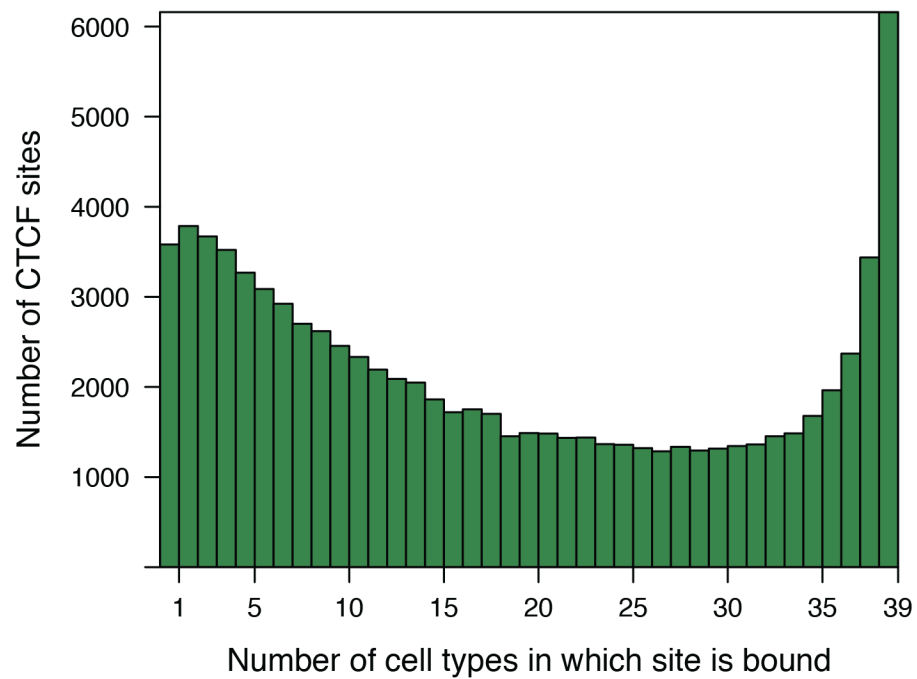


Figure 7-2 Genome-wide analysis of CTCF binding in 40 cell types.
Variable sites exhibit a wide degree of cell-type selectivity across 40 cell types.

Table 7-2 Summary of CTCF ChIP-seq experiments.

Cell type	# of non-redundant tags	SPOT Rep1	SPOT Rep2	SPOT Rep3	Published	Accession
A549	14061527	0.4586	0.4349		Novel	GSE30263
AG04449	29565255	0.6157	0.2732		Novel	GSE30263
AG04450	35133130	0.2939	0.3082		Novel	GSE30263
AG09309	22121486	0.3901	0.264		Wang 2012	GSE30263
AG09319	39296604	0.4307	0.3251		Wang 2012	GSE30263
AG10803	48734285	0.3229	0.2775		Wang 2012	GSE30263
AoAF	34080709	0.3891	0.3219		Wang 2012	GSE30263
BE2_C	38938541	0.3717	0.3043		Novel	GSE30263
BJ	20154780	0.5473	0.4401		Wang 2012	GSE30263
Caco_2	22633608	0.4291	0.2837		Wang 2012	GSE30263
GM06990	17564629	0.4851	0.446		Wang 2012	GSE30263
GM12878	19651161	0.445	0.3454		Novel	GSE30263
HAc	30983231	0.3286	0.2379		Novel	GSE30263
HAsp	26294223	0.3498	0.2369		Novel	GSE30263
HBMEC	31277391	0.542	0.2946		Wang 2012	GSE30263
HCM	27614074	0.4139	0.3096		Novel	GSE30263
HCPEpiC	41581449	0.6368	0.3573		Novel	GSE30263
HEEpiC	29763677	0.3467	0.3198		Wang 2012	GSE30263
HEK293	15353682	0.5978	0.4668		Novel	GSE30263
HeLa_S3	13695593	0.4769	0.4253		Wang 2012	GSE30263
HepG2	18174841	0.499	0.4195		Wang 2012	GSE30263
HFF_Myc	31402050	0.2936	0.2287		Novel	GSE30263
HMEC	27836859	0.566	0.3643		Novel	GSE30263
HMF	27854938	0.414	0.4151		Wang 2012	GSE30263
HPAF	46244156	0.4154	0.3409		Wang 2012	GSE30263
HPF	41607137	0.3115	0.2477		Wang 2012	GSE30263
HRE	12267228	0.5636	0.5548		Wang 2012	GSE30263
HRPEpiC	36393999	0.3228	0.3102		Novel	GSE30263
HUVEC	10749551	0.6721	0.5917		Novel	GSE30263
HVMF	22608466	0.3393	0.3958		Novel	GSE30263
K562	20398176	0.4664	0.4433		Wang 2012	GSE30263
MCF_7	38770153	0.3355	0.2505		Novel	GSE30263
NHDF_Neo	35771540	0.3492	0.2522		Novel	GSE30263
NHEK	12262582	0.6636	0.453		Novel	GSE30263
RPTEC	31992786	0.4334	0.3625		Novel	GSE30263
SAEC	13092335	0.6743	0.645		Wang 2012	GSE30263
Sk_N_Sh_RA	16616235	0.5992	0.6092		Wang 2012	GSE30263
WERI_Rb_1	18427416	0.571	0.4352		Wang 2012	GSE30263
WI_38	21034423	0.2201	0.2104		Novel	GSE30263
K562-Mock	38165687	0.2677	0.1259	0.3027	Novel	GSE50610
K562-aza	25113857	0.2298	0.2008	0.2111	Novel	GSE50610
HCT116	32270754	0.2999	0.2362	0.2531	Novel	GSE50610
DNMT1KO	9645811	0.196			Novel	GSE50610
DNMT3bKO	12618070	0.2367			Novel	GSE50610
DKO	16167225	0.2351	0.2393		Novel	GSE50610

Number of tags represents the total uniquely-mapped reads (excluding PCR duplicates) for all biological replicates of each cell type. Number of peaks refers to the number of CTCF binding sites considered present in a given cell type (Section 7.6). Pearson correlation coefficient (R) measures concordance between replicate normalized density tracks at peaks. Peak concordance between replicates refers to the average of the proportion of peaks in one replicate that are found in the second replicate and vice versa. 19 cell types have been previously published (Wang et al., 2012). SPOT score indicates data quality.

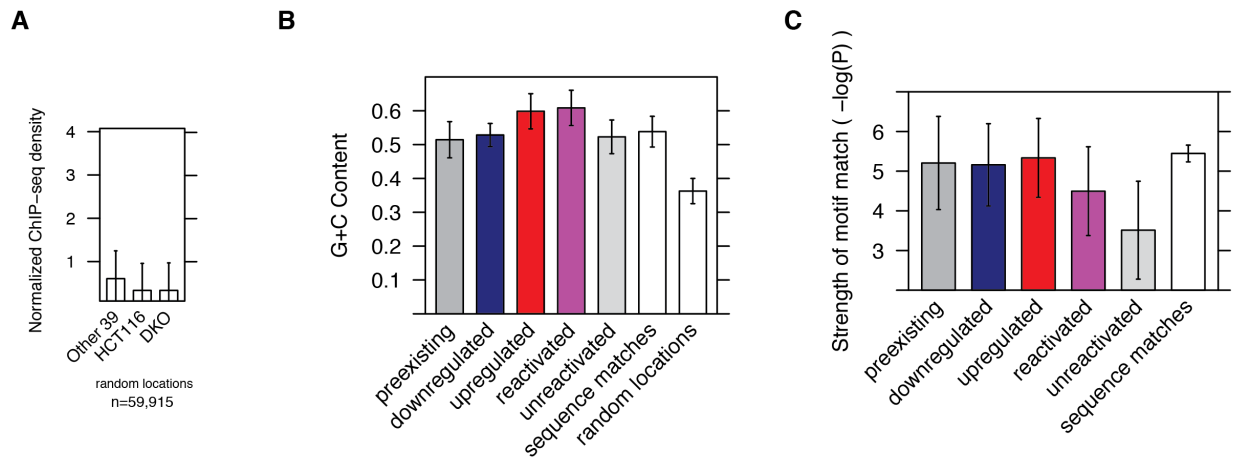


Figure 7-3 Sequence characteristics of CTCF sites.

(A) Informatically predicted sequence matches to the CTCF binding motif (Figure 7-1E) show no higher occupancy than random sequence.

(B) Strength of match to the consensus binding sequence by class of site. Note that sequence matches have as good or better matches when compared to in vivo determined binding sites.

(C) G+C content by class of site. Note increased G+C content for upregulated and reactivated sites. Note the selection of random locations without any match to the CTCF recognition sequence results in low G+C content.

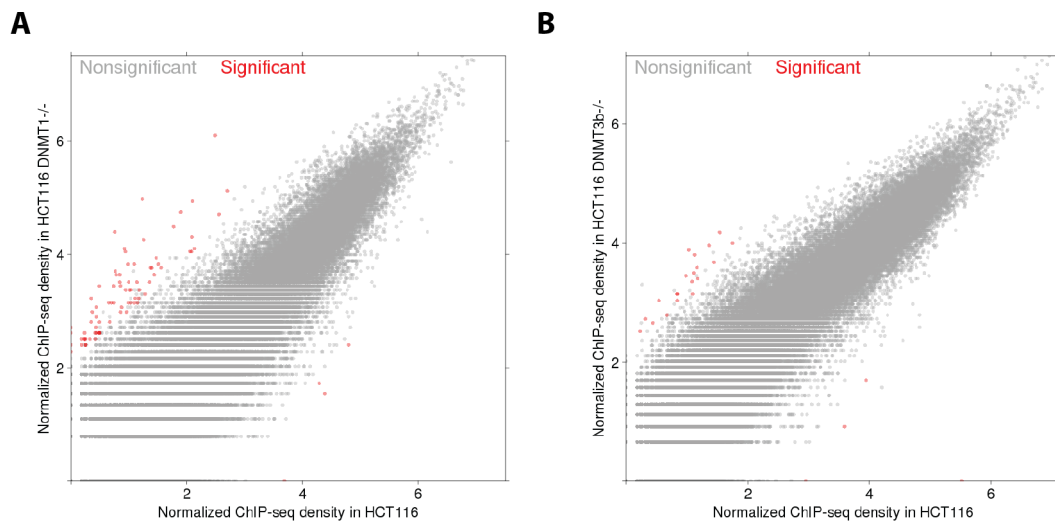


Figure 7-4 Lack of CTCF reactivation in HCT116 DNMT1 (A) and DNMT3B (B) ^{-/-} cells.

Red indicates significant difference in occupancy. Cells lacking DNMT3B demonstrated 20 upregulated sites and 4 downregulated sites, while cells lacking DNMT1 demonstrated 77 upregulated sites and 4 downregulated sites at 1% FDR.

CTCF class	HCT116 mean proportion of methylation	# sites >25% methylation in HCT116	# monitored sites	# sites with CpG in core 44 bp	# sites in CpG islands
preexisting	0.12	680	5065	4807	3190
downregulated	0.36	3	7	6	1
upregulated	0.55	96	138	137	89
reactivated	0.85	647	675	670	439
unreactivated	0.44	1688	3323	3129	2111
sequence matches	0.78	7460	8293	7417	2101
random locations	0.82	5	6	4	0

Table 7-3 Sites susceptible to reactivation.

Methylation levels and CpG content at potential CTCF binding sites. Shown are sites for which methylation data are available in HCT116.

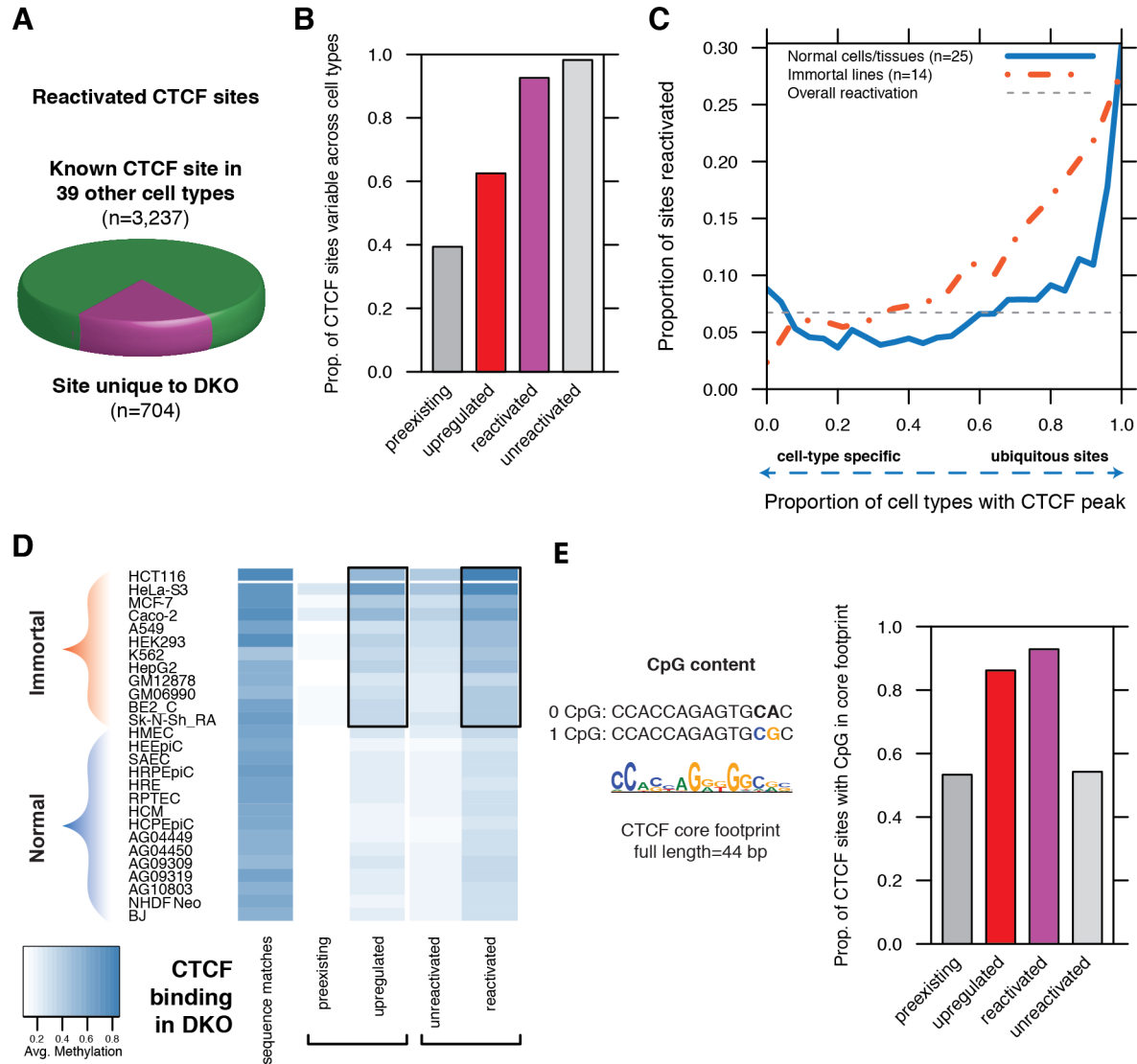


Figure 7-5 Reactivation of predetermined lineage-specific sites.

- (A) Most altered CTCF binding in DKO is at sites that vary in other cell types rather than novel binding to sequence matches.
- (B) Methylation-sensitive sites demonstrate high epigenetic instability across cell types (y axis; see Section 7.6).
- (C) Higher reactivation frequency (y axis, reactivated sites as a proportion of both reactivated and reactivated sites) at sites found in all normal cell types but silenced in HCT116 (solid line, right), as well as immortal-only sites not found in normal lines (solid line, left).
- (D) Reactivated and upregulated sites are distinguished by increased methylation in immortalized lines (boxes).
- (E) The vast majority of reactivated and upregulated sites overlap a CpG in their recognition sequence.

CTCF class	# sites	H3K4me3		H3K27ac		DNase I	
		HCT116	DKO	HCT116	DKO	HCT116	DKO
preexisting	59915	10219	7796	6144	6885	47755	42078
downregulated	158	13	2	7	2	148	100
upregulated	809	78	189	35	89	462	666
reactivated	3237	91	705	36	293	544	1789
unreactivated	43176	4752	4961	3111	4240	11106	13856
sequence matches	176630	1833	2769	1092	2333	13442	22754
random locations	7486	33	18	26	43	176	338

Table 7-4 Qualitative chromatin dynamics at CTCF sites.

For each class the total number of CTCF sites is shown, in addition to the subset of sites with H3K4me3, H3K27ac and DNase I peaks in HCT116 and/or DKO.

Cell type	Experiment	DS#	# of non-redundant tags	#Peaks	SPOT Score	Published?	Accession
HCT116	H3K4me3	DS16055	20347444	25747	0.6379	Thurman et al	GSE50610 (originally released in GSE35583)
HCT116	H3K4me3	DS16056	17659687	26088	0.6978	Thurman et al	GSE50610 (originally released in GSE35583)
DKO	H3K4me3	DS22869	8848852	25567	0.7665	Novel	GSE50610
HCT116	H3K27ac	DS22861	10270899	24658	0.3023	Novel	GSE50610
DKO	H3K27ac	DS22870	7700567	40502	0.465	Novel	GSE50610

Table 7-5 Summary of H3K4me3 and H3K27ac ChIP-seq experiments.

Experiment indicates the antibody used. Number of tags represents the total uniquely-mapped reads (excluding PCR duplicates) for all biological replicates of each cell type. Number of peaks refers to the number of peaks using hotspot. 2 K562 and all HCT116 H3K4me3 data has been published (Thurman 2012) and represents ENCODE clone. SPOT score indicates data quality.

Cell type	Experiment	# of non-redundant tags	#DHSs	SPOT Score	Published?	Accession
HCT116	DNaseI	23562622	114051	0.4056	Thurman et al	GSE50610 (originally released in GSE29692)
HCT116	DNaseI	19730809	113968	0.4544	Thurman et al	GSE50610 (originally released in GSE29692)
HCT116	DNaseI	26263533	104193	0.3274	Novel	GSE50610
DKO	DNaseI	14353151	174936	0.3294	Novel	GSE50610

Table 7-6 Summary of DNaseI experiments.

Number of tags represents the total uniquely-mapped reads (excluding PCR duplicates) for all biological replicates of each cell type. #DHSs represents 1% FDR hotspots in HCT116, 0.5% hotspots in DKO. SPOT score indicates data quality.

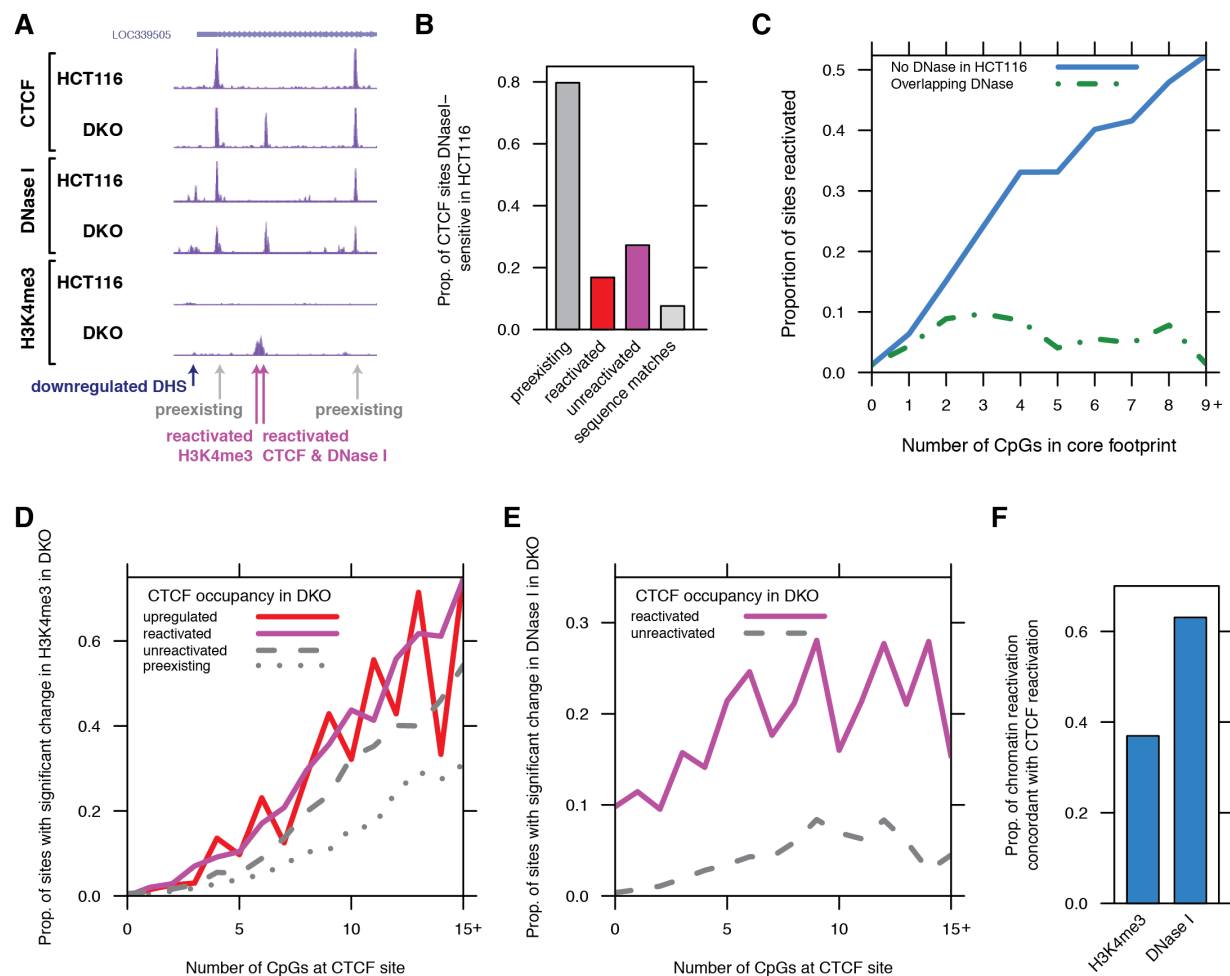


Figure 7-6 Chromatin dynamics at CTCF sites

(A) Profiling of CTCF, DNase I, and H3K4me3 reactivation in HCT116 and DKO.

(B) Proportion of CTCF sites overlapping DNase I hypersensitive sites (DHSs). Although preexisting CTCF sites are stereotypically hypersensitive to DNase I cleavage, reactivation upon demethylation occurs at sites inaccessible in HCT116.

(C) Reactivation frequency parallels local CpG content (solid line); while the presence of a preexisting DHS is associated with a near-total lack of reactivation regardless of CpG content (dashed line).

D-E) Increased CpG content (x axis) is strongly associated with reactivation of H3K4me3 (D) but not DNase I (E).

Shown are sites without H3K4me3 peak (D) or DHS and CTCF peak (E) in HCT116. Note strong reactivation of H3K4me3 at unreactivated CTCF sites, in contrast to lack of increase in DNase I at these sites.

(F) Reactivation of DNase I at CTCF sites coincides frequently with reactivation of CTCF occupancy, while H3K4me3 is more likely to increase without concordant CTCF reactivation.

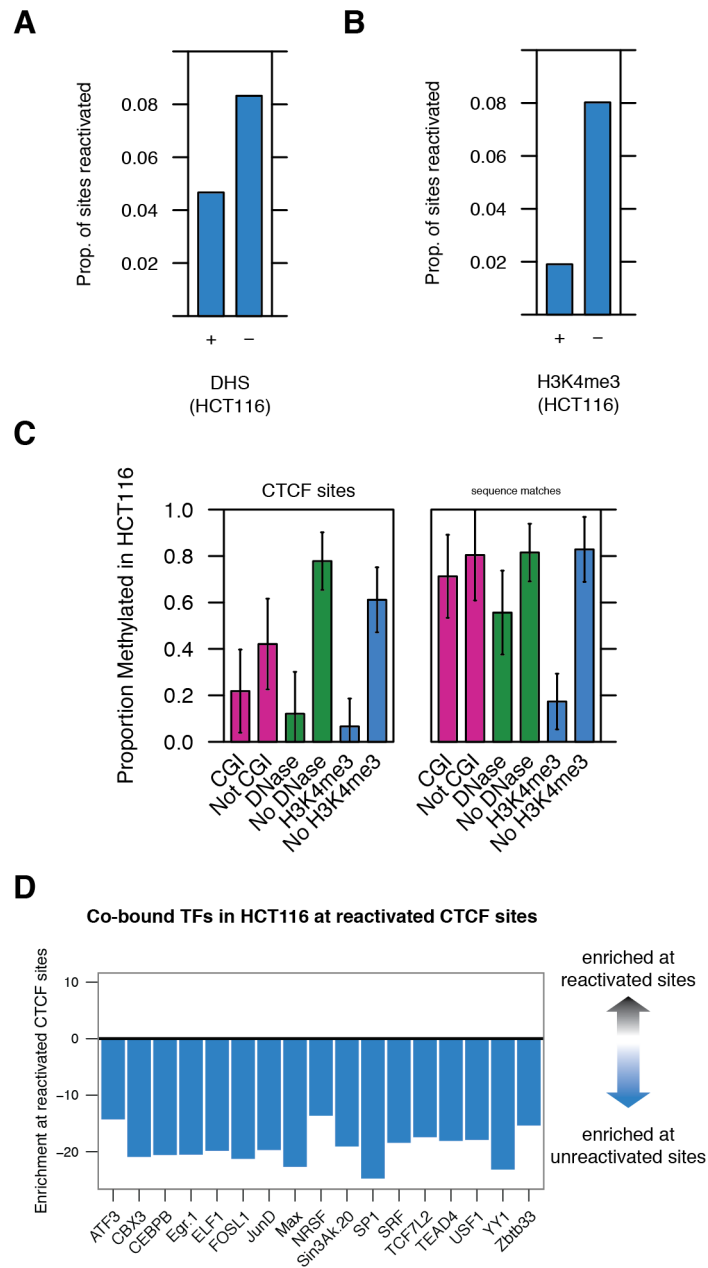


Figure 7-7 Chromatin characteristics of response to demethylation.

(A-B) Reactivation frequency (reactivated sites as a proportion of both reactivated and reactivated sites) of CTCF sites outside or inside a DHS (A) or H3K4me3 peak (B).

(C) Methylation levels at sites grouped by CTCF reactivation (panels) and genomic features (x axis). The presence of a CpG island, DHSs or H3K4me3 peak is associated with reduced levels of methylation, though the relationship is strongest for DNase I and H3K4me3.

(D) Reactivated CTCF sites are depleted for co-occupancy of common TFs.

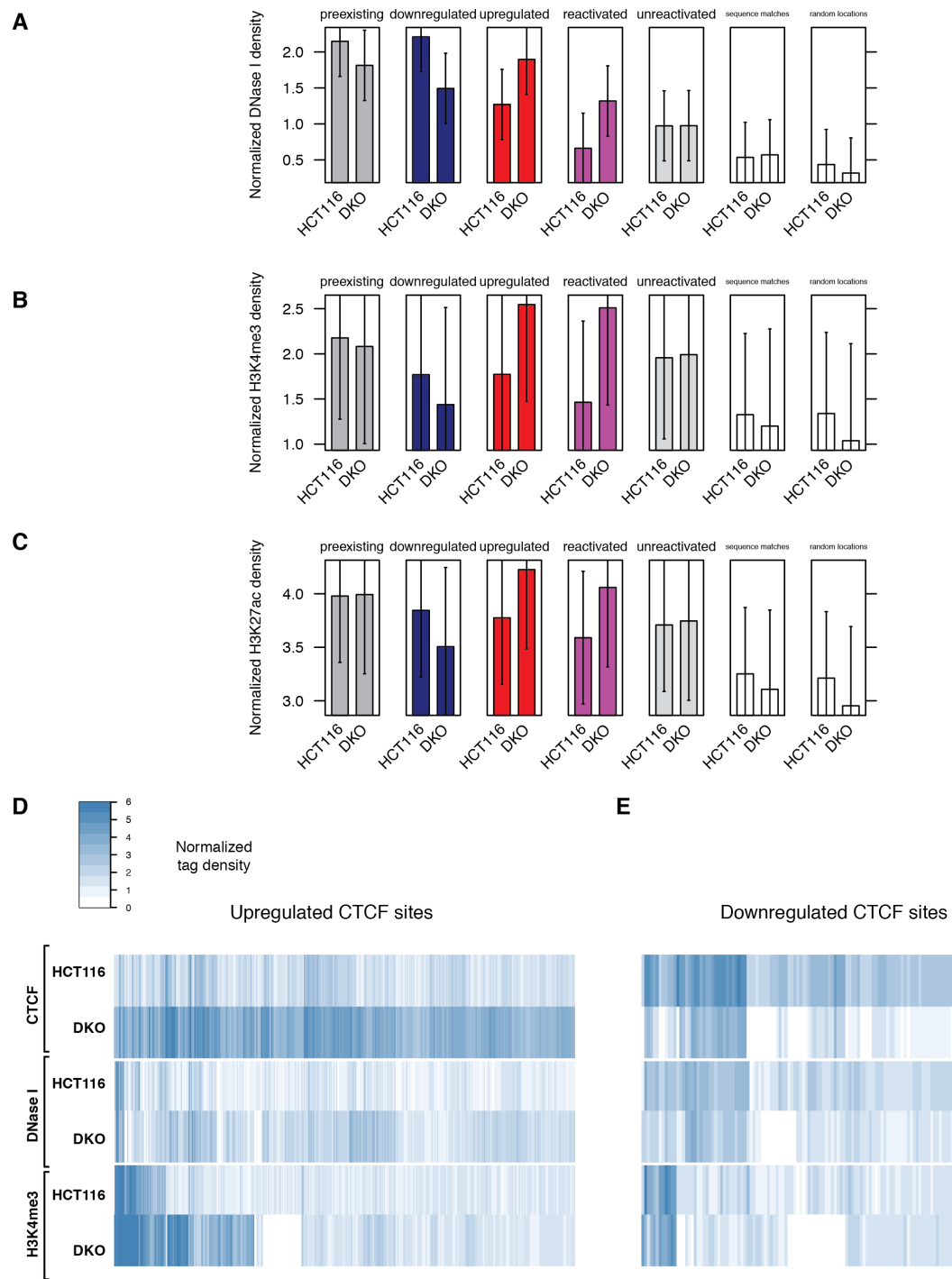


Figure 7-8 Chromatin dynamics at demethylated CTCF sites.

(A-C) Quantitative comparison of mean DNase I (A), H3K4me3 (B) and H3K27ac (C) density across each set of sites in HCT116 and DKO cells.

(D-E) Normalized CTCF, DNaseI and H3K4me3 signal at sites upregulated (D) or downregulated (E) for CTCF signal.

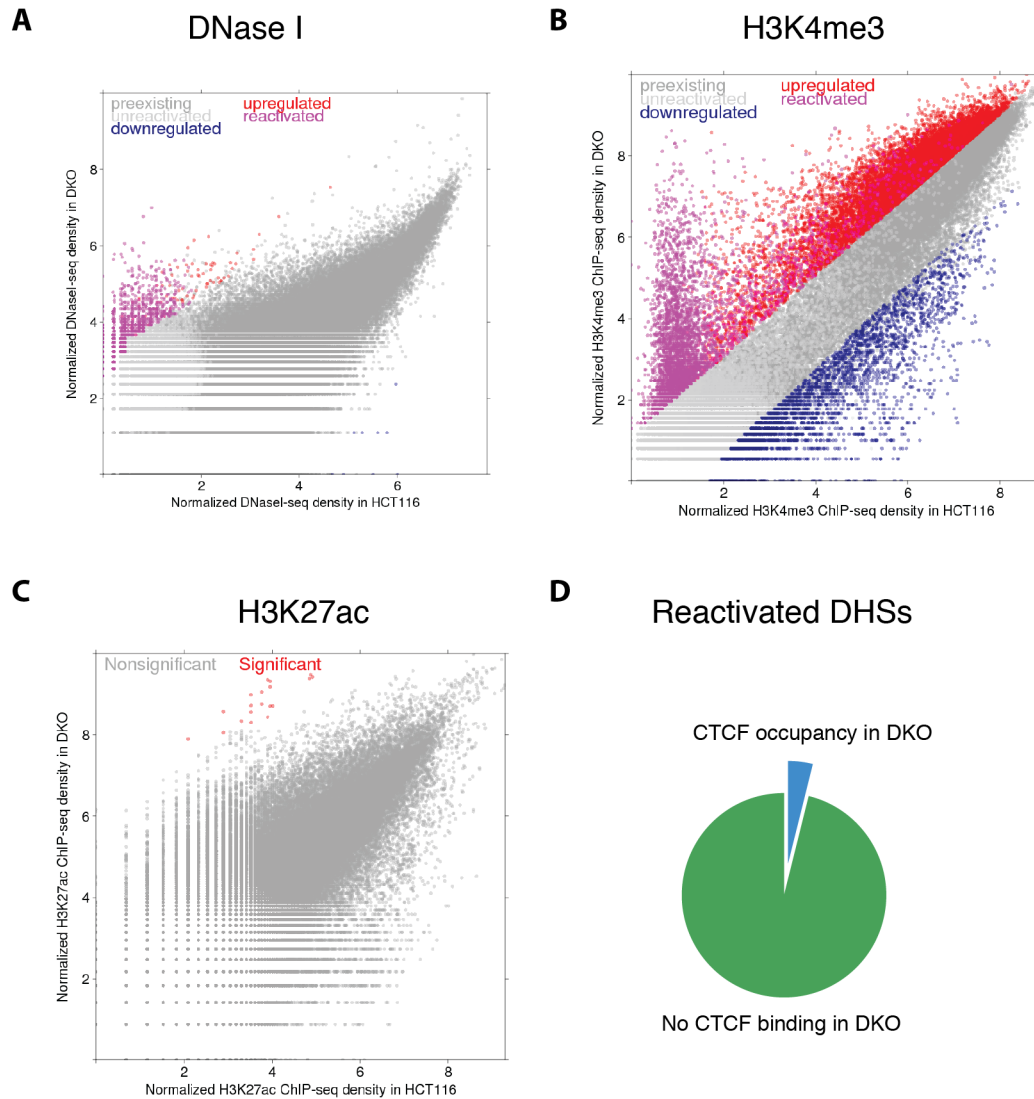


Figure 7-9 Global reactivation of chromatin structure.

(A-C) Comparison of genome-wide DNaseI (A), H3K4me3 (B) and H3K27ac (C) density between wild type HCT116 and DKO at DHSs. Red indicates significant difference in density.

(D) Broad extent of chromatin reactivation beyond CTCF sites.

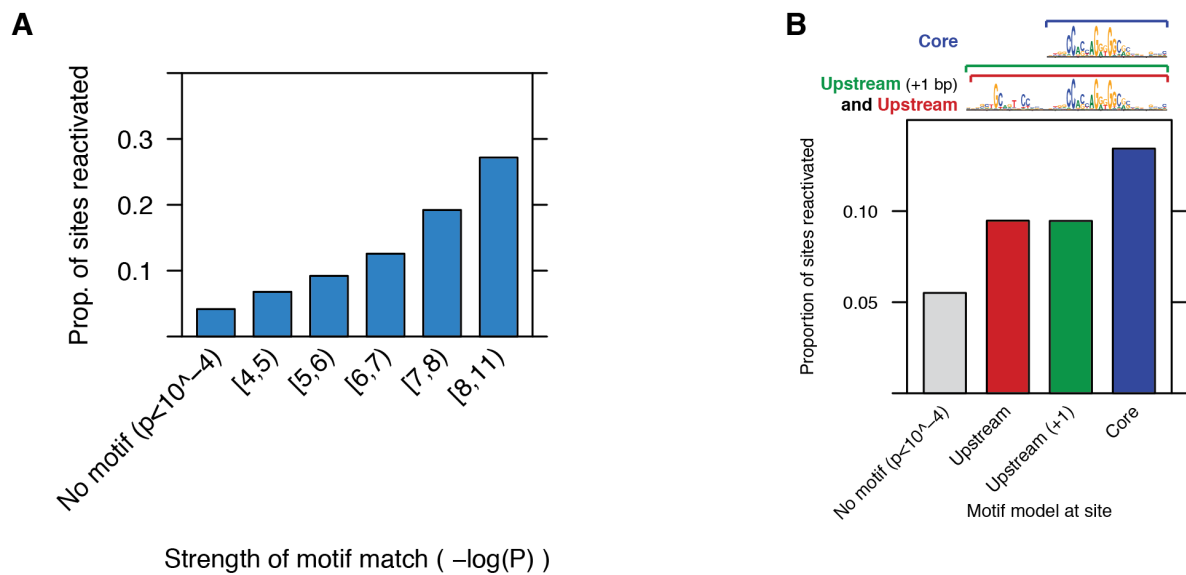


Figure 7-10 Reactivation frequency relates to sequence characteristics of the CTCF recognition sequence. Reactivation frequency by strength of motif match (A) binding mode (B).

Sample	# of tags	# of tags on target	# of non-redundant tags	#CpGs covered >8x	cov-	#CTCF sites covered >8x	Published?	Accession
K562-Mock	153040607	101382521	75931601	1665046		19775	Novel	GSE50610
K562-aza	136469310	98742701	29584174	813891		12860	Novel	GSE50610

Table 7-7 Summary of bisulfite experiments.

Number of tags refers to the total mapped tags; # of tags on target refers to the capture regions; # of non-redundant tags represents the total uniquely-mapped reads excluding PCR duplicates. Number of CTCF sites counts the number of overlapping at least one CpG with >8x read depth.

