Deciphering genetic associations using genome-wide epigenomics approaches

by

Xinchen Wang
B.Sc. Biochemistry
University of Toronto, 2011

Submitted to the Department of Biology
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

Signature of Author .....
Xinchen Wang
Department of Biology
May 26, 2017

Certified by .....
Manolis Kellis
Professor of Computer Science
Thesis Advisor

Certified by .....
Laurie A. Boyer
Associate Professor of Biology
Thesis Advisor

Accepted by .....
Amy E. Keating
Professor of Biology
Co-Chair, Biology Graduate Committee

Deciphering genetic associations using genome-wide epigenomics approaches

by

Xinchen Wang

Submitted to the Department of Biology on May 26, 2017 in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Biology

Abstract

Genetic mapping of the drivers of complex human phenotypes and disease through the genome-wide association study (GWAS) has identified thousands of causal genetic loci in the human population. However, genetic mapping approaches can often only reveal a particular causal locus, not the molecular mechanism through which it acts. Biological interpretation of these genetic results is thus a bottleneck for turning results from GWAS into meaningful biological insights for human biology.

Genetic mapping of complex human traits has revealed that most common variants influencing human phenotypes have weak effect sizes and reside outside protein-coding regions, complicating biological interpretation of their function. In this thesis we use computational and experimental approaches to study the non-coding genome. In particular, we focus on using epigenomic signatures to characterize non-coding transcriptional regulatory elements and predict regulatory segments of DNA disrupted by genetic variants. In Chapter 2, we describe how genome-wide maps of epigenomic modifications can be used to characterize and discover new GWAS loci. In Chapter 3, we outline an experimental method for the high-throughput assessment of putative transcriptional regulatory elements.

In summary, our research highlights the value of interpreting human genetics information through an epigenomic lens, and provides a glimpse into the possible biological insights that manifest from the intersection of these two areas of research.

Thesis Supervisor: Manolis Kellis
Title: Professor of Computer Science

Thesis Supervisor: Laurie A. Boyer
Title: Associate Professor of Biology

Acknowledgements

To my advisors, Manolis Kellis and Laurie Boyer, thank you for supporting and mentoring me over the past six years. Manolis, your enthusiasm for science is infectious. Our impromptu whiteboard talks have taught me so much about being creative and adventurous as a scientist. Laurie, thank you for helping me grow as a scientist. Your support and encouragement has been invaluable. To my informal advisor Melina Claussnitzer, you have been a terrific mentor to me, and I am grateful to have had the chance to work with you.

Thank you to my committee members Peter Reddien, David Page and David Altshuler. I deeply appreciate the scientific and career advice you have given me over the years and the many hours you have been willing to let me sit in your offices.

Thank you to my undergraduate research mentors, Ben Blencowe and Claudia dos Santos. Working with you is what inspired me to pursue scientific research and come to graduate school. To Chris Newton-Cheh and David Milan, thank you for the many conversations we had and your constant willingness to help me along the way.

I have had the fortune and privilege to meet some amazing friends and labmates over the past six years. To everyone from the Boyer lab, including Paul, Kunle, Arune, Vidya, Lauren, Sera, Joe, Gizem, Carla and Darlene, you made it a joy to come into work every day. Thank you for so many wonderful memories in and out of the lab. I am grateful to have known everyone in the Kellis lab, especially my wonderful officemates Abhishek, Kunal, Angela, Bob, Khoi, Carles and Atlas, as well as Gerald, Pouya, Andreas, and Yue, who have taught me so much over the years. Thank you to the nascent Claussnitzer lab, especially Sarah, Li, and Alham, for being great colleagues and friends. Finally, thank you to friends from outside lab and classmates I have had these years for drinks, weekend brunches, Portuguese dinners and recurring coffee trips, and a special thanks to my two roommates Dave and David who I have lived with for five years and have been a constant source of support and humor.

Lastly and most importantly, I want to express my deepest gratitude to my parents and sister for always being there for me and to my girlfriend Cheng for her endless optimism and joy. You have been my biggest supporters through a long graduate school process that has had its share of bumps and detours along the way. Thank you for your encouragement and for always believing in me.

Table of Contents

Chapter 1: Introduction

Motivation

Much of human disease is influenced by heritable variation in DNA that is transmitted between generations, suggesting that an understanding of human genetics will aid in the study of human biology and disease. Moreover, the knowledge that genetic mutations disrupting specific genes can modify disease risk hints that the perturbed biological pathways can be promising therapeutic targets[1]. Unbiased genetic mapping of disease variation has now implicated thousands of DNA regions in common human disease, painting a complex picture where common disease is influenced by hundreds or thousands of distinct loci that largely fall within non-coding DNA. Non-coding DNA makes up 98% of the genome and is largely not under evolutionary conservation, making it difficult to distinguish critical regulatory DNA sequences from their non-functional surroundings. The work presented in this thesis describes computational and experimental approaches to understand how genetic variants in non-coding DNA can affect molecular processes and ultimately cellular and organismal phenotypes associated with human traits and disease.

Background

Genetic architecture of human disease and the implications for therapeutics development

Tracing the inheritance of human genetic variation within families and populations has led to the discovery of thousands of regions of the human genome associated with a plethora of rare diseases[2]. In the cases of rare Mendelian diseases, the majority of pathogenic mutations are localized to protein-coding regions and these mutations are often individually capable of conferring a strong deleterious phenotype. These family-based "linkage" analyses use genetic markers, such as single nucleotide polymorphisms (SNPs), short insertions and deletions (indels) or larger structural variants, as a proxy to infer the inheritance of different segments of DNA from parents to children[3] (Fig. 1). The linkage, or statistical correlation between inheritance of a specific genetic marker and the phenotypic manifestation of a disease, has been used highly effectively to map mutations

contributing to Mendelian diseases, however this group of statistical approaches is not applicable to complex traits and diseases that do not have clear family clustering and inheritance patterns and are believed to be influenced by the collective effects of hundreds or thousands of genetic loci. Population genetics offers a solution to this problem through the design of a genetic association study, where instead of tracking allele transmissions within pedigrees through time, the frequencies of densely spaced polymorphic alleles spanning the entire human genome are compared within cohorts drawn from the human population[2] (Fig. 1). Unlike a family-based linkage study, the genome-wide association study (GWAS) uses cohorts of unrelated individuals to avoid artifacts from genetic relatedness and thus does not explicitly model a family structure. With increasingly large cohort sizes now in the hundreds of thousands of individuals, the genome-wide association study has identified thousands of genetic loci targeting a multitude of biological pathways associated with complex human traits[4].

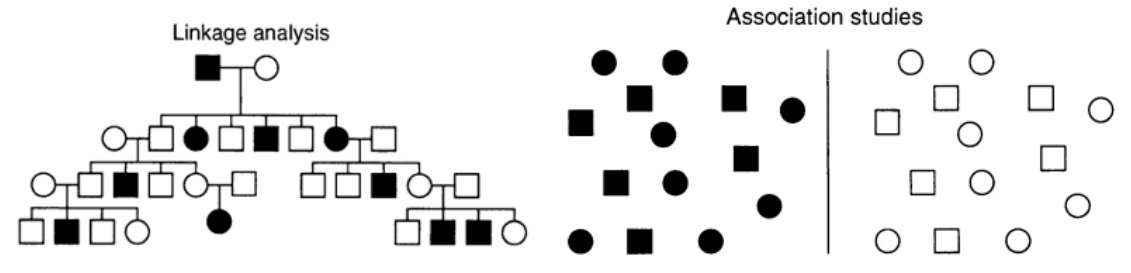


Figure 1: Overview of two approaches for genetic mapping of human disease. Adapted from Lander and Schork (1994)[3]. *Left*, Study design for genetic linkage analysis. A model is constructed to identify genetic variants inherited within pedigrees that follow a similar transmission pattern as the disease phenotype (shaded individuals). *Right*, Study design for a case-control genetic association study. For complex traits with hundreds or thousands of contributing genetic loci, allele frequencies are compared between sample groups with no consideration of family structure that would be too difficult to model under a linkage analysis.

Due to the presence of discrete recombination hotspots in the genome, individual alleles are often inherited with other nearby alleles, a phenomenon that is termed linkage disequilibrium (LD, Fig. 2). Thus, the genotype of any particular SNPs is often statistically correlated with those nearby. This correlation is measured using two metrics – r^2, the correlation between two genetic markers across individuals, and D', a metric inversely related to the number of recombination events that occurred between two markers in a

population[5]. The strength of both metrics drop by distance from the SNP it is calculated from, however it is not uncommon for any particular SNP to be in strong linkage disequilibrium (r^2>0.8) with dozens of nearby variants. This effect of linkage disequilibrium has two practical consequences on disease mapping. First, the correlation structure imposes a substantial reduction in the number of unique "haplotype blocks", groups of SNPs strongly correlated to each other in genotypes across individuals, that are present in the human genome[6] (Fig. 2). Thus, disease mapping can be accomplished by genotyping only a subset of all human genetic variants, and the remaining unobserved genotypes could be computationally inferred by comparison to a pre-existing reference panel of individuals. This has the effect of reducing experimental burden for performing an association study across large cohorts of individuals. Second, the haplotype architecture makes it difficult to identify the specific SNP(s) that are causally acting to influence a specific human disease or trait. The SNP in a particular locus with the lowest p-value for association may not be the causal variant contributing to an association signal, and on average it is statistically unlikely to be the case[7].

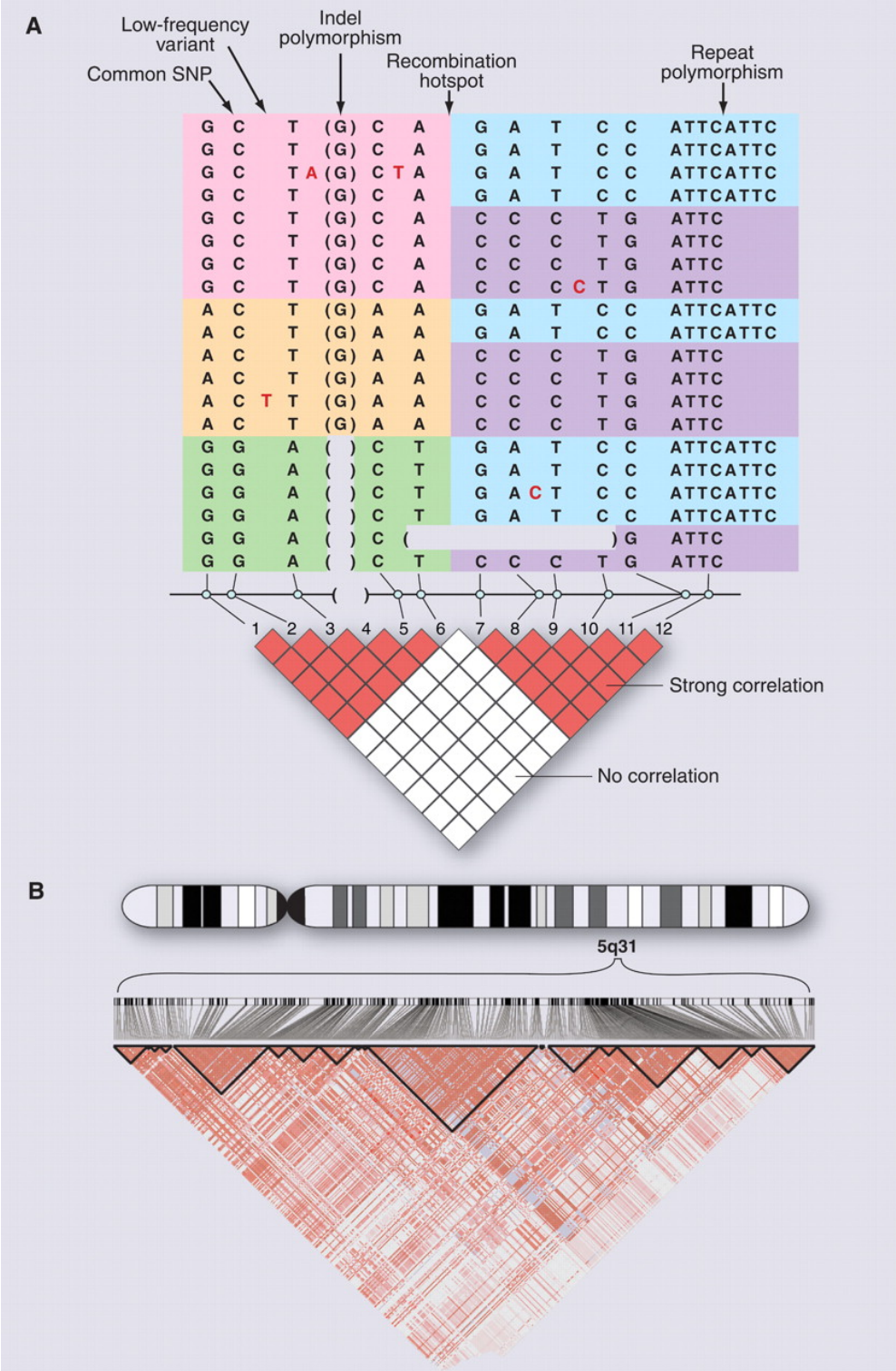


Figure 2: Haplotype structure of the human genome. Adapted from Altshuler *et al.* (2008). Genetic variants, such as SNPs, are densely spaced over an entire genome interval, and the possible allelic combinations that exist across individuals in a population are limited by the positioning of discrete recombination hotspots. Thus, SNP alleles are often correlated with each other across individuals (*red*=stronger correlation, *white*=weaker correlation).