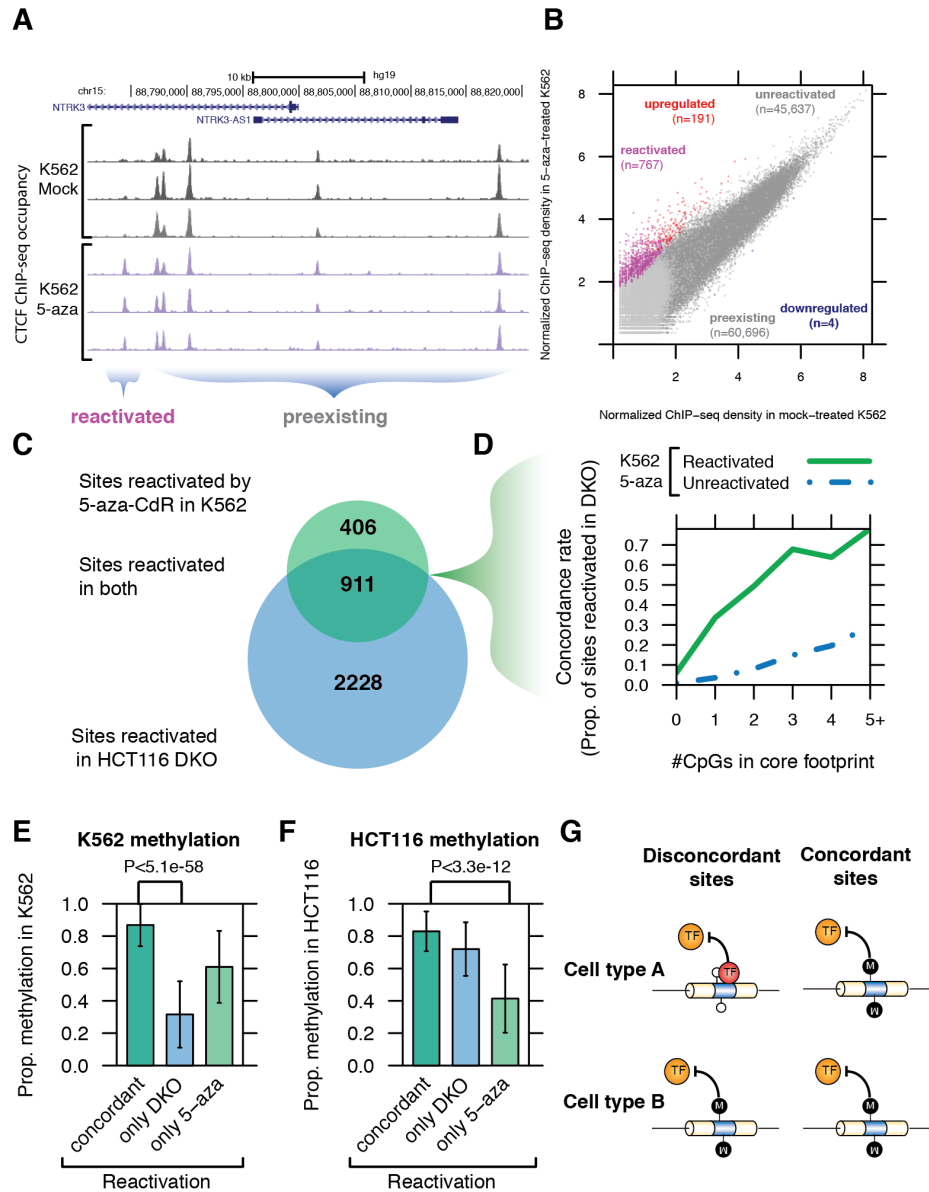


Figure 7-12 Cell type-specific reactivation highlights methylation-independent repression.

(A) Reactivation at NTRK3 locus showing reactivated and preexisting sites in a mock- and 5-aza-CdR-treated K562 cells.

(B) Limited but reproducible reactivation (FDR 1%). Sites were classified (labels) based on presence of peak in mock-treated K562 and significant differential occupancy in treated cells.

(C) Majority of 5-aza-CdR-reactivated sites are concordantly reactivated in DKO (of a potential 46,726 sites not occupied in both K562 and HCT116, FDR 5%).

(D) Sites reactivated in K562 (solid line) are more likely to be reactivated in DKO (y axis) than unreactivated sites (dashed line) regardless of CpG content.

(E) Reduced preexisting methylation (y axis) in both K562 (left) and HCT116 (right) distinguishes sites reactivated specifically by genetic or chemical means, suggesting that non-concordant sites are attributable to methylation-independent silencing. P-values by Wilcoxon signed-rank test.

(F) Concordant reactivation implies methylation-dependent silencing across cell types, while discordant reactivation implies methylation-independent silencing in the unreactivated cell type.

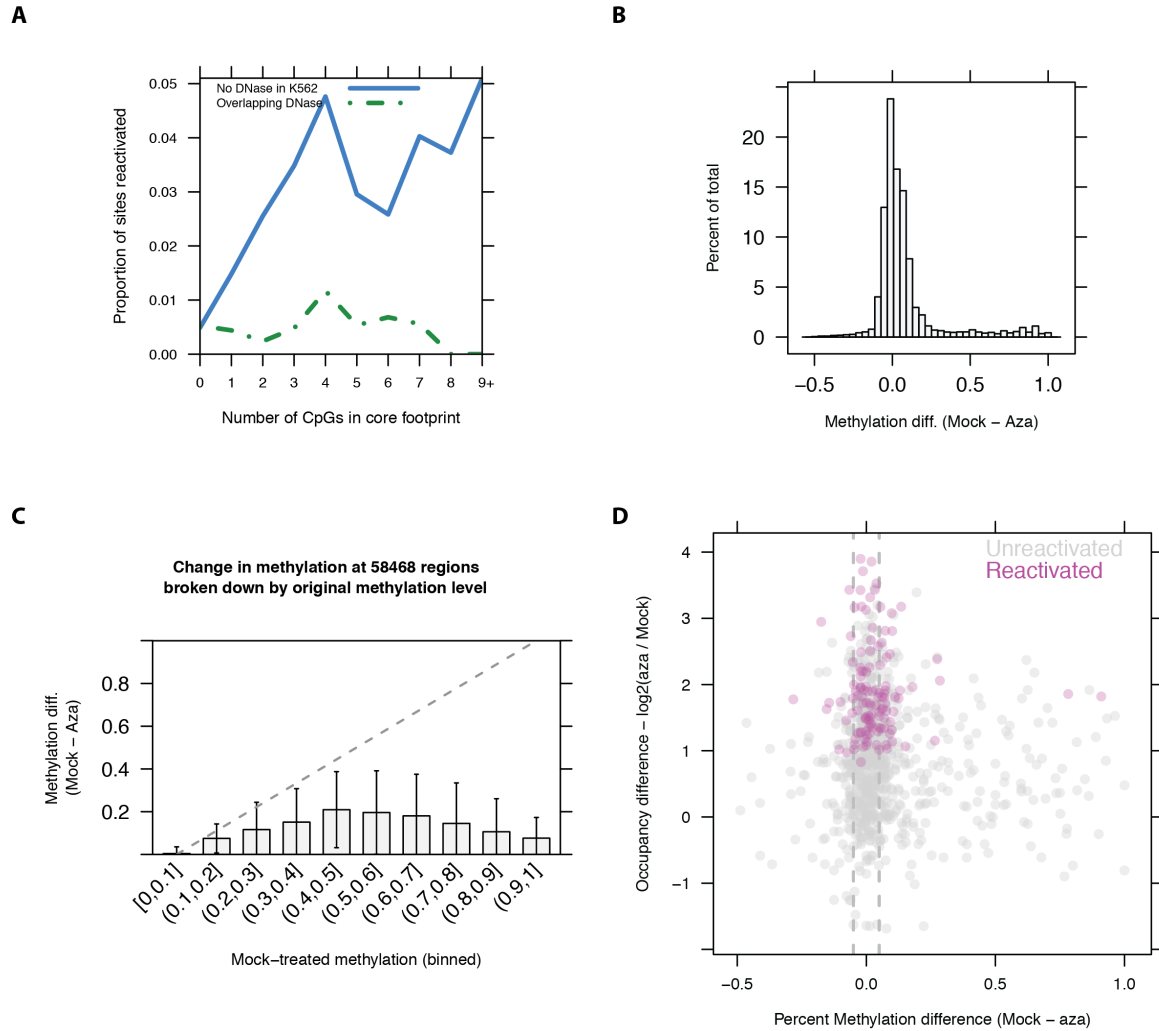


Figure 7-13 Transient depletion of DNA methylation using 5-aza-2'-deoxycytidine.

(A) Reactivation frequency parallels local CpG content.

(B) Difference in methylation upon treatment with 5-aza-CdR at 11456 regions of preexisting (>50%) methylation in K562.

(C) Change in methylation upon 5-aza-CdR treatment for sites with varying levels of existing methylation in K562. Reduction of methylation is nearly complete at moderately methylated CpGs (<50%), but absolute change in methylation is reduced for sites demonstrating >50% methylation.

(D) Comparison between occupancy differences and methylation differences. Dashed vertical lines represent ± 0.05 difference in proportion of methylation.

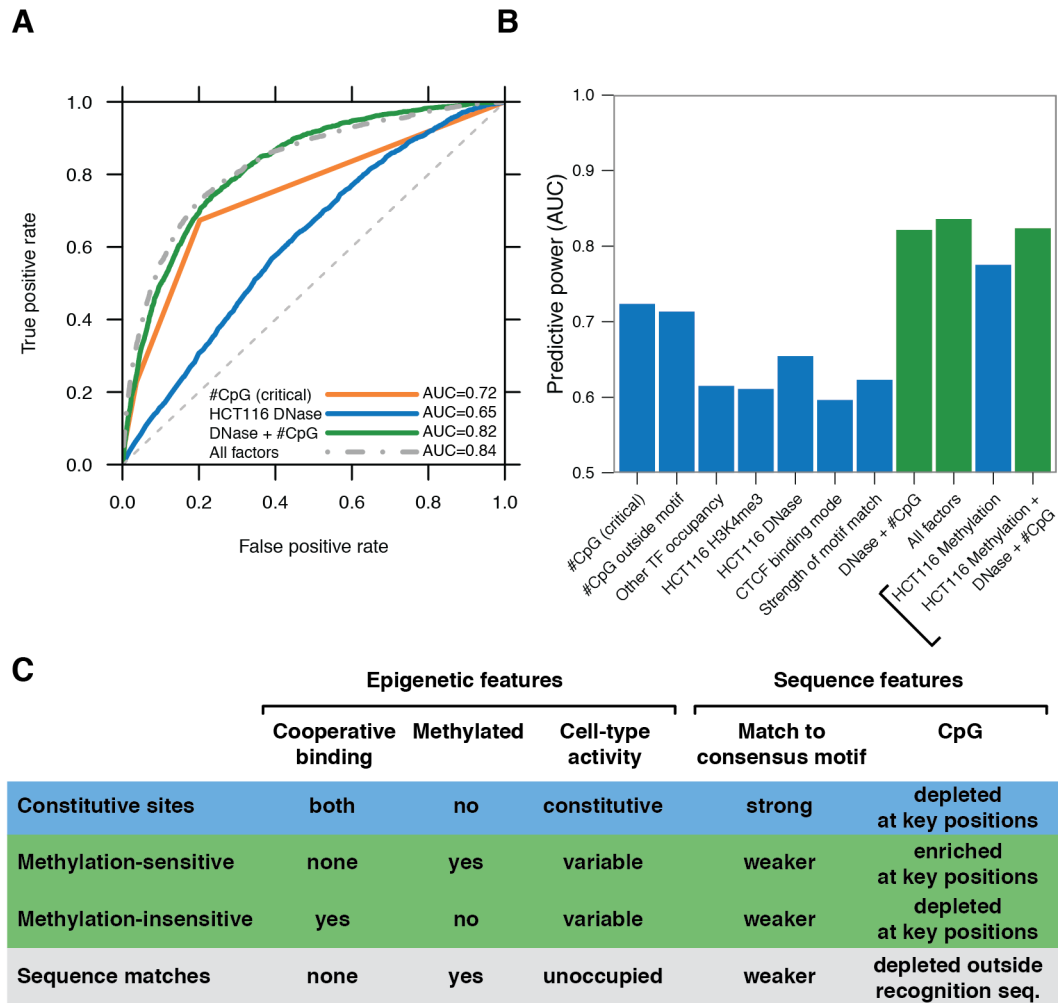


Figure 7-14 Recognition of labile methylation-sensitive sites

(A) Predictive power of features to distinguish reactivated from unreactivated sites, illustrated by receiver-operator curves (ROC). Area under the curve (AUC) summarizes overall predictive power. Dashed gray line indicates a random predictor and has an AUC of 0.5. A perfect predictor would be plotted as a right angle.

(B) Detailed predictive power (y axis) of selected genomic features. Models including methylation (bracketed) measured only at sites with RRBS data. Green bars indicate linear models combining multiple factors. Other TF occupancy refers to the sum of HCT116 ENCODE TF ChIP-seq track densities.

(C) CTCF sites exhibit classes of sensitivity to methylation.

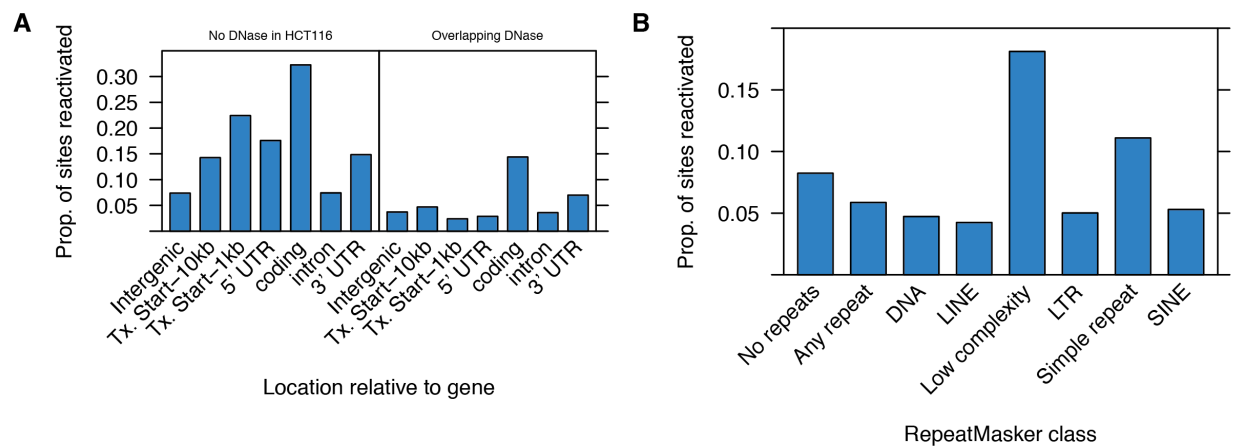


Figure 7-15 Genomic characteristics associated with reactivation.
 Reactivation frequency by genic location **(A)**, repeat class **(B)**.

7.6 Materials and Methods

Cell culture and ChIP-seq/DNaseI-seq

HCT116 and DKO cells have been previously described (Rhee et al., 2002). Cells were cultured in appropriate medium and maintained in a humidified incubator at 37°C in the presence of 5% CO₂. (Table S1). K562 cells were treated with 5-aza-2'-deoxycytidine (decitabine; Sigma, A3656) dissolved in DMSO to 10mM. Drug was administered at 1uM daily for 3 days. Control K562 cells were mock-treated with DMSO. Conditions for other cell types as described (Table 7-1).

Chromatin immunoprecipitations were performed as described for CTCF (Wang et al., 2012) and H3K4me3 and H3K27ac (Thurman et al., 2012). Briefly, cells were crosslinked for 10 min in 1% formaldehyde and quenched in 125 mM glycine. Chromatin was sheared by Bioruptor (Diagenode) and incubated with antibody conjugated to Dynabeads (M-280, Invitrogen). After reversing crosslinks, immunoprecipitated DNA was purified by phenol-chloroform-isoamyl alcohol extraction and ethanol precipitation. DNase was performed as described (John et al., 2011; Thurman et al., 2012), whereby small (300-1000 bp) fragments are isolated from lysed nuclei following DNase treatment. Libraries generated from immunoprecipitated or DNase-treated DNA were sequenced on an Illumina Genome Analyzer IIX or HiSeq 2000 by the High-Throughput Genomics Center (University of Washington) according to a standard protocol. Most experiments were performed on two independent biological replicates (Table 7-2, Table 7-5, Table 7-6).

Data processing

ChIP-seq and DNase data were mapped to the human genome (GRCh37/ hg19) using bowtie (Langmead et al., 2009) with the options “bowtie -n 2 -m 1 -e 70 --best” allowing up to three mismatches. Reads mapping to multiple locations were then excluded, and reads with identical 5' ends and strand were presumed to be PCR duplicates and were excluded using Picard MarkDuplicates. Smoothed density tracks were generated using bedmap (<http://code.google.com/p/bedops/>) to count the number of tags overlapping a sliding 150-bp window, with a resolution of 20 bp (Neph et al., 2012a). Density tracks for display were normalized for sequencing depth by a global linear scaling to 10 million tags.

CTCF data from 39 cell types was processed as previously (Wang et al., 2012). Briefly, a master peak list was established from ENCODE project 2% IDR-processed SPP peak calls (Kharchenko et al., 2008; Li et al., 2011)

(ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/peaks/june2012/spp/optimal/), with locations adjusted to center on matches to the nearest CTCF motif ($P < 10^{-4}$, fimo) if the motif was within 50 bp. Peak presence in a given cell type was established by the presence of a 1% FDR hotspot (John et al., 2011) (also <http://www.uwencode.org/proj/hotspot/>) in any replicate. Peaks only found in DKO and/or 5-aza-CdR-treated K562 cells were not included in total peak counts to maximize their generality. We identified possible further CTCF binding sequences using the program fimo ($P < 10^{-5}$), as well as random sequences without a CTCF binding sequence ($P > 10^{-4}$). Both sets were required to be further than 500 bp from any known *in vivo* CTCF site, to consist of base pairs 90% of which are uniquely mappable by 36 bp reads, and not to be on the ENCODE blacklist (ENCODE Project Consortium, 2012). We identified DNase I hypersensitive sites by the presence of a hotspot (FDR 1%) in any HCT116 replicate or a hotspot (FDR 0.5%) in DKO; novel sites in DKO were included in the analysis if they overlapped no unthresholded hotspot in any HCT116 replicate. H3K4me3 and H3K27ac regions were called using hotspot.

At both CTCF sites and DHSs separately, we measured CTCF occupancy or DNaseI accessibility by the number of tags overlapping the 150-bp region, and H3K4me3 and H3K27ac density by tags overlapping the 2150-bp region. We then used the package DESeq (Anders and Huber, 2010) to identify significant differences. CTCF ChIP-seq peaks and DHSs discordant between the two previously published replicates and the new replicate for HCT116 (FDR 10%) were removed. Pairwise differences were called between samples for each ChIP-seq and DNase I data set using `estimateDispersions(method="pooled", sharingMode="maximum", fitType="local")` and `nbinomTest`. After observing a near-complete lack of reactivation at sequence matches and random sites (<10 sites), we subsequently excluded these sites and recomputed pairwise differences. We applied the Benjamini-Hochberg procedure to control for multiple testing. To obtain normalized data at all sites, we used `estimateDispersions` as before but with `fitType="parametric"`, and then

applied a variance stabilizing transformation and averaged the occupancy of all replicates. These values were scaled to [0, 10] by subtracting the global minimum and dividing by the global maximum * 10.

We obtained location of CTCF sites relative to genes using RefSeq and relative to CpG islands and repetitive regions in RepeatMasker from the UCSC Genome Browser.

Classification of CTCF motif models

CTCF binds in a multivalent fashion, whereby three modes of binding are distinguished by the presence and position of an upstream motif (Nakahashi et al., 2013). At each site we chose the best-matching (by fimo P-value) of the three motif models.

Per-nt regression model

We used the `lm()` function in R to perform a binomial regression considering all preexisting, upregulated, unreactivated and reactivated CTCF sites with a recognizable motif:

$$\text{reactivation} \sim c + x_{-22} + \dots + x_{21} + \text{upstream} + (\text{\#CpGs in flanking region})$$

Upstream represents an indicator variable for the presence of either upstream motif. Multiple testing correction was performed using the Bonferroni method.

ROC and PR curves were computed using the R package ROCR. Binomial models for multiple factors were constructed using the `lm()` function in R.

Monitoring methylation

Purified DNA from mock- and 5-aza-CdR-treated K562 cells was fragmented following the Agilent SureSelect Methyl-Seq protocol with slight modifications (Table 7-6). 4 ug from each of the samples was fragmented in a Covaris S2 (Covaris) under the following conditions: 10% duty cycle at intensity 5 for 60 seconds with 200 cycles per burst for 6 cycles with mode set to sweeping. The 250 bp fragmented DNA was then end repaired and adenylated followed by ligation to adapters synthesized with 5' –methyl-cytosine instead of cytosine. The adaptor-ligated library was purified using AMPure XP beads (Beckman Coulter Genomics). 500 ng of each library was hybridized to Agilent SureSelect Methyl-Seq biotinylated RNA baits (84 Mbp) for 24 h at

65 °C. The biotinylated probe/target hybrids were captured on Dynal MyOne Streptavidin T1 (Invitrogen), washed, eluted, and desalted following purification on a MinElute PCR column (Qiagen) as described in the SureSelect protocol.

Bisulfite conversion of the purified captured library was performed using the EZ DNA Methylation Gold Kit (Zymo Research) as per manufacturer's instructions. The bisulfite converted captured library was amplified by PCR with minimal amount of PCR cycles, then purified using AMPure XP beads and quantified by Qubit dsDNA assay (Invitrogen) following SureSelect protocol.

Samples were diluted to a working concentration of 10 nm. Cluster generation was performed for each sample and loaded on to single lane of an Illumina HiSeq flowcell. Single end sequencing was performed for 36 cycles according to manufacturer's instructions.

Capture bisulfite reads (Table 7-7) were mapped to the human genome (hg19) using Bismark (Krueger and Andrews, 2011) and bowtie2 beta 6 (Langmead et al., 2009) with the options “-n 1” and excluding duplicate reads and combining reads from the top and bottom strands. Methylation was monitored at CpGs with ≥ 8 read coverage, and averaged for all CpGs within the 150 bp window at each CTCF site. We obtained $\geq 8x$ coverage in both samples for 658,010 CpGs.

RRBS data for 29 cell types is available from the UCSC Genome Browser (Varley et al., 2013), processed as described (Wang et al., 2012).

7.7 Notes

This work was performed in collaboration with Hao Wang, Sam John, Anthony Shafer, Theresa Canfield, Kristen Lee, and John A. Stamatoyannopoulos.

Data access

Data have been submitted to the NCBI Gene Expression Omnibus (GEO) (<http://www.ncbi.nlm.nih.gov/geo/>) under accession nos. GSE30263 and GSE50610 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=hjypvqoommwekjo&acc=GSE50610>). Data sets are also available for viewing in the UCSC Genome Browser (<http://genome.ucsc.edu/>).

Acknowledgements

We thank Bert Vogelstein for graciously sharing HCT116 DNMT1/3B DKO cells. We thank Bob Thurman for informatics support. This work was supported by NIH fellowship F31MH094073 to M.T.M. and grant U54HG004592 to J.A.S..

8 Systematic Localization Of Common Disease-Associated Variation In Regulatory DNA

8.1 Abstract

Genome-wide association studies (GWAS) have identified many non-coding variants associated with common diseases and traits. We show that these variants are concentrated in regulatory DNA marked by DNase I hypersensitive sites (DHSs). 88% of such DHSs are active during fetal development, and are enriched for gestational exposure-related phenotypes. We identify distant gene targets for hundreds of DHSs that may explain phenotype associations. Disease-associated variants systematically perturb transcription factor recognition sequences, frequently alter allelic chromatin states, and form regulatory networks. We also demonstrate tissue-selective enrichment of more weakly disease-associated variants within DHSs, and the *de novo* identification of pathogenic cell types for Crohn's disease, multiple sclerosis, and an electrocardiogram trait, without prior knowledge of physiological mechanisms. Our results suggest pervasive involvement of regulatory DNA variation in common human disease, and provide pathogenic insights into diverse disorders.

8.2 Results

Disease- and trait-associated genetic variants are rapidly being identified with genome-wide association studies (GWAS) and related strategies (Stranger et al., 2011). To date, hundreds of GWAS have been conducted, spanning diverse diseases and quantitative phenotypes (Hindorff et al.) (Figure 8-1A). However, the majority (~93%) of disease- and trait-associated variants emerging from these studies lie within non-coding sequence (Figure 8-1B), complicating their functional evaluation. Several lines of evidence suggest involvement of a proportion of such variants in transcriptional regulatory mechanisms, including modulation of promoter and enhancer elements (Cookson et al., 2009; Harismendy et al., 2011; Musunuru et al., 2010; Pomerantz et al., 2009), and enrichment within expression quantitative trait loci (eQTL).

Human regulatory DNA encompasses a variety of *cis*-regulatory elements within which the cooperative binding of transcription factors creates focal alterations in chromatin structure. DNase I hypersensitive sites (DHSs) are sensitive and precise markers of this actuated regulatory DNA, and DNase I mapping has been instrumental in the discovery and census of human *cis*-regulatory elements (Gross and Garrard, 1988). We performed DNase I mapping genome-wide (Thurman et al., 2012) in 349 cell and tissue samples including 85 cell types studied under the ENCODE Project (Thurman et al., 2012) and 264 samples studied under the Roadmap Epigenomics Program (Bernstein et al., 2010a). These encompass several classes of cell types including cultured primary cells with limited proliferative potential (n=55); cultured immortalized (n=6), malignancy-derived (n=18) or pluripotent (n=2) cell lines; and primary hematopoietic cells (n=4) as well as purified differentiated hematopoietic cells (n=11), and a variety of multipotent progenitor and pluripotent cells (n=19). We also surveyed regulatory DNA by generating DHS maps from 233 diverse fetal tissue samples across post-conception days ~60-160 (late-first to late-second trimester of gestation). We used a uniform processing algorithm to identify DHSs and the surrounding boundaries of DNase I accessibility (i.e., the nucleosome-free region harboring regulatory factors) (Section 8.5). We defined an average of 198,180 DHSs per cell type (range 89,526-369,920; Table 8-1) spanning on average ~2.1% of the genome. In total, we identified 3,899,693 distinct DHS positions along the genome (collectively spanning 42.2%), each of which was detected in one or more cell/tissue types (median= 5).

Disease- and trait-associated variants are concentrated in regulatory DNA

We examined the distribution of 5,654 non-coding genome-wide significant associations (5,134 unique SNPs; Figure 8-1) for 207 diseases and 447 quantitative traits (Hindorff et al.) with the deep genome-scale maps of regulatory DNA marked by DHSs. This revealed a collective 40% enrichment of GWAS SNPs in DHSs (Figure 8-1C, $P < 10^{-55}$, binomial, compared to the distribution of HapMap SNPs). Fully 76.6% of all non-coding GWAS SNPs either lie within a DHS (57.1%, 2,931 SNPs) or are in complete linkage disequilibrium (LD) with SNPs in a nearby DHS (19.5%, 999 SNPs) (Figure 8-2A) (Section 8.5). To confirm this enrichment, we sampled variants from the 1000 Genomes Project (1000 Genomes Project Consortium, 2010) with the same genomic feature localization (intronic vs. intergenic), distance from the nearest transcriptional start site, and allele frequency in individuals of European ancestry. We confirmed significant enrichment both for SNPs within DHSs ($P < 10^{-59}$, simulation) and also including variants in complete LD ($r^2=1$) with SNPs in DHSs ($P < 10^{-37}$, simulation) (Figure 8-3).

In total, 47.5% of GWAS SNPs fall within gene bodies (Figure 8-2B); however, only 10.9% of intronic GWAS SNPs within DHSs are in strong LD ($r^2 \geq 0.8$) with a coding SNP, indicating that the vast majority of non-coding genic variants are not simply tagging coding sequence. Analogously, only 16.3% of GWAS variants within coding sequences are in strong LD with variants in DHSs. We noted that SNPs on widely used genotyping arrays (e.g., Affymetrix) were modestly enriched within DHSs (Figure 8-3), possibly due to selection of SNPs with robust experimental performance in genotyping assays. However, we found no evidence for sequence composition bias (Table 8-2).

To further examine the enrichment of GWAS SNPs in regulatory DNA, we systematically classified all non-coding GWAS SNPs by the quality of their experimental replication. This disclosed 2,436 unreplicated SNPs; 2,374 'internally-replicated' SNPs (confirmed in a second population in the initial publication); and 324 'externally-replicated' SNPs (confirmed in an independent study) (Section 8.5). We observed a monotonic increase in the proportion of disease/trait variants localizing in DHSs with increasing quality of GWAS SNP experimental replication (Figure 8-2B), as well as with increasing strength of association and study sample size (Figure 8-4). For externally replicated non-coding SNPs, 69.8% lie within a DHS ($n=226$, $P < 10^{-14}$, simula-

tion, Figure 8-3). To exclude the influence of population stratification, we compared the fixation index in African and European populations between GWAS SNPs in DHSs and matched SNPs not in DHSs and found them to be nearly identical ($F_{ST}=0.0843$ vs. 0.0847 , respectively) (Section 8.5). The monotonic relationship between evidence for association and SNP concentration in DHSs strongly suggests that many variants are functional and that unreplicated or weaker associations may obscure the true degree of enrichment in DHSs.

GWAS variants localize in cell- and developmental stage-selective regulatory DNA

We observed selective localization within physiologically or pathogenically-relevant specific cell or tissue types, including affected tissues or known or posited effector cell types (Figure 8-2C). For a given disorder, cell-selective localization within physiologically or pathogenically-relevant cell types was repeatedly observed for multiple independently-associated SNPs distributed widely around the genome (Figure 8-5). These results suggest a tissue-specific regulatory role for many common variants, as well as the potential for comprehensive regulatory DNA maps to illuminate associations within disease-relevant cell types.

Many common disorders have been linked with early gestational exposures or environmental insults (Symonds et al., 2009). Because of the known role of the chromatin accessibility landscape in mediating responses to cellular exposures such as hormones (John et al., 2011), we examined if DHSs harboring GWAS variants were active during fetal developmental stages. Of 2,931 non-coding disease- and trait-associated SNPs within DHSs globally, 88.1% (2,583) lie within DHSs active in fetal cells and tissues. 57.8% of DHSs containing disease-associated variation are first detected in fetal cells and tissues and persist in adult cells ('fetal origin' DHSs), while 30.3% are fetal stage-specific DHSs (Figure 8-6). GWAS variants in adult stage-specific DHSs localize chiefly in mature hematopoietic cells, connective tissue, endothelial cells, and malignant cells (Figure 8-7).

We next analyzed the enrichment or depletion of replicated disease-specific GWAS variants in fetal stage DHSs relative to the proportion of total GWAS SNPs in these DHSs. We found the greatest enrichment in phenotypes for which gestational exposures or growth trajectory have been shown to play major roles, including menarche, cardiovascular disease, and body mass index (Figure 8-8) (Barker et al., 1993; Symonds et al.,

2009). By contrast, we observed relative depletion in fetal DHSs of aging-related diseases, cancer, and inflammatory disorders with presumed (postnatal) environmental triggers. These findings suggest a recurring connection between an exposure-responsive gestational chromatin landscape, regulatory genotype, and risk for specific classes of adult diseases and traits.

DHSs harboring GWAS variants control distant phenotype-relevant genes

Enhancers may lie at great distances from the gene(s) they control (Maston et al., 2006) and function through long-range regulatory interactions (Lettice et al., 2002; Pomerantz et al., 2009), complicating the identification of target genes of regulatory GWAS variants. Most DHSs display quantitative, cell-selective DNase I hypersensitivity patterns which may be systematically correlated with DNase I sensitivity patterns at *cis*-linked promoters. DHSs that are strongly correlated ($R > 0.7$) with specific promoters function as enhancers that physically interact with their target promoter as detected by chromosome conformation capture methods including 5C and ChIA-PET (Thurman et al., 2012).

To systematically identify the genic targets of DHSs harboring GWAS variants and thereby gain insights into disease mechanisms, we applied the approach described in (Thurman et al., 2012) to the much broader range of cell and tissue types in the present study (Section 8.5), and intersected the result sets with GWAS data. This analysis revealed 419 DHSs harboring GWAS variants that were strongly correlated ($R > 0.7$) with the promoter of a specific target gene within ± 500 kb of the DHS. Among these are numerous examples of target genes that plausibly explain the disease or trait association (Table 8-3). For example, a SNP (rs385893) associated with platelet count (Soranzo et al., 2009) lies in a DHS tightly correlated ($R = 0.97$) and physically interacting with the 222 kb distant promoter of JAK2, a cytokine-activated signal transducer linked with platelet coagulation and myeloproliferative disorders (Figure 8-9). Similarly, a SNP (rs3859192) associated with peripheral leukocyte counts (Crosslin et al., 2011) lies in the body of the GSDMA gene, yet is most strongly correlated with the retinoic acid receptor-alpha gene (RARA), a well-studied regulator of granulocytic and monocytic differentiation that lies 336 kb away (Figure 8-10B). Fully 40.8% of correlated DHS-gene pairs span >250 kb (Figure 8-11), and 79% represent pairings with distant promoters vs. those of the nearest gene

(Table 10-1, Table 10-2). Notably, these interactions typically extend beyond the range of LD (mean $r^2=0.06$; Table 10-1).

GWAS variants in DHSs frequently alter allelic chromatin state

Next, we examined how GWAS variants in DHSs were distributed with respect to transcription factor recognition sequences, defined using a scan for known motif models at a stringency of $P < 10^{-4}$ (Section 8.5). Of GWAS SNPs in DHSs, 93.2% (2,874) overlap a transcription factor recognition sequence Table 8-7. We partitioned GWAS variants into 10 disease/trait classes, and then determined the frequency of GWAS variants associated with a particular disease/trait class that localized within sites for transcription factors independently partitioned into the same classes based on gene ontology annotations (Figure 8-12) (Section 8.5). This analysis revealed that common variants associated with specific diseases or trait classes were systematically enriched in the recognition sequences of transcription factors governing physiological processes relevant to the same classes.

Functional variants that alter transcription factor recognition sequences frequently affect local chromatin structure. At heterozygous SNPs altering transcription factor recognition sequences, altered nuclease accessibility of the chromatin template manifests as an imbalance in the fraction of reads obtained from each allele (Knight et al., 2003) (see also Section 4). As the concentration of sequence reads and highly overlapping read coverage results in an effective re-sequencing of DHSs, we were able both to detect cell types heterozygous for common SNPs and to quantify the relative proportions of reads from each allele across all cell types (Section 8.5). This imbalance is indicative of the functional effect of a particular allele on local chromatin state. We detected 584 heterozygous GWAS SNPs with sufficient sequencing coverage, of which 120 showed significant allelic imbalance in chromatin state (at FDR 5%). We identified sites where regulatory variants were associated with allelic chromatin states, with the predicted higher-affinity allele exhibiting higher accessibility (Figure 8-13). In nearly 50% of cases, the magnitude of imbalance was $>2:1$ (Figure 8-14). The GWAS SNPs were the sole local sequence difference between haplotypes, indicating that disease-associated variants are responsible for modulating local chromatin accessibility. Further, at sites with very high sequencing depth ($>200\times$), 38.7% (53/137) show significant allelic imbalance (FDR $< 5\%$). As sensitivity to detect allelic imbal-

ance is governed by sequencing depth, this suggests that nearly 40% of GWAS variants in similarly-sequenced DHSs would be expected to show allelic imbalance.

Disease-associated variants cluster in transcriptional regulatory pathways

Transcriptional control of glucose homeostasis and beta cell genesis and function is mediated by a closely-knit transcriptional regulatory pathway defined by specific transcription factors. The Mendelian phenotypes of maturity-onset diabetes of the young (MODY) are caused by separate lesions disrupting the coding sequences of each of these transcription factors (Fajans et al., 2001). Interestingly, we observed clustering of common non-coding variants associated with abnormal glucose homeostasis, insulin and glycohemoglobin levels, and diabetic complications within recognition sites for the same six transcription factors ($P < 0.029$, binomial; 48% enrichment over random SNPs; Figure 8-15). This suggests that non-coding variants that predispose to dysregulation of glucose homeostasis perturb peripheral nodes of the same regulatory network responsible for Mendelian forms of Type 2 diabetes.

Using known interacting sets of transcription factors, we identified related disease-associated variants in the recognition sequences of a central target factor and its interacting partners (Figure 8-16, Figure 8-17, Figure 8-18, Figure 8-19) for factors involved in autoimmune disease, cancer and neurological development. IRF9 is a transcription factor associated with type I interferon induction (Tamura et al., 2008). Of 26 transcription factors in the IRF9-centered interaction network, 15 represent transcription factors with recognition sequences in multiple distinct DHSs that contain GWAS variants associated with a wide variety of autoimmune disorders ($P < 1.6 \times 10^{-13}$, binomial; 2.8-fold enrichment vs. random SNPs, Figure 8-16) (Section 8.5). Notably, 24.4% (64/262) of GWAS SNPs within DHSs of immune cells and associated with autoimmune disease alter one or more of the 15 transcription factor motifs from the IRF9-centered network. This example and those in Figure 8-18 and Figure 8-19 illustrate that disease-associated variants from the same or related disorders and traits repeatedly localize within the recognition sequences of transcription factors that form interacting regulatory networks.

Common networks for common diseases

The observation that GWAS variants associated with multiple distinct diseases within the same broader disease class (e.g., inflammation, cancer) repeatedly localize within the recognition sites of interacting transcription factors suggested that cohorts of such transcription factors might form shared regulatory architectures. To explore whether non-coding GWAS SNPs from related diseases perturb different recognition sequences of a common set of transcription factors, we tabulated all transcription factors for which at least 8 recognition sequences in DHSs were perturbed by GWAS SNPs associated with autoimmune diseases (Figure 8-20A). Among the 22 factors identified were canonical immune signaling regulators, such as STAT1 and STAT3, NF- κ B, and PPAR α and PPAR γ . These 22 transcription factors comprise a highly significant ($P < 9.8 \times 10^{-51}$, simulation vs. number of factors for random SNPs (Section 8.5), shared regulatory architecture that is repeatedly perturbed in a wide range of autoimmune disorders (Figure 8-20A).

The same analysis in the context of 17 different malignancies exposed a very different network of transcription factors connecting seemingly disparate cancer types ($P < 7.1 \times 10^{-11}$, simulation (Section 8.5)) including neoplastic regulatory relationships, linking FoxA1 and breast cancer, FoxO3 and colorectal cancer, and TP53 and melanoma, breast and prostate cancer (Figure 8-20B). We also analyzed six neuropsychiatric disorders, and identified 23 transcription factors whose recognition sequences were perturbed by at least 3 disease-associated variants (Figure 8-21). Collectively, these results support the hypothesis that shared genetic liability may underlie many related categories of disease (Cotsapas et al., 2011; International Schizophrenia Consortium et al., 2009).

De novo identification of pathogenic cell types

To provide insights into the cellular structure of disease and potentially highlight pathogenic cell types, we explored the selective localization of GWAS SNPs within the regulatory DNA of individual cell types. We further considered the enrichment of all tested variants, not just those with genome-wide significance, and performed serial determination of the cell/tissue-selective enrichment patterns of progressively more strongly associated variants to expose collective localization within specific lineages or cell types. We used all SNPs tested in GWAS meta-analyses of two common autoimmune disorders, Crohn's disease (Franke et al., 2010)

and multiple sclerosis (MS) (Patsopoulos et al., 2011), and a common continuous physiological trait, cardiac conduction measured by the electrocardiogram QRS duration (Sotoodehnia et al., 2010) ($n=938,703$, $2,465,832$, and $2,828,796$ SNPs, respectively). For SNPs meeting increasingly significant P -value cutoffs, we compared the proportion of SNPs in DHSs of each cell type to the proportion of all SNPs in DHSs of the same cell type (Figure 8-22). For all three studies, we observed enrichment of more weakly associated variants in regulatory DNA. This enrichment suggests that a large number of functional variants of small quantitative effect act through modulation of regulatory DNA. Additionally, it suggests that conditioning association analyses on regulatory DNA may ameliorate the stringent statistical correction for multiple testing required for genome-wide testing of unselected SNPs.

Furthermore, with progressively stringent P -value thresholds, we observed increasingly selective enrichment of disease-associated variants within specific cell types (Figure 8-22). Strikingly, in the case of Crohn's disease, the Th17 (12.0-fold enriched) and Th1 (8.87-fold enriched) T-cell subtypes have a concentration of the most-significant GWAS variants in their DHSs (Figure 8-22A). While Crohn's pathology has classically been associated with Th1 cytokine responses, an emerging consensus points to a defining role for IL17-producing Th17 cells (Brand, 2009). Notably, this analysis was accomplished without any prior knowledge about Crohn's disease pathology.

In the case of MS, sequential cell-selective enrichment analysis highlighted two cell types: CD3⁺ T-cells from cord blood, and CD19⁺/CD20⁺ B-cells (Figure 8-22B). While MS has long been thought to be T-cell mediated, a critical role for B-cells has only recently been recognized and has major therapeutic implications (Büdingen et al., 2011). It is notable that cord blood CD3⁺ cells – essentially a naïve population – garner the most highly selective enrichment, particularly in comparison with total adult CD3⁺ cells or other T-cell subsets, suggesting a role for variants influencing immune education. Also of note, DHSs active in brain tissue were moderately depleted (~10%) for MS-associated variants, suggesting that neural regulatory elements do not play a substantial role in MS pathogenesis, as proposed (International Multiple Sclerosis Genetics Consortium (IMSGC) et al., 2010a). Analogously, analysis of variants associated with the continuously varying trait of QRS duration revealed similarly specific enrichment within fetal heart DHSs (Figure 8-22C). Importantly, in

all three cases, the results were obtained without any prior knowledge of physiological mechanisms. These data suggest a generally applicable approach, and highlight the value of extensive maps of regulatory DNA for gaining insights into disease physiology and pathogenesis.

8.3 Discussion

Despite a long appreciation of the involvement of regulatory variants in human disease (Collins et al., 1985; King and Wilson, 1975; Rockman and Wray, 2002), difficulty in delineating regulatory DNA regions, particularly in a cell-specific context, has heretofore prevented comprehensive assessment of the relationship between gene regulation and common phenotypes. Our results indicating widespread and systematic localization of variants associated with a wide spectrum of common diseases and traits in regulatory DNA marked by DHSs have many implications for interpreting diverse genotype-phenotype association studies. The connection of numerous DHSs harboring GWAS SNPs with promoters of distant genes expands the genomic horizon of disease and trait associations, and provides a trove of plausible causal genes to explain those associations. The data also unify seemingly unconnected variants associated with related diseases by virtue of their convergent perturbation of common transcription factor networks. Tissue-selective enrichment of phenotype-associated variants raises the possibility of more focused genetic association studies that condition on the regulatory DNA of a known or hypothesized target tissue type. Further, selective enrichment of many more weakly-associated variants within regulatory DNA of pathogenic cell types points to the quantitative contribution of hundreds of variants of small effect size that modulate transcription factor binding characteristics, in contrast to Mendelian variants in transcription factor genes that may perturb entire networks. The results thus highlight a continuous quantitative spectrum of disordered gene regulation between common disease and Mendelian traits, and lend a new perspective on the genetic architecture of common human disease (Gibson, 2012).

8.4 Figures and Tables

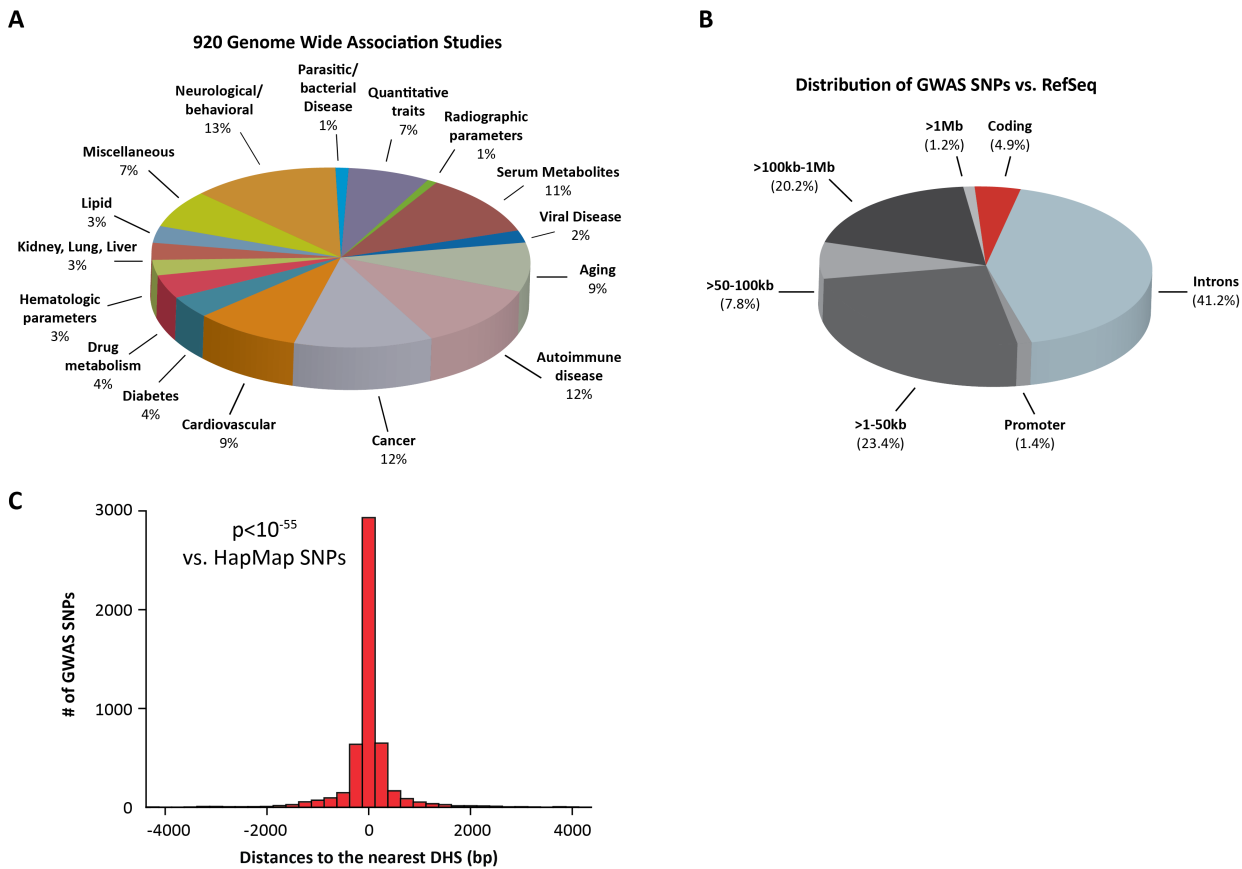


Figure 8-1 Diseases and traits studied by GWAS and distribution of GWAS variants

(A) Catalog of 6,011 trait-SNP associations (5,386 distinct SNPs) from 920 different studies. Chart shows percentage of GWAS SNPs by disease/trait class.

(B) Location of GWAS SNPs relative to genic features. Note only 4.9% of GWAS SNPs lie in coding sequence.

(C) Overlap of non-coding GWAS SNPs (5,134 distinct SNPs) and regulatory DNA. Horizontal axis: Binned distances from DHSs. Central "0" bin contains only GWAS SNPs within DHSs. The overlap is highly significant, even when corrected for a baseline enrichment of HapMap SNPs in DHSs.

Cell_line	Description	Fetal?	#DHSs	#SNPs
A549	Epithelial cell line derived from a lung carcinoma tissue	N	117,992	180
AG04449	Fetal buttock/thigh fibroblast	Y	174,802	202
AG04450	Fetal lung fibroblast	Y	150,114	187
AG09309	Adult human toe fibroblast	N	197,301	266
AG09319	Adult human gum tissue fibroblasts	N	137,192	190
AG10803	Adult human abdominal skin fibroblasts	N	171,903	224
AoAF	Normal Human Aortic Adventitial Fibroblast Cells	N	169,477	261
BE2_C	Human Neuroblastoma cell line	N	168,003	259
BJ	Skin fibroblasts	N	162,671	246
Caco-2	Colorectal adenocarcinoma	N	117,293	179
CD14+	Monocytes, CD14+	N	117,181	273
CD19+	B-lymphocytes, CD19+	N	75,086	225
CD20+	B-lymphocytes, CD20+	N	170,412	328
CD20+	B-lymphocytes, CD20+	N	86,908	268
CD3+	T-lymphocytes, CD3+	N	77,933	177
CD34+	Mobilized hematopoietic progenitor cells	N	134,718	230
CD3+_CordBlood	Cord blood, CD3+	N	74,992	176
CD4+	T helper cells, CD4+	N	94,881	239
CD56+	Lymphocytes, CD56+	N	105,724	277
CD8+	Cytotoxic T cells, CD8+	N	75,382	185
CMK	Acute megakaryocytic leukemia cells	N	123,561	210
GM06990	Lymphoblastoid	N	86,958	210
GM12864	Lymphoblastoid	N	132,370	262
GM12865	Lymphoblastoid	N	133,962	280
GM12878	Lymphoblastoid	N	109,419	240
H7-hESC	Undifferentiated embryonic stem cells	-	284,627	305
HAEPiC	Human Amniotic Epithelial Cells	Y	200,771	292
HAc	Human Astrocytes - cerebellar	Y	183,752	239
HAh	Human Astrocytes - hippocampal	Y	215,151	351
HAsp	Human Astrocytes - Spinal cord	Y	215,720	350
HBMEC	Human Brain Microvascular Endothelial Cells	Y	196,870	320
HCF	Human Cardiac Fibroblasts	Y	171,858	268
HCFaa	Human Cardiac Fibroblasts - adult atrial	N	184,810	323
HCM	Human Cardiomyocytes	Y	191,262	308
HCEPiC	Human Choroid Plexus Epithelial Cells	Y	209,492	304
HCT-116	Colon adenocarcinoma	N	104,196	170
HConF	Human Conjunctival Fibroblasts	Y	150,877	209
HEEPiC	Human Esophageal Epithelial Cells	Y	213,954	266
HepG2	Hepatocellular carcinoma	N	81,159	133
HESC	Human H1 Embryonic Stem Cell line	-	163,880	195
HFF	Human Foreskin Fibroblasts	N	189,148	329
HFF_Myc	Human Foreskin Fibroblasts_Myc Transgene	N	215,171	333
HGF	Human Gingival Fibroblasts	N	148,852	191
HIPEPiC	Human Iris Pigment Epithelial Cells	Y	231,963	304
HL-60	Human promyelocytic leukemia cells	N	153,865	296
HMEC	Human mammary epithelial cells	N	139,620	214
HMF	Human Mammary Fibroblasts	N	176,102	236
HMVEC-LBI	Human Lung Blood Microvascular Endothelial Cells	N	161,548	283
HMVEC-LLy	Human Lung Lymphatic Microvascular Endothelial Cells	N	130,544	235
HMVEC-dAd	Adult Human Dermal Microvascular Endothelial Cells	N	115,973	175
HMVEC-dBI-Ad	Adult Human Dermal Blood Microvascular Endothelial Cells	N	149,796	268
HMVEC-dBI-Neo	Neonatal Human Dermal Blood Microvascular Endothelial Cells	N	154,291	310
HMVEC-dLy-Ad	Adult Human Dermal Lymphatic Microvascular Endothelial Cells	N	115,834	194
HMVEC-dLy-Neo	Neonatal Human Dermal Lymphatic Microvascular Endothelial Cells	N	139,708	242
HMVEC-dNeo	Neonatal Human Dermal Microvascular Endothelial Cells	N	132,325	215
HNPCEPiC	Human Non-Pigment Ciliary Epithelial Cells	Y	217,558	296
HPAEC	Human Pulmonary Artery Endothelial Cells	N	125,462	170
HPAF	Human Pulmonary Artery Fibroblasts	Y	181,244	302
HPF	Human Pulmonary Fibroblasts	Y	147,153	225
HPdLF	Human Periodontal Ligament Fibroblasts	N	169,679	260
HRCEPiC	Human renal cortical epithelial cells (normal)	N	193,462	294
HRE	Human renal epithelial cells (normal)	N	197,779	257
HRGEC	Human Renal Glomerular Endothelial Cells	Y	143,319	188
HRPEPiC	Human Retinal Pigment Epithelial Cells	Y	229,606	298
HSMM	Human Skeletal Muscle Myoblasts	N	234,182	335
HSMM_D	Human Skeletal Muscle Myoblasts_differentiated	N	233,756	414
HUVEC	Human umbilical vein endothelial cells	N	115,081	229
HVMF	Human Villous Mesenchymal Fibroblasts	Y	170,308	296
HeLa-S3	Cervical carcinoma	N	119,081	247
IMR90	Fibroblasts	N	196,940	278

Cell_line	Description	Fetal?	#DHSs	#SNPs
Jurkat	T lymphoblastoid cell line derived from acute T cell leukemia	N	152,487	251
K562	Chronic myeloid leukemia	N	142,920	268
LNCaP	Prostate adenocarcinoma cell line	N	184,899	239
MCF-7	Mammary gland adenocarcinoma	N	133,229	168
NB4	Acute Promyelocytic Leukemia cell line	N	131,948	240
NH-A	Normal Human Astrocytes	Y	189,150	280
NHDF-Ad	Adult Human Dermal Fibroblasts	N	226,683	330
NHDF-neo	Neonatal Human Dermal Fibroblasts	N	184,888	269
NHEK	Normal Human Epidermal Keratinocytes	N	145,886	216
NHLF	Normal Human Lung Fibroblasts	N	204,839	296
NPC	H1 derived neuroprogenitor cells	-	93,396	148
NT2-D1	Human malignant pluripotent embryonal cancer cell line - Induced by RA to neuronal cells	N	187,959	259
PANC-1	Pancreatic carcinoma cell line	N	117,169	203
PrEC	Human Prostate Epithelial Cell Line	N	176,183	220
RPTEC	Human Renal Proximal Tubule Cells	N	171,601	293
SAEC	Small airway epithelial cells	N	195,662	279
SK-N-SH_RA	Neuroblastoma cell lines differentiated with retinoic acid	N	78,279	99
SK_N_MC	Neuroepithelioma cell line derived from a metastatic supra-orbital human brain tumor	N	154,275	177
SKMC	Human skeletal muscle cells	Y	208,844	274
WERI-Rb1	Retinoblastoma cell line	N	190,883	257
WI-38	Embryonic lung fibroblasts immortalized hTERT	Y	164,321	252
WI-38_TAM	Embryonic lung fibroblasts immortalized hTERT_Tamoxifen treated	Y	206,929	358
fAdrenal	Fetal adrenal tissue, 5 samples, ages 7-12 weeks	Y	282,181	480
fBrain	Fetal brain tissue, 12 samples, ages 12-20 weeks	Y	441,136	621
fHeart	Fetal heart tissue, 12 samples, ages 13-21 weeks	Y	393,615	743
fIntestine_Lg	Fetal large-intestine tissue, 15 samples, ages 12-16 weeks	Y	439,553	839
fIntestine_Sm	Fetal small-intestine tissue, 13 samples, ages 12-16 weeks	Y	360,316	735
fKidney	Fetal kidney tissue, 47 samples, ages 12-21 weeks	Y	666,350	1124
fLung	Fetal lung tissue, 34 samples, ages 10-17 weeks	Y	442,491	917
fMuscle	Fetal muscle tissue, 48 samples, ages 12-18 weeks	Y	632,517	1176
fPlacenta	Placenta tissue, 4 samples, ages 12-15 weeks	Y	281,754	553
fSkin	Fetal fibroblasts, 17 samples, ages 12-14 weeks	Y	392,999	591
fSpinal_cord	Fetal spinal-cord tissue, 3 samples, ages 12-16 weeks	Y	320,476	554
fSpleen	Fetal spleen tissue, age 16 weeks	Y	175,572	334
fStomach	Fetal stomach tissue, 11 samples, ages 13-21 weeks	Y	346,348	658
fTestes	Fetal testicle tissue, age 16 weeks	Y	170,843	309
fThymus	Fetal thymus tissue, 10 samples, ages 12-21 weeks	Y	341,548	658
Th1	Human primary T helper 1 cells	N	70,474	141
Th1	Human primary T helper 1 cells	N	73,754	190
Th17	Human primary T helper 17 cells	N	78,543	130
Th2	Human primary T helper 2 cells	N	111,450	220
Th2	Human primary T helper 2 cells	N	80,196	201

Table 8-1 Mapping of DHSs in 349 cell and tissue samples.

DNase I mapping of cell types and tissues used in the study, including the shorthand name for the tissue, a description of the tissue, whether the tissue is of fetal origin, the total number of DHSs observed, the number of GWAS SNPs with- in the DHSs.

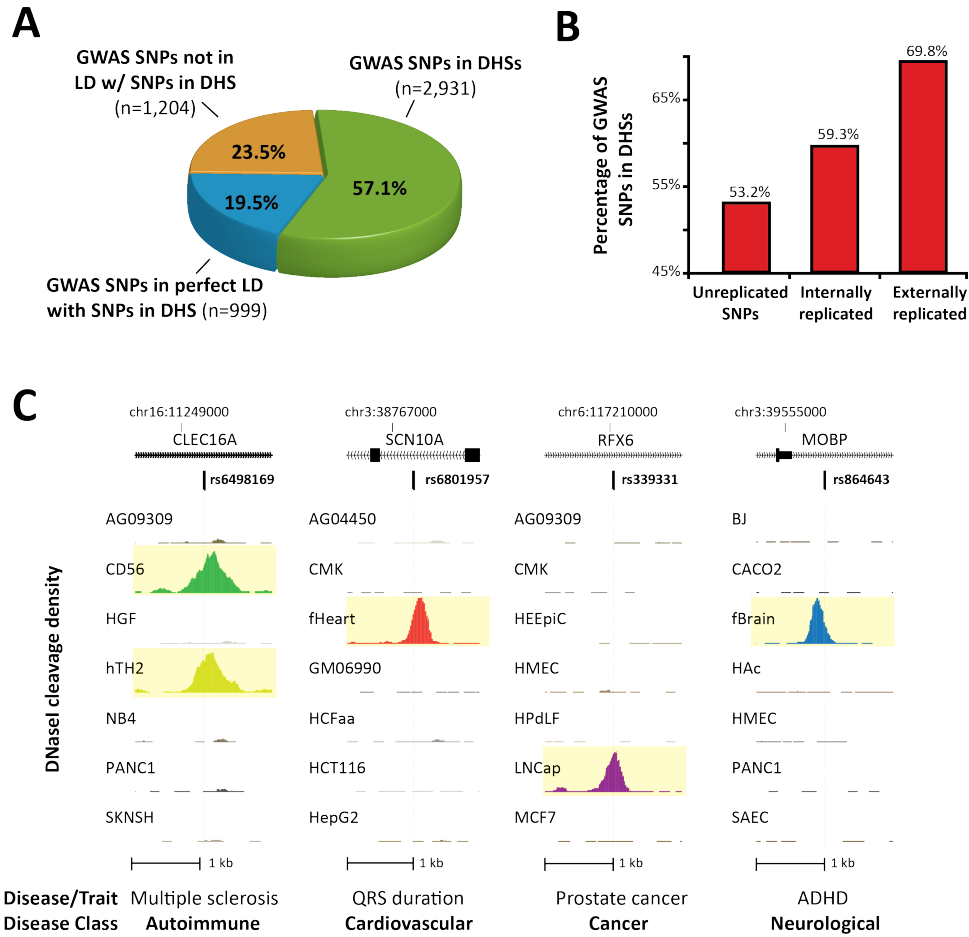


Figure 8-2 Disease-associated variation is concentrated in DNase I hypersensitive sites.

(A) Proportions of non-coding GWAS SNPs localizing within DHSs (green); in complete linkage disequilibrium ($r^2 = 1$) with a SNP in a DHS (blue); or neither (yellow). Note that 76.5% of GWAS SNPs are either within or in perfect LD with DHSs.

(B) Proportions of GWAS SNPs overlapping DHSs after partitioning by degree of replication.

(C) Representative DNase I hypersensitivity (tag density) patterns at diverse disease-associated variants.

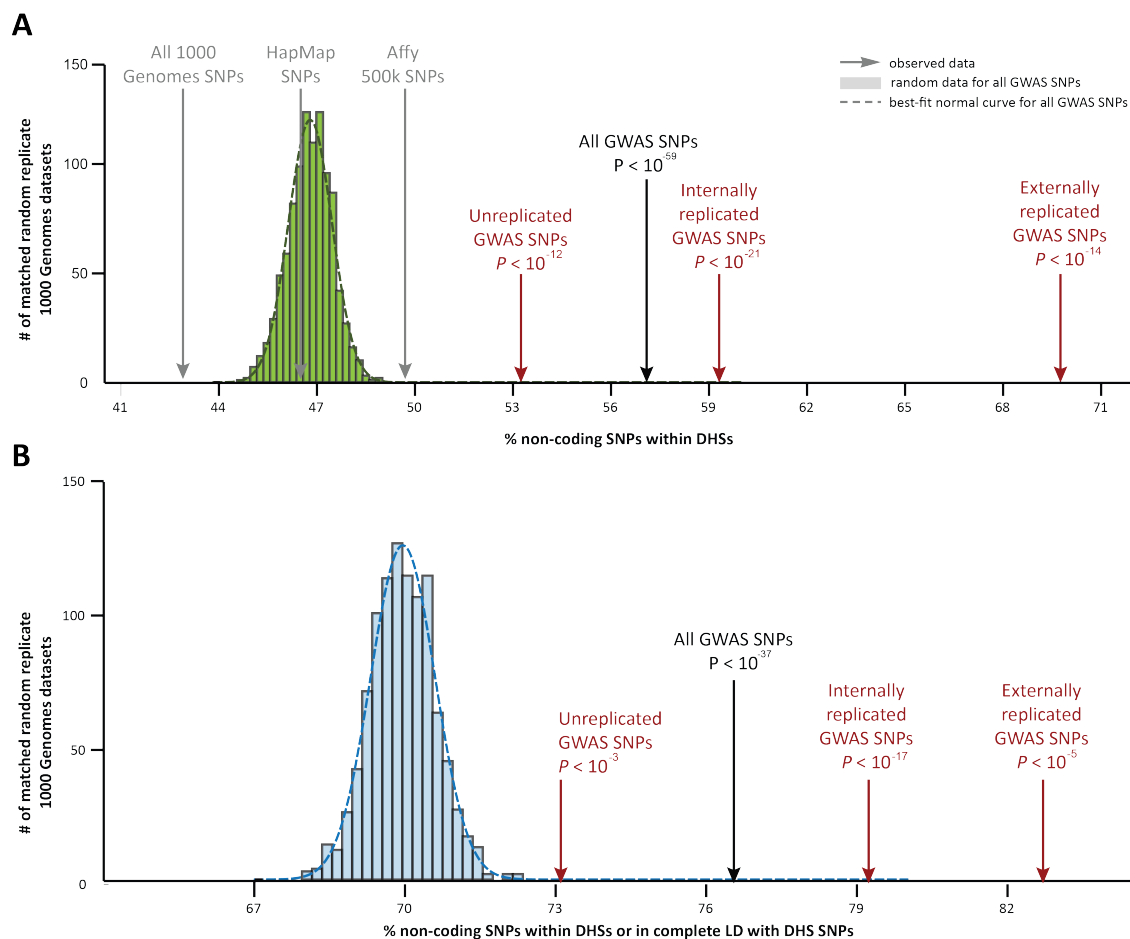


Figure 8-3 Overlap of non-coding GWAS SNPs and regulatory DNA marked by DHSs.

(A) Best-fit normal distribution of 1000 independent replicates of randomly-sampled SNPs matching all non-coding GWAS SNPs in genomic feature localization (intronic vs. intergenic), distance from the nearest TSS, and MAF in northwestern European populations; P -value (black arrow) was estimated as the area under the corresponding portion of the tail. Note the monotonic increase in the enrichment of disease/trait variants in DHSs with increasing quality of GWAS SNP experimental replication (red arrows); P -values shown in red were estimated from distributions (not shown) matched to each class of SNPs. Control sets consisting of all non-coding 1000 Genomes, HapMap CEU SNPs and Affymetrix 500K SNPs are shown for comparison.

(B) Same as (A), but measuring the percentage of GWAS SNPs within or in complete LD with 1000 Genomes SNPs in DHSs.

Sites	% in DHSs	Enrichment for DHSs	Mean MAF	Median distance to nearest TSS (kb)	Mean %CG (± 10 bp)	% in introns	%Intronic sites in DHSs
Random 36mer-mappable sites	41.1%	1.00	NA	71.4	40.16	36.7%	47.5%
1000 Genomes CEU SNPs	42.9%	1.04	0.18	77.7	41.01	35.9%	49.6%
HapMap CEU SNPs	46.5%	1.13	0.22	84.1	40.95	37.7%	52.9%
Affymetrix 500k SNPs	49.7%	1.21	0.23	82.2	40.85	37.5%	55.9%
Unreplicated GWAS SNPs	53.2%	1.29	0.24	57.0	41.66	40.2%	60.3%
Internally replicated GWAS SNPs	59.5%	1.45	0.28	31.4	42.97	45.2%	61.9%
Externally replicated GWAS SNPs	69.8%	1.70	0.28	24.1	43.21	48.1%	71.2%
All GWAS SNPs	57.1%	1.39	0.26	40.7	42.36	43%	62%

Table 8-2 Enrichment of GWAS and control sites for DHSs.

Listing of values of factors that can contribute to enrichment of sites within DHSs. Mean minor allele frequency (MAF) within the CEU population was computed using 1000 Genomes data for all except HapMap. Standard deviation (SD) of the MAF was ~ 0.14 in all cases; SD was $\sim 1\%$ of the mean for each reported %CG value. Only non-coding sites were surveyed for this table. Note that HapMap SNPs are not distinguished by G+C content. Further, although they are not enriched for introns, within introns, HapMap SNPs are enriched for DHSs.

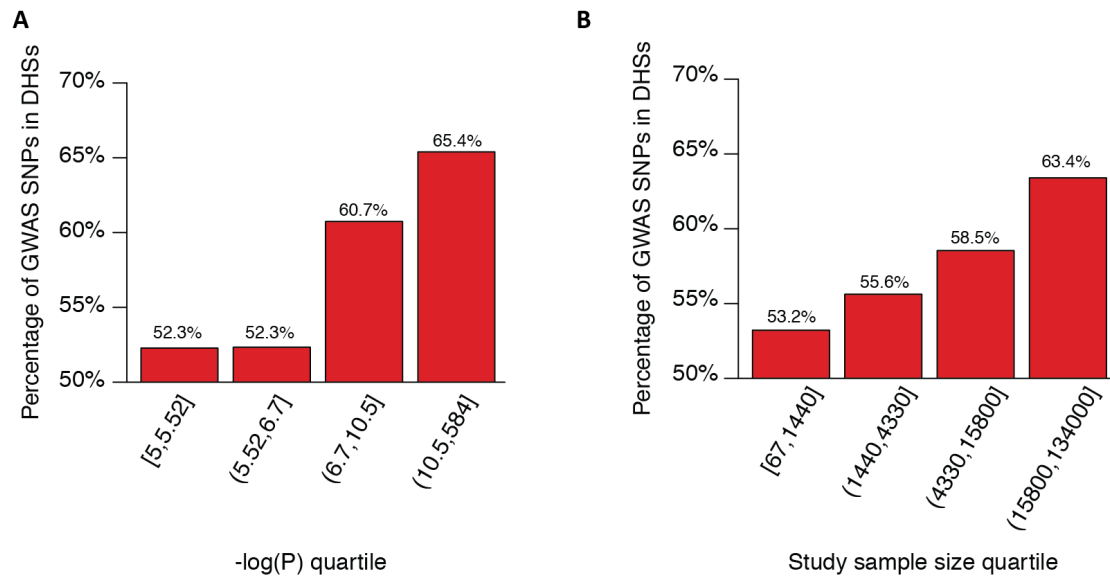


Figure 8-4 Enrichment for regulatory DNA increases with strength of association.

Overlap of GWAS SNPs in specified P -value (A) and sample size (B) quartiles with DHSs. Note the progressive enrichments parallel Figure 8-2B.

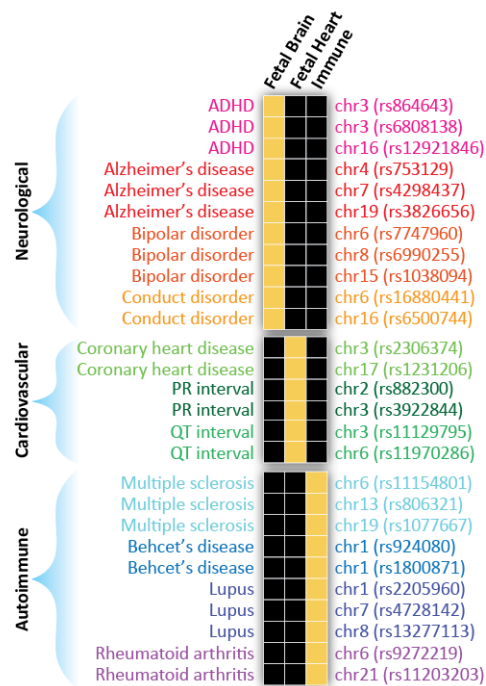


Figure 8-5 Multiple distinct genomic disease associations repeatedly localize within relevant cell-selective DHSs.

Each cell represents the presence or absence of a DHS at the location of the given GWAS SNP. Yellow = DHS present in that cell/tissue class; black = absent.

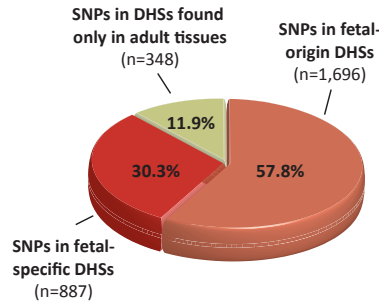


Figure 8-6 Disease-associated variation impinges on fetal regulatory elements

Proportion of GWAS SNPs localizing in DHSs active in fetal tissues that persist in adult cells (salmon); fetal stage-specific DHSs (red); and adult stage DHSs (green).

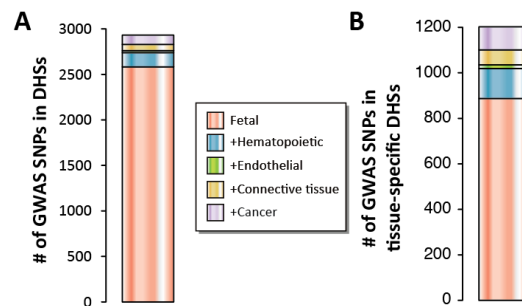


Figure 8-7 Localization of GWAS SNPs in DHSs of fetal and adult tissue classes.

(A) Cumulative tally of GWAS SNPs by DHS tissue category. Each color denotes SNPs overlapping DHSs in that tissue type but not in preceding categories. Note that the vast majority of adult-stage DHSs with GWAS variants derive from either differentiated hematopoietic cells or cancer lines.

(B) Same as (A) except for DHSs specific to a tissue class.

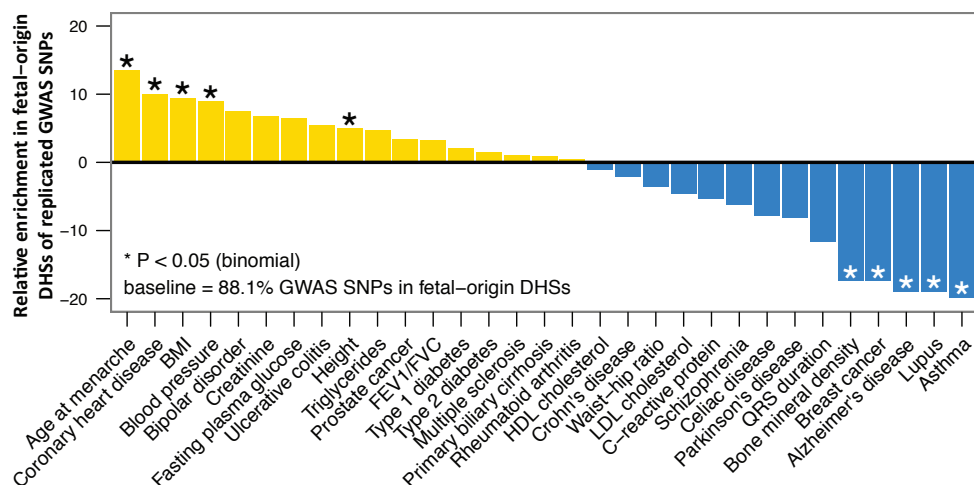


Figure 8-8 Phenotype-specific enrichment of regulatory GWAS SNPs for fetal DHSs.

Shows the magnitude of enrichment or depletion of replicated GWAS SNPs in DHS for fetal regulatory elements (Y-axis) for a given disease/trait (X-axis), relative to the background prevalence of all GWAS SNPs in DHSs in fetal elements (88.1%). Asterisks indicate the significance of the enrichment ($P < 0.05$, binomial). Only traits with >15 internal-ly- or externally-replicated associations are shown.

Disease or trait	R	Target gene	Distance
Amyotrophic lateral sclerosis	1	SYNGAP1* - Axon formation; component of NMDA complex	411 kb
Crohn's disease	1	TRIB1* - NFkB regulation	95 kb
Time to first primary tooth	0.99	PRDM1* - Craniofacial development	452 kb
C-reactive protein	0.99	NLRP3 - Response to bacterial pathogens	20 kb
Multiple sclerosis	0.98	AHI1* - White matter abnormalities	149 kb
QRS duration	0.96	SCN10A* - Sodium channel involved in cardiac conduction	181 kb
Breast cancer	0.96	TACC2* - Tumor suppressor	411 kb
Schizophrenia/brain imaging	0.95	KIF1A* - Neuron-specific kinesin involved in axonal transport	428 kb
Brain structure	0.94	CXCR6* - Chemokine receptor involved in glial migration	357 kb
Rheumatoid arthritis	0.94	CTSB* - Cysteine proteinase linked to articular erosion	359 kb
Ovarian cancer	0.93	HSPG2* - Ovarian tumor suppressor	268 kb
Multiple sclerosis	0.93	ZP1* - Known autoantigen	153 kb
ADHD	0.93	PDLIM5* - Neuronal calcium signaling	328 kb
Breast cancer	0.88	MAP3K1* - Response to growth factors	158 kb
Amyotrophic lateral sclerosis	0.88	CNTN4 - Neuronal cell adhesion	306 kb
Schizophrenia	0.81	FXR1* - Cognitive function	120 kb
Type 1 diabetes	0.75	ACAD10* - Mitochondrial oxidation of fatty acids	343 kb
Lupus	0.74	STAT4 - Mediates IL12 immune response and Th1 differentiation	113 kb

Table 8-3 Target genes of distal DHSs harboring GWAS variants.

Examples of distal DHSs-to-promoter connections that highlight candidate genes potentially underlying the association.

*' indicates that highest correlated gene is not the nearest gene. R, Pearson's correlation coefficient.