The success of GWAS has led to a greater understanding of the architecture of complex human traits, but also the realization that in contrast to Mendelian disease, arriving at a mechanistic understanding of how complex disease loci influence phenotype will be substantially more difficult. This is for four main reasons: first, as described above there are many candidate causal variant(s) at each locus. Second, unlike in Mendelian disease, the vast majority of loci associated with complex traits reside outside protein-coding regions, complicating the identification of a causal gene, if any, and increasing the number of possible molecular mechanisms through which the locus can act. Third, the collection of current GWAS hits collectively often explain only a small proportion of the total heritability for any disease[8,9]. With few exceptions, GWAS loci have very modest effect sizes. In contrast to Mendelian disease, where there is often a singular large-effect perturbation of a pathway, for any complex trait there appear to be orders of magnitude more loci that each act to subtly perturb many different pathways[2]. Fourth, the observed distribution of modest effect sizes, which may have been caused by evolutionary pressure acting against strongly deleterious variants, combined with a large multiple hypothesis-testing burden necessitates that GWAS be performed with very large cohorts to achieve sufficient statistical power to implicate any individual locus[8] (Fig. 3). These four problems compel the development of computational and experimental tools to "fine map" individual GWAS loci to identify the causal variants and identify therapeutically important genes and pathways involved in complex human disease.

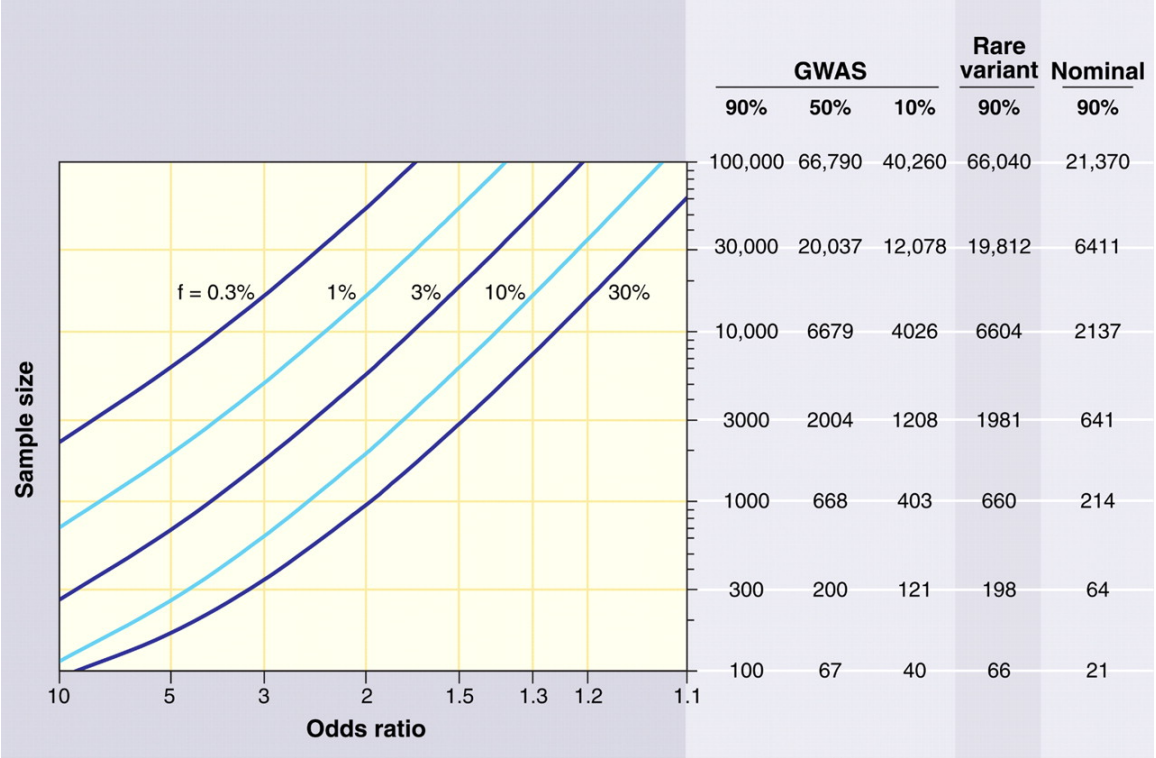


Figure 3: Sample sizes required in genome-wide association studies to detect loci with varying effect size. Adapted from Altshuler *et al.* (2008)[2]. Columns correspond to varying degrees of statistical power to reach genome-wide significance ($p<5x10^{-8}$). Curves correspond to different frequencies (f) of disease-causing allele in the cohort.

Genetic mapping of Mendelian diseases often implicate genes encoding proteins with a direct functional role in a critical cellular process, such as the CFTR gene for cystic fibrosis that acts as a channel for chloride ions in the lung, or the phenylalanine hydroxylase enzyme that when lost leads to an inability to metabolize phenylalaine and the development of phenylketonuria**.** However, genetic mapping of Mendelian diseases occasionally implicates mutations in regulators of gene expression, such as transcription factors or chromatin remodelers, as the culprits for a particular disease, indicating that pathogenic mutations are not confined to only proteins with a direct biochemical role in a disease phenotype but can also target upstream transcriptional regulators of such proteins[10,11]. Remarkably, there is also recent evidence indicating that mutations in the binding sites of transcription factors on DNA can similarly be causal for human disease[12]. As non-coding DNA represents 98% of the human genome and harbors many such transcriptional regulatory elements, we hypothesized that mutations and genetic variation within these non-coding regions that influence gene expression, termed transcriptional

enhancers, could contribute to the pathogenesis of complex diseases as well. In particular, as the majority of variants associated with complex traits identified through GWAS do not appear to affect protein-coding sequence, we believe that a detailed investigation of genomic variation in enhancers is critical for understanding the etiology of complex human traits.

Regulation of gene transcription

Advances in genome research have provided an unprecedented opportunity to investigate the function of transcriptional enhancers – noncoding DNA regulatory regions that control transcription. The recognition that transcriptional enhancer elements contribute to gene activation was first illustrated in the early 1980's based on experiments using SV40 viral enhancers to drive transcription *in vitro*[13]. These distal enhancers have subsequently been shown to function at variable distances from the genes they regulate (up to 1Mb away), and can function in an orientation-independent manner to promote expression of reporter genes both *in vitro* and *in vivo*[14-16]. Enhancer trapping experiments in *Drosophila* provided critical evidence that these non-coding *cis*-regulatory regions mediate the precise spatial and temporal control of gene expression during metazoan development[17]. However, pinpointing the location of distal enhancers in the genome has historically been difficult because these elements reside at variable distances from the genes they regulate and lack conspicuous sequence features similar to those leveraged to annotate the genome for protein-coding genes. In the past decade, high-throughput analyses of the binding of transcription factors (TFs) and chromatin regulators as well as histone modification patterns and DNase I accessible sites have led to the identification of hundreds of thousands of enhancers across a large number of different cell types and tissues in flies, mice, and humans[18-24]. As transcription factors bind to enhancer elements and recruit co-activators and chromatin regulators to facilitate transcription, the experimentally-derived binding sites of these TFs and chromatin remodelers has become a widely used proxy for enhancer identification.

Epigenetic landscapes predict enhancer activity

Large-scale studies using chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) have recently identified hundreds of thousands of distal enhancer elements in the human genome at various tissues and at different development stages. Their discovery has revealed new insights into the mechanistic details of how tissue-specific gene expression patterns are established and maintained during development. Emerging evidence indicates that tissue-specific transcription factors and chromatin regulators coordinate the activation of distal enhancers to ensure robust control of gene expression programs in a cell type-specific manner. The recent ability to identify enhancers on a global scale has provided investigators with new opportunities to dissect how *cis*-regulatory elements control gene expression programs in normal and disease states.

Genome-wide profiling of post-translational histone modifications has similarly become a powerful method for identifying tissue specific regulatory elements (Figure 4). In particular, mono-methylation of histone H3 lysine 4 (H3K4me1) and acetylation of histone H3 lysine 27 (H3K27ac) can identify active enhancer regions in the genome[19,23,24]. H3K4me1 marks various classes of enhancer elements as well as promoters depending on the co-occurrence of H3K4me1 with other histone modifications. Specifically, the combination of H3K4me1 and H3K27ac marks the class of "active enhancers", designated as these elements strongly correlate with transcriptional activation of target genes[19,23,24]. In contrast, the presence of H3K4me1 alone has been used to predict enhancer elements and is generally indicative of a "weak enhancer" that shows weaker but still statistically significant correlation with transcriptional activation of nearby genes. The set of enhancers identified by histone modification patterns is highly unique in each cell type and active enhancers often neighbor genes with cell type specific functions. Collectively, combinations of histone modifications can distinguish enhancer states and can be used to predict non-coding regions of the genome important in transcriptional regulation.

Additional histone modifications have been implicated in the regulation of enhancer regions and may allow for further delineation of different subclasses of enhancer regions. For example, the repressive mark H3K27me3 and the active marks H3K4me3 and H3K4me2, appear to be enriched at subsets of these *cis*-regulatory elements. H3K27me3 in combination with activating modifications has been shown to mark "poised" enhancer regions in embryonic stem cells and implicates Polycomb activity in regulating enhancer states prior to transcriptional activation and also in deactivation of enhancer elements, similar to its role at gene promoters[19,23,25]. Conversely, enrichment of several other marks including H3K9ac and H3K18ac is associated with active enhancer regions[23,25]. In Drosophila, H3K79me3, a mark of active transcription, has also been implicated at active enhancers occupied by RNAPII[26]. Thus, future investigations aimed at determining how combinations of histone modifications distinguish regulatory regions and impact transcription of target genes may reveal additional insights into enhancer biology.

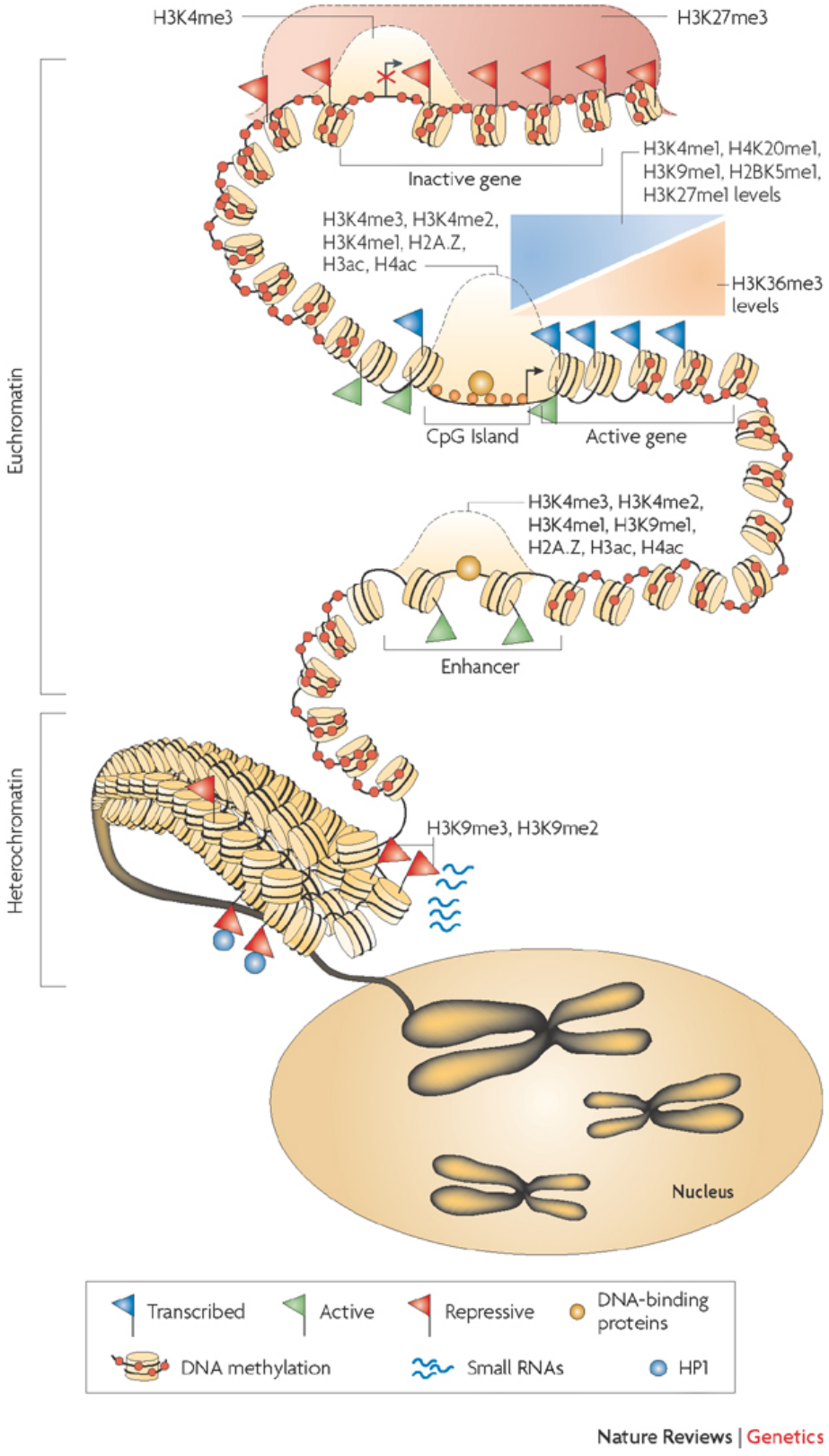


Figure 4: Overview of different histone modifications present at different classes of DNA elements. Adapted from Schones and Zhao (2008)[27].

Combinations of epigenomic marks, including histone modifications, can be integrated in a formal machine learning framework to identify chromatin "states". One approach taken by Ernst and Kellis (2010) uses an unsupervised machine learning approach, the multivariate Hidden Markov Model (HMM), to discover over-represented combinations of epigenomic marks in the genome[23,28] (Fig. 5). Specifically, the multivariate HMM models multiple "observed" input data tracks, such as the genome-wide presence of various histone modifications, as being generated by unobserved "hidden" states that each individually represent a class of genomic element, such as an active enhancer or heterochromatic region. This model, termed ChromHMM, has been applied to dozens of marks, however a core set of 5 and 7 histone modifications are now commonly used to annotate the genome into 15 or 18 informative chromatin states, respectively, encompassing both active and repressive regulatory regions (Fig. 5). This approach was recently expanded to profile histone modifications across 127 different human tissues and cell lines under the Epigenomics Roadmap project and predict both 15 and 18-state ChromHMM models in these regions (Fig. 6, 7).