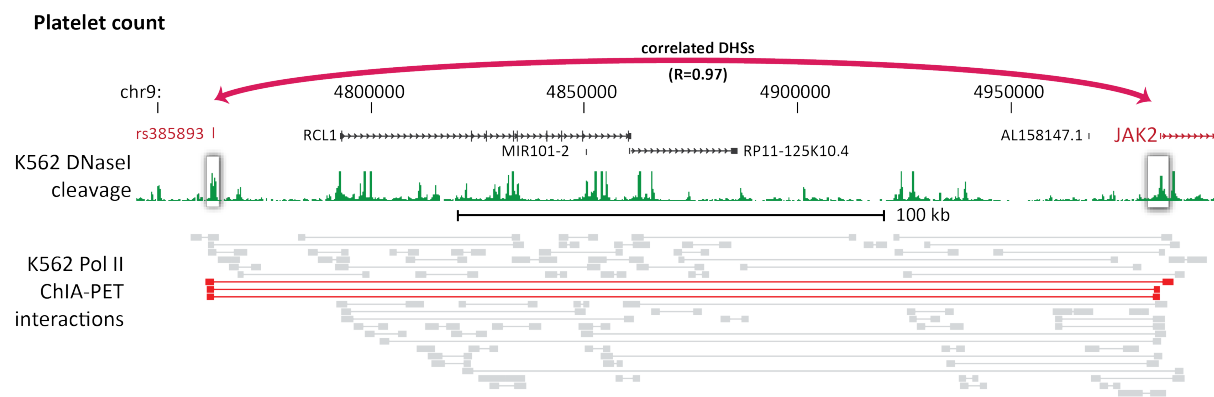


Figure 8-9 Correlation analysis links platelet count-associated SNP and JAK2 gene

GWAS variant associated with platelet count is connected with the JAK2 gene (myeloproliferative disorders) 222 kb away. Below, ChIA-PET tags (Li et al., 2012) validate direct chromatin interactions between this DHS and the JAK2 promoter; red tags demonstrate an interaction between these DHSs.

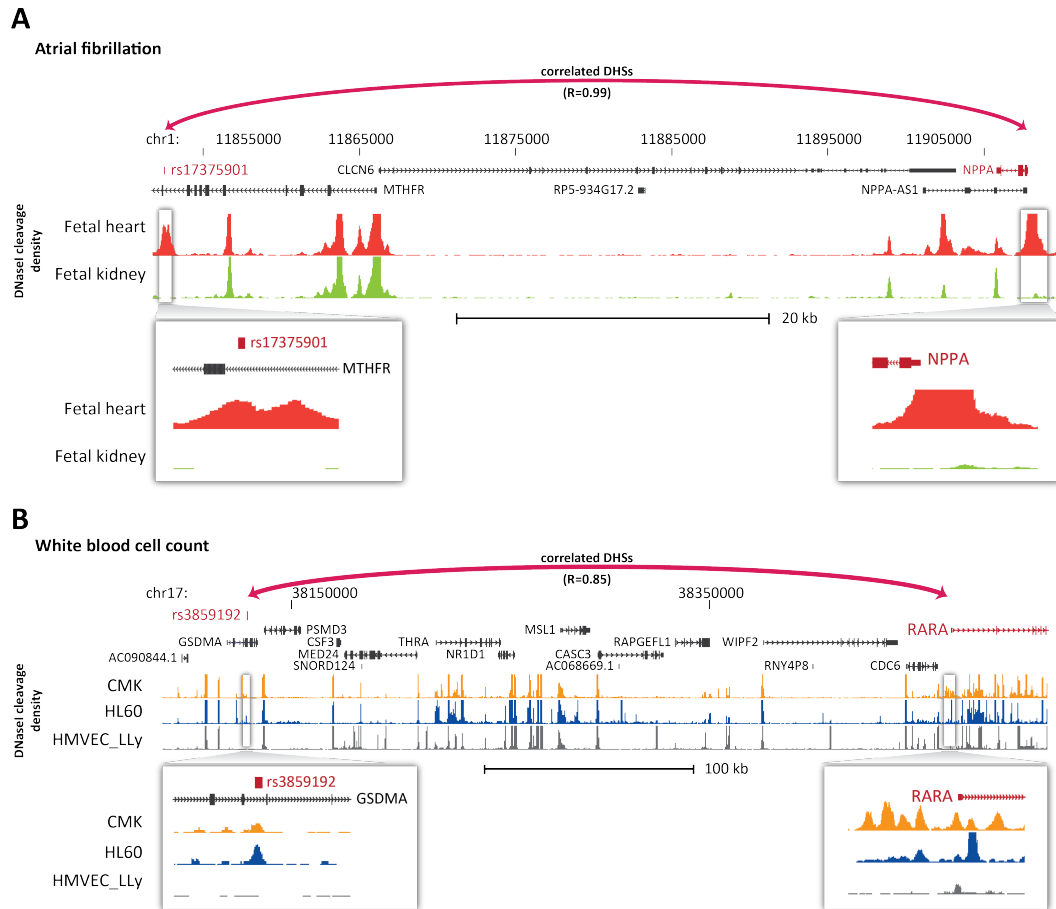


Figure 8-10 Regulatory GWAS variants are linked to distant target genes.

(A) A DHS specific to fetal heart, connected with the gene encoding atrial natriuretic peptide A (NPPA), harbors a GWAS variant associated with atrial fibrillation.

(B) Distant (336 kb) DHS connected with retinoic acid receptor- α (RARA), a nuclear receptor involved in myeloid differentiation, harbors a GWAS variant associated with white blood cell count.

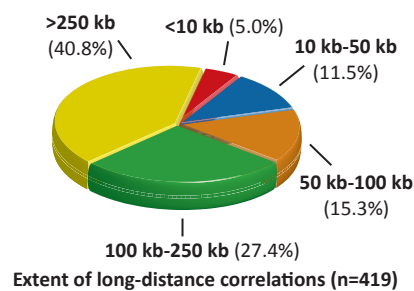


Figure 8-11 Distance from correlated DHSs harboring GWAS variants to target promoters.

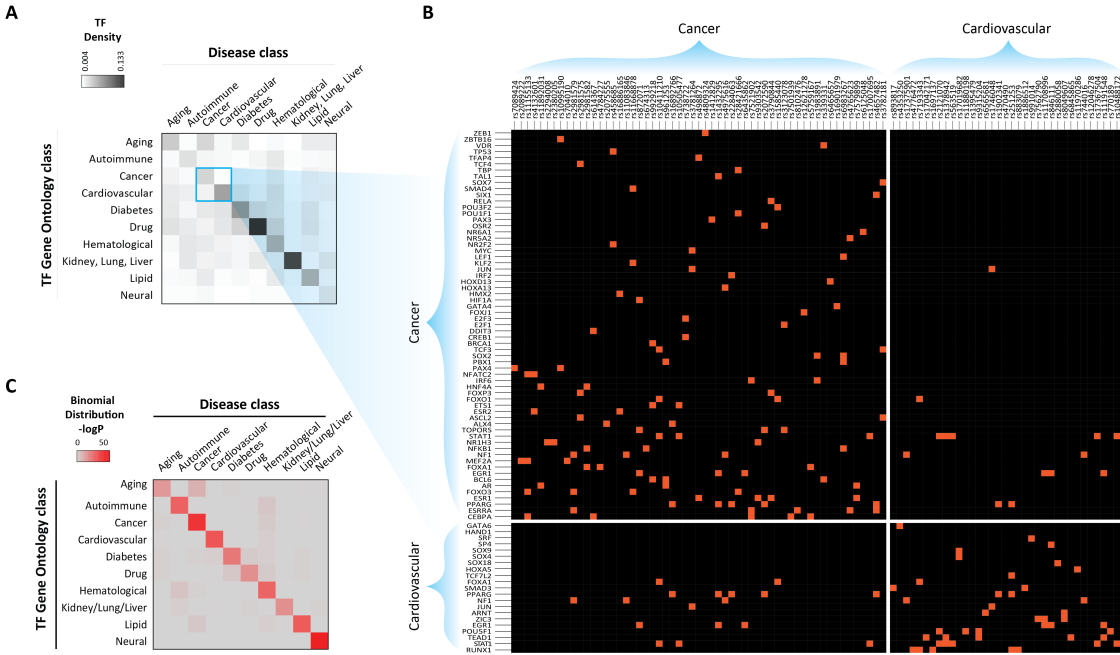


Figure 8-12 GWAS variants in DHSs localize within physiologically relevant TF binding sites.

(A) GWAS variants in DHSs localize within physiologically relevant TF binding sites. Columns: Categorization of GWAS SNPs by disease area. Rows: Transcription factor categories from unbiased mapping into disease/pathophysiologic categories with GO. Each cell shows the proportion (grayscale bar) of all GWAS SNPs from a given disease category (column) that fall into the binding sites of TFs within each TF category (row).

(B) Close-up of cancer and cardiovascular diseases showing the presence (red) or absence (black) of a recognition sequence for a particular recognition sequence (rows) at the location of a GWAS SNP in the indicated disease category (columns).

(C) Significance of proportions in (A), with high significance (binomial test) along the diagonal indicating systematic localization of GWAS variants from a given disease category within binding sites of pathophysiologically-related TFs

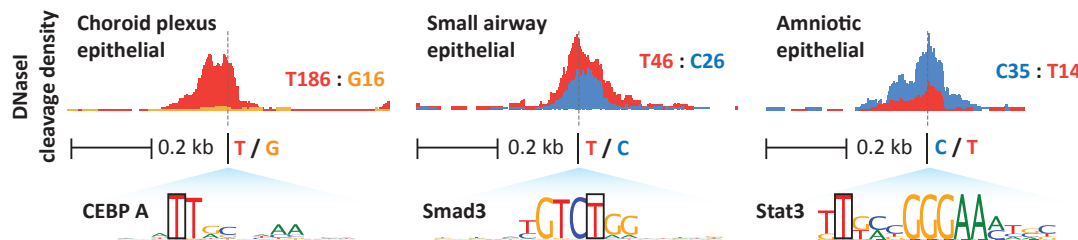


Figure 8-13 Examples of allele-specific DNase I sensitivity in cell types derived from heterozygous individuals for GWAS variants that alter TF recognition motifs within DHSs.

Each cell type track shows DNase I cleavage density scaled by allelic imbalance at the GWAS variant and colored by variant nucleotide (blue = C, green = A, yellow = G, red = T). Total reads from each allele are also shown.

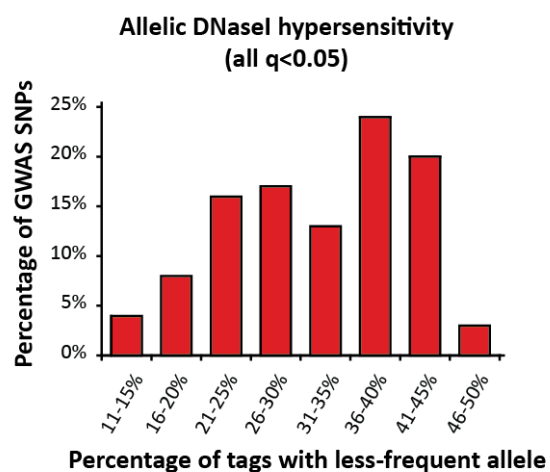


Figure 8-14 Allelic imbalance distribution.

Distribution of the proportion of reads from the less-frequent allele at DHSs with significant (FDR < 5%) imbalance in DNase I hypersensitivity.

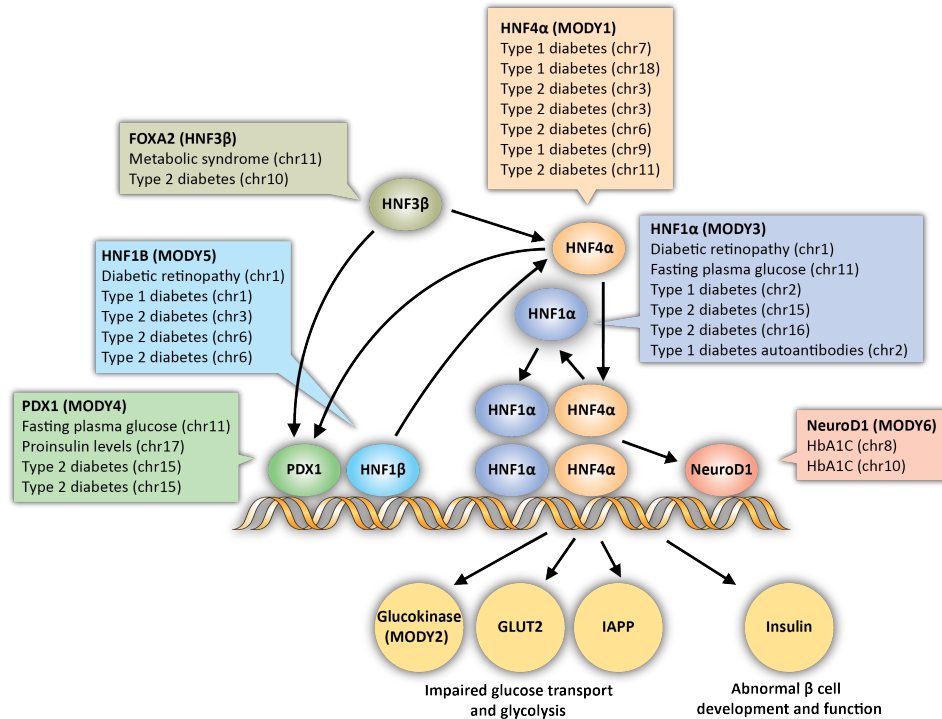


Figure 8-15 Diabetes-associated SNPs cluster in Mendelian diabetes network.

SNPs in DHSs associated with diabetes (Type I and Type II), diabetic complications, and glucose homeostasis localize in recognition sites of transcriptional regulators (labeled ellipses) controlling glucose transport, glycolysis, and beta cell function that are structurally disrupted in the Mendelian phenotypes of maturity-onset diabetes of the young (MODY). Chromosome of each SNP associated with the indicated phenotype is listed.

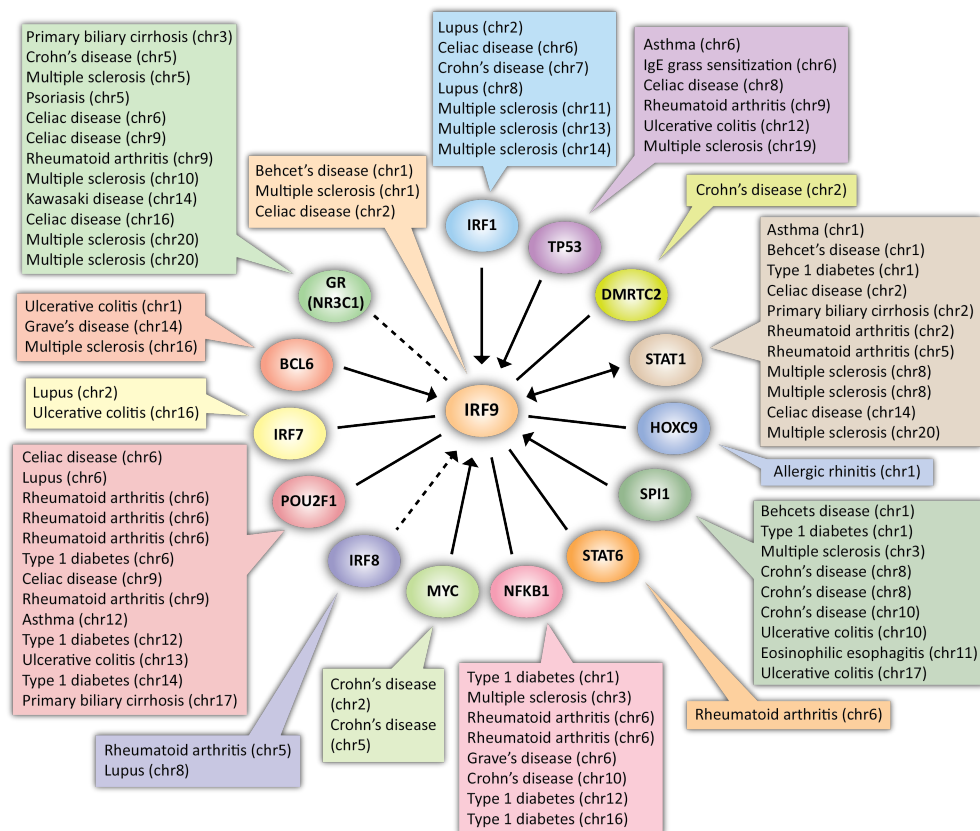


Figure 8-16 Autoimmune disease-associated SNPs cluster in IRF9 interaction network

24.4% of SNPs associated with autoimmune disorders that fall within DHSs localize in recognition sequences of TFs that interact with IRF9. Arrows indicate directionality of relationship, dotted lines represent indirect interactions (Section 8.5). The complete network is shown in Figure 8-17.

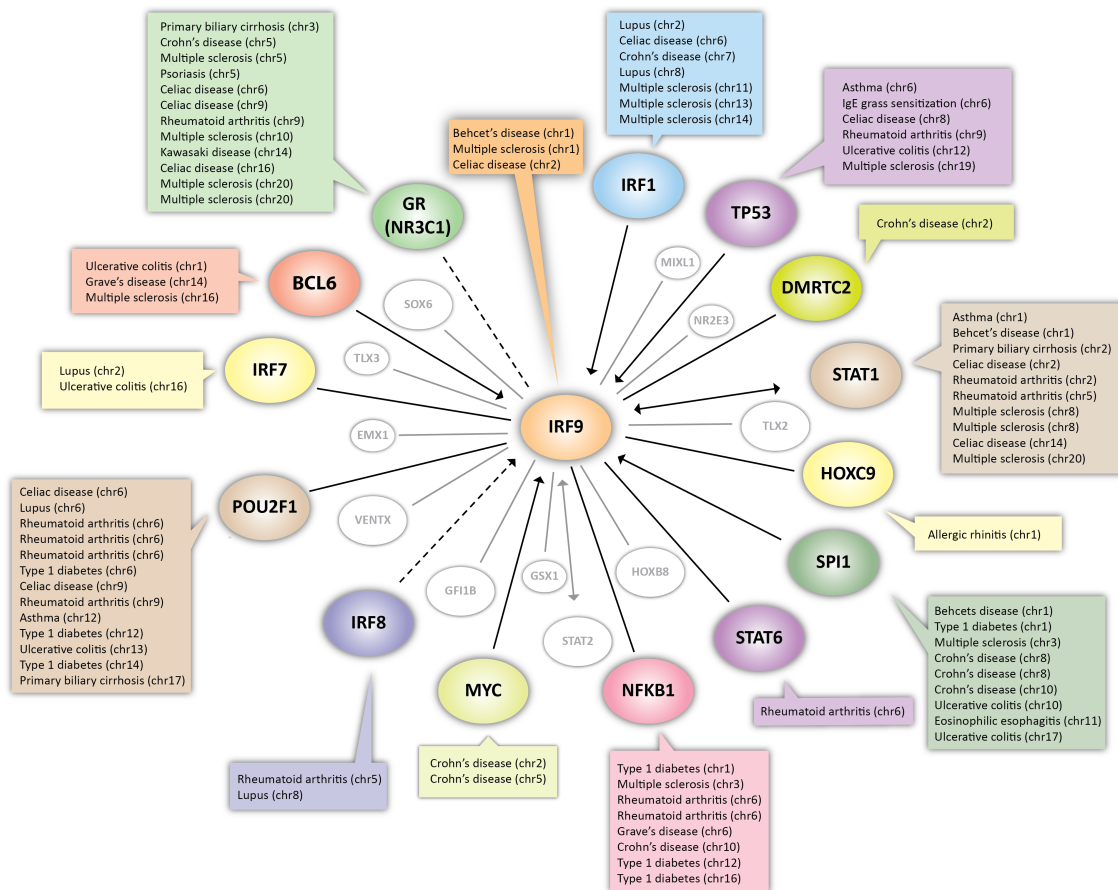


Figure 8-17 Autoimmune disease-associated SNPs cluster in IRF9 interaction network (full).

Detail of Figure 8-16 showing factors for which no position weight matrices were available (small gray ellipses, $n=6$) or for which no GWAS SNPs were in their binding sites (large gray ellipses, $n=5$). SNPs in DHSs associated with autoimmune diseases repeatedly localize in recognition sequences for transcriptional regulators (labeled ellipses) that interact with IRF9. Arrows indicate directionality of relationship; dotted lines represent indirect interactions (Section 8.5).

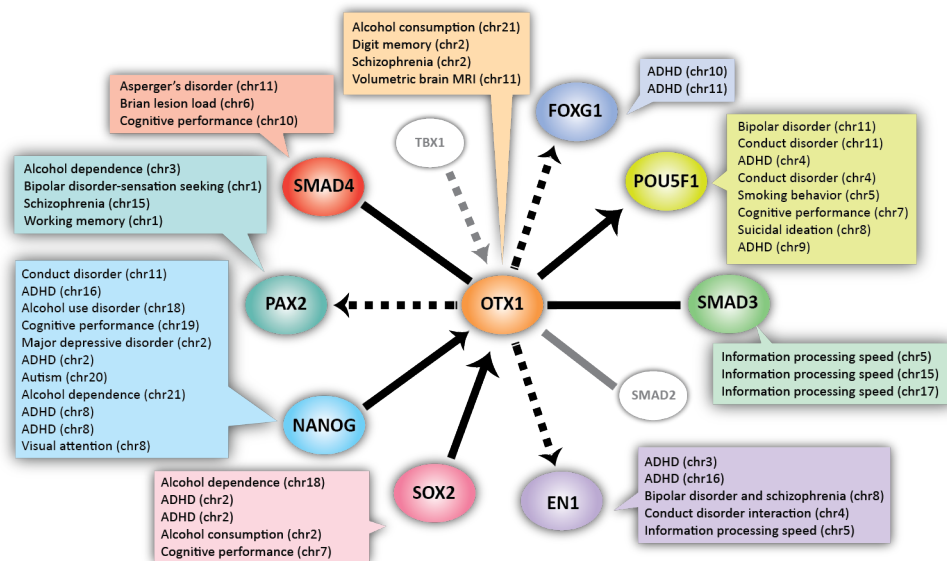


Figure 8-18 Repeated involvement of the OTX1 pathway in neuropsychiatric diseases and traits.

SNPs in DHSs associated with diverse neuropsychiatric diseases and traits repeatedly localize within recognition sequences of TFs that interact with the brain morphogenic factor OTX1. SNPs in DHSs associated with cancer repeatedly localize in recognition sequences for transcriptional regulators (labeled ellipses) that interact with ESRRA, significant at $P < 0.049$ (binomial; 1.3x enrichment vs. proportion of random SNPs) (Section 8.5). Arrows indicate directionality of relationship; dotted lines represent indirect interactions (Section 8.5). Large gray ellipses indicate factors for which position weight matrices were available, but no recognition sequences overlapped GWAS SNP.

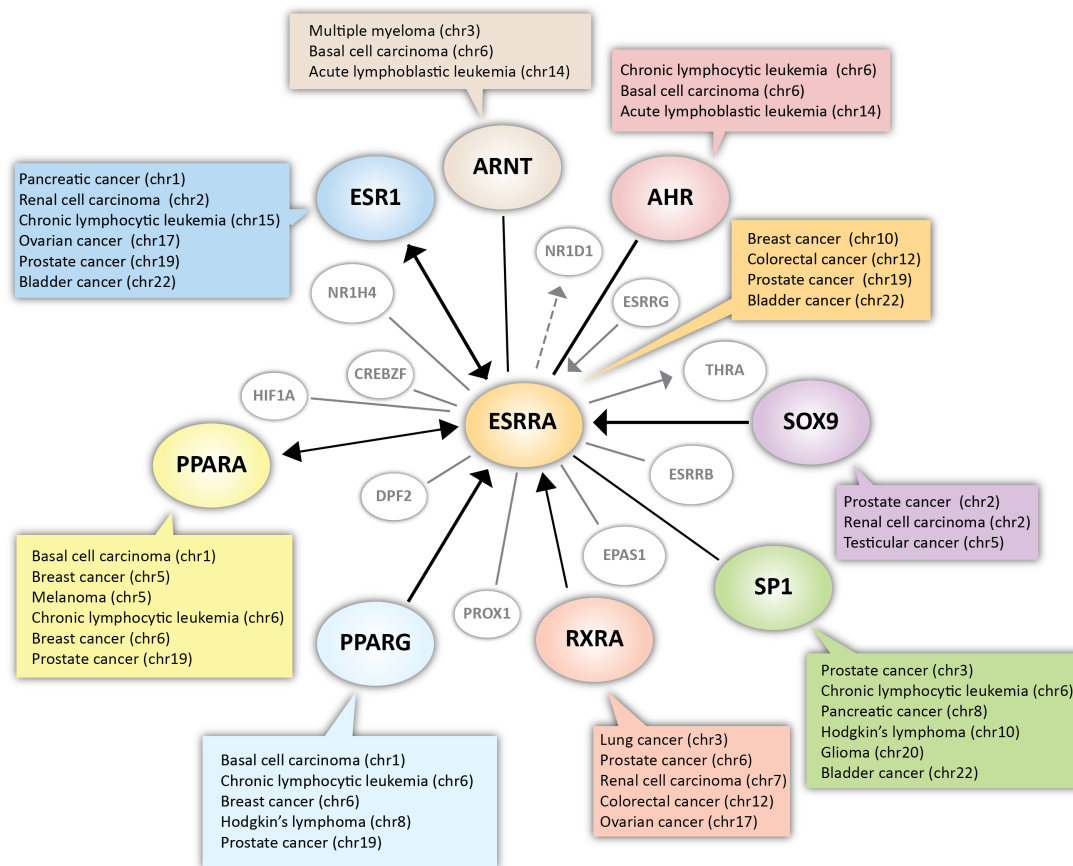


Figure 8-19 Cancer-associated SNPs cluster in orphan nuclear receptor ESRRA network.

SNPs in DHSs associated with cancer repeatedly localize in recognition sequences for transcriptional regulators (labeled ellipses) that interact with ESRRA, significant at $P < 0.010$ (binomial; 1.5x enrichment vs. proportion of random SNPs) (Section 8.5). Arrows indicate directionality of relationship, dotted lines represent indirect interactions (Section 8.5). Large gray ellipses indicate factors for which position weight matrices were available, but no recognition sequences overlapped GWAS SNPs; small gray ellipses indicate that a position weight matrix for this factor is unavailable. The 28 distinct SNPs in the ESRRA network represent 12.7% of 220 cancer GWAS SNPs overlapping DHSs.

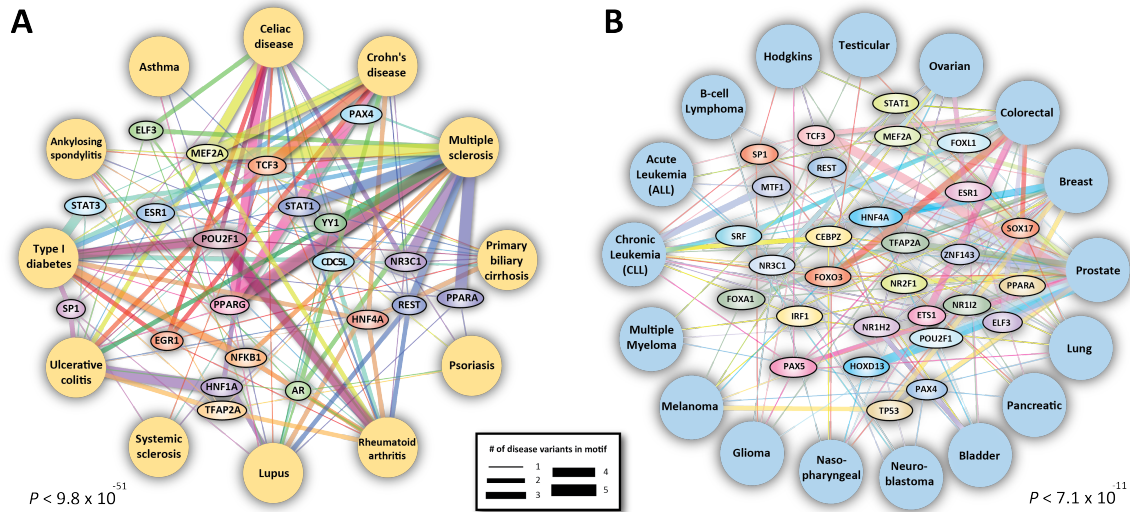


Figure 8-20 Common disease networks.

GWAS SNPs from related diseases repeatedly perturb recognition sequences of common transcription factors. Shown are factors whose recognition sequences harbor ≥ 8 or ≥ 6 GWAS SNPs in inflammatory/autoimmune diseases (A) and cancer (B), respectively. Edge thickness represents number of associations between TF and disease in DHSs in relevant tissues. Both networks are significantly enriched for overlap with disease-relevant GWAS SNPs, and include many well-studied regulators.

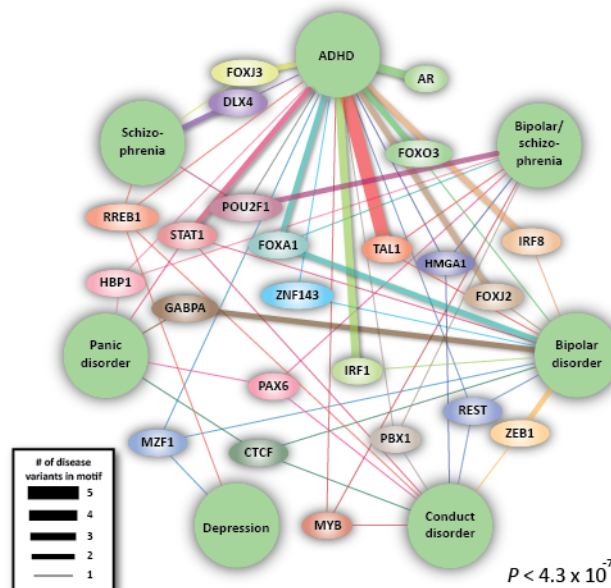


Figure 8-21 Psychiatric disease network.

GWAS SNPs from related diseases repeatedly perturb recognition sequences of common transcription factors. Shown are factors whose recognition sequences harbor ≥ 3 GWAS SNPs in psychiatric diseases. Edge thickness represents number of associations between TF and disease in DHSs in relevant tissues.

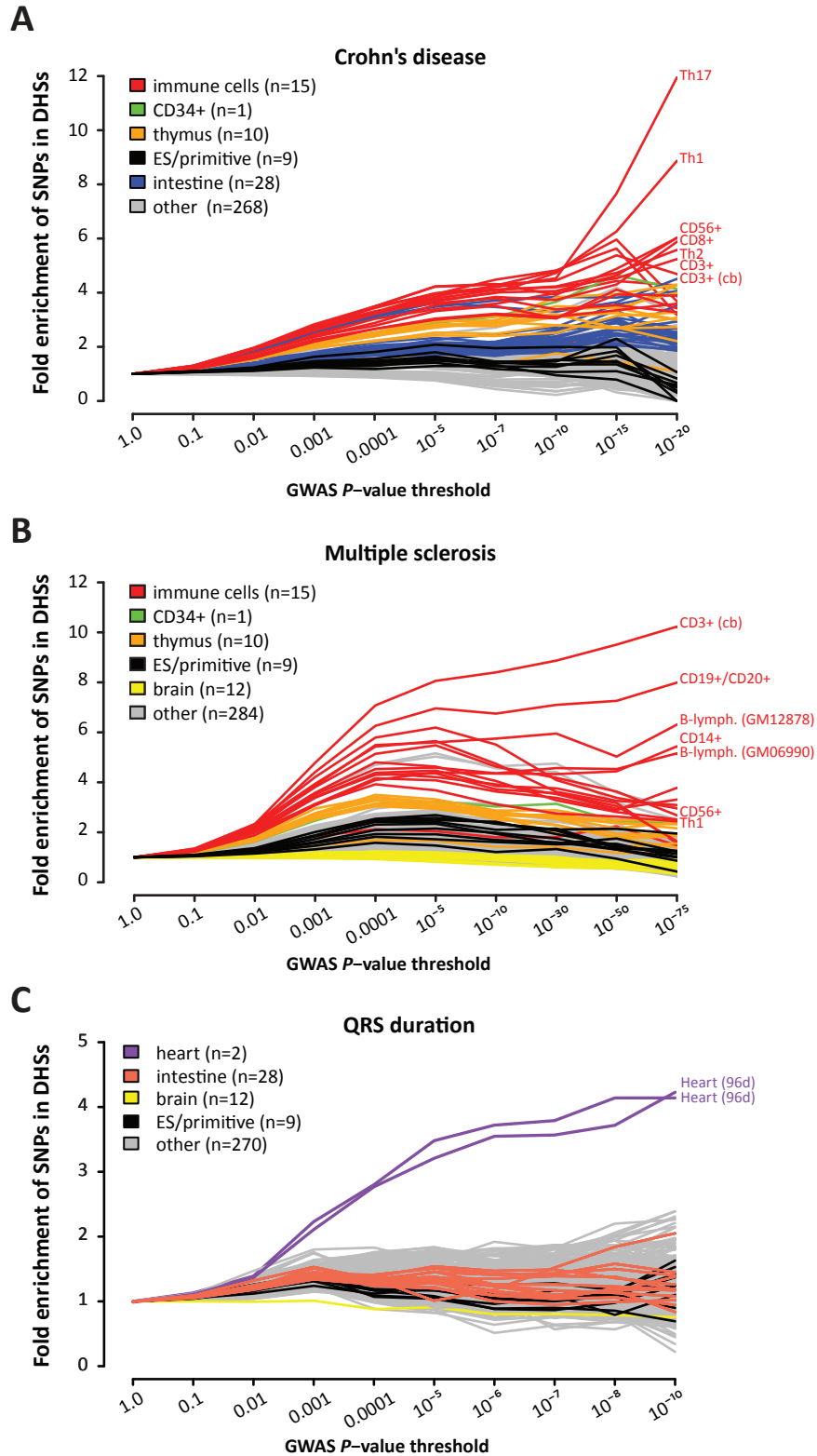


Figure 8-22 Identification of pathogenic cell types.

GWAS SNPs are systematically enriched in the regulatory DNA of disease-specific cell types throughout the full range of significance. Shown are SNPs tested for association with the autoimmune disorders Crohn's disease (A), multiple sclerosis (B) and QRS duration (C).

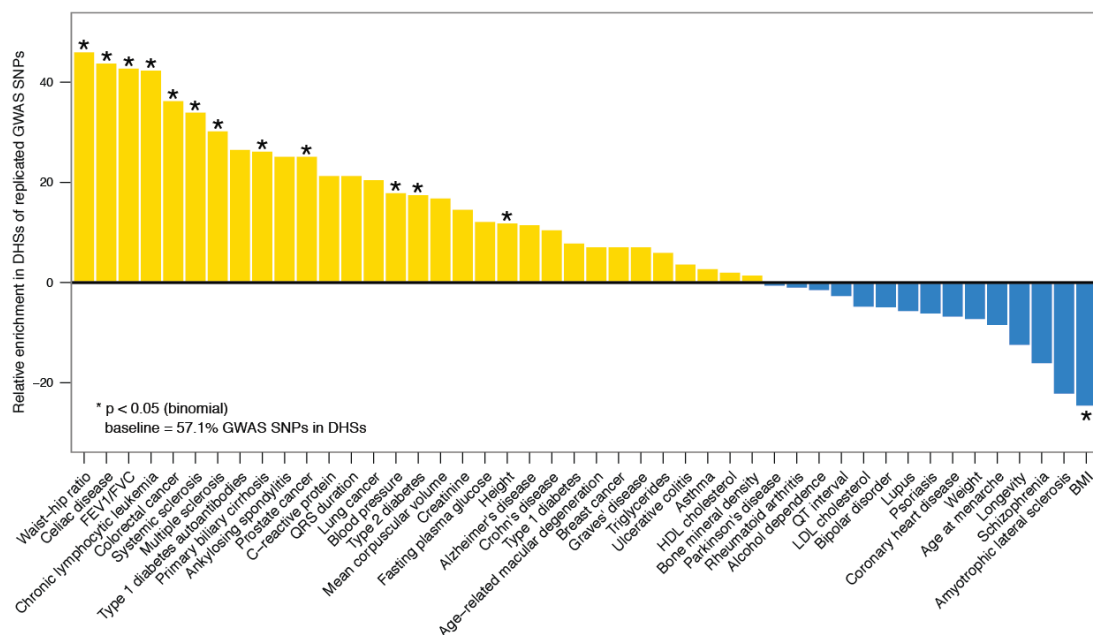


Figure 8-23 Enrichment of GWAS SNPs for DHSs by disease/trait.

Shows the magnitude of enrichment or depletion of replicated GWAS SNPs in DHSs (Y-axis) for a given disease/trait (X-axis), relative to the background prevalence of all GWAS SNPs in DHSs (57.1%). Asterisks indicate the significance of the enrichment ($P < 0.05$, binomial). Only traits with >15 internally- or externally-replicated associations are shown.

Category	Cell types represented
1	WERI_Rb1
2	BE_2_C
3	CACO2, HEPG2, SKNSH
4	HESC, hESCT0
5	A549, HCT116, HeLa, PANC1
6	LNCap, MCF7
7	CD56, CD4, hTH1, hTH2
8	GM06990, GM12864, GM12865, GM12878
9	CD34, Jurkat
10	K562, CMK
11	NB4, HL60, CD14, HRGEC
12	HMVEC_LBI, HMVEC_dLyNeo, HMVEC_dBIAd, HMVEC_dBINeo, HUVEC
13	HMVEC_dLyAd, HMVEC_dNeo
14	NHLF, NHA
15	HAc
16	HAsp
17	HVMF
18	HAEpiC
19	WI_38, AG04450, IMR90
20	SkMC
21	HCFaa
22	HBMEC, HCPEpiC, HIPEpiC, HNPCEpiC
23	HSMM, HSMM_D
24	HCF, HCM, HPAF
25	AG04449, AG09309, AG10803, BJ, HFF
26	NHDF_Ad, HDF_Neo
27	AoAF, HConF, HMF, HPF
28	AG09319, HGF, HPdLF
29	HRCE, HRE, RPTEC
30	HRPEpiC
31	HMEC, NHEK
32	HEEpiC, SAEC

Table 8-4 Cell types used in DHS correlation analysis from Thurman et al.