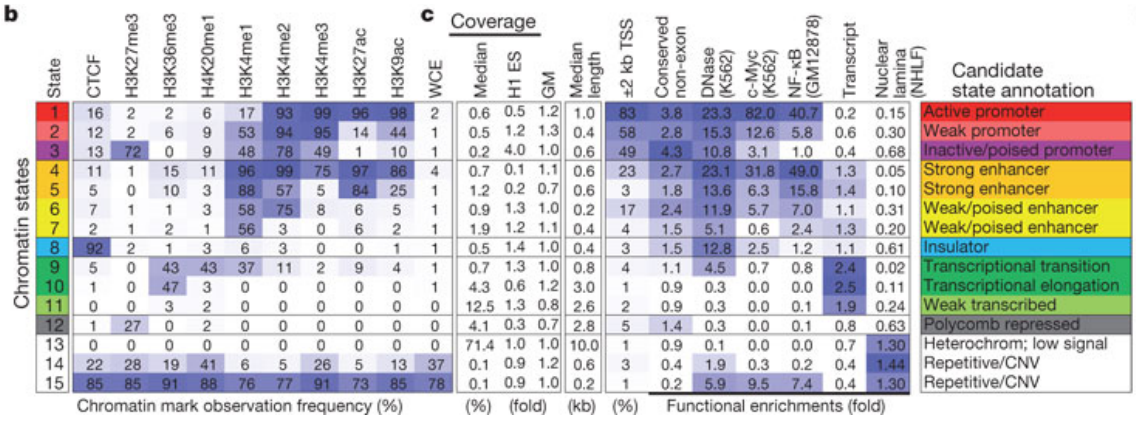


Figure 5: Systematic unbiased annotation of the human genome using epigenomic marks. Adapted from Ernst *et al.* (2011). ChromHMM uses an unbiased machine learning algorithm to scan the genome and identify recurring combinations of histone modifications. These combinations of histone modifications correspond to distinct "chromatin states" that are enriched for a variety of genomic elements, including promoters, enhancers and repressed regions[28].

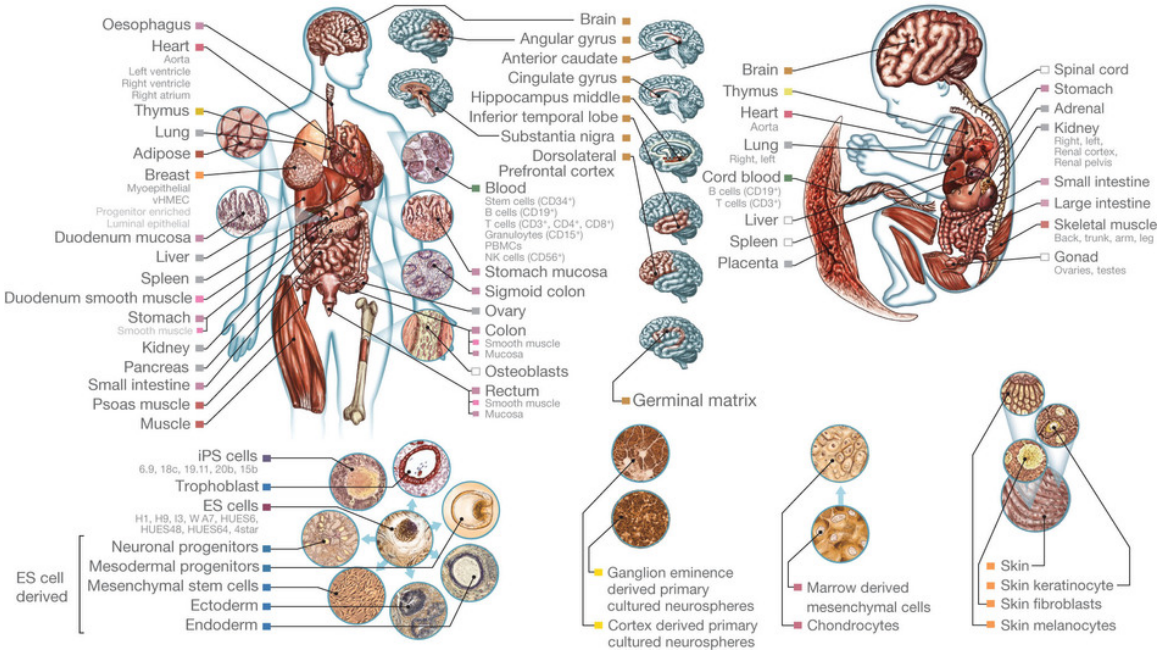


Figure 6: Overview of human tissues and cell lines used profiled by the Roadmap Epigenomics Consortium. Adapted from the Roadmap Epigenomics Consortium (2015)[20].

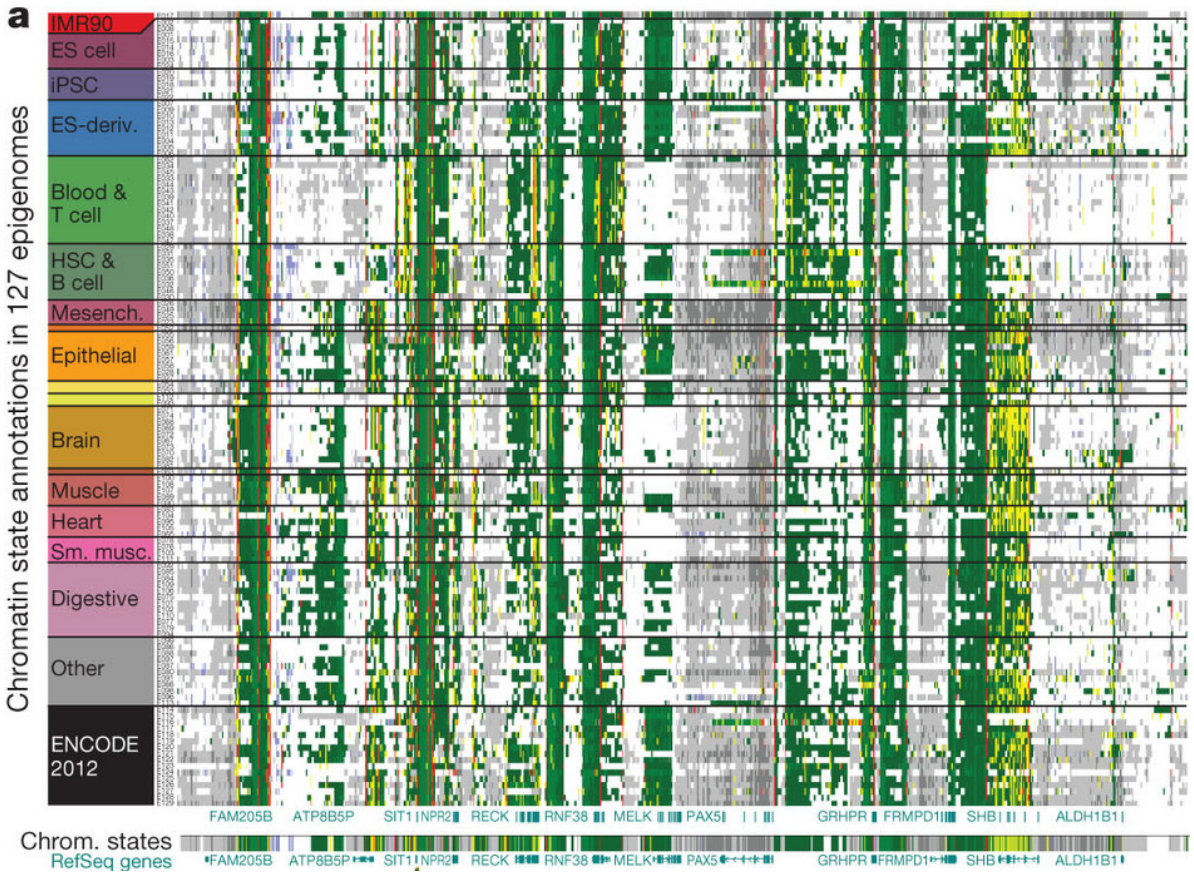


Figure 7: ChromHMM-predicted chromatin states across epigenomics of 127 human tissues and cell lines. Adapted from the Roadmap Epigenomics Consortium (2015)[20]. Coloring of chromatin states correspond to coloring used in Figure 5. Some regions of the genome (e.g. promoters and transcribed regions) maintain the same chromatin state across cell types, while others are highly variable and tissue-specific.

Regulating enhancer activity by modulating DNA accessibility

Access to DNA plays a critical role in the ability of the transcriptional machinery to recognize sequence features at promoter regions and at the numerous distal regulatory elements required for precise control of gene expression. Studies using DNaseI or the Tn5 transposon, both enzymes that preferentially cleave DNA in regions of open chromatin, combined with high-throughput sequencing have proved successful in identifying regulatory regions on a genome-wide level[29-31]. These DNaseI hypersensitive sites (DHSs) or ATAC-seq (Assay for Transposase-Accessible Chromatin with high-throughput sequencing) peaks discretely mark TF binding throughout the genome and can identify a broad set of regulatory elements, including enhancers, silencers, boundary elements and

promoters in a relatively unbiased manner[31]. Thus, mapping open chromatin can capture the shifting regulatory landscape that occurs during developmental transitions.

Exploiting this principle, punctate regions of open chromatin have now been profiled in a wide array of human tissues and cell types. Among the cell types tested by the Roadmap Epigenomics Consortium, DHS signatures could be used to identify the developmental identity of the cell[20]. The importance of DNA accessibility in regulatory regions throughout the genome in a cell type specific manner suggests that certain factors must locate, define, and initiate the process of opening chromatin in these regions. A class of transcription factors, termed "pioneer factors", possess the ability to bind regions of silent chromatin and initiate the cascade of events leading to recruitment of chromatin remodelers and downstream transcription factors[32]. Emerging evidence suggests that pioneer factors bind regulatory regions early in development to facilitate rapid transcriptional responses. The binding of pioneer factors at enhancer elements provides a molecular explanation for the persistence of a subset of DHSs across regulatory regions as cells differentiate. Although the exact mechanism by which pioneer factors alter overall chromatin structure to prime DNA for future regulatory events is not fully understood, these factors clearly play important roles in establishing the early connection between lineage specific transcription factors and the enhancer network.

Association mapping to identify disease loci

Emerging evidence suggests that the perturbation of enhancer activity by disease-associated single nucleotide polymorphisms (SNPs) is a common phenomenon[33-36]. For example, a recent genome-wide study identified enhancers in seven cell types and demonstrated that trait-associated SNPs were specifically enriched in enhancers from biologically relevant cell types[28]. Similar observations were reported by comparing the localization of disease SNPs to DHSs[37]. These two studies suggest that SNPs can contribute to phenotypes by affecting enhancer activity.

Disruption of enhancer activity may occur through a mechanism where the SNP either adds or removes a TF binding site and alters the activity of enhancer and expression of a nearby gene. Indeed, recent targeted studies demonstrated that individual disease-associated loci identified through genome wide association studies (GWAS) could affect TF binding. For example, the minor allele at rs12740374, a GWAS SNP within a non-coding region on chromosome 1p13 that affects plasma LDL cholesterol levels and myocardial infarction risk, resulted in the creation of a novel C/EBP binding site that increased expression of a nearby gene, SORT1, that is involved in controlling VLDL secretion from the liver[33]. Another recent study found that the rs1421085 at the FTO locus in obesity disrupted an ARID5B binding site that controls IRX3 and IRX5 expression, altering the balance between adipocyte thermogenesis and lipid storage[38] (Fig. 8). Collectively, these studies strongly support the idea that non-coding regulatory region such as enhancers are disrupted by complex trait variants.

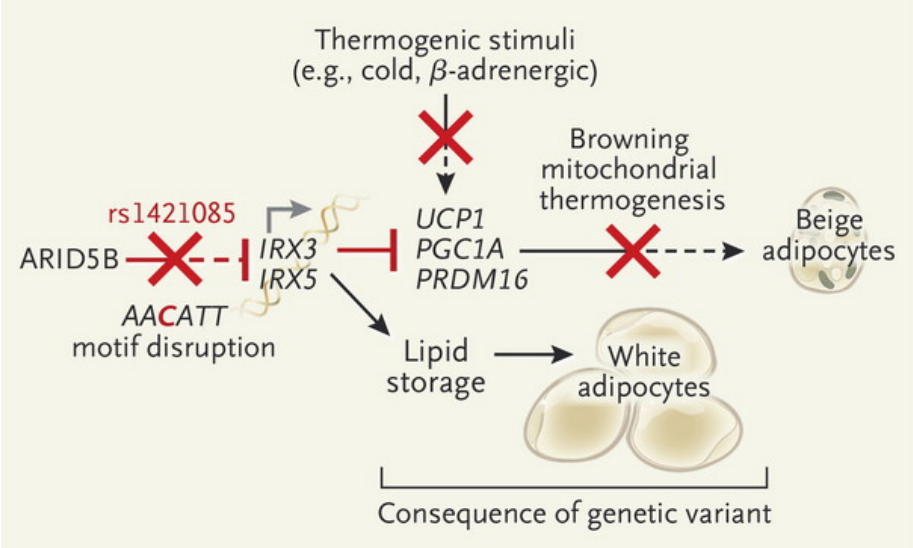


Figure 8: **Overview of FTO/IRX3/IRX5 locus molecular mechanism in obesity**. Adapted from Claussnitzer *et al.* (2015)[38]. A non-coding SNP, rs1421085, in the FTO-obesity locus alters ARID5B binding at an enhancer that regulates IRX3/5 expression and the balance between adipocyte thermogenesis and lipid storage.

The identification of specific causal variants at individual GWAS loci remains another outstanding problem in the human genetics field. While initial studies focused primarily on the SNPs with the most significant p-value at each locus, these "lead" SNPs are often in linkage disequilibrium with dozens of other SNPs and variants[2]. Therefore, at most

GWAS loci, it is statistically unlikely that the lead SNP is responsible for the discovered association signal. Epigenetics may play a critical role in resolving this issue by providing annotations for the regions surrounding SNPs, which help prioritize candidate causal SNPs for experimental validation. Demonstrating that putative causal SNPs can cause molecular or cellular phenotypes will be an important next step. Recent advances in genome engineering, including CRISPR-Cas9-based systems, have the potential to make this step a possibility[39,40]. Thus, strategies using epigenomics to interpret association signals from human genetics can be used to uncover mechanistic insights into the molecular pathways underlying complex human traits and the role of enhancers in this process.

References

1. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat Rev Drug Discov* **12,** 581–594 (2013).

2. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322,** 881–888 (2008).

3. Lander, E. S. & Schork, N. J. Genetic dissection of complex traits. *Science* **265,** 2037–2048 (1994).

4. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106,** 9362–9367 (2009).

5. Raychaudhuri, S. Mapping Rare and Common Causal Alleles for Complex Human Diseases. *Cell* **147,** 57–69 (2011).

6. Gabriel, S. B. The Structure of Haplotype Blocks in the Human Genome. *Science* **296,** 2225–2229 (2002).

7. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* (2014). doi:10.1038/nature13835

8. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases : Article : Nature. *Nature* **461,** 747–753 (2009).

9. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456,** 18–21 (2008).

10. Zaidi, S. *et al.* De novo mutations in histone-modifying genes in congenital heart disease. *Nature* **498,** 220–223 (2013).

11. Li, N. *et al.* Mutations in the Histone Modifier PRDM6 Are Associated with Isolated Nonsyndromic Patent Ductus Arteriosus. *The American Journal of Human Genetics* **98,** 1082–1091 (2016).

12. Smemo, S. *et al.* Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum. Mol. Genet.* **21,** 3255–3263 (2012).

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

14. Bulger, M. & Groudine, M. Functional and Mechanistic Diversity of Distal Transcription Enhancers. *Cell* **144,** 327–339 (2011).

15. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet* **7,** 29–59 (2006).

16. Atchison, M. L. Enhancers: Mechanisms of Action and Cell Specificity. *Annu. Rev. Cell. Biol.* **4,** 127–153 (1988).

17. O'Kane, C. J. & Gehring, W. J. Detection in situ of genomic regulatory elements in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.* **84,** 9123–9127 (1987).

18. Barski, A. *et al.* High-resolution profiling of histone methylations in the human genome. *Cell* **129,** 823–837 (2007).

19. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–283 (2010).

20. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

21. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457,** 854–858 (2009).

22. Nord, A. S. *et al.* Rapid and Pervasive Changes in Genome-wide Enhancer Usage during Mammalian Development. *Cell* **155,** 1521–1531 (2013).

23. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28,** 817–825 (2010).

24. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences* **107,** 21931–21936 (2010).

25. Zentner, G. E., Tesar, P. J. & Scacheri, P. C. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions.

26. Taylor, G. C. A., Eskeland, R., Hekimoglu-Balkan, B., Pradeepa, M. M. & Bickmore, W. A. H4K16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction. *Genome Res.* **23,** 2053–2065 (2013).

27. Schones, D. E. & Zhao, K. Genome-wide approaches to studying chromatin modifications. *Nat Rev Genet* **9,** 179–191 (2008).

28. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).

29. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10,** 1213–1218 (2013).

30. Neph, S. *et al.* Circuitry and Dynamics of Human Transcription Factor Regulatory Networks. *Cell* **150,** 1274–1286 (2012).

31. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).

32. Smale, S. T. Pioneer factors in embryonic stem cells and differentiation. *Curr. Opin. Genet. Dev.* **20,** 519–526 (2010).

33. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466,** 714–719 (2000).

34. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507,** 371–375 (2014).

35. Bezzina, C. R. *et al.* Common variants at SCN5A-SCN10A and HEY2 are associated with Brugada syndrome, a rare disease with high risk of sudden cardiac death. *Nature Genetics* **45,** 1044–1049 (2013).

36. van den Boogaard, M. *et al.* Genetic variation in T-box binding element functionally affects SCN5A/SCN10A enhancer. *J. Clin. Invest.* **122,** 2519–2530 (2012).

37. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337,** 1190–1195 (2012).

38. Claussnitzer, M. *et al.* FTOObesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med* **373,** 895–907 (2015).

39. Sanjana, N. E. *et al.* High-resolution interrogation of functional elements in the noncoding genome. *Science* **353,** 1545–1549 (2016).

40. Ran, F. A. *et al.* Double Nicking by RNA-Guided CRISPR Cas9 for Enhanced Genome Editing Specificity. *Cell* **154,** 1380–1389 (2013).

Chapter 2

Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures

A similar version of this work has been published

Wang, X. *et al.* Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. *eLife* **5,** e10557 (2016).

Abstract

Genetic variants identified by genome-wide association studies explain only a modest proportion of heritability, suggesting that meaningful associations lie "hidden" below current thresholds. Here, we integrate information from association studies with epigenomic maps to demonstrate that enhancers significantly overlap known loci associated with the cardiac QT interval and QRS duration. We apply functional criteria to identify loci associated with QT interval that do not meet genome-wide significance and are missed by existing studies. We demonstrate that these "sub-threshold" signals represent novel loci, and that epigenomic maps are effective at discriminating true biological signals from noise. We experimentally validate the molecular, gene-regulatory, cellular and organismal phenotypes of these sub-threshold loci, demonstrating that most sub-threshold loci have regulatory consequences and that genetic perturbation of nearby genes causes cardiac phenotypes in mouse. Our work provides a general approach for improving the detection of novel loci associated with complex human traits.

Introduction

Genome-wide association studies (GWAS) hold the promise of identifying genetic loci that drive complex disease, however realizing this goal has been challenging due to the modest effect sizes of most common variants that require extremely large cohorts to detect with significance. The recent demonstration that disease-associated single nucleotide polymorphisms (SNPs) reside preferentially in enhancer elements provides a unique opportunity to leverage epigenomic maps of regulatory elements for understanding the function of known GWAS loci and for prioritizing new loci missed in current studies[1-4]. Despite increasingly large GWAS cohort sizes, the current catalog of genome-wide significant loci still explains only a modest proportion of the heritability for any given trait, with an excess of low p-value loci still below the genome-wide significance threshold[5]. These observations suggest that many more signals with "sub-threshold" significance remain to be identified, however, the recognition of biologically relevant sub-threshold loci is hindered by a higher false positive rate[6-8]. Thus, new computational approaches that integrate genetic data with genome-wide epigenomic profiles are needed to use existing cohorts to discover new loci and genes that influence complex traits and diseases.

Here, we use epigenomic maps of 127 tissues from the Roadmap Epigenomics Project as a guide to systematically identify biologically relevant sub-threshold variants[9]. As proof of concept, we focused on two cardiac traits with clinical significance: electrocardiographic QT interval reflecting myocardial repolarization and QRS duration reflecting cardiac conduction. These two traits have a clear tissue of origin and published GWASs have reported over a hundred QT/QRS loci, making these traits ideal for testing variants with sub-threshold significance[6-8]. In particular, variation within QT interval length plays an important role in human disease, where extreme QT prolongation is associated with sudden cardiac death and can occur as an unintended side effect of many non-cardiac medications[10,11]. We combine genome-wide maps of cardiac enhancer activity with the results from a large study of QT interval duration to identify dozens of novel QT loci with sub-threshold statistical significance. We provide multiple lines of evidence to show that

these sub-threshold loci can alter enhancer activity, and we implicate specific genes through which these loci act to influence QT interval length. Importantly, we demonstrate that epigenetic signals can distinguish true biological signals from noise, thus bypassing the higher false positive rate that has previously hindered study of sub-threshold loci. We expect our work will uncover new genes involved in cardiac electrophysiology, aid in the identification of patients at risk for sudden cardiac death, and enable development of new treatments for susceptible individuals. More broadly, our work demonstrates the power of integrating epigenomics with existing GWAS to discover sub-threshold genetic loci and novel genes associated with complex human disease.

Results

QT/QRS-associated variants are enriched in cardiac enhancers

We compiled a list of 112 QT/QRS loci from the NHGRI GWAS database (accessed July 2013) and identified SNPs in strong linkage disequilibrium (r^2>0.8) using genotype data from the 1000 Genomes Project (Phase 1, CEU population)[12]. We also collected GWAS loci from a later meta-analysis of QT interval studies, published in June 2014 by Arking *et al.*, which we held out from the aforementioned 112 QT/QRS loci as a validation dataset for subsequent analyses[5]. While Mendelian disease mutations often alter the function of proteins and are thus found in protein-coding regions of the genome, the majority of common variants (those where the frequency of the minor allele in the population is greater than 5%) linked to common human traits and diseases are reported to lie within non-coding regions of the genome. In line with this, we find that only 22 of 112 loci (20%) associated with QT interval length and QRS duration harbor SNPs that overlap exons. Thus, we hypothesized that many common genetic loci associated with QT/QRS and other traits act on the regulation of gene transcription, rather than on protein function. To test if this is the case, we first computationally examined whether QT/QRS variants are present in predicted transcriptional enhancer elements more often than expected. As specific histone modifications such as H3K27ac and H3K4me1 are [enriched at enhancer elements], we used the combination of histone modifications maps to

predicted cardiac enhancers genome-wide. We used chromatin maps across 127 tissues generated by the Roadmap Epigenomics Project including adult left ventricle (LV), adult right ventricle (RV), fetal heart (FH) and adult right atrium (RA), first focusing on enhancers defined by five histone modifications – H3K4me1, H3K4me3, H3K9ac, H3K36me3, and H3K27me3, chosen to maximize the number of distinct human tissues with available enhancer predictions[9]. QT/QRS variants have greatest overlap with predicted enhancers (as defined by high levels of H3K4me1 and low H3K4me3 using ChromHMM) from the four cardiac tissues compared to the other 123 non-cardiac tissues (red circles, Fig. 1b, Supplementary File 1)[1]. To assess the statistical significance of this overlap, we sampled a background distribution of "control" loci in the genome that have genetic characteristics similar to the 112 QT/QRS loci but do not show statistical association with cardiac electrophysiology. We matched control loci for genetic properties including minor allele frequency, number of SNPs in LD, distance to nearest gene, number of nearby genes, and presence on an Affymetrix 660W genotyping array (Fig. 1a, Methods). Of the 127 tissues, we observed that enhancers from the four cardiac tissues were most enriched for QT/QRS loci compared to the sampled background of control loci. In particular, enhancers from the LV showed the strongest enrichment of any tissue (z-score=7.67, empirical $p<1x10^{-5}$, 10^5 permutations), demonstrating that an unbiased analysis can resolve the causal tissue with high precision, as QT interval and QRS duration are primarily reflective of myocardial repolarization in the ventricles.

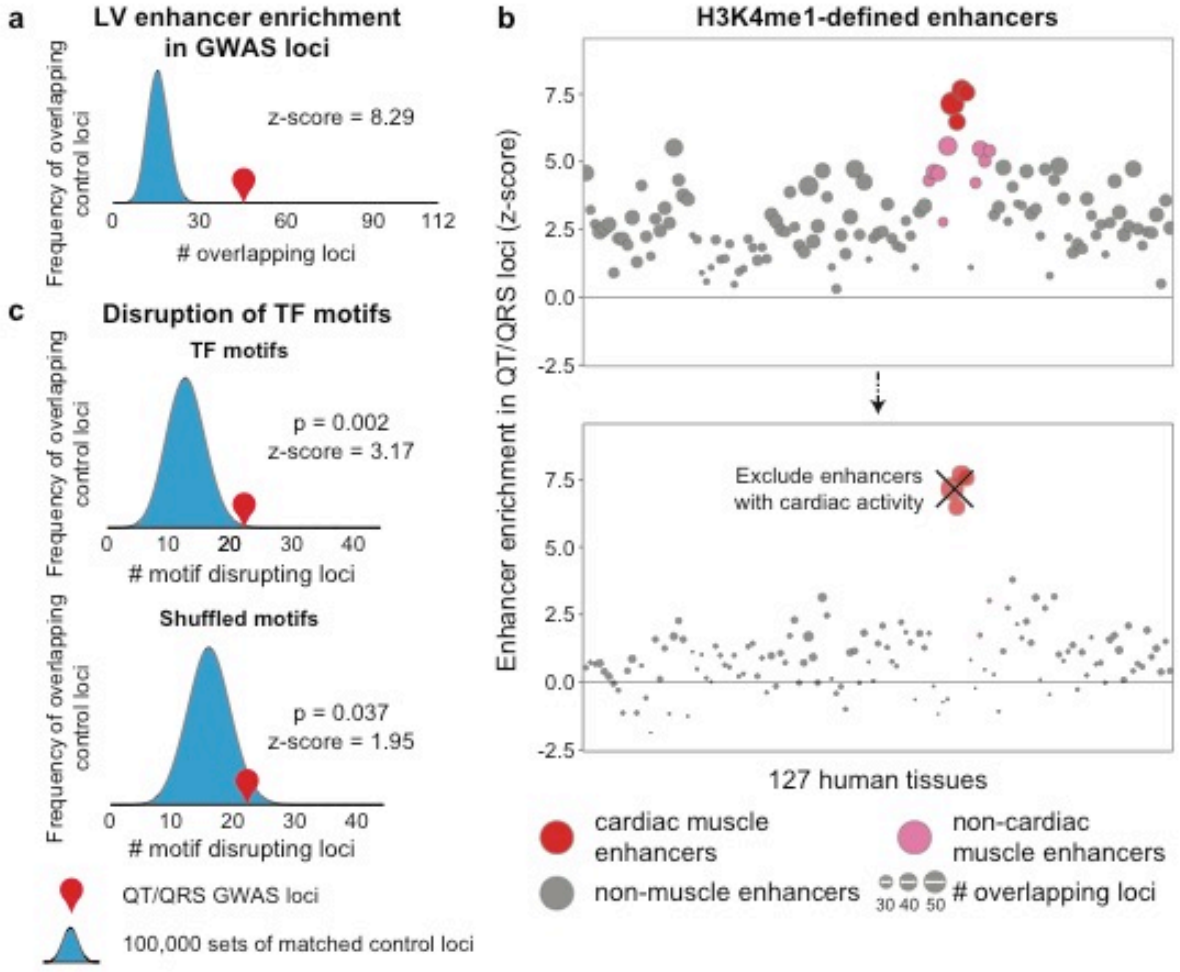


Figure 1. GWAS repolarization loci preferentially overlap cardiac enhancers. a, Enrichment of human left ventricle enhancers in 112 QT/QRS loci. The number of loci that contain a SNP overlapping an enhancer are computed for the 112 QT/QRS loci, and compared against 100,000 permutations of randomly sampled control loci matched for LD block size (number of SNPs), MAF, distance to nearest gene, number of nearby genes, and presence on genotyping array. **b,** *Top,* Enrichment of enhancers from 127 human tissues in QT/QRS loci. *Bottom,* Enrichment of enhancers from non-cardiac tissues for QT/QRS loci is substantially weaker following removal of enhancers active in any of the four cardiac tissues. **c,** *Top,* QT/QRS SNPs are more likely to disrupt motifs corresponding to expressed TFs compared to 100,000 sets of matched control loci. *Bottom*, Weaker enrichment was observed between repolarization and matched control loci when the sequence of the TF motif was randomly shuffled and re-mapped to the genome (10,000 permutations).