The 79 cell types in previous work (Thurman et al., 2012) were assigned to 32 numbered cell-type categories.

Category	Cell types represented
1	fAdrenal-DS12528
2	fBrain-DS11872
3	fHeart-DS16582
4	fIntestine_Lg-DS17313
5	fIntestine_Sm-DS17092
6	fKidney-DS21096
7	fLung-DS12817
8	fMuscle_leg-DS20239
9	fPlacenta-DS20793
10	fSkin_fibro_scalp-DS19449
11	fSpinal_cord-DS20351
12	fSpleen-DS17448
13	fStomach-DS17878
14	fTestes-DS18942
15	fThymus-DS20341
16	Skin_Fibroblasts-DS18252
17	Skin_Keratinocytes-DS18692
18	Skin_Melanocytes-DS18590
19	CD20-DS17371
20	HAh-DS15192
21	HFF_MyC-DS15079
22	HPAEC-DS12916
23	Mesendoderm-DS19310
24	NPC-DS18739
25	NT2_D1-DS14575
26	PrEC-DS12098
27	SK_N_MC-DS14408
28	Trophoblast-DS19317
29	vHMEC-DS18438
30	WI_38_TAM-DS14323
31	HMVEC_dAd-DS12957, HMVEC_dNeo-DS12937
32	iPS_19_7-DS15148, iPS_19_11-DS15153, iPS_4_7-DS15169
33	iPS_6_9-DS15164, H1_P18-DS18873, H9_P42-DS18522
34	CD3_CordBlood-DS17706, CD19-DS17281, CD20-DS17541
35	CD8-DS17203, CD3-DS17198, hTH1-DS17592, hTH17-DS11039

Table 8-5 New cell types used in DHS correlation analysis

For the correlation analysis, 46 additional cell and tissue types (choosing a representative single time point from each fetal tissue vs. all developmental time points) were condensed to 35 numbered categories.

Disease class	Controlled vocabulary used for GO queries
Aging	"aging", "death"
Autoimmune	"immune", "interleukin", "B cell", "T-cell", "B-cell", "T-cell", "cytokine", "NF-kappa", "leukocyte", "lymphocyte", "interferon"
Cancer	"proliferation", "tumor", "P53", "apopto", "death", "RAS"
Cardiovascular	"heart", "cardio", "angio", "artery", "cardiac", "circulat", "vascu"
Diabetes	"glucose", "insulin", "glycolysis", "glycogen". "hexose", "gluconeogenesis", "pancreatic B cell", "keto"
Drug	"xenobiotic", "drug"
Hematological	"blood", "hema", "hemo", "myeloid"
Kidney, Lung, Liver	"kidney", "neph", "lung", "hepatic", "liver", "hepatico", "bile", "urogen", "renal", "urete", "pulmon"
Lipid	"lipid", "sterol", "glyceride", "phosphatidyl", "sphingo", "acylglycerol", "icosanoid", "steroid", "adipose", "fat"
Neural	"neur", "brain", "action potential", "astrocyte", "axon", "hindbrain", "hippocampus", "spinal", "nervous", "oligodendrocyte", "cerebellar", "perception", "nerve", "glial", "glio"

Table 8-6 Search terms for association of GO Biological Processes with Disease/Trait classes

GO search terms were used to associate transcription factors with disease/trait.

SNP	Target celltype	GO factor name	GWAS disease classes	GWAS diseases	Motif logo
chr9:117,568,766	GM06990;GM12864;GM12865; Jurkat;hTH1;hTH2	HOXA7	Autoimmune_disease	Inflammatory bowel disease	 V_HOXA7_01 -
chr5:150,458,146	GM12864;GM12865	ARNT	Autoimmune_disease	Systemic lupus erythematosus	 V_ARNT_01 +
chr1:117,100,957	CD4;GM12864;GM12865; GM12878	IRF9	Autoimmune_disease	Multiple sclerosis	 V_ISRE_01 +
chr8:11,352,541	GM12864;GM12865;GM12878	IRF8	Autoimmune_disease	Systemic lupus erythematosus	 V_ICSBP_Q6 -
chr5:110,435,490	CD56	ETV4	Hematological_parameters	Plasma eosinophil count	 V_PEA3_Q6 +
chr4:77,420,784	GM06990;GM12864	AHR	Autoimmune_disease	Eosinophilic esophagitis (pediatric)	 V_AHR_Q5 -
chr4:27,744,718	fBrain	MEF2A	Neurological_behavioral	Brain imaging	 V_MEF2_Q4 -
chr4:27,744,718	fBrain	ELK1	Neurological_behavioral	Brain imaging	 V_ELK1_Q1 -
chr22:30,592,487	CACO2;GM06990	ATF1	Autoimmune_disease	Crohns disease	 V_ATF_Q1 -
chr6:135,426,573	K562	SIX1	Hematological_parameters	Mean corpuscular volume WBC count	 Six1_Q935.2 -
chr7:50,269,672	GM12864;GM12865	IRF3	Autoimmune_disease	Crohns disease	 V_IRF_Q6_Q1 +
chr8:11,352,541	GM12864;GM12865;GM12878	IRF9	Autoimmune_disease	Systemic lupus erythematosus	 Isgf3g_primary -
chr20:8,114,704	BE_2_C;fBrain	VSX1	Neurological_behavioral	Colorword	 Vsx1_1728.1 +
chr14:72,883,874	BE_2_C	EP300	Neurological_behavioral	Smoking_cessation	 V_P300_Q1 -
chr11:131,320,070	hESCT0	MAX	Neurological_behavioral	Asperger disorder	 Max_primary -
chr11:10,669,228	CD34;K562HL60	SMAD3	Hematological_parameters	Platelet aggregation-ADP	 Smad3_primary -
chr19:34,168,273	BE_2_C;SKNSH	POU3F2	Neurological_behavioral	Working memory	 V_BRN2_Q1 +
chr15:36,293,605	fBrain	REST	Neurological_behaviora	ADHD	 V_NRSF_Q1 -
chr12:90,013,089	HBMEC;HCM	E2F4	Cardiovascular	Systolic blood pressure	 V_E2F_Q6_Q1 +
chr1:206,939,904	CACO2;CD14;CD34;CD4;CD56; CMK;GM06990;GM12864; GM12865;GM12878;hTH1;hTH2	EGR1	Autoimmune_disease Diabetes	Crohns disease Ulcerative colitis Type_1_diabetes	 Egr1_secondary +
chr3:39,555,580	fBrain	SOX14	Neurological_behavioral	ADHD	 Sox14_secondary -
chr11:131,320,070	hESCT0	GCM1	Neurological_behavioral	Asperger disorder	 Gcm1_primary -
chr14:69,408,697	hESCT0	GLI1	Neurological_behavioral	Smoking behavior	 V_GLI_Q2 -
chr1:117,100,957	CD4;GM12864;GM12865; GM12878	IRF9	Autoimmune_disease	Multiple sclerosis	 Isgf3g_primary -
chr4:15,397,906	fBrain	PBX1	Neurological_behavioral	Conduct disorder (case_status)	 V_PBX1_Q2 +
chr8:34,126,948	fBrain	FOXA1	kidney_lung_liver neural	Neurological behavioral	 V_HNF3ALPHA_Q6 +

Table 8-7 GWAS SNPs at high-information positions in TF recognition sequences.

Transcription factor binding motifs overlapping GWAS SNPs were selected which had a large predicted difference in binding affinity between alleles, had a transcription factor consistent with the disease/trait class according to Gene Ontology annotation and which occurred in a DHS in a cell type compatible with the disease. Sequence logos derived from position weight matrices from the TRANSFAC and Jaspas databases. Box indicates position of GWAS SNP.

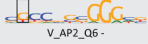
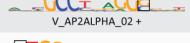
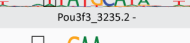
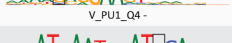
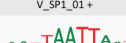
SNP	Target celltype	GO factor name	GWAS disease classes	GWAS diseases	Motif logo
chr4:15,397,906	fBrain	PBX1	Neurological_behavioral	Conduct disorder (case_status)	 Pbx1_3203.1 +
chr3:60,771,108	BE_2_C	TFAP2A	Neurological_behavioral	ADHD	 V_AP2_Q6 -
chr3:32,842,101	hESCT0	NR2F2	Neurological_behavioral	ADHD	 Nr2f2_secondary +
chr6:139,839,423	CMK;fHeart	EGR1	Hematological_parameters	Mean corpuscular volume	 V_EGR_Q6 -
chr2:12,568,996	HASp	FOXA2	Neurological_behavioral	Bipolar disorder and schizophrenia	 Foxa2_secondary +
chr21:17,828,291	HASp	PPARG	Neurological_behavioral	ADHD	 V_PPARG_Q2 +
chr15:68,198,911	BE_2_C;HAc	STAT3	Neurological_behavioral	Trails B	 V_STAT3STAT3_Q3 +
chr2:11,220,562	fBrain	TFAP2A	Neurological_behavioral	PC1	 V_AP2ALPHA_Q2 +
chr1:11,852,516	fHeart	NF1	Cardiovascular	Atrial fibrillation	 V_NF1_Q6 +
chr1:76,484,014	BE_2_C	POU3F3	Neurological_behavioral	Novelty seeking	 Pou3f3_3235.2 -
chr17:38,051,348	CD56;hTH2	SPI1	Autoimmune_disease	Ulcerative colitis	 V_PU1_Q4 -
chrX:44,400,186	fBrain	POU3F2	Neurological_behavioral	Personality dimensions	 V_POU3F2_Q1 -
chr7:50,269,672	GM12864;GM12865	GATA4	Autoimmune_disease	Crohns disease	 V_GATA4_Q3 +
chr6:142,750,516	SAEC	RORA	Kidney_lung_liver	FEV1/FVC	 V_RORA_Q4 -
chr11:131,320,070	hESCT0	GCM1	Neurological_behavioral	Asperger disorder	 V_GCM_Q2 -
chr16:20,367,690	fKidney	SP1	Kidney_lung_liver Serum_metabolite	Chronic kidney disease Creatinine	 V_SP1_Q1 +
chr1:93,401,837	fBrain	PAX6	Neurological_behavioral	Bipolar disorder and schizophrenia	 Pax6_3838.3 -
chr17:72,268,208	fBrain	KLF7	Neurological_behavioral	Information processing speed (inspection_time)	 Klf7_secondary +

Table 8-7 GWAS SNPs at high-information positions in transcription factor motifs. (cont'd)

8.5 Materials and Methods

Disease- and trait-associated variants from GWAS

The GWAS SNP set used for analysis was derived from the NHGRI GWAS Catalog, downloaded on January 4, 2012 (Hindorff et al.). The catalog is a continually-updated compendium of GWAS which lists the single SNP from each gene or region with the strongest disease association identified by the studies. Each study attempted to assay at least 100,000 SNPs across the genome. The catalog contained 6,896 entries at the time of download. We excluded SNPs mapping outside the main chromosome contigs, including the random chromosome fragments, SNPs without coordinates in the GRCh37/hg19 human genome assembly, SNPs without dbSNP ID, and records which were a combination of multiple SNPs associated with a disease or trait. The catalog contained data from 920 publications mapping 679 total diseases or traits. There were 6,011 unique SNP-disease/trait combinations; as some SNPs have been associated with more than one disease or trait, these represent 5,386 unique dbSNP IDs. Of these, 5,654 associations and 5,134 SNPs were in non-coding regions. Coding regions were defined by the CCDS Project (downloaded from the UCSC genome browser at <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/ccdsGene.txt.gz> on March 5, 2011) (Pruitt et al., 2009).

For some analyses, we grouped SNPs into classes of similar disease or trait, namely, aging-related; autoimmune disease; cancer; cardiovascular diseases and traits; diabetes-related; drug metabolism; hematological; kidney, lung, or liver; lipids, miscellaneous, neurological/behavioral; parasitic or bacterial disease; quantitative traits; radiographic (primarily bone density); serum metabolites; and viral disease.

Identification of replicated GWAS associations.

Not all reported associations from GWAS studies are replicated when tested in subsequent studies of the same disease or trait. We examined whether associations with stronger evidence were more likely to map to DNase I hypersensitive sites (DHSs). Data in the GWAS catalog was tabulated and the SNPs divided into three overlapping classes whose associations had varying levels of experimental support. SNPs were classified as “internally replicated” if the association was confirmed in a second replication population within the study as noted in the NHGRI GWAS Catalog. An association was classed as “externally replicated” if an associa-

tion was observed in a second publication linking the same disease or trait to the same SNP. Associations which were not yet replicated by a second sample population within the study or by an independent study were classed as “unreplicated”. A SNP could be included in both the “internally replicated” and “externally replicated” class; in such cases it was treated as externally replicated for the purpose of analysis.

DNase I mapping.

DNase I mapping was conducted on cultured cells, primary hematopoietic cells, and isolated fetal tissues using appropriate nuclei isolation protocols (Table 8-1).

Isolation of nuclei from cultured cells.

Cells were grown in accordance with protocols obtained from the source (Table 8-1). Freshly grown cells were centrifuged at 500g for 5 minutes (4°C) in an Eppendorf Centrifuge 5810R, and washed in cold PBS (Cellgro/Mediatech Inc.). Cell pellets were resuspended in Buffer A (15 mM Tris-Cl pH 8.0, 15 mM NaCl, 60 mM KCl, 1 mM EDTA (Ambion/Life Technologies Corp) pH 8.0, 0.5 mM EGTA (Boston BioProducts) pH 8.0, 0.5 mM spermidine (MP Biomedicals, LLC) and 0.15 mM spermine (MP Biomedicals, LLC) to a final concentration of 2×10^6 cells/mL. Nuclei were obtained by dropwise addition of an equal volume of Buffer A containing 0.04% IGEPAL CA-630 (Sigma-Aldrich) to the cells, followed by incubation on ice for 10 min. Nuclei were centrifuged at 1,000g for 5 min and then resuspended and washed with 25 mL of cold Buffer A. Nuclei were resuspended in 2 mL of Buffer A at a final concentration of 1×10^7 nuclei/mL.

Isolation of nuclei from hematopoietic cells.

Lymphocyte subclasses were isolated by immunomagnetic separation. Cells were pelleted by centrifugation for 5 minutes at 500g at 4 °C. Cells were washed in ice-cold PBS, then resuspended to 5 million cells per mL in Buffer A. An equal volume of ice-cold 2X IGEPAL CA-630 solution (ranging from 0.02%-0.06%) was added and the tube was incubated for 5 - 6 minutes on ice to lyse the cells. Nuclei were pelleted by centrifugation for 5 minutes at 500g at 4 °C, resuspended in Buffer A and counted with a hemocytometer.

Isolation of nuclei from fetal tissues.

Tissue was minced, resuspended in cold 250 mM sucrose, 1 mM MgCl₂, 10 mM Tris-Cl pH 7.5, with added EDTA Protease Inhibitor Cocktail (Roche Applied Science Corp.). Resuspended tissue from fetal brain, fetal lung, fetal kidney, and fetal adrenal was dissociated by slowly homogenizing with a Dounce homogenizer. Resuspended tissue from fetal heart or fetal intestine was dissociated in a gentleMACS Dissociator (Miltenyi Biotech Inc.). Following dissociation, all fetal tissues were filtered through a 100 µm filter, and nuclei pelleted by centrifugation 600g for 10 minutes. Pelleted nuclei were washed with Buffer A, resuspended in Buffer A and counted in a hemocytometer.

DNase I mapping from isolated nuclei.

Briefly, DNase I digestion was performed as described in an earlier publication (John et al., 2011), with minor modifications. Isolated nuclei (2×10^6) from suspension cells or dissociated tissue were washed with 15 mM Tris-Cl pH 8.0, 15 mM NaCl, 60 mM KCl, 1 mM EDTA pH 8.0, 0.5 mM EGTA pH 8.0, 0.5 mM spermidine and 0.15 mM spermine then subjected to DNase I digestion for 3 min at 37 °C in 13.5 mM Tris-HCl pH 8.0, 87 mM NaCl, 54 mM KCl, 6 mM CaCl₂, 0.9 mM EDTA, 0.45 mM EGTA, 0.45 mM Spermidine. Digestion was stopped by addition of 50 mM Tris-HCl pH 8.0, 100 mM NaCl, 0.1% SDS, 100 mM EDTA pH 8.0, 1 mM spermidine, 0.3 mM spermine. A range of DNase I (Sigma-Aldrich), 10–80 U/mL) concentrations was used for each preparation of nuclei and the sample giving the optimum difference between DNase I treated and untreated was used for sequencing library construction. DNase I double-hit fragments were collected by ultra-centrifugation and gel-purified. Adaptors were ligated to the ends of purified fragments, and the resulting libraries sequenced on an Illumina Genome Analyzer IIx according to a standard protocol.

Processing of DNase I-seq data.

For the ENCODE cell lines, the primary replicate was used for analysis. For the NIH Roadmap Epigenomics Consortium samples, data sets obtained from the tissues of fetal heart (12 developmental timepoint samples), fetal brain (12 developmental timepoint samples), fetal lung (34 developmental timepoint samples), fetal kidney (47 developmental timepoint samples), fetal intestine (15 developmental timepoint samples), fetal muscle

(48 developmental timepoint and anatomical localization samples), fetal placenta (4 developmental timepoint samples), fetal skin (17 samples, 14 of which correspond to 7 replicate pairs from the same individual, different anatomical locations, 2 of which correspond to 1 replicate pair from a different individual and timepoint, and one sample from a third individual), fetal spinal cord (3 developmental timepoint samples), fetal stomach (11 developmental timepoint samples), fetal thymus (10 developmental timepoint samples), fetal adrenal (5 developmental timepoint samples), neonatal skin fibroblasts (4 samples corresponding to 2 replicate pairs from 2 different individuals), and neonatal skin keratinocytes (4 samples corresponding to 2 replicate pairs from 2 different individuals), we pooled the data following hotspot calculation from all timepoints and samples into a single DNase I hypersensitivity profile for each tissue. 36-base reads with up to two mismatches were mapped to the human genome (GRCh37/hg19) using the sequence aligner BOWTIE (Langmead et al., 2009). DHSs were identified using the Hotspot algorithm (John et al., 2011) at a false discovery rate (FDR) threshold of 5%. Genomic feature overlaps and distance calculations were performed using the BEDOPS suite of software tools available at <http://code.google.com/p/bedops/> (Neph et al., 2012a).

Data availability

The DNase I data used in this study have been released as part of the ENCODE Project (Thurman et al., 2012) or the NIH Roadmap Epigenomics Mapping Consortium (Bernstein et al., 2010a). Data released through both projects and available (Table 8-1) include mapped reads and hotspots that have not been filtered for FDR thresholding. These data have been deposited in GEO under accession numbers GSE29692 and GSE18927. Data are also available for download through www.uwencode.org/data and through www.epigenomebrowser.org.

Enrichment of GWAS SNPs within DHSs relative to genomic space occupied.

The P -values for the enrichment of GWAS SNPs in DHSs, and various classes of DHSs, were computed using the binomial cumulative distribution function $b(x; n, p)$, the probability of x or more successes in n Bernoulli trials, with probability of success p . We used the R (R Development Core Team) function `pbinom` for calculating $b(x; n, p)$. We set the parameter n of the binomial equal to the total number of GWAS SNPs under

consideration. For a given class of DHS we set the parameter p equal to the fraction of the 36-mer uniquely-mappable GRCh37/hg19 genome occupied by the DHS class (using 2,630,301,437 uniquely mappable bp), and parameter x equal to the number of the SNPs overlapped by the DHSs.

For comparison of the overlap of GWAS SNPs and DHSs to the overlap of HapMap (International HapMap 3 Consortium et al., 2010) SNPs and DHSs, we obtained 4,029,798 CEPH population (Utah residents with ancestry from northern and western Europe, CEU) HapMap SNPs from the UCSC Genome Browser (release 27, merged Phase II + Phase III genotypes, lifted over from hg18 to hg19, downloaded from genome.ucsc.edu using the Table Browser). To compute the enrichment of GWAS SNPs in DHSs relative to the enrichment of HapMap SNPs in DHSs (Figure 8-2C), we set the expectation p equal to the fraction of HapMap SNPs overlapped by DHSs, we set n equal to the total number of GWAS SNPs, and we set x equal to the number of GWAS SNPs overlapped by DHSs.

Enrichment of GWAS SNPs in LD with SNPs in DHSs relative to randomly chosen 1KG SNPs.

We used CEU population genotype data from the 1000 Genomes Project (1000 Genomes Project Consortium, 2010) to compute the linkage disequilibrium (LD) measure r^2 between GWAS SNPs and SNPs in the DHSs near them. We converted the September 2010 release from GRCh36/hg18 to GRCh37/hg19 genomic coordinates using the UCSC Genome Browser liftOver tool. We then excluded SNPs for which a phased genotype was not available for all 60 CEU individuals sampled, or more than two alleles were present within the genotypes, or the minor allele frequency (MAF) was under 2/120. We then obtained the subset of these that were GWAS SNPs lying within intronic and intergenic regions ($n = 4,885$), using the CCDS gene definitions (Pruitt et al., 2009). We computed r^2 between each such GWAS SNP lying outside a DHS and every SNP within a 125 kb radius lying within a DHS. We partitioned the overall results into three categories: GWAS SNPs within DHSs, GWAS SNPs achieving $r^2 = 1$ with a SNP lying within a DHS within a 125 kb radius, and all GWAS SNPs not belonging to the first two categories.

For each of 4,885 non-coding GWAS SNPs meeting our filtering criteria, we drew a SNP at random from the subset of 1000 Genomes non-coding SNPs having the same MAF, approximate distance from the transcrip-

tion start site (TSS) of the nearest gene, and status of intronic or intergenic. This triple-matching procedure effectively accounts for any positional bias that might have been present in the SNP arrays. In addition to these three matching criteria, we also verified that the G+C content was the same between the GWAS SNPs and the matched control SNPs (Table 8-2).

We obtained 1,000 independent, randomly-drawn replicate data sets of 4,885 SNPs, each set matched to the non-coding GWAS SNPs. For each replicate data set, we performed the r^2 calculations and categorization of results as we had done for the GWAS SNPs. We tallied the percentages of SNPs falling into these categories within each random data set and fit a normal distribution to these data (Figure 8-3). To estimate the P -value for observing as many of the GWAS SNPs as we did within the first two categories, we computed the area of the upper tail of this distribution that exceeded the percentage of GWAS SNPs falling into these categories ($\sim 78\%$). The upper tail had no detectable area in the range beyond 100%. The percentage of non-coding GWAS SNPs within DHSs or achieving $r^2 = 1$ with a SNP in a nearby DHS is significant at the level $P < 10^{-37}$.

To verify that the DHSs showing such strong associations with possibly-functional GWAS SNPs are not merely surrogates for coding exons, we then removed from consideration any DHS overlapping any coding exon by at least 1 bp, and re-measured the percentages of GWAS and random-matched SNPs falling within a DHS. This only removed $\sim 4\%$ of the DHSs, covering ~ 45 Mbp, from the pool, and hence had a negligible effect. $\sim 77\%$ of non-coding GWAS SNPs were found to lie within these DHSs or be in complete LD with them ($P < 10^{-28}$).

Calculation of F_{ST} for GWAS SNPs

We identified all noncoding autosomal sites for which 1000 Genomes (1000 Genomes Project Consortium, 2010) had fully-phased genotypes in both the CEU and Yoruba from Nigeria (YRI) populations, and partitioned these into sites within DHSs and sites outside of DHSs. We then chose 150,000 of these DHS sites at random, in the same proportion of intergenic to intronic sites that we observed in all noncoding 1000 Genomes CEU data across the autosomes (70.8% intergenic, 29.2% intronic). Next, for each intergenic DHS

SNP, we chose an intergenic non-DHS SNP with the same minor allele frequency in CEU located at approximately the same distance from its nearest TSS, and did likewise for the intronic DHS SNPs. We filtered out any site at which the MAF pooled across the populations' genotypes fell below 10%, leaving us with 122,648 SNPs in the within-DHSs set and 122,810 SNPs in the non-DHS set. We computed F_{ST} and obtained values of 0.08433 and 0.08455 for these two SNP sets, respectively. Relaxing the restriction of matching on distance to the nearest TSS did not yield a significantly different result (0.08468). We also observed virtually no difference in F_{ST} between the two SNP sets when relaxing the constraint on MAF to 5% and 0%.

Disease-specific enrichment of GWAS SNPs in DHSs and fetal-origin DHSs

We computed the enrichment of GWAS SNPs from particular diseases or traits in DHSs (Figure 8-23) by dividing the proportion of GWAS SNPs in DHSs by the overall proportion of GWAS SNPs in DHSs (57.1%). Enrichments are reported as percentage enrichment or depletion. The individual significances of these enrichments was computed using the binomial distribution $b(x; n, p)$, setting the parameter x to the number of GWAS SNPs of a given disease or trait in DHSs, n to the number of GWAS SNPs for the disease or trait, and p to 0.571.

We computed the enrichment of GWAS SNPs from particular diseases or traits in fetal-origin DHSs (Figure 8-8) by dividing the proportion of GWAS SNPs in fetal-origin DHSs by the overall proportion of GWAS SNPs in fetal-origin DHSs (88.1%). Enrichments are reported as percentage enrichment or depletion. The individual significances of these enrichments was computed using the binomial distribution $b(x; n, p)$, setting the parameter x to the number of GWAS SNPs of a given disease or trait in fetal-origin DHSs, n to the number of GWAS SNPs for the disease or trait, and p to 0.881. To compensate for the overall enrichment or depletion of disease categories in DHSs in general, GWAS SNPs not in any DHS were excluded.

DHS-to-promoter assignments based on cross-cell-type hypersensitivity correlations.

We previously measured DHSs genome-wide across 79 diverse cell types, and performed correlation analyses on the patterns of DNase I occupancy across the cell types (Thurman et al., 2012). Briefly, we first collapsed the 79 cell types into 32 categories, based on the similarities and differences of their DHS profiles genome-

wide (Table 8-4). Then for each DHS, we formed a 32-element vector of DNase I tag counts to represent the occupancy pattern within those cell types at that DHS. Then for each promoter DHS representing a Gencode TSS, we computed the correlation between its occupancy pattern vector and the vector for each non-promoter DHS distal to it within a 500 kb radius. We defined a distal/promoter DHS pair to be “connected” if its Pearson correlation coefficient R was at least 0.7. We identified 578,905 connected distal DHSs genome-wide (mean separation = 266 kb), 429,283 (74%) of which hop over an adjacent gene to find its highest correlation with a different gene farther away within a 500-kb radius.

Here we used this correlation map to obtain a set of 296 unique non-coding GWAS SNPs lying within distal DHSs achieving $R \geq 0.7$ with a promoter DHS within 500 kb (Table 10-1). We also repeated this analysis using DHSs found in 46 cell types that were used for other analyses in this paper but not included among the 79 used for the above (Table 8-5). This correlation map identified an additional 123 unique non-coding GWAS SNPs lying within distal DHSs achieving $R \geq 0.7$ with a promoter DHS within 500 kb (Table 10-2).

To establish the extent of LD between the distal and promoter DHSs, we computed r^2 between all pairs of 1000 Genomes SNPs fully phased in the CEU population and with minor allele frequency $\geq 5\%$ lying within 2 kb of the DHS containing the GWAS SNP and lying within 2 kb of the promoter DHS. For a typical DHS pair, we computed ~ 127 r^2 values, between ~ 14 SNPs at one DHS and ~ 9 SNPs at the other.

Two replicates of PolII ChIA-PET data (Li et al., 2012) in K562 cells were obtained from the UCSC Genome Browser (<http://hgdownload-test.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGisChiaPet/>) and processed with awk (Aho et al., 1988).

Transcription factor motif data.

Potential sites of transcription factor binding were identified by scanning relevant regions utilizing position weight matrices from three major transcription factor binding motif repositories: TRANSFAC (Matys et al., 2006), JASPAR (Portales-Casamar et al., 2010), and UniPROBE (Newburger and Bulyk, 2009). To avoid ascertainment bias for motifs better matching the reference allele of common polymorphisms, we created an alternate genome to complement the reference GRCh37/hg19 human genome. This alternate genome incor-

porates the non-reference allele at the location of each SNP identified in the CEU population of the 1000 Genomes Project (1000 Genomes Project Consortium, 2010). Regions in the vicinity of GWAS or control SNPs were then scanned for motifs in both the reference and alternate genomes with a threshold $P \leq 10^{-4}$ using the program FIMO (Grant et al., 2011).

Mapping transcription factors to GWAS disease/trait classes.

We used information from the Gene Ontology (GO) (Ashburner et al., 2000) to identify potentially relevant motif matches. All GO biological processes for 282 transcription factors were extracted from the Gene Ontology MySQL database. For each disease/trait class, a collection of key terms which could identify factors potentially involved in the class was developed and used to search the list of GO biological processes associated with each transcription factor for which a position weight matrix was available (Table 8-6). Many transcription factors were found to be consistent with multiple disease/trait classes. The set of transcription factor motifs detected ($P < 10^{-4}$), with at least one Gene Ontology Biological Process matching search terms for the disease/trait class and which overlapped GWAS SNPs in a DHS was identified and used for subsequent pathway/interaction analyses.

For our measurements of GWAS SNP enrichment within transcription factor motif groups, we began by forming a matrix of potential associations between transcription factor GO groups (e.g., aging) and disease classes (e.g., cancer). We computed the relative frequency with which GWAS SNPs from a particular disease class localized within the recognition sequence of a transcription factor annotated with related physiological processes, and derived a P -value using the binomial distribution $b(x; n, p)$, setting the first parameter to the number of GWAS SNPs present in the given factor group, and the second parameter to the proportion of GWAS SNPs belonging to the given disease class.

Allelic imbalance in chromatin accessibility.

We first called heterozygous SNPs directly from the DNase I reads. At each of the 5,386 unique GWAS SNPs (coding and non-coding), reads were extracted from DNase I alignments using SAMtools (Li et al., 2009), and compared to the GRCh37/hg19 human reference sequence. To reduce the risk of false positives

due to sequencing errors, only GWAS SNPs identified either in the 1000 Genomes Project's low-coverage CEU population data, or Complete Genomics' 54-individual sample were considered. To correct for mapping bias caused by the extra mismatch in reads containing the non-reference allele a less-stringent mismatch threshold was applied. Reads containing the reference allele were only counted if they contained zero or one base mismatches (over the entire read length) to the reference sequence; reads with the non-reference allele were counted if they had one or two base mismatches (one of which is the SNP). Any SNPs located within one read-length (36 bp) of another known SNP, represented by more than one chromosome in either sample from 1000 Genomes or Complete Genomics, were excluded from this analysis. Samples were called heterozygous at a SNP if each known allele was represented by reads aligned to at least three distinct positions (unique genomic coordinate and strand).

We identified 872 heterozygous SNPs, and pooled allele counts from all heterozygous samples. Confirming our strategy for avoiding reference mapping bias, we observed 412 SNPs with more reads from the reference allele, 416 SNPs with more reads containing the non-reference allele, and 44 SNPs with an equal amount of reads. Sites with fewer than 21 reads were excluded for lack of power to test for allelic imbalance. We then tested the remaining 584 sites for imbalance using a two-tailed binomial test. We calculated a false discovery rate using the R package qvalue (Storey and Tibshirani, 2003). To set an overall cutoff for significantly imbalanced sites, we simulated 200 random sets of read counts at 584 sites using the binomial distribution, with the ratios at imbalanced sites sampled from the actual data. We tested the power of our method to correctly discover imbalanced sites, and measured the actual false discovery rate to be $< 5\%$ for a cutoff of $P < 0.025$.

Transcription factor-centered networks.

We obtained factors involved in maturity onset diabetes of the young (MODY, Figure 8-15), as well as those interacting with OTX1 (Figure 8-18), IRF9 (Figure 8-16), and ESRRA (Figure 8-19) using Ingenuity Pathways Analysis (Ingenuity Systems, www.ingenuity.com). The transcription factors in each network with known sequence specificities were examined for overlap with non-coding GWAS SNPs in DHSs in all cell types (the IRF9 network was restricted to cell types related to immune function: CD3+, CD3+_Cord_Blood, CD4+, CD8+, CD14+, CD19+, CD20+, CD34+, CD56+, fThymus, GM06990, GM12864, GM12865, GM12878,

Th1, Th2, Th17, Jurkat). MODY factors were examined for GWAS SNPs associated with Type 1 or Type 2 diabetes or glucose metabolism-related traits. OTX1-interacting, IRF9-interacting, and ESRRA-interacting factors were examined for GWAS SNPs associated with neurological, autoimmune, and cancer classes, respectively. We tested the significance of the enrichment of disease-relevant GWAS SNPs in the TFs in these networks against the enrichment of random SNPs in the network. We used a binomial distribution with the parameter p set to the proportion of non-coding SNPs in the Affymetrix 500K genotyping array overlapping motifs of the same TFs.

Disease networks.

For the autoimmune network (Figure 8-20A), we used SNPs of the autoimmune class plus SNPs associated with Type 1 diabetes. We examined only GWAS SNPs in DHSs from cell types related to immune function and tallied the number of GWAS SNPs associated with autoimmune disease. Transcription factors overlapping 8 or more GWAS SNPs are shown.

For the cancer network (Figure 8-20B), we used a set of GWAS SNPs associated with cancer in DHSs from all tissue types. Transcription factors overlapping 6 or more GWAS SNPs are shown.

For the psychiatric network (Figure 8-21), we used a set of GWAS SNPs associated with psychiatric diseases which were present in DHSs of fetal brain. Transcription factors overlapping 3 or more GWAS SNPs are shown, except for FOXP1 and FOXP3, which were removed from the network due to lack of hypersensitivity at their promoter DHSs.

For each network, we computed the significance of finding a set of TFs whose recognition sequences overlap such a high number of GWAS SNPs by comparing to random equally-sized samples of non-coding SNPs from the Affymetrix 500K genotyping array (10,000 replicates). P -values were estimated using a fitted Poisson distribution.

Cell-type selective GWAS variant-DHS enrichment analysis.

At a given P -value threshold, enrichment in a cell type's DHSs was calculated as the fraction of SNPs with a P -value below that threshold that overlap DHSs, divided by the fraction of all non-coding SNPs in the study

that overlap DHSs. Malignancy-derived cell lines were excluded. Enrichments were tested at P -value thresholds from 1.0 to 10^{-75} . The thresholds were chosen as powers of ten which approximately halved the number of additional SNPs included at each successively-lower threshold. The smallest threshold was chosen to retain sufficient sample size (>100 SNPs). The statistical significance of each enrichment was measured with a one-sided Fisher's exact test, implemented in R's "fisher.test" function.

8.6 Notes

This chapter has been published as **Maurano MT***, Humbert R*, Rynes E*, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kuttyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, and Stamatoyannopoulos JA. "Systematic Localization of Common Disease-Associated Variation in Regulatory DNA." **Science** 337: 1190–1195 (2012). * *co-first authors*

Acknowledgments

We thank many colleagues for their insightful comments and critical readings of the manuscript. We also thank many colleagues who provided individual cell samples for DNase I analysis. This work was supported by National Institutes of Health grants U54HG004592 and U01ES01156 (J.A.S.); P30 DK056465 (S.H.); R01HL088456 (N.S.); and R24HD000836-47 (I.G.). We would like to acknowledge the generous sharing of results from the International MS Genetics, International IBD Genetics, and CHARGE QRS Consortia. DNase I data have been deposited in GEO under accession numbers GSE29692 and GSE18927, and are also available for viewing and download at genome.ucsc.edu, www.uwencode.org/data and www.epigenomebrowser.org.

9 Conclusion

9.1 Summary

Understanding individual variation in gene expression is a critical challenge in modern human genetics. Work in the last decade has linked natural polymorphism with hereditary differences in mRNA levels, and extended this line of analysis to modeling their effect on disease states with Bayesian networks. It has since become clear that natural genetic variation has dramatic effects on cellular machinery. This leads to a number of fundamental questions addressed herein: What of this variation is directly functional? What causes differences in gene expression? What general regulatory mechanisms are there? Can high-throughput methods extend the lessons of model organism genetics to complex natural populations? These questions are fundamental to understanding the regulation of living organisms. Yet the key to understanding lies in carefully-designed analyses which distill the overwhelming precision of catalogs of human genetic diversity into relevant insight.

Along these lines, my thesis first addressed the broad question of how transcription factors select their genomic binding sites, and additionally how human polymorphism affects transcription factor binding *in vivo*. I have built an empirical model of binding specificity for the transcription factor CTCF which goes beyond the canonical motif representation and identifies different modes of binding, as well as using the genetic variation segregating in a three-generation pedigree to study systematic patterns of inter-individual variation in binding. I further studied variation in binding among different cell types, showing a great diversity in regulatory programs. My work showing a link between cell-type specific binding and methylation patterns suggests that methylation plays a stabilizing rather than dominant role in TF occupancy. In the second part of my thesis, I applied this framework to the study of human disease-associated variation, showing that this variation is concentrated in regulatory DNA. Importantly, the cell-type specific nature of these binding elements can be used to identify involved tissue types. This work establishes a strong basis for the global involvement of regulatory elements in common human disease, and suggests manifold avenues of follow-up research.

9.2 Discussion And Future Directions

Genetic architecture of human disease

The identification of a shared regulatory basis for hundreds of small-effect-size variants lends new perspective to the debate regarding the genetic architecture of human disease frequency (Gibson, 2012; Park et al., 2010; Pritchard, 2001; Reich and Lander, 2001). Indeed, recent work uncovering a deep landscape of low-effect size associations in LD with common SNPs with the potential to explain a substantial proportion of the trait's heritability (International Multiple Sclerosis Genetics Consortium (IMSGC) et al., 2010b; Lango Allen et al., 2010; Lee et al., 2011; Purcell et al., 2007; Yang et al., 2010). Our results suggest the quantitative contribution of hundreds of small effect size transcription factor binding sites, in contrast with Mendelian variants that perturb entire networks through modulation of transcription factor specificity, and highlights a continuous quantitative spectrum of disordered gene regulation between common disease and Mendelian traits.

Importance of relevant cell types, studied in vivo

As shown in Section 4, although functional polymorphism in metazoan regulatory elements is amenable to study, this is only possible with relevant *in vivo* data. Notably, sequence-based approaches such as phylogenetic conservation are expected to continue to provide inadequate power for comprehensive identification of regulatory elements. Equally, while position weight matrices or other models representing *in vitro* determined binding affinity are capable of recognizing functional variation, they can only be applied within the actively bound regulatory component of a given cell type – which again must be determined *in vivo*. Likewise, *in vivo* attempts to study the effect of polymorphism on regulatory elements hinge on the use of a nuclear context including the appropriate trans-acting factors and programmed chromatin landscape. DNase I mapping is expected to remain the only universal approach for the quantitative determination of active regulatory DNA (Guertin et al., 2012; Kaplan et al., 2011). Thus, a comprehensive catalog of precisely isolated and minimally cultured primary cell types is imperative for future studies of human regulatory genetics.

Focused regulatory association studies

The DNase I-hypersensitive landscapes of the cell types studied herein mark a precise and highly specific genomic compartment enriched in the recognition sequences of the range of cell-type specific regulatory factors. Although the combined set of DHSs from 136 cell types covers 42.2% of the genome, the active regulatory DNA in an individual cell type covers only 2.9% of the genome on average, offering potential for a 100-fold enrichment. Even within this narrow compartment, transcription factor recognition sequences are 2-fold concentrated in DNase I footprints (Neph et al., 2012b). Functional genomics now offers the potential to interpret the regulatory sequence code of these elements (Section 4), enabling true genome-wide analysis of sequence data from disease samples. Alternatively, the systematic sequencing of regulatory regions offers a promising alternative to monumental sample sizes for common disease, much as exome studies have accelerated the genetics of rare disease.

Functional exploration of disease-associated regulatory sites

Although we have demonstrated the utility of a survey of regulatory elements in a wide variety of cell types, the next challenge is to identify their regulatory interactions. Here and in other work we have made progress in identifying both patterns of co-occupancy at individual sites (Neph et al., 2012b), as well as genome-wide correlations of chromatin accessibility indicating global co-regulation (Section 8). These methods will need to be extended and refined to establish the global context of regulatory elements genome-wide.

The choice of model system used for functional follow-up of the disease-associated regulatory variants identified herein is critical. Reverse genetics in an isogenic setting (e.g., knockout mouse lines) is the gold standard for determination of function. However, until recently the ability to perform validation at this level has not been possible in the human genome. The availability of designed site-specific nucleases (Soldner et al., 2011) offers a potent approach for achieving this in human cells. Zinc finger nucleases (ZFNs) are now being used extensively in the lab to perform genome editing of regulatory DNA, ranging from individual DNase I hypersensitive sites to complex regulatory regions. By generating knockouts of regulatory DNA and then performing genome-wide DNase I mapping, it is possible to assay the functional consequences of a specific regulatory region, and to discover systematically its target gene(s). Because the isogenetic setting allows for pairwise

comparison with a control sample, downstream effects of the deletion can be identified as differences in DNase I hypersensitivity and gene expression with high sensitivity. This approach offers both validation of the functional consequence of the disease variant as well as clues to its mechanistic involvement in neurological physiology through its role in the control of nearby or distant genes, and its impact on global gene expression profiles.

Conversely, our results in Section 8.2 highlight the complementary perturbations associated with common and Mendelian variants on transcriptional networks. While this thesis has discussed the perturbations in transcription factor binding sites associated with common variants, recent work has identified frameshift mutations in transcription factor coding sequence associated with many diseases, including autism (O'Roak et al., 2011). Some of these associated mutations might be expected to demonstrate a more subtle effect, maintaining protein expression and stability but altering DNA sequence specificity (Filippova et al., 2002). In turn, this altered specificity may affect only a subset of binding sites. Such limited consequences are more likely to be compatible with life, yet may cause developmental defects manifest later on. High-throughput methods such as SELEX are now available to characterize binding specificity (Bulyk et al., 2001; Johnson et al., 2007; Jolma et al., 2010), allowing the identification of the direct downstream consequences of coding variants in transcription factors, amongst a background of indirect differences.

References

- 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073.
- Abecasis, G.R., Cherny, S.S., Cookson, W.O., and Cardon, L.R. (2001). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* **30**, 97–101.
- Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249.
- Aho, A.V., Kernighan, B.W., and Weinberger, P.J. (1988). *The AWK Programming Language* (Reading, MA: Addison Wesley).
- Akalin, A., Garrett-Bakelman, F.E., Kormaksson, M., Busuttill, J., Zhang, L., Khrebtukova, I., Milne, T.A., Huang, Y., Biswas, D., Hess, J.L., et al. (2012). Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. *PLoS Genet.* **8**, e1002781.
- Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits* (Boca Raton, FL: Chapman and Hall/CRC).
- Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106.
- Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29.
- Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723.
- 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–W208.
- Baniahmad, A., Steiner, C., Köhne, A.C., and Renkawitz, R. (1990). Modular structure of a chicken lysozyme silencer: involvement of an unusual thyroid hormone receptor binding site. *Cell* **61**, 505–514.
- Barker, D.J., Gluckman, P.D., Godfrey, K.M., Harding, J.E., Owens, J.A., and Robinson, J.S. (1993). Fetal nutrition and cardiovascular disease in adult life. *Lancet* **341**, 938–941.
- Baubec, T., Ivanek, R., Lienert, F., and Schübeler, D. (2013). Methylation-dependent and -independent genomic targeting principles of the MBD protein family. *Cell* **153**, 480–492.
- Bell, A.C., and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the *Igf2* gene. *Nature* **405**, 482–485.
- Benos, P.V., Bulyk, M.L., and Stormo, G.D. (2002). Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res.* **30**, 4442–4451.
- Berg, J., Willmann, S., and Lässig, M. (2004). Adaptive evolution of transcription factor binding sites. *BMC Evol. Biol.* **4**, 42.

- Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., et al. (2010a). The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol.* **28**, 1045–1048.
- Bernstein, I., Byun, H.-M., Mohrbacher, A., Douer, D., Gorospe, G., Hergesheimer, J., Groshen, S., O'Connell, C., and Yang, A.S. (2010b). A phase I biological study of azacitidine (VidazaTM) to determine the optimal dose to inhibit DNA methylation. *Epigenetics* **5**, 750–757.
- Berry, M., Grosveld, F., and Dillon, N. (1992). A single point mutation is the cause of the Greek form of hereditary persistence of fetal haemoglobin. *Nature* **358**, 499–502.
- Blow, M.J., McCulley, D.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2010). ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* **42**, 806–810.
- Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M., and Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. *Science* **317**, 815–819.
- Bowers, S.R., Mirabella, F., Calero-Nieto, F.J., Valeaux, S., Hadjur, S., Baxter, E.W., Merckenschlager, M., and Cockerill, P.N. (2009). A conserved insulator that recruits CTCF and cohesin exists between the closely related but divergently regulated interleukin-3 and granulocyte-macrophage colony-stimulating factor genes. *Mol. Cell. Biol.* **29**, 1682–1693.
- Boyle, A.P., Song, L., Lee, B.-K., London, D., Keefe, D., Birney, E., Iyer, V.R., Crawford, G.E., and Furey, T.S. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Res.* **21**, 456–464.
- Brand, S. (2009). Crohn's disease: Th1, Th17 or both? The change of a paradigm: new immunological and genetic insights implicate Th17 cells in the pathogenesis of Crohn's disease. *Gut* **58**, 1152–1167.
- Brandeis, M., Frank, D., Keshet, I., Siegfried, Z., Mendelsohn, M., Nemes, A., Temper, V., Razin, A., and Cedar, H. (1994). Sp1 elements protect a CpG island from de novo methylation. *Nature* **371**, 435–438.
- Brem, R.B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755.
- Bulyk, M.L., Huang, X., Choo, Y., and Church, G.M. (2001). Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 7158–7163.
- Butcher, D.T., Mancini-DiNardo, D.N., Archer, T.K., and Rodenhiser, D.I. (2004). DNA binding sites for putative methylation boundaries in the unmethylated region of the BRCA1 promoter. *Int. J. Cancer* **111**, 669–678.
- Büdingen, von, H.-C., Bar-Or, A., and Zamvil, S.S. (2011). B cells in multiple sclerosis: connecting the dots. *Current Opinion in Immunology* **23**, 713–720.
- Chadwick, B.P. (2008). DXZ4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive X-specific role involving CTCF and antisense transcripts. *Genome Res.* **18**, 1259–1269.
- Chernukhin, I., Shamsuddin, S., Kang, S.Y., Bergstrom, R., Kwon, Y.W., Yu, W., Whitehead, J., Mukhopad-

- yay, R., Docquier, F., Farrar, D., et al. (2007). CTCF Interacts with and Recruits the Largest Subunit of RNA Polymerase II to CTCF Target Sites Genome-Wide. *Mol. Cell. Biol.* 27, 1631–1648.
- Chung, J.H., Whiteley, M., and Felsenfeld, G. (1993). A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in *Drosophila*. *Cell* 74, 505–514.
- Collings, C.K., Waddell, P.J., and Anderson, J.N. (2013). Effects of DNA methylation on nucleosome stability. *Nucleic Acids Res.* 41, 2918–2931.
- Collins, F.S., Metherall, J.E., Yamakawa, M., Pan, J., Weissman, S.M., and Forget, B.G. (1985). A point mutation in the A gamma-globin gene promoter in Greek hereditary persistence of fetal haemoglobin. *Nature* 313, 325–326.
- Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* 10, 184–194.
- Cooper, G.M., Goode, D.L., Ng, S.B., Sidow, A., Bamshad, M.J., Shendure, J., and Nickerson, D.A. (2010). Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods* 7, 250–251.
- Cotsapas, C., Voight, B.F., Rossin, E., Lage, K., Neale, B.M., Wallace, C., Abecasis, G.R., Barrett, J.C., Behrens, T., Cho, J., et al. (2011). Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.* 7, e1002254.
- Coventry, A., Bull-Otterson, L.M., Liu, X., Clark, A.G., Maxwell, T.J., Crosby, J., Hixson, J.E., Rea, T.J., Muzny, D.M., Lewis, L.R., et al. (2010). Deep resequencing reveals excess rare recent variants consistent with explosive population growth. *Nature Commun.* 1, 131–136.
- Crooks, G.E., Hon, G., Chandonia, J.-M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190.
- Crosslin, D.R., McDavid, A., Weston, N., Nelson, S.C., Zheng, X., Hart, E., de Andrade, M., Kullo, I.J., McCarty, C.A., Doheny, K.F., et al. (2011). Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. *Hum. Genet.*
- Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.Y., Cui, K., and Zhao, K. (2008). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res.* 19, 24–32.
- Dávalos-Salas, M., Furlan-Magaril, M., González-Buendía, E., Valdes-Quezada, C., Ayala-Ortega, E., and Recillas-Targa, F. (2011). Gain of DNA methylation is enhanced in the absence of CTCF at the human retinoblastoma gene promoter. *BMC Cancer* 11, 232.
- De Gobbi, M., Viprakasit, V., Hughes, J.R., Fisher, C., Buckle, V.J., Ayyub, H., Gibbons, R.J., Vernimmen, D., Yoshinaga, Y., de Jong, P., et al. (2006). A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* 312, 1215–1217.
- Deaton, A.M., and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes Dev.* 25, 1010–1022.
- 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., et al. (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.

Dermitzakis, E.T., and Clark, A.G. (2002). Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* 19, 1114–1121.

Dickson, J., Gowher, H., Strogantsev, R., Gaszner, M., Hair, A., Felsenfeld, G., and West, A.G. (2010). VEZF1 elements mediate protection from DNA methylation. *PLoS Genet.* 6, e1000804.

Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., et al. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* 325, 1246–1250.

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.

Docquier, F., Farrar, D., D'Arcy, V., Chernukhin, I., Robinson, A.F., Loukinov, D., Vatolin, S., Pack, S., Mackay, A., Harris, R.A., et al. (2005). Heightened expression of CTCF in breast cancer cells is associated with resistance to apoptosis. *Cancer Res.* 65, 5112–5122.

Donohoe, M.E., Silva, S.S., Pinter, S.F., Xu, N., and Lee, J.T. (2009). The pluripotency factor Oct4 interacts with Ctfc and also controls X-chromosome pairing and counting. *Nature* 460, 128–132.

Donohoe, M.E., Zhang, L.-F., Xu, N., Shi, Y., and Lee, J.T. (2007). Identification of a Ctfc Cofactor, Yy1, for the X Chromosome Binary Switch. *Mol. Cell* 25, 43–56.

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

Eddy, S.R. (2005). A Model of the Statistical Power of Comparative Genome Sequence Analysis. *PLoS Biol.* 3, e10.

Eldar, A., and Elowitz, M.B. (2010). Functional roles for noise in genetic circuits. *Nature* 467, 167–173.

Emery, D.W., Yannaki, E., Tubb, J., and Stamatoyannopoulos, G. (2000). A chromatin insulator protects retrovirus vectors from chromosomal position effects. *Proc. Natl. Acad. Sci. U.S.A.* 97, 9150–9155.

Emery, D.W. (2011). The use of chromatin insulators to improve the expression and safety of integrating gene transfer vectors. *Hum. Gene Ther.* 22, 761–774.

ENCODE Project Consortium (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* 9, e1001046.

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

Engel, N., West, A.G., Felsenfeld, G., and Bartolomei, M.S. (2004). Antagonism between DNA hypermethylation and enhancer-blocking activity at the H19 DMD is uncovered by CpG mutations. *Nat. Genet.* 36, 883–888.

- Enver, T., Zhang, J.W., Papayannopoulou, T., and Stamatoyannopoulos, G. (1988). DNA methylation: a secondary event in globin gene switching? *Genes Dev.* *2*, 698–706.
- Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shores, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* *473*, 43–49.
- Fajans, S.S., Bell, G.I., and Polonsky, K.S. (2001). Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young. *N. Engl. J. Med.* *345*, 971–980.
- Farnham, P.J. (2009). Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.* *10*, 605–616.
- Feingold, E. (2002). Regression-based quantitative-trait-locus mapping in the 21st century. *Am. J. Hum. Gen.* *71*, 217–222.
- Filippova, G.N., Fagerlie, S., Klenova, E.M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P.E., Collins, S.J., and Lobanenko, V.V. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Mol. Cell. Biol.* *16*, 2802–2813.
- Filippova, G.N., Lindblom, A., Meincke, L.J., Klenova, E.M., Neiman, P.E., Collins, S.J., Doggett, N.A., and Lobanenko, V.V. (1998). A widely expressed transcription factor with multiple DNA sequence specificity, CTCF, is localized at chromosome segment 16q22.1 within one of the smallest regions of overlap for common deletions in breast and prostate cancers. *Genes Chromosomes Cancer* *22*, 26–36.
- Filippova, G.N., Thienes, C.P., Penn, B.H., Cho, D.H., Hu, Y.J., Moore, J.M., Klesert, T.R., Lobanenko, V.V., and Tapscott, S.J. (2001). CTCF-binding sites flank CTG/CAG repeats and form a methylation-sensitive insulator at the DM1 locus. *Nat. Genet.* *28*, 335–343.
- Filippova, G.N., Qi, C.-F., Ulmer, J.E., Moore, J.M., Ward, M.D., Hu, Y.J., Loukinov, D.I., Pugacheva, E.M., Klenova, E.M., Grundy, P.E., et al. (2002). Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter its DNA-binding specificity. *Cancer Res.* *62*, 48–52.
- Franke, A., McGovern, D.P.B., Barrett, J.C., Wang, K., Radford-Smith, G.L., Ahmad, T., Lees, C.W., Balschun, T., Lee, J., Roberts, R., et al. (2010). Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* *42*, 1118–1125.
- Frazer, K.A., Narla, G., Zhang, J.L., and Rubin, E.M. (1995). The apolipoprotein(a) gene is regulated by sex hormones and acute-phase inducers in YAC transgenic mice. *Nat. Genet.* *9*, 424–431.
- 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., et al. (2009). An oestrogen-receptor- α -bound human chromatin interactome. *Nature* *462*, 58–64.
- 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.
- Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., et al. (2010). A map of open chromatin in human pancreatic islets. *Nat. Genet.* *42*, 255–259.
- Gebhard, C., Benner, C., Ehrlich, M., Schwarzfischer, L., Schilling, E., Klug, M., Dietmaier, W., Thiede, C., Holler, E., Andreesen, R., et al. (2010). General Transcription Factor Binding at CpG Islands in Normal Cells

Correlates with Resistance to De novo DNA Methylation in Cancer Cells. *Cancer Res.* *70*, 1398–1407.

Gelinas, R., Endlich, B., Pfeiffer, C., Yagi, M., and Stamatoyannopoulos, G. (1985). G to A substitution in the distal CCAAT box of the A gamma-globin gene in Greek hereditary persistence of fetal haemoglobin. *Nature* *313*, 323–325.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nat. Rev. Genet.* *13*, 135–145.

Gieger, C., Radhakrishnan, A., Cvejic, A., Tang, W., Porcu, E., Pistis, G., Serbanovic-Canic, J., Elling, U., Goodall, A.H., Labrune, Y., et al. (2011). New gene functions in megakaryopoiesis and platelet formation. *Nature* *480*, 201–208.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., and Lieb, J.D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* *17*, 877–885.

Goecks, J., Nekrutenko, A., Taylor, J., Galaxy Team (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.* *11*, R86.

Grant, C.E., Bailey, T.L., and Noble, W.S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics* *27*, 1017–1018.

Grice, E.A., Rochelle, E.S., Green, E.D., Chakravarti, A., and McCallion, A.S. (2005). Evaluation of the RET regulatory landscape reveals the biological relevance of a HSCR-implicated enhancer. *Hum. Mol. Gen.* *14*, 3837–3845.

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

Groudine, M., and Conkin, K.F. (1985). Chromatin structure and de novo methylation of sperm DNA: implications for activation of the paternal genome. *Science* *228*, 1061–1068.

Guertin, M.J., Martins, A.L., Siepel, A., and Lis, J.T. (2012). Accurate prediction of inducible transcription factor binding intensities in vivo. *PLoS Genet.* *8*, e1002610.

Gumucio, D.L., Shelton, D.A., Zhu, W., Millinoff, D., Gray, T., Bock, J.H., Slightom, J.L., and Goodman, M. (1996). Evolutionary strategies for the elucidation of cis and trans factors that regulate the developmental switching programs of the beta-like globin genes. *Mol. Phylogenet. Evol.* *5*, 18–32.

Hagemann, S., Heil, O., Lyko, F., and Brueckner, B. (2011). Azacytidine and decitabine induce gene-specific and non-random DNA demethylation in human cancer cell lines. *PLoS ONE* *6*, e17388.

Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W.H., Ye, C., Ping, J.L.H., Mulawadi, F., et al. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. *Nat. Genet.* *43*, 630–638.

Harismendy, O., Notani, D., Song, X., Rahim, N.G., Tanasa, B., Heintzman, N., Ren, B., Fu, X.-D., Topol, E.J., Rosenfeld, M.G., et al. (2011). 9p21 DNA variants associated with coronary artery disease impair interferon- γ signalling response. *Nature* *470*, 264–268.

Hark, A.T., Schoenherr, C.J., Katz, D.J., Ingram, R.S., Levorse, J.M., and Tilghman, S.M. (2000). CTCF medi-

ates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* 405, 486–489.

He, B.Z., Holloway, A.K., Maerkl, S.J., and Kreitman, M. (2011). Does Positive Selection Drive Transcription Factor Binding Site Turnover? A Test with *Drosophila* Cis-Regulatory Modules. *PLoS Genet.* 7, e1002053.

Heap, G.A., Yang, J.H.M., Downes, K., Healy, B.C., Hunt, K.A., Bockett, N., Franke, L., Dubois, P.C., Mein, C.A., Dobson, R.J., et al. (2009). Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. *Hum. Mol. Gen.* 19, 122–134.

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

Hellman, L.M., and Fried, M.G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc.* 2, 1849–1861.

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

Hindorff, L., MacArthur, J., Wise, A., Junkins, H., Hall, P., Klemm, A., and Manolio, T. A Catalog of Published Genome-Wide Association Studies.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. U.S.A.* 106, 9362–9367.

Hou, C., Zhao, H., Tanimoto, K., and Dean, A. (2008). CTCF-dependent enhancer-blocking by alternative chromatin loop formation. *Proc. Natl. Acad. Sci. U.S.A.* 105, 20398–20403.

Huppi, K., Pitt, J.J., Wahlberg, B.M., and Caplen, N.J. (2012). The 8q24 gene desert: an oasis of non-coding transcriptional activity. *Front Genet* 3, 69.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324, 218–223.

International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58.

International Multiple Sclerosis Genetics Consortium (IMSGC), Booth, D.R., Heard, R.N., Stewart, G.J., Cox, M., Scott, R.J., Lechner-Scott, J., Goris, A., Dobosi, R., Dubois, B., et al. (2010a). Lack of support for association between the KIF1B rs10492972[C] variant and multiple sclerosis. *Nat. Genet.* 42, 469–470; authorreply470–authorreply471.

International Multiple Sclerosis Genetics Consortium (IMSGC), Bush, W.S., Sawcer, S.J., De Jager, P.L., Oksenberg, J.R., McCauley, J.L., Pericak-Vance, M.A., and Haines, J.L. (2010b). Evidence for polygenic susceptibility to multiple sclerosis--the shape of things to come. *Am. J. Hum. Gen.* 86, 621–625.

International Schizophrenia Consortium, Purcell, S.M., Wray, N.R., Stone, J.L., Visscher, P.M., O'Donovan, M.C., Sullivan, P.F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748–752.

- 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.
- Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Silanpää, M.J., et al. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.* *20*, 861–873.
- Jolma, A., Yan, J., Whittington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G., et al. (2013). DNA-Binding Specificities of Human Transcription Factors. *Cell* *152*, 327–339.
- Jones, P.A., and Taylor, S.M. (1980). Cellular differentiation, cytidine analogs and DNA methylation. *Cell* *20*, 85–93.
- Jones, P.A. (2012). Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* *13*, 484–492.
- Jones, P.A., and Baylin, S.B. (2007). The Epigenomics of Cancer. *Cell* *128*, 683–692.
- Kadota, M., Yang, H.H., Hu, N., Wang, C., Hu, Y., Taylor, P.R., Buetow, K.H., and Lee, M.P. (2007). Allele-Specific Chromatin Immunoprecipitation Studies Show Genetic Influence on Chromatin State in Human Genome. *PLoS Genet.* *3*, e81.
- Kanduri, C., Pant, V., Loukinov, D., Pugacheva, E., Qi, C.F., Wolffe, A., Ohlsson, R., and Lobanenko, V.V. (2000). Functional association of CTCF with the insulator upstream of the H19 gene is parent of origin-specific and methylation-sensitive. *Curr. Biol.* *10*, 853–856.
- Kaplan, T., Li, X.-Y., Sabo, P.J., Thomas, S., Stamatoyannopoulos, J.A., Biggin, M.D., and Eisen, M.B. (2011). Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early *Drosophila* development. *PLoS Genet.* *7*, e1001290.
- Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S.M., Habegger, L., Rozowsky, J., Shi, M., Urban, A.E., et al. (2010). Variation in transcription factor binding among humans. *Science* *328*, 232–235.
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* *423*, 241–254.
- Kharchenko, P.V., Tolstorukov, M.Y., and Park, P.J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol.* *26*, 1351–1359.
- Kheradpour, P., Stark, A., Roy, S., and Kellis, M. (2007). Reliable prediction of regulator targets using 12 *Drosophila* genomes. *Genome Res.* *17*, 1919–1931.
- Kim, J., He, X., and Sinha, S. (2009). Evolution of Regulatory Sequences in 12 *Drosophila* Species. *PLoS Genet.* *5*, e1000330.
- Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenko, V.V., and Ren, B. (2007). Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human

Genome. *Cell* 128, 1231–1245.

King, M.C., and Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. *Science* 188, 107–116.

Kinney, J.B., Murugan, A., Callan, C.G., and Cox, E.C. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. U.S.A.* 107, 9158–9163.

Kleinjan, D.A., and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Gen.* 76, 8–32.

Klenova, E.M., Chernukhin, I.V., El-Kady, A., Lee, R.E., Pugacheva, E.M., Loukinov, D.I., Goodwin, G.H., Delgado, D., Filippova, G.N., Leon, J., et al. (2001). Functional phosphorylation sites in the C-terminal region of the multivalent multifunctional transcriptional factor CTCF. *Mol. Cell. Biol.* 21, 2221–2234.

Knight, J.C. (2005). Regulatory polymorphisms underlying complex disease traits. *J. Mol. Med.* 83, 97–109.

Knight, J.C., Keating, B.J., Rockett, K.A., and Kwiatkowski, D.P. (2003). In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nat. Genet.* 33, 469–475.

Kochi, Y., Okada, Y., Suzuki, A., Ikari, K., Terao, C., Takahashi, A., Yamazaki, K., Hosono, N., Myouzen, K., Tsunoda, T., et al. (2010). A regulatory variant in CCR6 is associated with rheumatoid arthritis susceptibility. *Nat. Genet.* 42, 515–519.

Komashko, V.M., and Farnham, P.J. (2010). 5-azacytidine treatment reorganizes genomic histone modification patterns. *Epigenetics* 5.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27, 1571–1572.

Kryukov, G.V., Pennacchio, L.A., and Sunyaev, S.R. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Gen.* 80, 727–739.

Kumar, V., Westra, H.-J., Karjalainen, J., Zhernakova, D.V., Esko, T., Hrdlickova, B., Almeida, R., Zhernakova, A., Reinmaa, E., Vösa, U., et al. (2013). Human disease-associated genetic variation impacts large intergenic non-coding RNA expression. *PLoS Genet.* 9, e1003201.

Lai, A.Y., Fatemi, M., Dhasarathy, A., Malone, C., Sobol, S.E., Geigerman, C., Jaye, D.L., Mav, D., Shah, R., Li, L., et al. (2010). DNA methylation prevents CTCF-mediated silencing of the oncogene BCL6 in B cell lymphomas. *J. Exp. Med.* 207, 1939–1950.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25.

Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832–838.

Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Gen.* 88, 294–305.

Lefevre, P., Witham, J., Lacroix, C.E., Cockerill, P.N., and Bonifer, C. (2008). The LPS-Induced Transcrip-

tional Upregulation of the Chicken Lysozyme Locus Involves CTCF Eviction and Noncoding RNA Transcription. *Mol. Cell* 32, 129–139.

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Gen.* 12, 1725–1735.

Lettice, L.A., Horikoshi, T., Heaney, S.J.H., van Baren, M.J., van der Linde, H.C., Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B.A., Endo, N., et al. (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc. Natl. Acad. Sci. U.S.A.* 99, 7548–7553.

Lettice, L.A., Williamson, I., Wiltshire, J.H., Peluso, S., Devenney, P.S., Hill, A.E., Essafi, A., Hagman, J., Mort, R., Grimes, G., et al. (2012). Opposing functions of the ETS factor family define Shh spatial expression in limb buds and underlie polydactyly. *Dev. Cell* 22, 459–467.

Leung, T.H., Hoffmann, A., and Baltimore, D. (2004). One Nucleotide in a κ B Site Can Determine Cofactor Specificity for NF- κ B Dimers. *Cell* 118, 453–464.

Lewis, J.D., Meehan, R.R., Henzel, W.J., Maurer-Fogy, I., Jeppesen, P., Klein, F., and Bird, A. (1992). Purification, sequence, and cellular localization of a novel chromosomal protein that binds to methylated DNA. *Cell* 69, 905–914.

Li, E., Bestor, T.H., and Jaenisch, R. (1992). Targeted mutation of the DNA methyltransferase gene results in embryonic lethality. *Cell* 69, 915–926.

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

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079.

Li, Q., Brown, J.B., Huang, H., and Bickel, P.J. (2011). Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* 5, 1752–1779.

Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., Clark, T.A., Schweitzer, A.C., Blume, J.E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 456, 464–469.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326, 289–293.

Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., and Schübeler, D. (2011). Identification of genetic elements that autonomously determine DNA methylation states. *Nat. Genet.* 43, 1091–1097.

Lin, I.G., Tomzynski, T.J., Ou, Q., and Hsieh, C.L. (2000). Modulation of DNA binding protein affinity directly affects target site demethylation. *Mol. Cell. Biol.* 20, 2343–2349.

- Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M.F., Parker, B.J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., et al. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478, 476–482.
- Lister, R., Pelizzola, M., Downen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462, 315–322.
- Liu, Z., Scannell, D.R., Eisen, M.B., and Tjian, R. (2011). Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell* 146, 720–731.
- Macleod, D., Charlton, J., Mullins, J., and Bird, A.P. (1994). Sp1 sites in the mouse *aprt* gene promoter are required to prevent methylation of the CpG island. *Genes Dev.* 8, 2282–2292.
- MacPherson, M.J., Beatty, L.G., Zhou, W., Du, M., and Sadowski, P.D. (2009). The CTCF Insulator Protein Is Posttranslationally Modified by SUMO. *Mol. Cell. Biol.* 29, 714–725.
- Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., and Turner, D.J. (2010). Target-enrichment strategies for next-generation sequencing. *Nat. Methods* 7, 111–118.
- Martin, D., Pantoja, C., Fernández Miñán, A., Valdes-Quezada, C., Moltó, E., Matesanz, F., Bogdanović, O., la Calle-Mustienes, de, E., Domínguez, O., Taher, L., et al. (2011). Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nat. Struct. Mol. Biol.* 18, 708–714.
- Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* 7, 29–59.
- Matsuo, K., Silke, J., Georgiev, O., Marti, P., Giovannini, N., and Rungger, D. (1998). An embryonic demethylation mechanism involving binding of transcription factors to replicating DNA. *Embo J.* 17, 1446–1453.
- Matsuzaki, H., Okamura, E., Fukamizu, A., and Tanimoto, K. (2010). CTCF binding is not the epigenetic mark that establishes post-fertilization methylation imprinting in the transgenic H19 ICR. *Hum. Mol. Gen.* 19, 1190–1198.
- Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., et al. (2006). TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 34, D108–D110.
- Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012a). Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337, 1190–1195.
- Maurano, M.T., Wang, H., Kuttyavin, T., and Stamatoyannopoulos, J.A. (2012b). Widespread site-dependent buffering of human regulatory polymorphism. *PLoS Genet.* 8, e1002599.
- McDaniell, R., Lee, B.-K., Song, L., Liu, Z., Boyle, A.P., Erdos, M.R., Scott, L.J., Morken, M.A., Kucera, K.S., Battenhouse, A., et al. (2010). Heritable individual-specific and allele-specific chromatin signatures in humans. *Science* 328, 235–239.
- Meijsing, S.H., Pufall, M.A., So, A.Y., Bates, D.L., Chen, L., and Yamamoto, K.R. (2009). DNA binding site

sequence directs glucocorticoid receptor structure and activity. *Science* *324*, 407–410.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., et al. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* *454*, 766–770.

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

Melzer, D., Perry, J.R.B., Hernandez, D., Corsi, A.-M., Stevens, K., Rafferty, I., Lauretani, F., Murray, A., Gibbs, J.R., Paolisso, G., et al. (2008). A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.* *4*, e1000072.

Mirny, L.A., and Gelfand, M.S. (2002). Structural analysis of conserved base pairs in protein-DNA complexes. *Nucleic Acids Res.* *30*, 1704–1711.

Moffatt, M.F., Kabesch, M., Liang, L., Dixon, A.L., Strachan, D., Heath, S., Depner, M., Berg, von, A., Bufe, A., Rietschel, E., et al. (2007). Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature* *448*, 470–473.

Monks, S.A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., Phillips, J.W., Sachs, A., and Schadt, E.E. (2004). Genetic Inheritance of Gene Expression in Human Cell Lines. *Am. J. Hum. Gen.* *75*, 1094–1105.

Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S., and Cheung, V.G. (2004). Genetic analysis of genome-wide variation in human gene expression. *Nature* *430*, 743–747.

Moses, A.M. (2009). Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites. *BMC Evol. Biol.* *9*, 286.

Moses, A.M., Chiang, D.Y., Kellis, M., Lander, E.S., and Eisen, M.B. (2003). Position specific variation in the rate of evolution in transcription factor binding sites. *BMC Evol. Biol.* *3*, 19.

Mouse Genome Sequencing Consortium, Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., et al. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* *420*, 520–562.

Mukhopadhyay, R., Yu, W., Whitehead, J., Xu, J., Lezcano, M., Pack, S., Kanduri, C., Kanduri, M., Ginjala, V., Vostrov, A., et al. (2004). The binding sites for the chromatin insulator protein CTCF map to DNA methylation-free domains genome-wide. *Genome Res.* *14*, 1594–1602.

Mustonen, V., Kinney, J., Callan, C.G., and Lässig, M. (2008). Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proc. Natl. Acad. Sci. U.S.A.* *105*, 12376–12381.

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N.E., Ahfeldt, T., Sachs, K.V., Li, X., Li, H., Kuperwasser, N., Ruda, V.M., et al. (2010). From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* *466*, 714–719.

Nakahashi, H., Kwon, K.-R.K., Resch, W., Vian, L., Dose, M., Stavreva, D., Hakim, O., Pruett, N., Nelson, S., Yamane, A., et al. (2013). A genome-wide map of CTCF multivalency redefines the CTCF code. *Cell Rep* *3*, 1678–1689.

- Neph, S., Kuehn, M.S., Reynolds, A.P., Haugen, E., Thurman, R.E., Johnson, A.K., Rynes, E., Maurano, M.T., Vierstra, J., Thomas, S., et al. (2012a). BEDOPS: high-performance genomic feature operations. *Bioinformatics* 28, 1919–1920.
- Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., et al. (2012b). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489, 83–90.
- Newburger, D.E., and Bulyk, M.L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* 37, D77–D82.
- Ng, P.C., and Henikoff, S. (2001). Predicting Deleterious Amino Acid Substitutions. *Genome Res.* 11, 863–874.
- Ng, S.B., Bigham, A.W., Buckingham, K.J., Hannibal, M.C., McMillin, M.J., Gildersleeve, H.I., Beck, A.E., Tabor, H.K., Cooper, G.M., Mefford, H.C., et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat. Genet.* 42, 802–805.
- Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E., et al. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461, 272–276.
- Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* 6, e1000895.
- 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.
- Noureddine, M.A., Menendez, D., Campbell, M.R., Bandele, O.J., Horvath, M.M., Wang, X., Pittman, G.S., Chorley, B.N., Resnick, M.A., and Bell, D.A. (2009). Probing the Functional Impact of Sequence Variation on p53-DNA Interactions Using a Novel Microsphere Assay for Protein-DNA Binding with Human Cell Extracts. *PLoS Genet.* 5, e1000462.
- O'Geen, H., Squazzo, S.L., Iyengar, S., Blahnik, K., Rinn, J.L., Chang, H.Y., Green, R., and Farnham, P.J. (2007). Genome-wide analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs. *PLoS Genet.* 3, e89.
- O'Roak, B.J., Deriziotis, P., Lee, C., Vives, L., Schwartz, J.J., Girirajan, S., Karakoc, E., MacKenzie, A.P., Ng, S.B., Baker, C., et al. (2011). Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat. Genet.* 43, 585–589.
- Ohlsson, R., Renkawitz, R., and Lobanenko, V. (2001). CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet.* 17, 520–527.
- Ohlsson, R., Lobanenko, V., and Klenova, E. (2010). Does CTCF mediate between nuclear organization and gene expression? *Bioessays* 32, 37–50.
- Olson, J.M., and Wijsman, E.M. (1993). Linkage between quantitative trait and marker loci: methods using all relative pairs. *Genet. Epidemiol.* 10, 87–102.
- Ooi, S.K.T., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H., Tempst, P., Lin, S.-P.,

Allis, C.D., et al. (2007). DNMT3L connects unmethylated lysine 4 of histone H3 to de novo methylation of DNA. *Nature* **448**, 714–717.

Ott, C.J., Suszko, M., Blackledge, N.P., Wright, J.E., Crawford, G.E., and Harris, A. (2009). A complex intronic enhancer regulates expression of the CFTR gene by direct interaction with the promoter. *J. Cell. Mol. Med.* **13**, 680–692.

Pandiyan, K., You, J.S., Yang, X., Dai, C., Zhou, X.J., Baylin, S.B., Jones, P.A., and Liang, G. (2013). Functional DNA demethylation is accompanied by chromatin accessibility. *Nucleic Acids Res.* **41**, 3973–3985.

Pant, V., Kurukuti, S., Pugacheva, E., Shamsuddin, S., Mariano, P., Renkawitz, R., Klenova, E., Lobanenko, V., and Ohlsson, R. (2004). Mutation of a single CTCF target site within the H19 imprinting control region leads to loss of Igf2 imprinting and complex patterns of de novo methylation upon maternal inheritance. *Mol. Cell. Biol.* **24**, 3497–3504.

Pant, V., Mariano, P., Kanduri, C., Mattsson, A., Lobanenko, V., Heuchel, R., and Ohlsson, R. (2003). The nucleotides responsible for the direct physical contact between the chromatin insulator protein CTCF and the H19 imprinting control region manifest parent of origin-specific long-distance insulation and methylation-free domains. *Genes Dev.* **17**, 586–590.

Parelho, V., Hadjur, S., Spivakov, M., Leleu, M., Sauer, S., Gregson, H.C., Jarmuz, A., Canzonetta, C., Webster, Z., Nesterova, T., et al. (2008). Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms. *Cell* **132**, 422–433.

Park, J.-H., Wacholder, S., Gail, M.H., Peters, U., Jacobs, K.B., Chanock, S.J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575.

Pastinen, T. (2010). Genome-wide allele-specific analysis: insights into regulatory variation. *Nat. Rev. Genet.* **11**, 533–538.