Enhancers are the most predictive class of elements for QT/QRS-associated SNPs

Because the left ventricle showed the strongest enrichment for QT/QRS loci, we focused on this tissue to determine the relative enrichment of diverse classes of annotations and to

identify the annotations that are most strongly associated with disease SNPs. We analyzed the enrichment of both coding annotations using GENCODE and non-coding annotations using individual chromatin marks and chromatin states defined by ChromHMM as well as DNase I hypersensitivity (DHS) maps available in heart tissue[1,13,14]. We observed that intergenic enhancers are the most strongly enriched annotated genomic region (z-score > 7.5) in QT/QRS loci, followed by gene transcription regions (z-score between 3 and 6) (Fig. 2 and 3). This enrichment increased significantly (z-score from 7.67 to 9.31 for left ventricle) when restricting the analysis to "strong" enhancers (H3K4me1 enhancers that are also marked by H3K27ac). Our results indicate that predicted enhancers are highly informative for annotating trait-associated variants compared to other classes of genomic regions.

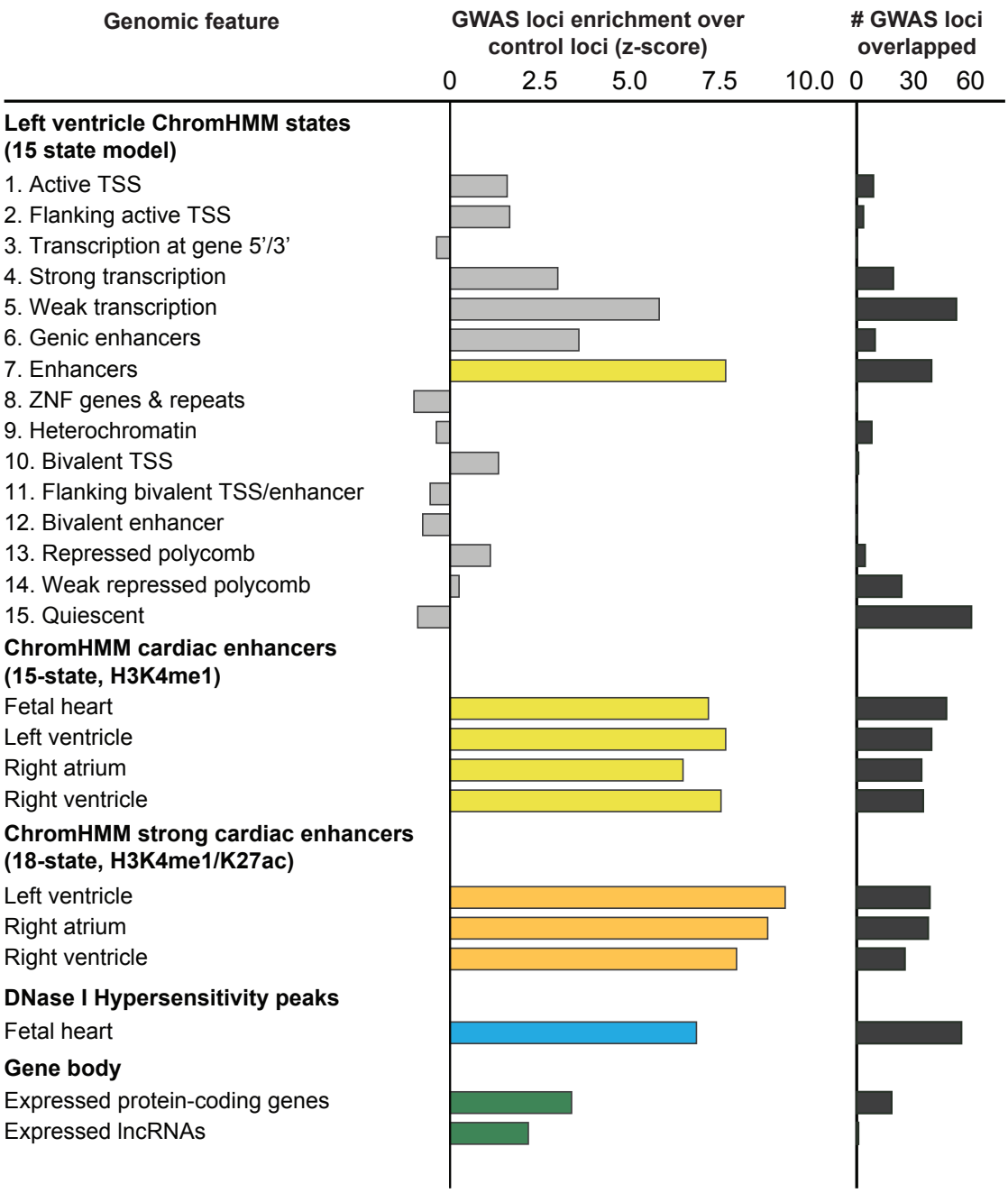


Figure 2: 112 QT/QRS loci overlap enhancers more significantly than other genomic regions in adult left ventricle. Comparison of H3K4me1-enhancers defined by a 15-state model of ChromHMM against other ChromHMM states including protein-coding and non-coding genes and their promoters as well as DNase I hypersensitive (DHS) peaks that broadly mark regulatory regions. The left panel shows the enrichment of features in the 112 GWAS loci compared to randomly sampled control loci; the right panel shows the total number of the 112 GWAS loci overlapped by each feature.

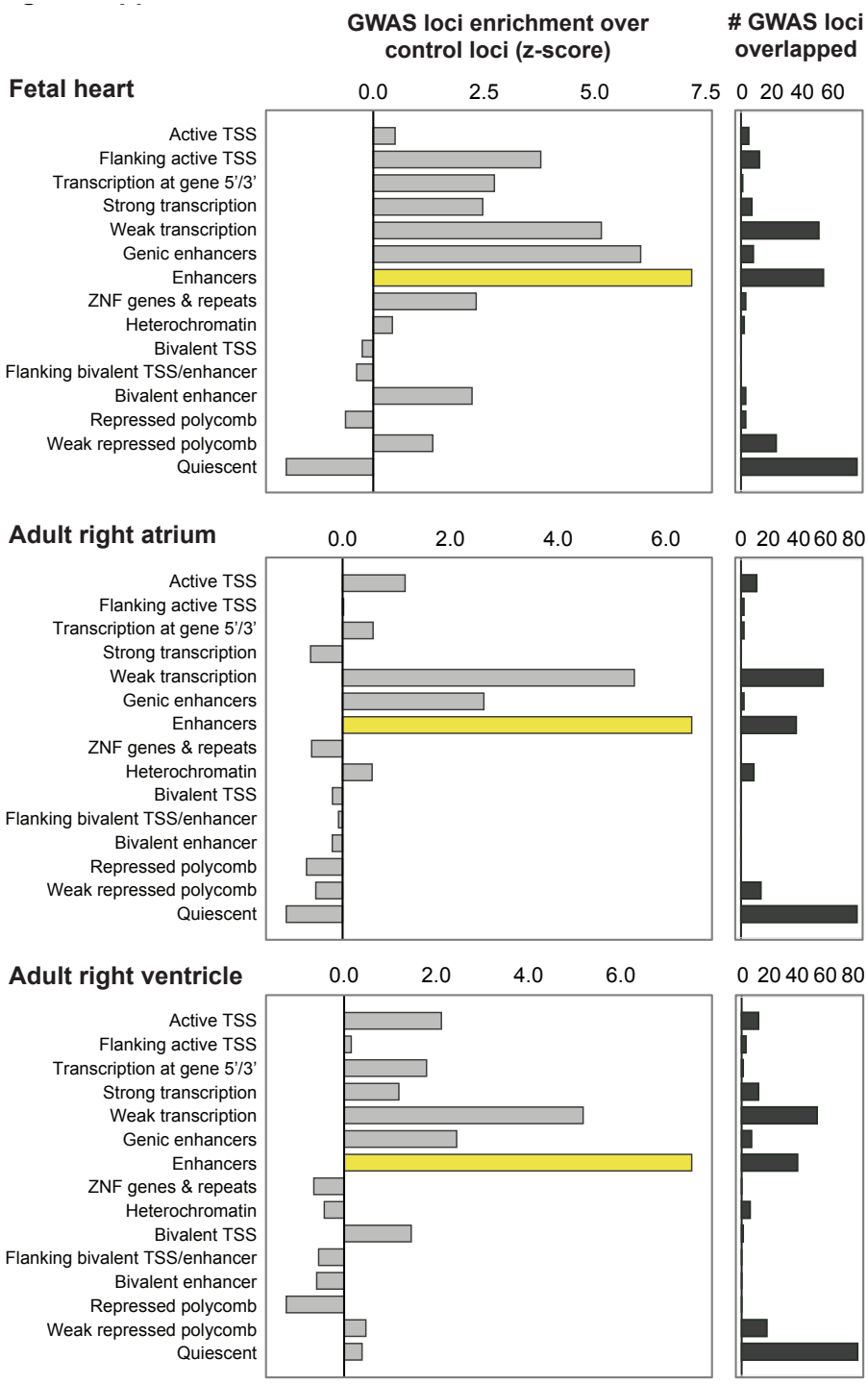


Figure 3: QT/QRS loci overlap enhancers more significantly than other genomic regions in non-LV cardiac tissue. Comparison of enhancers defined by H3K4me1 against other ChromHMM states in fetal heart, adult right ventricle and adult right atrium. The left panel shows the enrichment of features in the 112 GWAS loci compared to randomly sampled control loci; the right panel shows the total number of the 112 GWAS loci overlapped by each feature.

We next asked whether LV enhancers that overlap QT/QRS loci have features that distinguish them from putative LV enhancers identified by ChromHMM that do not overlap QT/QRS loci (Fig. 4). First, we considered the density of H3K27ac marks, as the co-enrichment of H3K4me1 and H3K27ac correlates with strong enhancer activity[15,16]. We found that the 65 enhancers overlapping 45 QT/QRS loci have a 3.1-fold higher H3K27ac density compared to non-GWAS LV enhancers ($p=1.54x10^{-4}$). In fact, incorporating H3K27ac into ChromHMM enhancer predictions resulted in substantially greater enrichment of QT/QRS loci (z-score = 10.10 vs. 8.29 for left ventricle); 44 of the 45 QT/QRS loci overlap an H3K27ac-defined "strong" enhancer. QT/QRS LV enhancers are also more likely to be marked by either H3K4me1 or H3K27ac in at least one of the other three heart tissues (fetal, right atrium, right ventricle) compared to non-GWAS LV enhancers (p-values between 0.008 and 0.07, Fig. 4) and less likely to be active in non-cardiac tissues ($p=9x10^{-3}$, Fig. 4).

Left ventricular QT/QRS enhancers are significantly more hypomethylated than predicted LV enhancers not overlapping QT/QRS loci (hypomethylation $p=1.07x10^{-6}$, hypermethylation p=0.60, Fig. 4). Similar to H3K27ac, CpG hypomethylation correlates with increased enhancer activity, possibly through modulation of TF binding site accessibility[17,18]. Consistent with this idea, 22 of the 45 GWAS loci contain an enhancer SNP that alters a predicted motif for a cardiac-expressed TF (empirical p=0.002, 10^5 permutations) (Fig. 1c). Moreover, QT/QRS GWAS enhancers are enriched for DHS and Cap Analysis Gene Expression (CAGE) signals in human fetal heart, both of which are marks of greater enhancer activity (Fig. 4) [14,19]. Finally, QT/QRS left ventricular enhancers show significant evolutionary conservation across the primate lineage compared to non-GWAS LV enhancers ($p=6.82x10^{-5}$ compared to 10^5 size-matched sets of LV enhancers), suggesting that perturbation of these enhancers is under stronger negative selection. Taken together, QT/QRS loci preferentially overlap conserved enhancers that show cardiac-restricted activity, suggesting that common variants associated with these loci play roles in regulating cardiac functions that drive human phenotypes.