Patsopoulos, N.A., Bayer Pharma MS Genetics Working Group, Steering Committees of Studies Evaluating IFN β -1b and a CCR1-Antagonist, ANZgene Consortium, GeneMSA, International Multiple Sclerosis Genetics Consortium, de Bakker, P.I.W., Esposito, F., Reischl, J., Lehr, S., et al. (2011). Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Ann. Neurol.* **70**, 897–912.

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., et al. (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.

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

Peterson, K.R., Clegg, C.H., Huxley, C., Josephson, B.M., Haugen, H.S., Furukawa, T., and Stamatoyannopoulos, G. (1993). Transgenic mice containing a 248-kb yeast artificial chromosome carrying the human beta-globin locus display proper developmental control of human globin genes. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 7593–7597.

- Peterson, K.R., Li, Q.L., Clegg, C.H., Furukawa, T., Navas, P.A., Norton, E.J., Kimbrough, T.G., and Stamatoyannopoulos, G. (1995). Use of yeast artificial chromosomes (YACs) in studies of mammalian development: production of beta-globin locus YAC mice carrying human globin developmental mutants. *Proc. Natl. Acad. Sci. U.S.A.* *92*, 5655–5659.
- Phillips, J.E., and Corces, V.G. (2009). CTCF: master weaver of the genome. *Cell* *137*, 1194–1211.
- Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* *20*, 110–121.
- Pomerantz, M.M., Ahmadiyeh, N., Jia, L., Herman, P., Verzi, M.P., Doddapaneni, H., Beckwith, C.A., Chan, J.A., Hills, A., Davis, M., et al. (2009). The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.* *41*, 882–884.
- Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., Lenhard, B., Wasserman, W.W., and Sandelin, A. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res.* *38*, D105–D110.
- Price, A.L., Helgason, A., Thorleifsson, G., McCarroll, S.A., Kong, A., and Stefansson, K. (2011). Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet.* *7*, e1001317.
- Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Gen.* *69*, 124–137.
- Pruitt, K.D., Harrow, J., Harte, R.A., Wallin, C., Diekhans, M., Maglott, D.R., Searle, S., Farrell, C.M., Loveland, J.E., Ruef, B.J., et al. (2009). The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.* *19*, 1316–1323.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Gen.* *81*, 559–575.
- Qiu, X., Hother, C., Ralfkiær, U.M., Sogaard, A., Lu, Q., Workman, C.T., Liang, G., Jones, P.A., and Grønbaek, K. (2010). Equitoxic doses of 5-azacytidine and 5-aza-2'-deoxycytidine induce diverse immediate and overlapping heritable changes in the transcriptome. *PLoS ONE* *5*.
- R Development Core Team R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).
- Ramirez-Carrozzi, V.R., Braas, D., Bhatt, D.M., Cheng, C.S., Hong, C., Doty, K.R., Black, J.C., Hoffmann, A., Carey, M., and Smale, S.T. (2009). A unifying model for the selective regulation of inducible transcription by CpG islands and nucleosome remodeling. *Cell* *138*, 114–128.
- Rasko, J.E., Klenova, E.M., Leon, J., Filippova, G.N., Loukinov, D.I., Vatin, S., Robinson, A.F., Hu, Y.J., Ulmer, J., Ward, M.D., et al. (2001). Cell growth inhibition by the multifunctional multivalent zinc-finger factor CTCF. *Cancer Res.* *61*, 6002–6007.
- Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. *Trends Genet.* *17*, 502–510.
- Renda, M., Baglivo, I., Burgess-Beusse, B., Esposito, S., Fattorusso, R., Felsenfeld, G., and Pedone, P.V. (2007). Critical DNA binding interactions of the insulator protein CTCF: a small number of zinc fingers me-

diate strong binding, and a single finger-DNA interaction controls binding at imprinted loci. *J. Biol. Chem.* **282**, 33336–33345.

Rhee, H.S., and Pugh, B.F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**, 1408–1419.

Rhee, I., Bachman, K.E., Park, B.H., Jair, K.-W., Yen, R.-W.C., Schuebel, K.E., Cui, H., Feinberg, A.P., Lengauer, C., Kinzler, K.W., et al. (2002). DNMT1 and DNMT3b cooperate to silence genes in human cancer cells. *Nature* **416**, 552–556.

Richards, A.L., Jones, L., Moskvina, V., Kirov, G., Gejman, P.V., Levinson, D.F., Sanders, A.R., Molecular Genetics of Schizophrenia Collaboration (MGS), International Schizophrenia Consortium (ISC), Purcell, S., et al. (2012). Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. *Mol. Psychiatry* **17**, 193–201.

Rockman, M.V., and Wray, G.A. (2002). Abundant raw material for cis-regulatory evolution in humans. *Mol. Biol. Evol.* **19**, 1991–2004.

Rodriguez, C., Borgel, J., Court, F., Cathala, G., Forné, T., and Piette, J. (2010). CTCF is a DNA methylation-sensitive positive regulator of the INK/ARF locus. *Biochem. Biophys. Res. Commun.* **392**, 129–134.

Rowntree, R.K., Vassaux, G., McDowell, T.L., Howe, S., McGuigan, A., Phylactides, M., Huxley, C., and Harris, A. (2001). An element in intron 1 of the CFTR gene augments intestinal expression in vivo. *Hum. Mol. Gen.* **10**, 1455–1464.

Rubio, E.D., Reiss, D.J., Welch, P.L., Disteche, C.M., Filippova, G.N., Baliga, N.S., Aebersold, R., Ranish, J.A., and Krumm, A. (2008). CTCF physically links cohesin to chromatin. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 8309–8314.

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

Savic, D., Bell, G.I., and Nobrega, M.A. (2012). An in vivo cis-Regulatory Screen at the Type 2 Diabetes Associated TCF7L2 Locus Identifies Multiple Tissue-Specific Enhancers. *PLoS ONE* **7**, e36501.

Savic, D., Ye, H., Aneas, I., Park, S.-Y., Bell, G.I., and Nobrega, M.A. (2011). Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism. *Genome Res.* **21**, 1417–1425.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., et al. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036–1040.

Schoenherr, C.J., Levorse, J.M., and Tilghman, S.M. (2002). CTCF maintains differential methylation at the Igf2/H19 locus. *Nat. Genet.* **33**, 66–69.

Seitan, V.C., Hao, B., Tachibana-Konwalski, K., Lavagnoli, T., Mira-Bontenbal, H., Brown, K.E., Teng, G., Carroll, T., Terry, A., Horan, K., et al. (2011). A role for cohesin in T-cell-receptor rearrangement and thymocyte differentiation. *Nature* **476**, 467–471.

Sekimata, M., Pérez-Melgosa, M., Miller, S.A., Weinmann, A.S., Sabo, P.J., Sandstrom, R., Dorschner, M.O., Stamatoyannopoulos, J.A., and Wilson, C.B. (2009). CCCTC-binding factor and the transcription factor T-bet orchestrate T helper 1 cell-specific structure and function at the interferon-gamma locus. *Immunity* **31**, 551–

- Selker, E.U. (1990). DNA methylation and chromatin structure: a view from below. *Trends Biochem. Sci.* *15*, 103–107.
- Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., and Segal, E. (2012). Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol.* *30*, 521–530.
- Shukla, S., Kavak, E., Gregory, M., Imashimizu, M., Shutinoski, B., Kashlev, M., Oberdoerffer, P., Sandberg, R., and Oberdoerffer, S. (2011). CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* *479*, 74–79.
- Siegfried, Z., Eden, S., Mendelsohn, M., Feng, X., Tsuberi, B.Z., and Cedar, H. (1999). DNA methylation represses transcription in vivo. *Nat. Genet.* *22*, 203–206.
- Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.* *38*, 1348–1354.
- Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCr: visualizing classifier performance in R. *Bioinformatics* *21*, 3940–3941.
- Smith, C.L., and Hager, G.L. (1997). Transcriptional regulation of mammalian genes in vivo. A tale of two templates. *J. Biol. Chem.* *272*, 27493–27496.
- Soldner, F., Laganière, J., Cheng, A.W., Hockemeyer, D., Gao, Q., Alagappan, R., Khurana, V., Golbe, L.I., Myers, R.H., Lindquist, S., et al. (2011). Generation of isogenic pluripotent stem cells differing exclusively at two early onset Parkinson point mutations. *Cell* *146*, 318–331.
- Soranzo, N., Spector, T.D., Mangino, M., Kühnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., et al. (2009). A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.* *41*, 1182–1190.
- Soto-Reyes, E., and Recillas-Targa, F. (2010). Epigenetic regulation of the human p53 gene promoter by the CTCF transcription factor in transformed cell lines. *Oncogene* *29*, 2217–2227.
- Sotoodehnia, N., Isaacs, A., de Bakker, P.I.W., Dörr, M., Newton-Cheh, C., Nolte, I.M., van der Harst, P., Müller, M., Eijgelsheim, M., Alonso, A., et al. (2010). Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction. *Nat. Genet.* *42*, 1068–1076.
- Splinter, E. (2006). CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev.* *20*, 2349–2354.
- Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., Tiwari, V.K., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* *480*, 490–495.
- Stamatoyannopoulos, J.A. (2004). The genomics of gene expression. *Genomics* *84*, 449–457.
- Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* *100*, 9440–9445.

- Stormo, G.D. (2000). DNA binding sites: representation and discovery. *Bioinformatics* 16, 16–23.
- Stranger, B.E., Stahl, E.A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* 187, 367–383.
- Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z., and Cedar, H. (2009). Developmental programming of CpG island methylation profiles in the human genome. *Nat. Struct. Mol. Biol.* 16, 564–571.
- Stresemann, C., Bokelmann, I., Mählkecht, U., and Lyko, F. (2008). Azacytidine causes complex DNA methylation responses in myeloid leukemia. *Mol. Cancer Ther.* 7, 2998–3005.
- Sunyaev, S.R. (2012). Inferring causality and functional significance of human coding DNA variants. *Hum. Mol. Gen.* 21, R10–R17.
- Sur, I.K., Hallikas, O., Vähärautio, A., Yan, J., Turunen, M., Enge, M., Taipale, M., Karhu, A., Aaltonen, L.A., and Taipale, J. (2012). Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* 338, 1360–1363.
- Suzuki, R., and Shimodaira, H. (2006). Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22, 1540–1542.
- Symonds, M.E., Sebert, S.P., Hyatt, M.A., and Budge, H. (2009). Nutritional programming of the metabolic syndrome. *Nat Rev Endocrinol* 5, 604–610.
- Tamura, T., Yanai, H., Savitsky, D., and Taniguchi, T. (2008). The IRF family transcription factors in immunity and oncogenesis. *Annu. Rev. Immunol.* 26, 535–584.
- Tate, P.H., and Bird, A.P. (1993). Effects of DNA methylation on DNA-binding proteins and gene expression. *Curr. Opin. Genet. Dev.* 3, 226–231.
- Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 64–69.
- Thomson, J.P., Skene, P.J., Selfridge, J., Clouaire, T., Guy, J., Webb, S., Kerr, A.R.W., Deaton, A., Andrews, R., James, K.D., et al. (2010). CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature* 464, 1082–1086.
- Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. *Nature* 489, 75–82.
- Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B.E., Liu, X.S., and Raychaudhuri, S. (2013). Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* 45, 124–130.
- Tsai, H.-C., Li, H., Van Neste, L., Cai, Y., Robert, C., Rassool, F.V., Shin, J.J., Harbom, K.M., Beaty, R., Pappou, E., et al. (2012). Transient low doses of DNA-demethylating agents exert durable antitumor effects on hematological and epithelial tumor cells. *Cancer Cell* 21, 430–446.
- Tuupanen, S., Turunen, M., Lehtonen, R., Hallikas, O., Vanharanta, S., Kivioja, T., Björklund, M., Wei, G., Yan, J., Nüttymäki, I., et al. (2009). The common colorectal cancer predisposition SNP rs6983267 at chromo-

some 8q24 confers potential to enhanced Wnt signaling. *Nat. Genet.* **41**, 885–890.

Varley, K.E., Gertz, J., Bowling, K.M., Parker, S.L., Reddy, T.E., Pauli-Behn, F., Cross, M.K., Williams, B.A., Stamatoyannopoulos, J.A., Crawford, G.E., et al. (2013). Dynamic DNA methylation across diverse human cell lines and tissues. *Genome Res.* **23**, 555–567.

Veprintsev, D.B., and Fersht, A.R. (2008). Algorithm for prediction of tumour suppressor p53 affinity for binding sites in DNA. *Nucleic Acids Res.* **36**, 1589–1598.

Vernot, B., Stergachis, A.B., Maurano, M.T., Vierstra, J., Neph, S., Thurman, R.E., Stamatoyannopoulos, J.A., and Akey, J.M. (2012). Personal and population genomics of human regulatory variation. *Genome Res.* **22**, 1689–1697.

Veyrieras, J.-B., Kudaravalli, S., Kim, S.Y., Dermitzakis, E.T., Gilad, Y., Stephens, M., and Pritchard, J.K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS Genet.* **4**, e1000214.

Visel, A., Minovitsky, S., Dubchak, I., and Pennacchio, L.A. (2007). VISTA Enhancer Browser--a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92.

Visel, A., Zhu, Y., May, D., Afzal, V., Gong, E., Attanasio, C., Blow, M.J., Cohen, J.C., Rubin, E.M., and Pennacchio, L.A. (2010). Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature* **464**, 409–412.

Vostrov, A.A., and Quitschke, W.W. (1997). The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation. *J. Biol. Chem.* **272**, 33353–33359.

Walsh, C.P., and Bestor, T.H. (1999). Cytosine methylation and mammalian development. *Genes Dev.* **13**, 26–34.

Wang, H., Maurano, M.T., Qu, H., Varley, K.E., Gertz, J., Pauli, F., Lee, K., Canfield, T., Weaver, M., Sandstrom, R., et al. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res.* **22**, 1680–1688.

Ward, L.D., and Kellis, M. (2012). HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934.

Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5**, 276–287.

Weirauch, M.T., and Hughes, T.R. (2010). Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. *Trends Genet.* **26**, 66–74.

Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiho, T., et al. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451**, 796–801.

Wiench, M., John, S., Baek, S., Johnson, T.A., Sung, M.-H., Escobar, T., Simmons, C.A., Pearce, K.H., Bidie, S.C., Sabo, P.J., et al. (2011). DNA methylation status predicts cell type-specific enhancer activity. *Embo J.* **30**, 3028–3039.

- Wilson, M.D., Barbosa-Morais, N.L., Schmidt, D., Conboy, C.M., Vanes, L., Tybulewicz, V.L.J., Fisher, E.M.C., Tavaré, S., and Odom, D.T. (2008). Species-specific transcription in mice carrying human chromosome 21. *Science* *322*, 434–438.
- Witcher, M., and Emerson, B.M. (2009). Epigenetic Silencing of the p16INK4a Tumor Suppressor Is Associated with Loss of CTCF Binding and a Chromatin Boundary. *Mol. Cell* *34*, 271–284.
- Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., and Romano, L.A. (2003). The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* *20*, 1377–1419.
- Wunderlich, Z., and Mirny, L.A. (2009). Different gene regulation strategies revealed by analysis of binding motifs. *Trends Genet.* *25*, 434–440.
- Xiang, H., Noonan, E.J., Wang, J., Duan, H., Ma, L., Michie, S., and Boxer, L.M. (2011). The immunoglobulin heavy chain gene 3' enhancers induce Bcl2 deregulation and lymphomagenesis in murine B cells. *Leukemia* *25*, 1484–1493.
- Xie, X., Mikkelsen, T.S., Gnirke, A., Lindblad-Toh, K., Kellis, M., and Lander, E.S. (2007). Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites. *Proc. Natl. Acad. Sci. U.S.A.* *104*, 7145–7150.
- Yan, P., Frankhouser, D., Murphy, M., Tam, H.-H., Rodriguez, B., Curfman, J., Trimarchi, M., Geyer, S., Wu, Y.-Z., Whitman, S.P., et al. (2012). Genome-wide methylation profiling in decitabine-treated patients with acute myeloid leukemia. *Blood* *120*, 2466–2474.
- Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* *42*, 565–569.
- Yang, W.S., Nevin, D.N., Peng, R., Brunzell, J.D., and Deeb, S.S. (1995). A mutation in the promoter of the lipoprotein lipase (LPL) gene in a patient with familial combined hyperlipidemia and low LPL activity. *Proc. Natl. Acad. Sci. U.S.A.* *92*, 4462–4466.
- Yu, W., Ginjala, V., Pant, V., Chernukhin, I., Whitehead, J., Docquier, F., Farrar, D., Tavoosidana, G., Mukhopadhyay, R., Kanduri, C., et al. (2004). Poly(ADP-ribosyl)ation regulates CTCF-dependent chromatin insulation. *Nat. Genet.* *36*, 1105–1110.
- Yusufzai, T.M., Tagami, H., Nakatani, Y., and Felsenfeld, G. (2004). CTCF tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Mol. Cell* *13*, 291–298.
- Zhao, Y., and Stormo, G.D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nat Biotechnol.* *29*, 480–483.
- Zheng, W., Zhao, H., Mancera, E., Steinmetz, L.M., and Snyder, M. (2010). Genetic analysis of variation in transcription factor binding in yeast. *Nature* *464*, 1187–1191.
- Zhong, H., Beaulaurier, J., Lum, P.Y., Molony, C., Yang, X., MacNeil, D.J., Weingarth, D.T., Zhang, B., Greenawalt, D., Dobrin, R., et al. (2010). Liver and adipose expression associated SNPs are enriched for association to type 2 diabetes. *PLoS Genet.* *6*, e1000932.
- Zhu, J., Zhang, B., Smith, E.N., Drees, B., Brem, R.B., Kruglyak, L., Bumgarner, R.E., and Schadt, E.E. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks.

Nat. Genet. *40*, 854–861.

10 Appendix: Supplementary Tables for Section 8

Table 10-1 Genes correlated with distal DHSs harboring GWAS SNPs (Part I)

347 trait-SNP associations (296 unique SNPs) overlapping predicted long-distance interactions established by correlation of chromatin accessibility (R) (Thurman et al., 2012). LD measures the mean extent of LD between the correlated DHSs (r^2); NA means filtered 1000 Genomes SNPs with MAF $\geq 5\%$ in the CEU population were not found within 2 kb of both DHSs. Cor_gene_name represents the most-correlated gene; Dist, distance to gene in kb. "Adjacent?", whether the highest-correlated gene is an adjacent gene.

SNP	Disease or trait	Trait Category	R	LD	Cor_gene_name	Dist_cor	Adjacent?
rs10202497	Aging_traits-age_free_from_disease	Aging	0.75	0.05	COL6A3	52	Y
rs10781380	Hippocampal_atrophy	Aging	0.93	0.01	FOXB2	226	N
rs11006923	Alzheimers_disease	Aging	0.70	0.06	RP11-351M16.2	128	N
rs1172822	Menopause	Aging	0.95	0.01	NLRP2	-347	N
rs12203592	Progressive_supranuclear_palsy	Aging	0.85	0.01	RP3-416J7.3	-241	N
rs1532278	Alzheimers_disease_late_onset	Aging	0.81	0.01	RP11-16P20.2	89	N
rs1561570	Pagets_disease	Aging	0.71	0.03	RP11-730A19.2	-53	N
rs1564282	Parkinsons_disease	Aging	0.88	0.05	CPLX1	-54	N
rs1569476	Alzheimers_Total_ventricular_volume	Aging	0.98	0.03	SCYL3	250	N
rs157580	Alzheimers-AB1-42	Aging	0.85	0.00	CEACAM19	-221	N
rs157580	Alzheimers_disease	Aging	0.85	0.00	CEACAM19	-221	N
rs16938437	Menarche	Aging	0.89	0.29	PHF21A	-71	Y
rs1695739	Longevity	Aging	0.78	0.01	DDX25	-405	N
rs1934951	Osteonecrosis_of_the_jaw	Aging	0.88	0.16	RP11-310E22.4	192	N
rs2104362	Amyotrophic_lateral_sclerosis-age_of_onset	Aging	1.00	0.02	SYNGAP1	-412	N
rs2121433	Alzheimers-t-tau	Aging	0.76	0.01	AC105402.1	65	Y
rs2244621	Longevity	Aging	0.92	0.01	RASGRP2	484	N
rs2687729	Menarche	Aging	0.73	0.01	DNAJB8	291	N
rs2899472	Alzheimers-AB1-42	Aging	0.74	0.03	GLDN	153	N
rs3825776	Amyotrophic_lateral_sclerosis	Aging	0.71	0.02	AQP9	-316	N
rs4955755	Menopause	Aging	0.87	0.01	CLDN11	-339	N
rs5937496	Amyotrophic_lateral_sclerosis	Aging	0.86	N/A	SAR1AP4	-288	N
rs6031882	Hippocampal_atrophy	Aging	0.72	0.01	BLCAP	348	N
rs6602175	Alzheimers-Whole-brain_volume	Aging	0.77	0.02	TRDMT1	101	N
rs6701713	Alzheimers_disease_late_onset	Aging	0.85	0.00	PLXNA2	428	N
rs9652490	Essential_tremor	Aging	0.77	0.01	SH2D7	421	N
rs9871760	Alzheimers-Whole-brain_volume	Aging	0.82	0.12	RP11-305F5.2	-52	Y
rs10036748	Systemic_lupus_erythematosus	Autoimmune	0.78	0.01	ANXA6	63	N
rs10737562	Systemic_lupus_erythematosus	Autoimmune	0.83	0.65	RP11-398M15.1	21	Y
rs1077667	Multiple_sclerosis	Autoimmune	0.80	0.01	GTF2F1	-277	N
rs10931468	Primary_biliary_cirrhosis	Autoimmune	0.81	0.03	TMEM194B	-147	N
rs11154801	Multiple_sclerosis	Autoimmune	0.90	0.15	AHI1	-95	Y
rs11616188	Ankylosing_spondylitis	Autoimmune	0.91	0.01	NCAPD2	122	N
rs11742570	Crohns_disease	Autoimmune	0.77	0.03	PTGER4	271	Y
rs12212193	Multiple_sclerosis	Autoimmune	0.73	0.01	GJA10	-394	N
rs12580100	Psoriasis	Autoimmune	0.73	0.05	SLC39A5	185	N
rs1295686	Asthma	Autoimmune	0.91	0.24	AC004041.2	4	Y
rs1335532	Multiple_sclerosis	Autoimmune	0.92	0.01	CD101	443	N
rs1551398	Crohns_disease	Autoimmune	1.00	0.00	TRIB1	-95	N
rs17035378	Celiac_disease	Autoimmune	0.87	0.02	ARHGAP25	363	N
rs17582416	Crohns_disease	Autoimmune	0.71	0.51	RP11-324I22.2	232	N
rs17716942	Psoriasis	Autoimmune	0.71	0.01	AC007740.1	365	N
rs1790100	Multiple_sclerosis	Autoimmune	0.98	0.01	RP11-324E6.4	-422	N
rs1800871	Behcets_disease	Autoimmune	0.88	0.01	PFKFB2	297	N
rs2056626	Systemic_sclerosis	Autoimmune	0.82	0.01	ILDR2	-476	N
rs2075726	Ankylosing_spondylitis	Autoimmune	0.88	0.01	IL22NC01-81G9.3	50	N
rs2104286	Multiple_sclerosis	Autoimmune	0.70	0.11	IL2RA	-29	Y
rs2187668	Systemic_lupus_erythematosus	Autoimmune	0.88	0.01	BRD2	335	N
rs2205960	Systemic_lupus_erythematosus	Autoimmune	0.91	0.01	SLC9A11	379	N
rs2233287	Systemic_sclerosis	Autoimmune	0.86	0.02	ANXA6	68	N
rs2273017	Graves_disease	Autoimmune	0.72	0.03	CFB	-424	N
rs2431697	Systemic_lupus_erythematosus	Autoimmune	0.85	0.01	PWWP2A	-334	N
rs2546890	Multiple_sclerosis	Autoimmune	0.87	0.02	EBF1	-313	N
rs2546890	Psoriasis	Autoimmune	0.87	0.02	EBF1	-313	N
rs2618476	Systemic_lupus_erythematosus	Autoimmune	0.93	0.11	BLK	50	Y
rs2734583	Stevens-Johnson_syndrome_and_necrolysis	Autoimmune	0.80	0.04	WASF5P	-249	N
rs2836878	Inflammatory_bowel_disease	Autoimmune	0.80	0.03	LCA5L	352	N
rs2836878	Ulcerative_colitis	Autoimmune	0.80	0.03	LCA5L	352	N
rs3024505	Crohns_disease	Autoimmune	0.79	0.33	IL10	6	Y
rs3024505	Ulcerative_colitis	Autoimmune	0.79	0.33	IL10	6	Y
rs3129763	Systemic_sclerosis	Autoimmune	0.98	N/A	HLA-DRA	-183	N
rs3821236	Systemic_sclerosis	Autoimmune	0.74	0.02	STAT4	113	Y
rs3821236	Systemic_lupus_erythematosus	Autoimmune	0.74	0.02	STAT4	113	Y
rs4075958	Multiple_sclerosis	Autoimmune	0.90	0.02	GRK6	74	N
rs4129267	Asthma	Autoimmune	0.88	0.39	IL6R	-19	Y
rs4349859	Ankylosing_spondylitis	Autoimmune	0.92	0.02	POU5F1	-229	N
rs4409764	Crohns_disease	Autoimmune	0.90	0.02	DNMBP	390	N

SNP	Disease or trait	Trait Category	R	LD	Cor_gene_name	Dist_cor	Adjacent?
rs4639966	Systemic_lupus_erythematosus	Autoimmune	0.99	N/A	NLRX1	473	N
rs4781011	Ulcerative_colitis	Autoimmune	0.75	0.01	AC007014.1	144	N
rs4845783	Asthma	Autoimmune	0.81	0.04	LCE3A	103	N
rs485499	Primary_biliary_cirrhosis	Autoimmune	0.71	0.01	SMC4	392	N
rs5754217	Systemic_lupus_erythematosus	Autoimmune	0.86	N/A	PI4KAP2	-73	N
rs6074022	Multiple_sclerosis	Autoimmune	0.96	0.00	SLC35C2	252	N
rs610604	Psoriasis	Autoimmune	0.84	0.29	TNFAIP3	-11	Y
rs6806528	Celiac_disease	Autoimmune	0.89	0.18	FRMD4B	-3	Y
rs6859219	Rheumatoid_arthritis	Autoimmune	0.75	0.09	RPL17P22	-13	Y
rs6941421	Multiple_sclerosis	Autoimmune	0.96	0.22	RP1-190J20.2	18	N
rs734999	Ulcerative_colitis	Autoimmune	0.84	0.00	C1orf86	-369	N
rs7579944	Rheumatoid_arthritis_celiac_disease	Autoimmune	0.84	0.01	AC016907.2	-146	N
rs794185	Multiple_sclerosis- Brain_Glutamate_Concentrations	Autoimmune	0.75	0.01	ITPR1	419	N
rs806321	Multiple_sclerosis	Autoimmune	1.00	0.10	DLEU1	43	Y
rs881375	Rheumatoid_arthritis	Autoimmune	0.97	0.07	RP11-2711.2	-44	N
rs924080	Behcets_disease	Autoimmune	0.87	0.01	IL12RB2	13	Y
rs943072	Ulcerative_colitis	Autoimmune	0.92	0.01	SLC35B2	429	N
rs987870	Asthma	Autoimmune	0.92	0.06	MYL8P	261	N
rs987870	Systemic_sclerosis	Autoimmune	0.92	0.06	MYL8P	261	N
rs9888739	Systemic_lupus_erythematosus	Autoimmune	0.97	0.03	STX4	-265	N
rs10510102	Breast_cancer	Cancer	0.92	0.01	RPS15AP5	-152	N
rs11892031	Bladder_cancer	Cancer	0.92	0.03	AC019221.4	-302	N
rs12653946	Prostate_cancer	Cancer	0.92	0.04	RP11-259Q2.3	72	N
rs13397985	Chronic_lymphocytic_leukemia	Cancer	0.97	0.05	AC009950.1	-23	N
rs1432295	Hodgkins_lymphoma	Cancer	0.96	N/A	AC007381.3	-486	N
rs16886165	Breast_cancer	Cancer	0.88	0.23	MAP3K1	158	N
rs2157719	Glioma	Cancer	0.97	0.00	RP11-344A7.1	-398	N
rs2456449	Chronic_lymphocytic_leukemia	Cancer	0.87	0.05	RP11-255B23.2	-98	N
rs28421666	Nasopharyngeal_carcinoma	Cancer	0.85	0.02	BRD2	348	N
rs339331	Prostate_cancer	Cancer	0.97	0.03	FAM26D	-335	N
rs402710	Lung_cancer	Cancer	0.88	0.39	CLPTM1L	18	Y
rs4132601	Acute_lymphoblastic_leukemia_childhood	Cancer	0.75	0.02	AC020743.3	-226	N
rs4487645	Multiple_myeloma	Cancer	0.96	0.01	DNAH11	-344	Y
rs4975616	Lung_cancer	Cancer	0.96	0.35	CLPTM1L	16	Y
rs4980785	Renal_cell_carcinoma	Cancer	0.77	0.01	RP11-300I6.5	322	N
rs498872	Glioma	Cancer	0.71	0.02	SLC37A4	423	N
rs674313	Chronic_lymphocytic_leukemia	Cancer	0.92	0.05	PSMB8	232	N
rs7579899	Renal_cell_carcinoma	Cancer	0.72	0.02	RP11-417F21.2	188	N
rs961253	Colorectal_cancer	Cancer	0.96	0.01	RP5-859D4.3	312	N
rs10765792	Sudden_cardiac_arrest	Cardiovascular	0.90	0.01	FAM76B	-350	N
rs11710077	QRS_duration	Cardiovascular	0.96	0.02	SCN10A	181	N
rs12046278	Systolic_blood_pressure	Cardiovascular	0.87	0.01	MASP2	308	N
rs12576239	QT_interval	Cardiovascular	0.98	N/A	ASCL2	-210	N
rs1378942	Diastolic_blood_pressure	Cardiovascular	0.84	0.09	SEMA7A	-351	N
rs1378942	Systolic_blood_pressure	Cardiovascular	0.84	0.09	SEMA7A	-351	N
rs1378942	Blood_pressure	Cardiovascular	0.84	0.09	SEMA7A	-351	N
rs16857031	QT_interval	Cardiovascular	0.92	0.01	OLFML2B	-157	N
rs16933812	Blood_pressure	Cardiovascular	0.82	0.02	RP11-397D12.7	465	N
rs17259784	Cardiac_hypertrophy	Cardiovascular	0.71	0.04	RP11-565N2.2	49	N
rs1746048	Coronary_heart_disease	Cardiovascular	0.98	0.01	RP11-733D4.1	327	N
rs1746048	Myocardial_infarction	Cardiovascular	0.98	0.01	RP11-733D4.1	327	N
rs17672135	Coronary_heart_disease	Cardiovascular	0.76	0.06	FMN2	-47	Y
rs17691394	Carotid_atherosclerosis_in_HIV_infection	Cardiovascular	0.92	0.00	GRM8	430	N
rs190759	Sudden_cardiac_arrest	Cardiovascular	0.84	0.01	TFAP2B	-198	N
rs2074238	QT_interval	Cardiovascular	0.73	0.01	AC013791.2	307	Y
rs4638289	Atherosclerosis	Cardiovascular	0.98	0.01	TSG101	222	N
rs4687718	QRS_duration	Cardiovascular	0.89	N/A	TMEM110- MUSTN1	-405	N
rs54211	Sudden_cardiac_arrest	Cardiovascular	0.73	0.01	CTA-150C2.16	-293	N
rs6801957	QRS_duration	Cardiovascular	0.85	0.04	SCN10A	72	Y
rs7808424	Coronary_heart_disease	Cardiovascular	0.86	0.37	AC003045.1	15	N
rs789852	QT_interval	Cardiovascular	0.93	0.01	ATP13A3	-146	N
rs8049607	QT_interval	Cardiovascular	0.70	0.01	PRM1	-314	N
rs880315	Diastolic_blood_pressure	Cardiovascular	0.97	N/A	RP11-340B24.3	157	N
rs880315	Systolic_blood_pressure	Cardiovascular	0.97	N/A	RP11-340B24.3	157	N
rs9298506	Intracranial_aneurysm	Cardiovascular	0.95	0.35	RP11-53M11.3	28	Y
rs944260	Sudden_cardiac_arrest	Cardiovascular	0.77	0.01	RP11-429E11.3	52	Y
rs9470361	QRS_duration	Cardiovascular	0.84	0.01	RP1-90K10.3	272	N
rs9581094	Sudden_cardiac_arrest	Cardiovascular	0.73	0.23	PARP4	4	Y
rs964184	Coronary_heart_disease	Cardiovascular	0.90	0.04	SIK3	96	N
rs11867934	Diabetic_retinopathy	Diabetes	0.96	N/A	FLCN	195	N
rs17696736	Type_1_diabetes	Diabetes	0.75	0.08	ACAD10	-343	N
rs2237897	Type_2_diabetes	Diabetes	0.72	0.01	OSBPL5	253	N
rs3007729	Diabetic_retinopathy	Diabetes	0.82	0.02	IGSF21	-95	N
rs3024505	Type_1_diabetes_autoantibodies	Diabetes	0.79	0.33	IL10	6	Y
rs3024505	Type_1_diabetes	Diabetes	0.79	0.33	IL10	6	Y
rs5753037	Type_1_diabetes	Diabetes	0.77	0.26	HORMAD2	-63	N
rs7111341	Type_1_diabetes	Diabetes	0.95	0.02	IGF2	-43	N
rs7171171	Type_1_diabetes_autoantibodies	Diabetes	0.74	0.04	C15orf53	82	Y
rs10202231	Response_to_antipsychotic_therapy_perphenazi	Drug_metabolism	0.99	0.01	RP11-416L21.1	438	N