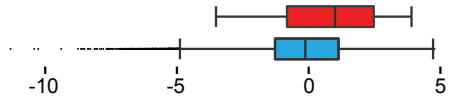
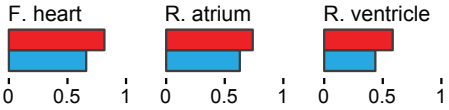
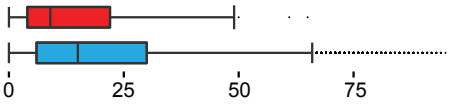
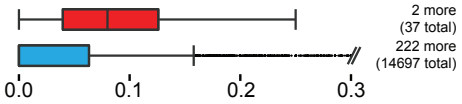
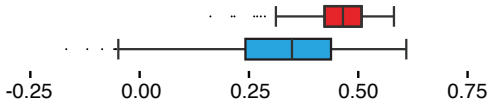
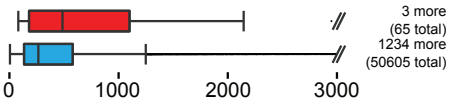
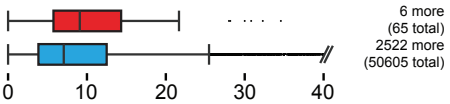
Enhancer characteristic	GWAS vs. non-GWAS LV enhancers	Fold difference	p-value
H3K27ac density log(H3K27ac density)		3.10	1.53×10^{-4}
Activity in cardiac tissues (Proportion overlap)		F. heart 1.24 R. atrium 1.18 R. ventricle 1.34	4.36×10^{-3} 4.11×10^{-2} 1.14×10^{-2}
Activity in non-cardiac tissues (# non-cardiac tissues with activity)		0.60	8.96×10^{-3}
LV-specific hypomethylation (Proportion of LV-specific hypomethylated CpGs)		undef.	9.99×10^{-7}
LV-specific hypermethylation (Proportion of LV-specific hypermethylated CpGs)		NA	0.60
Primate conservation (Average conservation best 100nt window)		1.14	6.82×10^{-5}
DNase I hypersensitivity fetal heart (DNase reads / kb)		1.94	5.17×10^{-4}
CAGE-seq fetal heart (CAGE-seq reads / kb)		1.30	1.43×10^{-3}

Figure 4: Enhancers overlapping QT/QRS loci differ in functional characteristics from all enhancers. Several functional characteristics were compared between enhancers overlapping QT/QRS loci (red) and non-GWAS left ventricle enhancers (blue). Fold change represents fold change between median values for the two groups, and p-values were calculated using the Mann-Whitney U test. See Methods for comparison methodology between GWAS QT/QRS enhancers and non-GWAS enhancers for each functional or epigenomic feature. For primate conservation, LV enhancers (blue) were size-matched (+/-1kb) to GWAS enhancers to control for skewed enrichments driven by larger GWAS enhancer size.

Applicability of enhancer enrichment for studying the genetics of other human traits and diseases

Quantifying the enrichment of GWAS loci in enhancers identified in a diverse panel of tissues has many potential uses within the human genetics field. For traits and diseases driven by known tissues-of-origin, observing enrichment of GWAS loci specifically in

enhancers identified from these tissues can act as an independent measure of validation to ensure that the loci did not reach genome-wide significance due to either technical artifacts from genotyping array usage, or biological cofounders such as population stratification[20]. For diseases where the tissue-of-origin is less clear, calculating enhancer enrichments by tissue can serve as an initial unbiased computational screen for putative biologically relevant tissues for downstream experimental follow-up.

To demonstrate the applicability of computing enhancer enrichments for a variety of complex traits, we first performed similar enrichment analyses for a GWAS of myocardial mass[21]. This GWAS was performed on 73,518 individuals and identified 52 loci associated with any of four electrophysiological traits associated with the QRS complex, a measure of ventricular depolarization. Using epigenomic information from the human adult left ventricle, we first calculated enrichments of overlap between these GWAS loci and six histone modifications that include both activating (e.g. H3K27ac, H3K4me1/3) and repressive modifications (e.g. H3K27me3, H3K9me3). We observed a substantial difference in enrichment of these six modifications: H3K27ac, the activating histone modification enriched at enhancers and promoters showed greatest enrichment for QRS-associated GWAS loci, while the two inactivating marks, H3K9me3 and H3K27me3, showed no enrichment for these GWAS loci, consistent with a role of these loci in promoting activation of gene expression. We also calculated the enrichment of overlap for the 53 voltage-associated GWAS loci and H3K4me1/H3K27ac-predicted enhancers identified in four cardiac developmental time points from an *in vitro* mouse cardiomyocyte differentiation system. We observe a weak enrichment for enhancers from embryonic stem cell (ESC), mesoderm (MES) and cardiac progenitor (CP) cells, but a substantially stronger enrichment for enhancers identified at the cardiomyocyte (CM) stage, consistent with a role for the QRS-associated GWAS loci in altering cardiomyocyte depolarization.

Similar to the enrichment plots presented in Figure 1 for QT/QRS GWAS loci, we also computed enhancer enrichments for genome-wide significant loci identified from a

variety of other complex human traits and diseases including LDL cholesterol levels, type 1 diabetes, platelet count and ulcerative colitis[22-25]. For these four complex traits, we observed that GWAS loci were enriched in specific tissues and cell types that match the known pathology of each trait or disease. For example, LDL cholesterol-associated GWAS loci are enriched in predicted active enhancers from adult liver tissue, while GWAS loci associated with platelet count are enriched in enhancers predicted from hematopoetic stem cells, the progenitor cell type that ultimately gives rise to platelets.

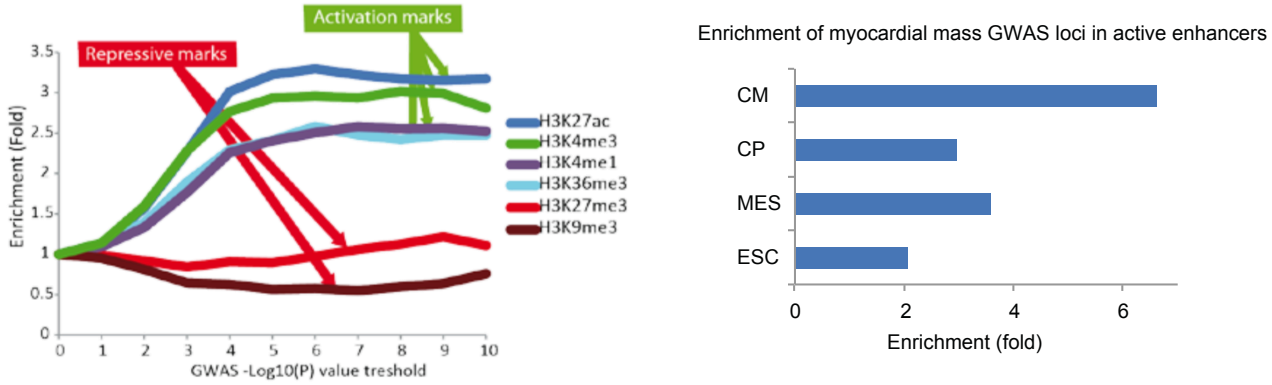


Figure 5: Enrichment of SNPs associated with myocardial mass in activating enhancers from human left ventricle and mouse cardiomyocytes. *Left,* Enrichment of GWAS SNPs associated with QRS complex and myocardial mass in ChIP-seq peaks of histone modifications from the adult human left ventricle. *Right*, Enrichment of GWAS loci associated with the QRS complex and myocardial mass in active enhancers from four *in vitro* developmental timepoints in the mouse cardiomyocyte differentiation system. *ESC – embryonic stem cells, MES – mesoderm cells, CP – cardiac progenitor cells, CM – cardiomyocytes.*

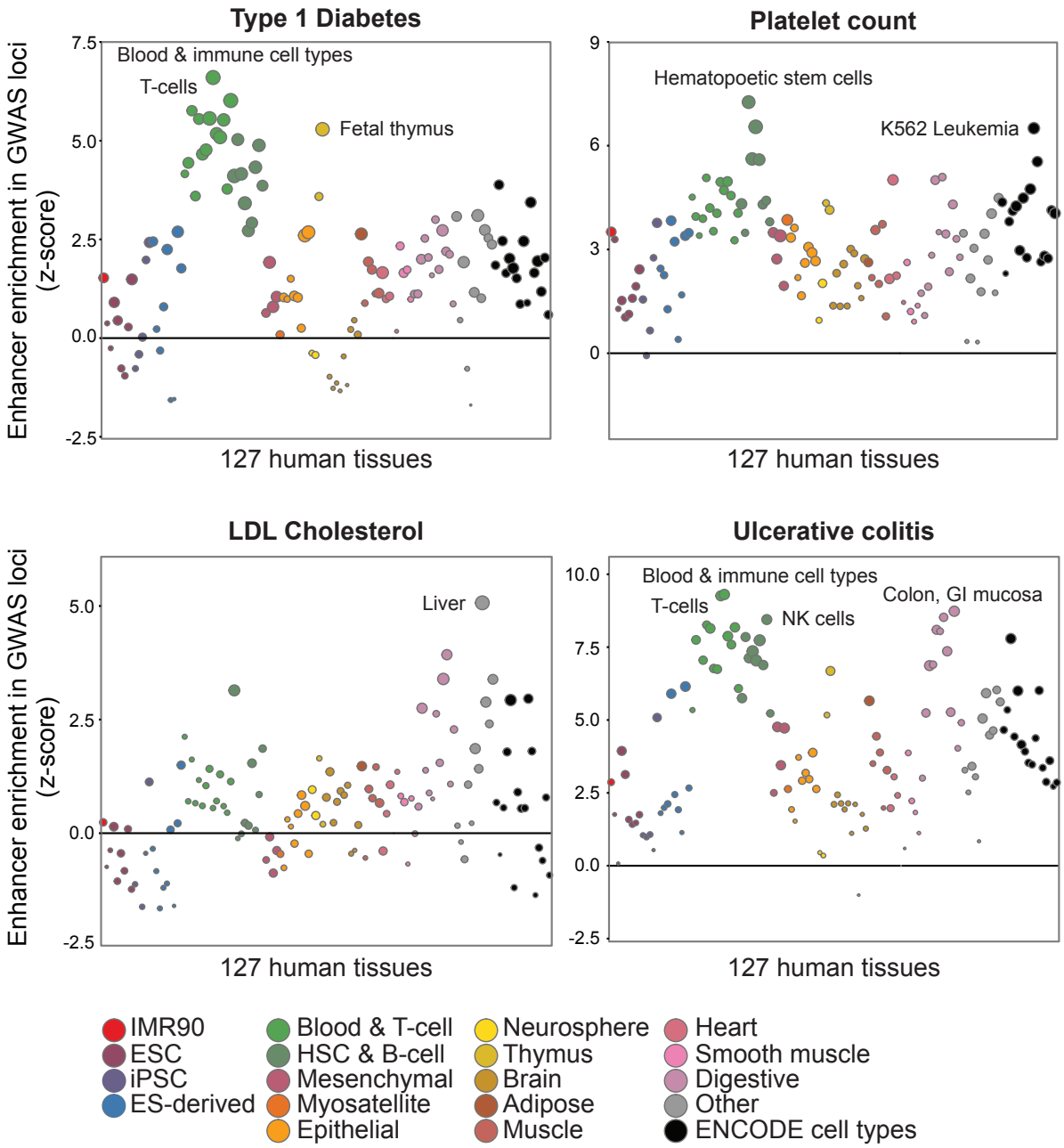


Figure 6: Unbiased enhancer enrichment analysis distinguishes known causal tissue and cell types of origin for a variety of complex human traits and diseases. Enrichment and plots are generated as described in Figure 1 for QT/QRS interval GWAS loci, with size of circles corresponding to number of GWAS loci that overlap enhancers from each tissue.

Common features in GWAS cardiac enhancers identify novel sub-threshold loci

Current GWAS loci collectively explain only a small fraction of the estimated heritability of a complex trait in part due to strict Bonferroni thresholds for multiple hypothesis

testing ($p<5x10^{-8}$) and the limited statistical power of existing studies to discover variants with modest effect sizes[7,26]. We hypothesized that knowledge of the genomic properties associated with existing GWAS loci can guide the search for additional genetic signals that cannot be detected without increasing GWAS cohort sizes, and that these loci with weaker "sub-threshold" p-values (i.e. $0.05>p>5x10^{-8}$) might reveal novel genes and biological pathways that contribute to complex disease. To test this idea, we used SNP summary statistics from the Arking *et al.* (2014) QT interval GWAS study we had earlier held out as a validation dataset[5]. These summary statistics include the 112 QT/QRS loci identified by prior GWASs (red dots, *bottom,* Fig. 7), as well as loci that reach genome-wide significance in the larger meta-analysis cohort but were not discovered in any previous GWAS (and therefore were not included in the 112 QT/QRS loci used for enrichment analyses above, gold dots, *bottom,* Fig. 7). We observed that active LV enhancers are strongly enriched for loci harboring SNPs with p-values between $1x10^{-4}$ and $5x10^{-8}$ (Fig. 7a, black line). Furthermore, the combination of functional features identified for above-threshold QT/QRS enhancers (Fig. 4) substantially improves sub-threshold locus enrichment across a wide range of p-value thresholds (Fig. 7a, colored lines, Fig. 8).

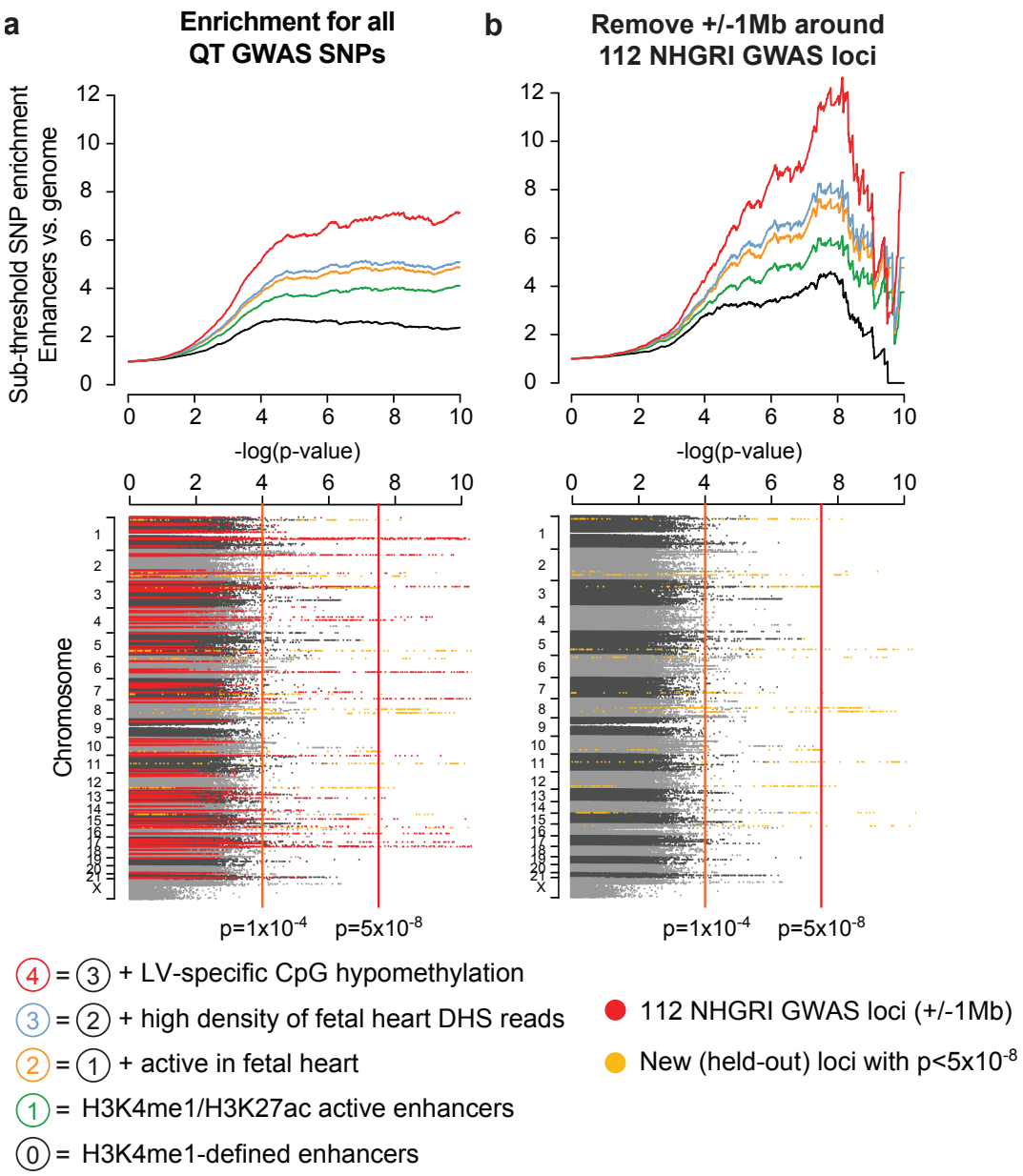


Fig. 7. Cardiac enhancers harbor additional sub-threshold QT loci. a, *Top*, Enhancer characteristics learned on above-threshold QT/QRS loci from Fig. 2 are predictive for additional sub-threshold loci (colored lines). Each point on a curve represents the fold difference in proportion of SNPs with p-value below the cutoff in the enhancer set versus the whole genome. *Bottom*, Manhattan plot of p-values for all SNPs from Arking *et al.* (2014) QT interval GWAS. 112 QT/QRS loci and all SNPs within 1Mb are highlighted in red. Genome-wide significant loci newly discovered by Arking *et al.* and not in the 112 QT/QRS loci are highlighted in yellow. **b,** *Top,* Enrichment signals for sub-threshold SNPs in left ventricle enhancers persists following removal of the 112 GWAS loci and nearby SNPs (+/- 1Mb). *Bottom*, Manhattan plot of p-values for all SNPs from Arking et al. (2014) QT interval GWAS following removal of 112 QT/QRS loci and all SNPs within 1Mb. Genome-wide significant loci newly discovered by Arking *et al.* and not in the 112 QT/QRS loci are highlighted in yellow.

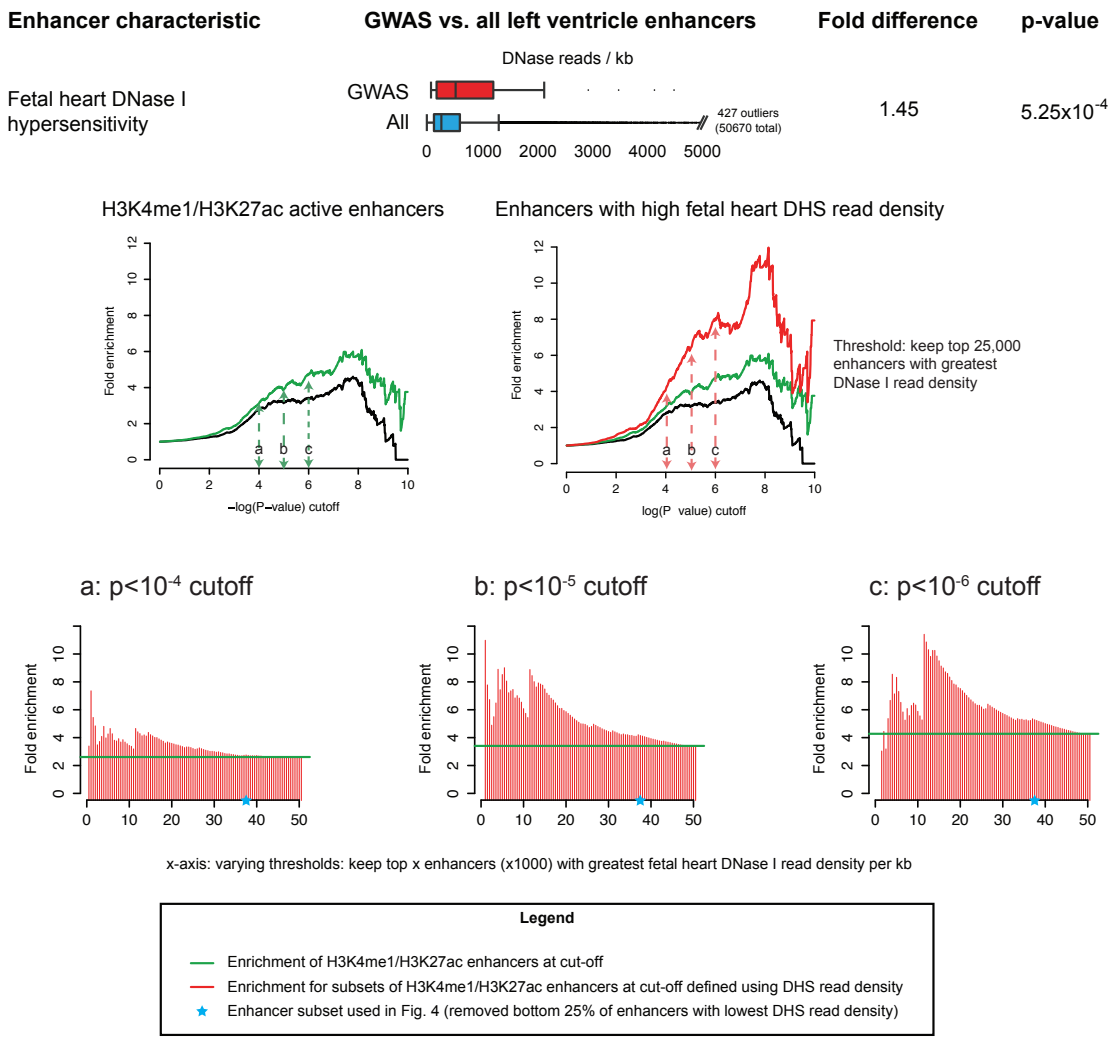


Figure 8: High density of fetal heart DNase I hypersensitivity reads in LV enhancers is robustly informative for identifying enriched sets of sub-threshold loci. *Top:* Enrichment of DHS reads in GWAS enhancers. *Middle:* Example comparison of sub-threshold locus enrichment in active LV enhancers vs. active LV enhancers with high DHS read density. *Bottom:* Y-axis of graphs corresponds to fold enrichment of sub-threshold loci in enhancers taken at three p-value cutoffs (10^{-4}, 10^{-5} and 10^{-6}). X-axis represents enrichments plotted for different subsets of enhancers chosen by varying DHS read density cutoffs.

Whether the enrichment of SNPs in the sub-threshold significance range represents linkage disequilibrium with existing above-threshold GWAS SNPs or novel biologically relevant loci remains an unresolved question[3]. In fact, an enrichment analysis using only SNPs nearby above-threshold GWAS loci produced a strong enrichment signature in the sub-threshold significance range (Fig. 9). To distinguish between the two possibilities, we

took a conservative approach and removed all SNPs within 1Mb of the initial 112 QT/QRS loci. Remarkably, the enrichment for LV enhancers persists and increases in the sub-threshold range (i.e. $p=1x10^{-4}$ to $5x10^{-8}$, Fig. 7b), likely due to removal of nominally significant SNPs that are in LD with above-threshold QT/QRS loci and do not represent true association signals. In total, we identified 2075 SNPs with $p<1x10^{-4}$ that are independent of the 112 published QT/QRS loci, of which 208 SNPs overlap LV enhancers.

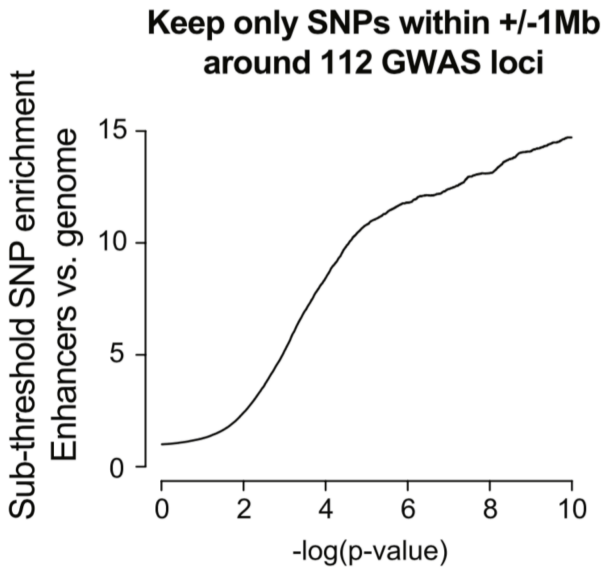


Figure 9: Enrichment in the sub-threshold significance range can be observed using only SNPs nearby known above-threshold loci. The foreground consisted of only SNPs within +/- 1Mb of the 112 QT/QRS loci, and was compared against a background of all SNPs in the genome. Enrichment analyses were performed as described for Figure 7.

Epigenomic prioritization identifies sub-threshold loci with molecular functions

Because the enrichment of sub-threshold SNPs in cardiac enhancers suggests that epigenetic prioritization can be used as a starting point for more in-depth investigations of sub-threshold signals from GWAS, we sought to directly test the molecular hypothesis that these sub-threshold loci impact the transcriptional regulation of cardiac genes (Fig. 10a). We grouped all 2075 sub-threshold SNPs using linkage disequilibrium data (minimum $r^2=0.2$) to identify 287 independent sub-threshold loci in the genome

(Methods). We prioritized loci where a sub-threshold SNP overlapped an active LV enhancer and either (i) also overlapped a fetal heart DNase I hypersensitivity peak or (ii) was an expression quantitative trait locus (eQTL) for a nearby gene. In total, we cloned allele-specific enhancer fragments from 22 cardiac enhancers that overlap SNPs from 18 independent sub-threshold loci, and performed quantitative luciferase assays in human iPSC-derived cardiomyocytes to determine whether the sub-threshold SNP genotypes influence enhancer activity (Methods). We observed that 13 of 18 sub-threshold loci (72.2%) contain an enhancer that drives luciferase activity in an allele-specific manner (Fig. 10b,d, Fig. 12, Fig. S1). Moreover, we estimate that between 51.1%-89.8% (95% Bayesian confidence interval) of prioritized sub-threshold loci show allele-specific activity on transcription, suggesting that the majority of sub-threshold loci identified by epigenomic prioritization do in fact have an impact on transcriptional enhancer activity.

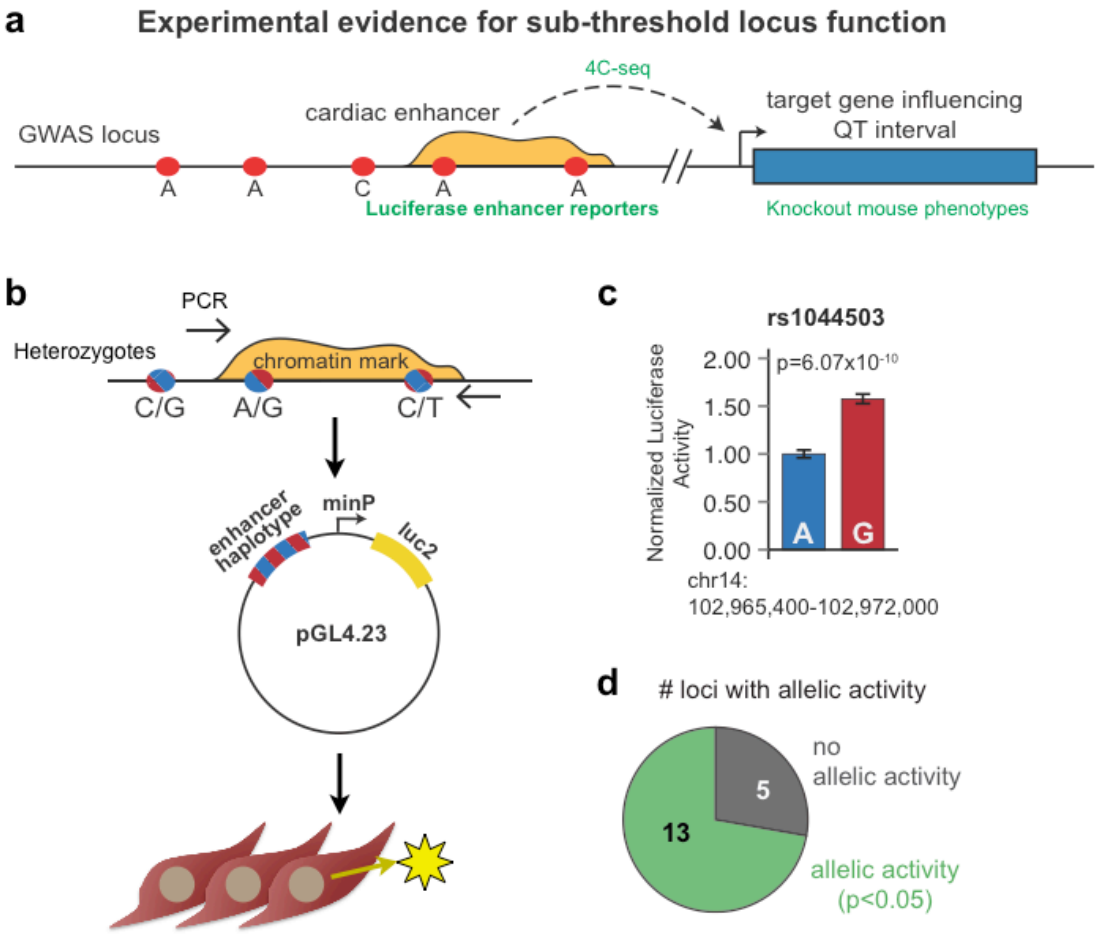


Figure 10: Sub-threshold loci prioritized by epigenomics alter enhancer activity. a, Model detailing how sub-threshold SNPs overlapping enhancers can affect QT interval. *Green text*: *methods used to test mechanistic step in model.* **b,** Summary of luciferase enhancer reporter experiments, cloning steps and luciferase enhancer reporter construct. **c,** Example luciferase assay on two haplotypes of containing the rs1044503 SNP shows allele-specific enhancer activity in human cardiomyocytes, n=24 per haplotype. Error bars represent standard error of the mean. **d,** Number of sub-threshold loci tested that exhibit significant allelic activity ($p<0.05$ between two haplotypes).