SNP	Disease or trait	Trait Category	R	LD	Cor_gene_name	Dist_cor	Adjacent?
rs1061235	ne-triglycerides						
rs10950821	Response_to_carbamazepine	Drug_metabolism	0.83	0.14	HLA-A	-3	N
rs12147450	Response_to_statin_therapy-acylcarnitine	Drug_metabolism	0.84	0.01	MACC1	-390	N
	Response_to_antipsychotic_therapy-extrapyr- idal_side_effects	Drug_metabolism	0.77	0.01	CCNB1IP1	-160	N
rs1535	Response_to_statin_therapy-braces	Drug_metabolism	0.92	0.01	C11orf66	-342	N
rs2163287	Response_to_antidepressants-bupropion	Drug_metabolism	0.97	N/A	SERAC1	499	N
rs2830840	Response_to_citalopram_treatment	Drug_metabolism	0.71	0.01	AP001601.2	-404	N
rs286913	Response_to_antipsychotic_therapy- FEV1/FVC	Drug_metabolism	0.96	0.01	ELF5	-120	N
rs2954038	Response_to_statin_therapy-Triglyceride_sum	Drug_metabolism	0.99	0.02	TRIB1	-62	N
rs3753242	Olanzapine_Schizophrenia_neurocognition	Drug_metabolism	0.91	0.45	PRKCZ	-3	Y
rs3795578	Acetaminophen_hepatotoxicity	Drug_metabolism	1.00	0.01	RP11-203F10.6	204	N
rs9658108	Response_to_antipsychotic_therapy_clozapine- glucose	Drug_metabolism	0.73	0.33	ZNF76	-105	N
rs1034566	Platelet_count	Hematological_param	0.99	0.23	ARVCF	-6	Y
rs10489087	Red_blood_cell_count	Hematological_param	0.94	0.01	RP11-341G5.1	23	Y
rs11628318	Platelet_count	Hematological_param	0.89	0.02	RAGE	-335	Y
rs12566888	Platelet_aggregation-ADP	Hematological_param	0.87	0.00	IQGAP3	-360	N
rs12566888	Platelet_aggregation-epinephrine	Hematological_param	0.87	0.00	IQGAP3	-360	N
rs12718597	Mean_corpuscular_volume	Hematological_param	0.72	0.01	AC020743.3	-184	N
rs1354034	Mean_platelet_volume	Hematological_param	0.92	N/A	CCDC66	-199	N
rs1354034	Platelet_count	Hematological_param	0.92	N/A	CCDC66	-199	N
rs1408272	Mean_corpuscular_hemoglobin	Hematological_param	0.70	0.06	TRIM38	124	N
rs1558324	Mean_platelet_volume	Hematological_param	0.87	0.13	VWF	-55	N
rs2336384	Platelet_count	Hematological_param	0.80	0.05	MIIP	34	N
rs385893	Platelet_count	Hematological_param	0.97	0.01	JAK2	221	N
rs3859192	WBC_count	Hematological_param	0.85	0.02	RARA	336	N
rs4148441	Platelet_count	Hematological_param	0.77	0.01	ABCC4	-188	Y
rs4660456	Platelet_count	Hematological_param	0.93	0.01	COL9A2	-456	N
rs4812048	Mean_platelet_volume	Hematological_param	0.98	0.00	EDN3	288	N
rs4895441	Mean_corpuscular_volume	Hematological_param	0.86	0.08	RP1-32B1.4	155	N
rs4895441	WBC_count	Hematological_param	0.86	0.08	RP1-32B1.4	155	N
rs6108011	Red_blood_cell_count	Hematological_param	0.73	0.02	RP5-836E8.1	264	Y
rs643381	Mean_corpuscular_volume	Hematological_param	0.81	0.00	RP11-15H7.1	100	Y
rs7775698	Mean_corpuscular_hemoglobin	Hematological_param	0.80	0.05	RP1-32B1.4	162	N
rs7775698	Mean_corpuscular_volume	Hematological_param	0.80	0.05	RP1-32B1.4	162	N
rs7775698	Red_blood_cell_traits	Hematological_param	0.80	0.05	RP1-32B1.4	162	N
rs7961894	Mean_platelet_volume	Hematological_param	0.96	0.01	CLIP1	395	N
rs7961894	Platelet_count	Hematological_param	0.96	0.01	CLIP1	395	N
rs8176746	Mean_corpuscular_hemoglobin	Hematological_param	0.88	0.05	C9orf7	193	N
rs9349205	Mean_corpuscular_hemoglobin	Hematological_param	0.79	0.03	TFEB	-221	N
rs9349205	Mean_corpuscular_volume	Hematological_param	0.79	0.03	TFEB	-221	N
rs9483788	Hematocrit	Hematological_param	0.81	N/A	HBS1L	-130	Y
rs10516526	FEV1	Kidney_lung_liver	0.83	0.08	NPNT	143	N
rs1529672	FEV1/FVC	Kidney_lung_liver	0.81	0.01	TOP2B	140	Y
rs1883414	IgA_nephropathy	Kidney_lung_liver	0.89	N/A	RXR8	82	N
rs2187668	idiopathic_membranous_nephropathy	Kidney_lung_liver	0.88	0.01	BRD2	335	N
rs2216228	NAFLD_histology	Kidney_lung_liver	0.90	0.06	RP11-268P4.2	383	N
rs2284746	FEV1/FVC	Kidney_lung_liver	0.91	0.43	MFAP2	0	Y
rs4129267	FEF	Kidney_lung_liver	0.88	0.39	IL6R	-19	Y
rs643608	NAFLD_histology	Kidney_lung_liver	0.83	0.01	CBS	-279	N
rs7632299	NAFLD_histology	Kidney_lung_liver	0.91	0.01	SLC9A9	360	N
rs10194115	Erectile_dysfunction_and_prostate_cancer_treat- ment	Miscellaneous	0.99	0.01	C2orf61	139	N
rs12045440	Goiter	Miscellaneous	0.95	N/A	UBR4	-270	N
rs12045440	Thyroid_volume	Miscellaneous	0.95	N/A	UBR4	-270	N
rs13208776	Vitiligo	Miscellaneous	0.89	0.01	FRMD1	-469	N
rs2280543	Uterine_fibroids	Miscellaneous	0.88	0.01	RNH1	299	N
rs2553268	Exercise_treadmill_test_traits	Miscellaneous	0.74	0.01	CTD-2373N4.1	-437	N
rs3796619	Recombination_rate_males	Miscellaneous	0.95	0.01	GAK	-246	N
rs6049375	Erectile_dysfunction_and_prostate_cancer_treat- ment	Miscellaneous	0.79	0.04	GAPDHP53	367	N
rs6847149	Exercise_treadmill_test_traits	Miscellaneous	0.90	0.03	AC004067.4	-205	N
rs735860	Glaucoma	Miscellaneous	0.86	0.01	RP1-214M20.3	-312	N
rs738322	Cutaneous_nevi	Miscellaneous	0.78	0.01	RP1-199H16.5	483	N
rs7567389	Self-rated_health	Miscellaneous	0.82	0.16	MAP3K2	119	N
rs10893366	Alcohol_dependence	Neurological_behavioral	0.71	0.02	EI24	271	N
rs1107592	Bipolar_disorder_and_schizophrenia	Neurological_behavioral	0.83	0.04	MAD1L1	215	Y
rs12290811	Bipolar_disorder	Neurological_behavioral	0.98	0.01	ODZ4	-102	Y
rs12807809	Schizophrenia	Neurological_behavioral	0.99	0.01	SLC37A2	327	N
rs1412115	Schizophrenia	Neurological_behavioral	0.80	0.07	RP11-490O24.1	64	Y
rs1449984	Major_depressive_disorder	Neurological_behavioral	0.95	0.02	AC016768.1	-158	Y
rs1550976	Asperger_disorder	Neurological_behavioral	0.93	0.00	AP002856.5	-197	N
rs16973500	ADHD	Neurological_behavioral	0.91	0.08	PMFBP1	240	N
rs17069122	Bipolar_disorder_and_schizophrenia	Neurological_behavioral	0.72	0.14	RP1-111B22.2	4	Y
rs1879248	Schizophrenia	Neurological_behavioral	0.81	0.21	FXR1	120	Y
rs2002030	Immediate_Story_Recall	Neurological_behavioral	0.81	0.04	BLK	75	N
rs2021722	Schizophrenia	Neurological_behavioral	0.70	0.11	KIAA1949	482	N
rs2070615	Bipolar_disorder	Neurological_behavioral	0.95	0.02	RPS10P20	-337	N
rs2268983	Smoking_behavior	Neurological_behavioral	0.76	0.03	EXD2	285	N

SNP	Disease or trait	Trait Category	R	LD	Cor_gene_name	Dist_cor	Adjacent?
rs2349775	Neuroticism	Neurological_behavioral	0.90	0.00	ICA1	-415	N
rs4307059	Autism	Neurological_behavioral	0.85	0.19	MSNP1	-60	Y
rs4380451	Bipolar_disorder	Neurological_behavioral	0.86	0.00	OSBPL10	-368	N
rs493187	Bipolar_disorder_and_schizophrenia	Neurological_behavioral	0.91	0.06	RP11-15J23.1	-327	N
rs6716455	Alcohol_dependence	Neurological_behavioral	0.83	0.10	AC113610.1	-10	Y
rs6716455	Alcohol_use_disorder	Neurological_behavioral	0.83	0.10	AC113610.1	-10	Y
rs6782029	Anorexia_nervosa	Neurological_behavioral	0.94	0.40	VGLL4	0	Y
rs6952808	Bipolar_disorder_and_schizophrenia	Neurological_behavioral	0.89	0.16	MAD1L1	-15	N
rs6968385	ADHD	Neurological_behavioral	0.93	0.10	AC003088.1	127	Y
rs702543	Neuroticism	Neurological_behavioral	0.82	0.00	PDE4D	-330	N
rs7045881	Schizophrenia	Neurological_behavioral	0.88	0.01	NCRNA00032	354	N
rs7178909	Common_traits_optimism	Neurological_behavioral	0.73	0.01	IDH2	198	N
rs7520258	Working_memory	Neurological_behavioral	0.92	0.01	LGALS8	391	N
rs7578035	Bipolar_disorder	Neurological_behavioral	0.95	0.08	YWHAQP5	-73	N
rs7581919	Conduct_disorder_case_status	Neurological_behavioral	0.99	0.01	RP11-120J4.1	345	N
rs7992643	ADHD	Neurological_behavioral	0.97	0.05	CLYBL	-32	Y
rs806276	ADHD	Neurological_behavioral	0.76	0.01	BACH2	-489	Y
rs933688	Smoking_behavior	Neurological_behavioral	0.95	0.21	RP11-414H23.2	-245	N
rs9810857	ADHD	Neurological_behavioral	0.80	0.01	RP11-372E1.4	-339	N
rs9845475	ADHD	Neurological_behavioral	0.76	0.10	CNOT10	-31	N
rs1451375	Malaria	Parasitic_bacterial_disease	0.84	0.11	GRB10	52	N
rs10514345	Hip_geometry	Quantitative_traits	0.91	0.03	RP11-414H23.2	94	Y
rs11989122	Height	Quantitative_traits	0.91	0.01	AC023590.1	475	N
rs12203592	Freckling	Quantitative_traits	0.85	0.01	RP3-416J7.3	-241	N
rs12203592	Hair_color-Black_vs._blond_hair_color	Quantitative_traits	0.85	0.01	RP3-416J7.3	-241	N
rs12203592	Hair_color-Black_vs._red_hair_color	Quantitative_traits	0.85	0.01	RP3-416J7.3	-241	N
rs12203592	Hair_color	Quantitative_traits	0.85	0.01	RP3-416J7.3	-241	N
rs1635852	Height	Quantitative_traits	0.76	0.00	CREB5	285	N
rs2054989	Hip_geometry	Quantitative_traits	0.98	0.25	C3orf63	133	N
rs2282978	Height	Quantitative_traits	0.76	0.02	KRIT1	-392	N
rs2284746	Height	Quantitative_traits	0.91	0.43	MFAP2	0	Y
rs2336725	Height	Quantitative_traits	0.95	0.01	PRKCD	71	N
rs2523178	Height	Quantitative_traits	0.87	0.04	DOT1L	-111	Y
rs2730245	Height	Quantitative_traits	0.84	0.10	NCAPG2	-227	N
rs291671	Hair_color-red_hair	Quantitative_traits	0.83	0.02	RP4-553F4.6	430	N
rs3782089	Height	Quantitative_traits	0.88	0.06	FIBP	317	N
rs3791950	Height	Quantitative_traits	0.72	0.01	PNKD	469	N
rs4072910	Height	Quantitative_traits	0.91	0.01	PRAM1	-87	N
rs4282339	Height	Quantitative_traits	0.98	0.18	SLIT3	15	Y
rs4823006	Waist-hip_ratio	Quantitative_traits	0.81	0.01	APIB1	315	N
rs4932217	Height	Quantitative_traits	0.83	0.23	POLG	-40	Y
rs619865	Freckling	Quantitative_traits	0.72	0.01	RBM39	457	N
rs6784615	Waist-hip_ratio	Quantitative_traits	0.73	0.22	BAP1	-64	N
rs6899976	Height	Quantitative_traits	0.71	0.01	RP1-69D17.4	-490	N
rs7007970	Height	Quantitative_traits	0.88	0.00	RP11-775B15.3	-152	N
rs7121446	Waist_circumference	Quantitative_traits	0.80	0.02	RP11-166D19.1	18	Y
rs7349332	Hair_curl	Quantitative_traits	0.72	0.18	AC097468.6	63	N
rs7349332	Hair_morphology	Quantitative_traits	0.72	0.18	AC097468.6	63	N
rs735854	Optic_disc_size_rim	Quantitative_traits	0.72	0.01	APOL3	-117	N
rs7466269	Height	Quantitative_traits	0.99	0.13	RP11-57C19.2	57	N
rs798497	Height	Quantitative_traits	0.95	0.00	EIF3B	-379	N
rs941873	Height	Quantitative_traits	0.79	0.43	RP11-342M3.5	2	Y
rs946053	Height	Quantitative_traits	0.95	0.01	AMBP	-210	N
rs228769	Bone_mineral_density-hip	Radiographic_parameter	0.71	0.07	MPP2	-215	Y
rs228769	Bone_mineral_density-spine	Radiographic_parameter	0.71	0.07	MPP2	-215	Y
rs4870044	Bone_mineral_density-hip	Radiographic_parameter	0.81	0.04	RP11-351K16.4	-164	N
rs4870044	Bone_mineral_density-spine	Radiographic_parameter	0.81	0.04	RP11-351K16.4	-164	N
rs1039302	C-reactive_protein	Serum_metabolites	0.94	0.01	RNF10	-265	N
rs10889353	Cholesterol	Serum_metabolites	0.96	0.02	RP5-1155K23.1	-463	N
rs10889353	LDL_cholesterol	Serum_metabolites	0.96	0.02	RP5-1155K23.1	-463	N
rs10889353	Triglycerides	Serum_metabolites	0.96	0.02	RP5-1155K23.1	-463	N
rs11597390	Alanine_aminotransferase	Serum_metabolites	0.86	N/A	ENTPD7	-442	N
rs11708067	Fasting_plasma_glucose	Serum_metabolites	0.98	0.03	PDIA5	-223	N
rs11708067	Insulin_resistance	Serum_metabolites	0.98	0.03	PDIA5	-223	N
rs11761528	Serum_dehydroepiandrosterone	Serum_metabolites	0.91	0.31	ARPC1B	-147	N
rs12239046	C-reactive_protein	Serum_metabolites	0.99	0.02	NLRP3	-20	Y
rs12239436	HDL_cholesterol	Serum_metabolites	0.70	0.02	RP11-101C11.1	-59	Y
rs12740374	LDL_cholesterol	Serum_metabolites	0.75	0.01	AMPD2	354	N
rs13022873	Triglycerides_waist_circumference	Serum_metabolites	0.72	0.16	IFT172	-131	N
rs13195786	Serum_calcium	Serum_metabolites	0.85	0.02	TFAP2A	256	N
rs1335645	Gamma_glutamyl_transferase	Serum_metabolites	0.74	N/A	DENND2D	59	N
rs1408272	Transferrin_receptor	Serum_metabolites	0.70	0.06	TRIM38	124	N
rs1535	HDL_cholesterol	Serum_metabolites	0.92	0.01	C11orf66	-342	N
rs1535	Serum_polyunsaturated_fatty_acids	Serum_metabolites	0.92	0.01	C11orf66	-342	N
rs157580	HDL_cholesterol	Serum_metabolites	0.85	0.00	CEACAM19	-221	N
rs157580	LDL_cholesterol	Serum_metabolites	0.85	0.00	CEACAM19	-221	N
rs1594468	Bilirubin	Serum_metabolites	0.74	0.03	ELMOD2	58	N
rs17319721	Creatinine	Serum_metabolites	0.77	N/A	CXCL11	-405	N
rs174536	Serum_polyunsaturated_fatty_acids	Serum_metabolites	0.88	0.00	DAK	-451	N
rs174546	HDL_cholesterol	Serum_metabolites	0.86	0.01	DDB1	-460	N
rs174546	LDL_cholesterol	Serum_metabolites	0.86	0.01	DDB1	-460	N

SNP	Disease or trait	Trait Category	R	LD	Cor_gene_name	Dist_cor	Adjacent?
rs174574	Serum_polyunsaturated_fatty_acids	Serum_metabolites	0.96	0.01	C11orf66	-344	N
rs1967017	Serum_urate	Serum_metabolites	0.76	0.03	RP11-458D21.2	-348	N
rs2052550	Ferritin	Serum_metabolites	0.86	0.08	DMGDH	54	N
rs2066219	Two_hour_glucose_challenge	Serum_metabolites	0.75	0.05	RPL12P34	-311	Y
rs2078267	Gout	Serum_metabolites	0.83	0.00	ARL2	450	N
rs2078267	Serum_urate	Serum_metabolites	0.83	0.00	ARL2	450	N
rs2153960	IGF1	Serum_metabolites	0.97	0.02	NR2E1	-491	N
rs2235302	P-selectin	Serum_metabolites	0.93	0.07	SELL	94	N
rs2650000	C-reactive_protein	Serum_metabolites	0.83	0.01	KDM2B	491	N
rs2650000	LDL_cholesterol	Serum_metabolites	0.83	0.01	KDM2B	491	N
rs2836878	C-reactive_protein	Serum_metabolites	0.80	0.03	LCA5L	352	N
rs2877716	Two-hour_glucose_challenge	Serum_metabolites	0.74	N/A	ADCY5	-56	Y
rs3093030	ICAM1	Serum_metabolites	0.94	0.08	ZGLP1	19	Y
rs3729639	HDL_cholesterol	Serum_metabolites	0.71	N/A	TRADD	-34	N
rs4129267	C-reactive_protein	Serum_metabolites	0.88	0.39	IL6R	-19	Y
rs4129267	IL6R	Serum_metabolites	0.88	0.39	IL6R	-19	Y
rs4273077	Protein-total	Serum_metabolites	0.75	0.01	AC055811.5	294	N
rs4516970	Ferritin	Serum_metabolites	0.76	0.01	RP1-249F5.3	182	N
rs4607517	Insulin_resistance	Serum_metabolites	0.85	0.01	POLD2	-74	N
rs4607517	Fasting_plasma_glucose	Serum_metabolites	0.85	0.01	POLD2	-74	N
rs4686760	Van_Wildebrand_factor_antibodies	Serum_metabolites	0.87	0.02	VPS8	122	Y
rs4737009	HbA1C	Serum_metabolites	0.94	0.03	AGPAT6	-197	N
rs4820599	Gamma_glutamyl_transferase	Serum_metabolites	0.80	0.01	C22orf13	-44	N
rs4963452	Serum_polyunsaturated_fatty_acids	Serum_metabolites	0.80	0.03	SCGB2A2	222	N
rs507666	ICAM1	Serum_metabolites	0.83	0.08	C9orf7	178	N
rs6442522	Serum_urate	Serum_metabolites	0.73	0.06	NR2C2	-451	N
rs6734238	C-reactive_protein	Serum_metabolites	1.00	0.18	IL1RN	44	Y
rs6984305	Alkaline_phosphatase	Serum_metabolites	0.85	0.02	RP11-115J16.2	118	N
rs7117404	Fibrin-D-dimer_levels	Serum_metabolites	0.79	0.01	ATG13	-446	N
rs7120118	HDL_cholesterol	Serum_metabolites	0.79	0.01	CELF1	225	N
rs7569328	LDL_cholesterol	Serum_metabolites	0.80	0.01	HS1BP3	-262	N
rs7778619	CD40_ligand	Serum_metabolites	0.82	0.01	AC060834.2	-385	N
rs8109578	Thyroid_stimulating_hormone	Serum_metabolites	0.92	0.01	PPAN	5	Y
rs911119	Cystatin_C	Serum_metabolites	0.75	0.01	RP4-737E23.4	-499	N
rs964184	Alpha-tocopherol	Serum_metabolites	0.90	0.04	SIK3	96	N
rs964184	HDL_cholesterol	Serum_metabolites	0.90	0.04	SIK3	96	N
rs964184	Hypertriglyceridemia	Serum_metabolites	0.90	0.04	SIK3	96	N
rs964184	Lipoprotein-associated_phospholipase_A2_activity_and_mass	Serum_metabolites	0.90	0.04	SIK3	96	N
rs964184	Triglycerides	Serum_metabolites	0.90	0.04	SIK3	96	N
rs9939224	HDL_cholesterol_fasting_glucose	Serum_metabolites	0.75	0.18	CETP	-7	Y
rs9992101	Creatinine	Serum_metabolites	0.85	0.02	ART3	-357	N
rs11697186	Response_to_hepatitis_C_treatment	Viral_disease	0.85	0.11	RP5-1187M17.10	-31	N
rs13394720	HIV_progression	Viral_disease	0.90	0.02	AC019221.4	-239	N
rs2885805	Cytomegalovirus_antibody_response	Viral_disease	0.87	0.00	C1orf88	471	N
rs9267665	Hepatitis-B_vaccine_response	Viral_disease	0.94	0.07	UQCRHP1	-289	N

Table 10-2 Genes correlated with distal DHSs harboring GWAS SNPs (Part II)

128 trait-SNP associations (123 unique SNPs) overlapping predicted long-distance interactions established by correlation of chromatin accessibility (R) in DHSs in 46 additional cell/tissue types. Cor_gene_name represents the most-correlated gene; Dist, distance to gene in kb. "Adjacent?", whether the highest-correlated gene is an adjacent gene.

SNP	Disease or trait	Trait Category	R	Cor_gene_name	Dist_cor	Adjacent?
rs62209	Alzheimers_disease_late_onset	Aging	0.94	RP11-271F18.2	496	N
rs4938933	Alzheimers_disease_late_onset	Aging	0.89	MS4A3	201	N
rs2619566	Amyotrophic_lateral_sclerosis-age_of_onset	Aging	0.88	CNTN4	307	N
rs9664222	Longevity	Aging	0.74	RP11-57C13.5	62	N
rs1036819	Longevity	Aging	0.87	RP11-513H8.1	279	N
rs947211	Parkinsons_disease	Aging	0.96	SLC26A9	147	N
rs6599388	Parkinsons_disease	Aging	0.84	PDE6B	294	N
rs6599389	Parkinsons_disease	Aging	0.84	PDE6B	294	N
rs11248060	Parkinsons_disease	Aging	0.99	SPON2	231	N
rs4698412	Parkinsons_disease	Aging	0.70	PROM1	349	N
rs10121009	Parkinsons_disease	Aging	0.83	RP11-156G14.4	180	N
rs10767971	Parkinsons_disease_age_of_onset	Aging	0.83	PIGCP1	202	N
rs17565841	Parkinsons_disease_age_of_onset	Aging	0.74	AC090696.2	23	N
rs13010713	Celiac_disease	Autoimmune_disease	0.74	AC104820.2	7	N
rs1819658	Crohns_disease	Autoimmune_disease	0.77	UBE2D1	211	N
rs762421	Crohns_disease	Autoimmune_disease	0.75	TRAPPC10	105	N
rs6596075	Crohns_disease	Autoimmune_disease	0.77	RAD50	188	N
rs212388	Crohns_disease	Autoimmune_disease	0.84	RP3-393E18.1	458	N
rs212388	Crohns_disease_celiac_disease	Autoimmune_disease	0.84	RP3-393E18.1	458	N
rs9355610	Graves_disease	Autoimmune_disease	0.73	RNASET2	12	N
rs6604026	Multiple_sclerosis	Autoimmune_disease	0.89	RP4-612C19.1	357	N
rs12025416	Multiple_sclerosis	Autoimmune_disease	0.89	RP4-655J12.5	62	N
rs7090512	Multiple_sclerosis	Autoimmune_disease	0.86	RP11-414H17.2	3	N
rs4939490	Multiple_sclerosis	Autoimmune_disease	0.93	ZP1	153	N
rs2248359	Multiple_sclerosis	Autoimmune_disease	0.75	CYP24A1	1	N
rs9321490	Multiple_sclerosis	Autoimmune_disease	0.98	AHI1	149	N
rs7779014	Multiple_sclerosis	Autoimmune_disease	0.87	POR	439	N
rs17149161	Multiple_sclerosis	Autoimmune_disease	0.80	AC005077.14	274	N
rs9303277	Primary_biliary_cirrhosis	Autoimmune_disease	0.72	IKZF3	44	N
rs842636	Psoriasis	Autoimmune_disease	0.90	AC007381.2	481	N
rs743777	Rheumatoid_arthritis	Autoimmune_disease	0.89	C1QTNF6	42	N
rs1600249	Rheumatoid_arthritis	Autoimmune_disease	0.94	CTSB	359	N
rs7329174	Systemic_lupus_erythematosus	Autoimmune_disease	0.82	ELF1	1	N
rs1317209	Ulcerative_colitis	Autoimmune_disease	0.86	TMCO4	75	N
rs3024493	Ulcerative_colitis	Autoimmune_disease	0.81	RP11-534L20.4	242	N
rs8067378	Ulcerative_colitis	Autoimmune_disease	0.86	IKZF3	31	N
rs11676348	Ulcerative_colitis	Autoimmune_disease	0.81	C2orf62	219	N
rs6017342	Ulcerative_colitis	Autoimmune_disease	0.71	WISP2	282	N
rs11978267	Acute_lymphoblastic_leukemia_childhood	Cancer	0.85	IKZF1	99	N
rs2380205	Breast_cancer	Cancer	0.92	GDI2	58	N
rs2981579	Breast_cancer	Cancer	0.96	TACC2	411	N
rs1219648	Breast_cancer	Cancer	0.79	FGFR2	7	N
rs10937405	Lung_adenocarcinoma	Cancer	0.84	TPRG1	451	N
rs7521902	Ovarian_cancer	Cancer	0.93	HSPG2	268	N
rs9311171	Prostate_cancer	Cancer	0.80	DLEC1	105	N
rs10263935	Aortic_root_size	Cardiovascular	0.75	RABGEF1	163	N
rs17375901	Atrial_fibrillation	Cardiovascular	0.99	NPPA	55	N
rs1320448	Cardiac_hypertrophy	Cardiovascular	0.98	COL17A1	0	N
rs216172	Coronary_heart_disease	Cardiovascular	0.78	MNT	178	N
rs7651039	Coronary_heart_disease	Cardiovascular	0.84	CAPN7	360	N
rs17577085	Coronary_heart_disease	Cardiovascular	0.83	ARHGAP26	423	N
rs17609940	Coronary_heart_disease	Cardiovascular	0.83	DEF6	241	N
rs6601530	Internal_carotid_intimal_medial_thickness	Cardiovascular	0.71	SOX7	83	N
rs499818	Major_CVD	Cardiovascular	0.81	GFOD1	78	N
rs11748327	Myocardial_infarction	Cardiovascular	0.88	CTD-2287N17.1	16	N
rs10757278	Myocardial_infarction	Cardiovascular	0.85	CDKN2B-AS1	11	N
rs3807989	PR_interval	Cardiovascular	0.82	ST7	469	N
rs17421627	Retinal_vascular_caliber	Cardiovascular	0.89	CTC-547D20.1	127	N
rs225717	Retinal_vascular_caliber	Cardiovascular	0.90	GPR126	75	N
rs4975709	Stroke	Cardiovascular	0.77	CTD-2194D22.2	20	N
rs10829156	Sudden_cardiac_arrest	Cardiovascular	0.71	RP11-288D15.2	390	N
rs16866933	Sudden_cardiac_arrest	Cardiovascular	0.85	AC092642.1	111	N
rs7042864	Tonometry	Cardiovascular	0.88	RP11-272G11.1	143	N
rs4948088	Type_1_diabetes	Diabetes	0.71	RP4-724E13.2	93	N
rs3788013	Type_1_diabetes_autoantibodies	Diabetes	0.86	AP001625.6	140	N
rs743777	Type_1_diabetes_autoantibodies	Diabetes	0.89	C1QTNF6	42	N
rs7901695	Type_2_diabetes	Diabetes	0.71	TCF7L2	157	N
rs2383208	Type_2_diabetes	Diabetes	0.77	RP11-70L8.3	329	N
rs2472297	Coffee_consumption	Drug_metabolism	1.00	CYP11A1	395	N
rs6588480	Response_to_statin_therapy-chol_sum	Drug_metabolism	0.76	HSPB11	417	N
rs9305406	Response_to_statin_therapy-SM	Drug_metabolism	0.71	KRTAP23-1	336	N
rs17135859	F-cell_distribution	Hematological_parameters	0.79	YTHDC2	71	N
rs17342717	Mean_corpuscular_hemoglobin	Hematological_parameters	0.74	RP3-501N12.3	140	N
rs131794	Mean_corpuscular_volume	Hematological_parameters	0.80	NCAPH2	17	N

SNP	Disease or trait	Trait Category	R	Cor_gene_name	Dist_cor	Adjacent?
rs172629	Mean_corpuscular_volume	Hematological_parameters	0.72	RP11-601I15.1	321	N
rs12485738	Mean_platelet_volume	Hematological_parameters	0.76	C3orf63	163	N
rs8022206	Platelet_count	Hematological_parameters	0.83	RAD51L1	233	N
rs441460	Platelet_count	Hematological_parameters	0.93	RP3-522P13.3	285	N
rs11611647	Red_blood_cell_count	Hematological_parameters	0.86	DYRK4	342	N
rs7805747	Chronic_kidney_disease	Kidney_lung_liver	0.86	RP13-452N2.1	96	N
rs10786284	ADHD	Neurological_behavioral	0.88	BLNK	147	N
rs1859156	ADHD	Neurological_behavioral	0.93	PDLIM5	329	N
rs12020569	Alcohol_dependence	Neurological_behavioral	0.88	RPS3AP44	437	N
rs12282742	Bipolar_disorder_and_schizophrenia	Neurological_behavioral	0.80	RP11-113D6.8	70	N
rs17197037	Bipolar_disorder	Neurological_behavioral	0.82	METTL3	248	N
rs6990255	Bipolar_disorder	Neurological_behavioral	0.90	RP1-273G13.3	256	N
rs1574192	Brain_imaging_in_schizophrenia_interaction	Neurological_behavioral	0.95	KIF1A	429	N
rs9442235	Cognitive_performance-PC1	Neurological_behavioral	0.84	RP11-169K16.4	315	N
rs16880441	Conduct_disorder_interaction	Neurological_behavioral	0.95	ACTBP8	31	N
rs332034	Conduct_disorder_interaction	Neurological_behavioral	0.92	RP11-115J16.1	389	N
rs3827730	Depression_and_alcohol_dependence	Neurological_behavioral	0.76	RP11-183G22.3	57	N
rs12042938	DISC1	Neurological_behavioral	0.91	C1orf131	455	N
rs1869901	Schizophrenia	Neurological_behavioral	0.70	IVD	111	N
rs8005962	Tuberculosis	Parasitic_bacterial_disease	0.84	TCL6	104	N
rs6545883	Tuberculosis	Parasitic_bacterial_disease	0.71	USP34	323	N
rs9990343	Brain_structure	Quantitative_traits	0.94	CXCR6	357	N
rs17646946	Hair_curl	Quantitative_traits	0.99	TCHHL1	1	N
rs17318596	Height	Quantitative_traits	0.71	BCKDHA	20	N
rs3791679	Height	Quantitative_traits	0.81	CCDC88A	470	N
rs6724465	Height	Quantitative_traits	0.94	NHEJ1	46	N
rs12658202	Height	Quantitative_traits	0.97	FAM114A2	430	N
rs6569648	Height	Quantitative_traits	0.74	ARHGAP18	430	N
rs6570507	Height	Quantitative_traits	0.79	GPR126	58	N
rs6611365	Optic_disc_size_disc	Quantitative_traits	0.70	CTD-2522E6.4	157	N
rs9386463	Primary_tooth_development_time_to_first_tooth_eruption	Quantitative_traits	0.99	PRDM1	453	N
rs1572050	Renal_sinus_fat	Quantitative_traits	0.80	RP11-23P11.2	30	N
rs9315632	Waist-hip_ratio	Quantitative_traits	0.73	C13orf23	70	N
rs1055144	Waist-hip_ratio	Quantitative_traits	0.70	CTA-242H14.1	141	N
rs9594738	Bone_mineral_density-hip	Radiographic_parameters	0.90	FABP3P2	8	N
rs9594738	Bone_mineral_density-spine	Radiographic_parameters	0.90	FABP3P2	8	N
rs4729260	Bone_mineral_density-spine	Radiographic_parameters	0.85	SHFM1	209	N
rs10492681	Alanine_aminotransferase	Serum_metabolites	0.76	RP11-518D7.1	52	N
rs2280401	Albumin	Serum_metabolites	0.71	AKT1S1	379	N
rs16856332	Alkaline_phosphatase	Serum_metabolites	0.77	AC007556.2	129	N
rs6742078	Bilirubin	Serum_metabolites	0.94	TRPM8	168	N
rs7953249	DG7_glycan	Serum_metabolites	0.75	HNF1A	12	N
rs17342717	Ferritin	Serum_metabolites	0.74	RP3-501N12.3	140	N
rs12029080	Fibrin-D-dimer_levels	Serum_metabolites	0.79	CNN3	340	N
rs1490453	Fibrinogen	Serum_metabolites	0.75	RP11-655B23.1	127	N
rs7998202	HbA1C	Serum_metabolites	0.72	MCF2L	409	N
rs7499892	HDL_cholesterol	Serum_metabolites	0.84	NLRCS	100	N
rs2083637	HDL_cholesterol	Serum_metabolites	0.77	CSGALNACT1	404	N
rs13702	HDL_cholesterol_triglycerides	Serum_metabolites	0.84	CSGALNACT1	363	N
rs9303029	IGF1-free	Serum_metabolites	0.85	FOXK2	69	N
rs7577642	IL6R	Serum_metabolites	0.81	SH2D6	43	N
rs591044	Insulin_resistance	Serum_metabolites	0.88	RP11-259P1.1	45	N
rs2280401	Protein-total	Serum_metabolites	0.71	AKT1S1	379	N
rs236918	Transferrin_receptor	Serum_metabolites	0.77	PAFAH1B2	64	N

CURRICULUM VITAE
Matthew Thomas Maurano

Address: University of Washington
Dept. of Genome Sciences
2211 Elliott Ave, Suite 600
Seattle, WA 98121
maurano@uw.edu

EDUCATION:

- 2013 Ph.D. Dept. of Genome Sciences, University of Washington, Seattle, WA, USA
 Thesis: *Regulatory variation and human disease*
 Advisor: Dr. John A. Stamatoyannopoulos
 (Thesis defense date June 2012; degree conferral electively delayed to December 2013)
- 2005 B.A. (Honors) Computer Science, University of California, Berkeley, CA, USA
 B.A. Molecular & Cell Biology

POSITIONS HELD:

- 2008-13 Doctoral research student, Dept. of Genome Sciences, University of Washington, Seattle, WA.
2005-08 Project Manager, Professional Services & Software Quality Engineer at Rosetta (Merck) in Seattle, WA.
2004 Research Intern, University of Trento, Italy. Advisors: Drs. Giuseppe Jurman and Cesare Furlanello.
 Thesis: *Microarray missing data reconstruction methods*.
2002-03 Undergraduate Researcher, Department of Plant and Microbial Biology, UC Berkeley.
 Advisor: Dr. Steven Brenner. Topic: *Homology of NUDIX proteins in D. Radiodurans*.

HONORS AND AWARDS:

- 2013 UW Graduate School Fund for Excellence and Innovation Travel Award
2011 UW Graduate School Fund for Excellence and Innovation Travel Award
2000-04 UC Berkeley Regents' scholarship
2000-04 Robert C. Byrd Honors scholarship

INVITED PRESENTATIONS:

- Invited speaker, American Heart Association annual meeting, Dallas, TX (11/17/2013).
Invited speaker, Enhancer Biology in Health and Disease, Banbury Center, Cold Spring Harbor, NY (10/28/2013).
Invited speaker, Dept. of Genetics, Albert Einstein College of Medicine, Bronx, NY (1/25/2013).

SPECIAL LOCAL RESPONSIBILITIES:

2011-12 Member, seminar committee, UW Dept. of Genome Sciences
2010 UW Genome Sciences graduate recruiting

EDITORIAL RESPONSIBILITIES:

Ad-hoc reviewer for *Cell*, *Human Mutation*, *European Journal of Human Genetics*, *PLoS Genetics*.

ORGANIZATIONS & AFFILIATIONS:

American Society of Human Genetics
New York Academy of Sciences
International Society for Computational Biology

TEACHING EXPERIENCE:

2011 Genome 372: Genomics & Proteomics. Graduate teaching assistant.
2011 Genome 361: Fundamentals of Genetics and Genomics. Graduate teaching assistant.
2007-08 Microarray data analysis using Rosetta Resolver – developed and taught 2-day course offered across Merck sites.

RESEARCH FUNDING:

CURRENT:

NIH (1F31MH094073)

Title: Regulatory Genomics of Human Neurodevelopment

Ruth L. Kirschstein National Research Service Award

PI: Matthew T. Maurano

Sponsors: John A. Stamatoyannopoulos and Evan E. Eichler

2012-15

PEER-REVIEWED PUBLICATIONS:

1. **Maurano MT***, Humbert R*, Rynes E*, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, Shafer A, Neri F, Lee K, Kuttyavin T, Stehling-Sun S, Johnson AK, Canfield TK, Giste E, Diegel M, Bates D, Hansen RS, Neph S, Sabo PJ, Heimfeld S, Raubitschek A, Ziegler S, Cotsapas C, Sotoodehnia N, Glass I, Sunyaev SR, Kaul R, and Stamatoyannopoulos JA. "Systematic Localization of

Common Disease-Associated Variation in Regulatory DNA." **Science** 337: 1190–1195 (2012). * *co-first authors*

Featured on the cover and **Perspective at Science**. 2012;337(6099):1179-80

Research highlight at Nature Genetics 2012;44(10):1079

2. **Maurano MT***, Wang H*, Kutayavin T, and Stamatoyannopoulos JA. "Widespread Site-Dependent Buffering of Human Regulatory Polymorphism." **PLoS Genetics** 8: e1002599 (2012). * *co-first authors*
Commentary: Nature Genetics 2012;44(5):481; **Nature Reviews Genetics** 2012;13(5):297.
3. Wang H*, **Maurano MT***, Qu H*, Varley KE, Gertz J, Pauli F, Lee K, Canfield T, Weaver M, Sandstrom R, Thurman RE, Kaul R, Myers RM, and Stamatoyannopoulos JA. "Widespread Plasticity in CTCF Occupancy Linked to DNA Methylation." **Genome Research** 22: 1680–1688 (2012). * *co-first authors*
4. Vernot B, Stergachis AB, **Maurano MT**, Vierstra J, Neph S, Thurman RE, Stamatoyannopoulos JA, and Akey JM. "Personal and Population Genomics of Human Regulatory Variation." **Genome Research** 22: 1689–1697 (2012).
5. Thurman RE, Rynes E, Humbert R, Vierstra J, **Maurano MT**, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutayavin T, Lajoie B, Lee B-K, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, and Stamatoyannopoulos JA. "The Accessible Chromatin Landscape of the Human Genome." **Nature** 489: 75–82 (2012).
6. Neph S, Vierstra J, Stergachis AB, Reynolds AP, Haugen E, Vernot B, Thurman RE, John S, Sandstrom R, Johnson AK, **Maurano MT**, Humbert R, Rynes E, Wang H, Vong S, Lee K, Bates D, Diegel M, Roach V, Dunn D, Neri J, Schafer A, Hansen RS, Kutayavin T, Giste E, Weaver M, Canfield T, Sabo P, Zhang M, Balasundaram G, Byron R, MacCoss MJ, Akey JM, Bender MA, Groudine M, Kaul R, and Stamatoyannopoulos JA. "An Expansive Human Regulatory Lexicon Encoded in Transcription Factor Footprints." **Nature** 489: 83–90 (2012).
7. ENCODE Project Consortium. "An Integrated Encyclopedia of DNA Elements in the Human Genome." **Nature** 489: 57–74 (2012).
8. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, **Maurano MT**, Vierstra J, Thomas S, Sandstrom R, Humbert R, and Stamatoyannopoulos JA. "BEDOPS: High-Performance Genomic Feature Operations." **Bioinformatics** 28: 1919–1920 (2012).

ABSTRACTS:

1. Maurano MT, Haugen E, Rynes E, Humbert R, and Stamatoyannopoulos JA. "Discovery and analysis of over 50,000 common functional regulatory variants in the human genome." American Society of Human Genetics, Boston, MA (10/2013).
2. Maurano MT, Wang H, Shafer A, John S, and Stamatoyannopoulos JA. "Reproducible but limited expansion of transcription factor binding landscape after depletion of methylation." Epigenetics and Chromatin: Interactions and processes, Boston, MA (3/2013).
3. Maurano MT, Wang H, Kutayavin T, and Stamatoyannopoulos JA. "Widespread Site-dependent Buffering of Human Regulatory Polymorphism." Cold Spring Harbor Personal Genomes (9/2011).