We also performed chromosome conformation capture combined with high-throughput sequencing (4C-seq) to experimentally test whether predicted enhancers in sub-threshold loci can form contacts with promoters, and to identify potential target genes of sub-threshold enhancers. We used 4C-seq to test ten predicted enhancers from eight sub-threshold loci in human iPSC-derived cardiomyocytes[27]. Eight enhancers in six loci

formed enhancer-promoter interactions in the proximal 500kb region (Fig. 11c, Fig. 12, Fig. S2). This analysis provides evidence that the novel QT loci enhancers have regulatory activity and that the sub-threshold SNPs identified in our analyses can alter the activity of cardiac enhancers.

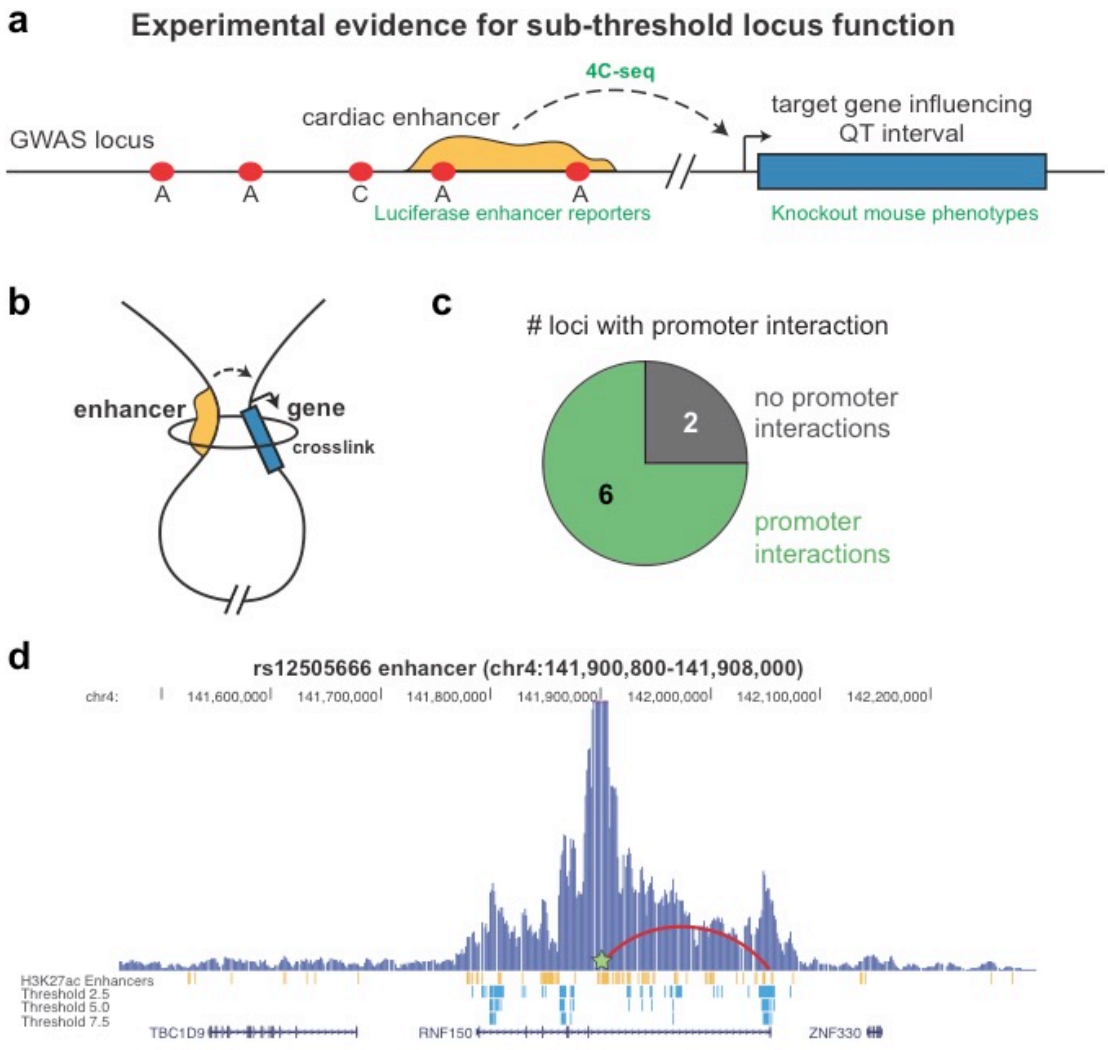


Figure 11: Predicted enhancers in sub-threshold loci form 3D interactions with nearby genes. a, Model detailing how sub-threshold SNPs overlapping enhancers can affect QT interval. *Green text*: *methods used to test mechanistic step in model.* **b,** Summary of luciferase enhancer reporter experiments, cloning steps and luciferase enhancer reporter construct. **c,** Example luciferase assay on two haplotypes of containing the rs1044503 SNP shows allele-specific enhancer activity in human cardiomyocytes, n=24 per haplotype. Error bars represent standard error of the mean. **d,** Number of sub-threshold loci tested that exhibit significant allelic activity ($p<0.05$ between two haplotypes).

Lead SNP tested	p-value,	fold below GWS	Enhancer chr,start,length			Luciferase reporter	Enhancer-promoter interactions
rs1044503	$5.13x10^{-7}$	10.3	chr14	102.96Mb	6.6kb	$6.07x10^{-10}$	CINP, RCOR1
rs10030238	$6.21x10^{-7}$	12.4	chr4	141.80Mb	1.8kb	$9.72x10^{-16}$	RNF150
			chr4	141.90Mb	7.2kb	Not tested	RNF150
rs4683418	$2.62x10^{-6}$	52.4	chr3	141.74Mb	1.4kb	$1.10x10^{-3}$	Not tested
rs2305054	$3.39x10^{-6}$	67.8	chr2	220.50Mb	5.4kb	0.064	Not tested
rs1275988	$4.22x10^{-6}$	84.4	chr2	26.91Mb	1.6kb	0.015	Not tested
rs1451509	$4.32x10^{-6}$	86.4	chr17	57.43Mb	27.0kb	$2.00x10^{-3}$	Not tested
rs4904569	$1.36x10^{-5}$	272	chr14	89.86Mb	0.8kb	$6.33x10^{-3}$	Not tested
			chr14	89.87Mb	2.0kb	$9.04x10^{-8}$	Not tested
rs6565060	$1.52x10^{-5}$	304	chr16	82.74Mb	4.4kb	$5.00x10^{-3}$	No promoters
rs3772570	$1.73x10^{-5}$	346	chr3	148.73Mb	5.4kb	0.40	Not tested
rs3734637	$2.23x10^{-5}$	446	chr6	126.08Mb	0.6kb	$1.96x10^{-4}$	HDDC2
rs10089107	$2.72x10^{-5}$	544	chr8	141.69Mb	2.0kb	0.29	Not tested
rs196067	$4.99x10^{-5}$	998	chr22	38.85Mb	1.6kb	0.016	Not tested
rs11119843	$5.05x10^{-5}$	1010	chr1	212.24Mb	1.0kb	0.29	Not tested
			chr1	212.26Mb	2.4kb	$2.68x10^{-3}$	Not tested
rs1743292	$6.48x10^{-5}$	1296	chr6	105.70Mb	3.6kb	$2.99x10^{-12}$	BVES, POPDC3
			chr6	105.72Mb	2.8kb	$1.79x10^{-5}$	BVES, POPDC3
rs11263841	$6.87x10^{-5}$	1374	chr1	35.30Mb	4.6kb	0.22	GJA4, DLGAP3
rs12613503	$7.37x10^{-5}$	1474	chr2	11.55Mb	3.4kb (two fragments tested)	0.54	ROCK2
						$9.75x10^{-8}$	
rs17779853	$7.73x10^{-5}$	1546	chr17	30.06Mb	3.0kb	$4.33x10^{-3}$	No promoters
rs7783216	$9.95x10^{-5}$	1990	chr7	103.25Mb	0.4kb	0.35	Not tested

Figure 12: Experimental evidence that sub-threshold SNPs alter enhancer activity and that sub-threshold enhancers interact with gene promoters. *Fold below GWS column* represents degree to which sub-threshold locus is below genome-wide significance ($5x10^{-8}$); *Luciferase reporter column* colored green if significant allelic difference in activity ($p<0.05$, Fig. S1); *Enhancer-promoter interactions column* colored green if there is a detectable enhancer-promoter interaction by 4C-seq (Fig. S2).

Epigenomic prioritization discriminates sub-threshold loci with stronger association statistics

We next tested whether epigenomic prioritization can distinguish statistically relevant sub-threshold loci by comparing the association statistics of sub-threshold loci that do or

do not overlap cardiac enhancers. From the 287 independent sub-threshold loci in the genome, we selected two subsets to compare: 60 loci that contain sub-threshold SNPs directly overlapping predicted active LV enhancers, and as a negative control, 129 sub-threshold loci that do not any SNPs (r^2>0.2) overlapping a cardiac enhancer.

Evidence from genome-wide association studies

We reasoned that if sub-threshold loci that overlap active cardiac enhancers represent true biological signals, they should have stronger GWAS association signals than the negative control set. We present three lines of evidence supporting this hypothesis below (Fig. 13a):

a) The 60 enhancer-overlapping sub-threshold loci have significantly stronger p-values than the 129 negative control loci, despite the application of the same $p=1x10^{-4}$ threshold for both sets ($p=1.95x10^{-5}$, *left*, Fig. 13a).
b) 9 of the 60 enhancer-overlapping sub-threshold loci are among the loci that reach genome-wide significance in the larger held out meta-analysis cohort (and not included in the 112 QT/QRS loci used for enrichment analyses in Fig. 1), compared to only 3 of the 129 sub-threshold loci that do not overlap enhancers (6.45-fold enrichment, $p=1.92x10^{-3}$, *middle*, Fig. 13a).
c) The 60 enhancer-overlapping sub-threshold loci are more likely to reach nominal significance ($p<0.05$) in a related GWAS study of QRS duration (see Methods for individuals shared between both studies)[28]. In the QRS duration GWAS, p-values are available for 56 of 60 enhancer-overlapping sub-threshold QT loci and 110 of 129 negative control sub-threshold loci. 31 of 56 (55.4%) enhancer-overlapping sub-threshold loci are nominally significant in the QRS GWAS, a rate 2.9-fold higher than the 129 negative control loci (21 of 110 loci, $p=3.28x10^{-6}$, *right*, Fig. 13a), suggesting that epigenetic prioritization is more likely to identify sub-threshold SNPs that replicate in subsequent GWASs.

Taken together, these analyses demonstrate that genome-wide maps of predicted enhancers can facilitate the detection of true sub-threshold loci.

Evidence from organismal phenotypes

Our identification of a high-confidence set of sub-threshold loci based on epigenomic signals provides a unique opportunity to discover new genes that contribute to cardiac electrophysiological traits. As enhancers can regulate genes up to 1Mb away, it is difficult to identify targets using a simple nearest gene approach[29]. To circumvent this limitation, we developed a computational enhancer-gene linking method that prioritizes gene targets based on correlated activity patterns between enhancer-gene pairs across 59 human tissues (Methods). Using this approach, we identified 106 candidate genes predicted to be regulated by the 60 enhancer-overlapping sub-threshold loci. Notably, 11 of the 15 observed 4C-seq interactions were predicted by our computational approach, compared to 3 of 15 by the commonly applied approach of assigning the enhancer target to the nearest gene.

We used the output of the enhancer-gene linking method to test whether these candidate genes have roles in QT interval. To this end, we studied mouse phenotypes for directed knockouts and genetic perturbations of the 106 predicted gene targets of sub-threshold enhancers. We identified 49 of the 106 genes where mouse mutant models were available with documented phenotypes[30]. Genetic perturbation in 11 of the 49 genes resulted in altered cardiac conduction or cardiac contractility: both processes that are also influenced by genes nearby above-threshold QT interval loci and genes implicated in the Mendelian Long QT syndrome. This represents a 4.11- fold enrichment compared to genes linked to all active LV enhancers (181 of 3311, $p=6.84x10^{-5}$, black bar, Fig. 13b). In contrast, phenotypes arising from genetic perturbation of LV-expressed genes nearby the 129 negative control sub-threshold loci outside enhancers are 7.30-fold less likely to result in altered cardiac conduction or contractility compared to our 60 prioritized sub-threshold loci ($p=1.92x10^{-3}$, perturbation of 2 of 65 genes nearby the negative control subset have relevant cardiac phenotypes, grey bar, Fig. 13b).

The study of biologically relevant sub-threshold loci has been hampered by a high false positive rate that makes the detailed investigation of any sub-threshold locus experimentally more difficult and less attractive than above-threshold loci. The data presented here provide multiple independent lines of evidence that epigenomic signatures can be used to prioritize sub-threshold GWAS loci with a significantly greater likelihood of being biologically relevant.

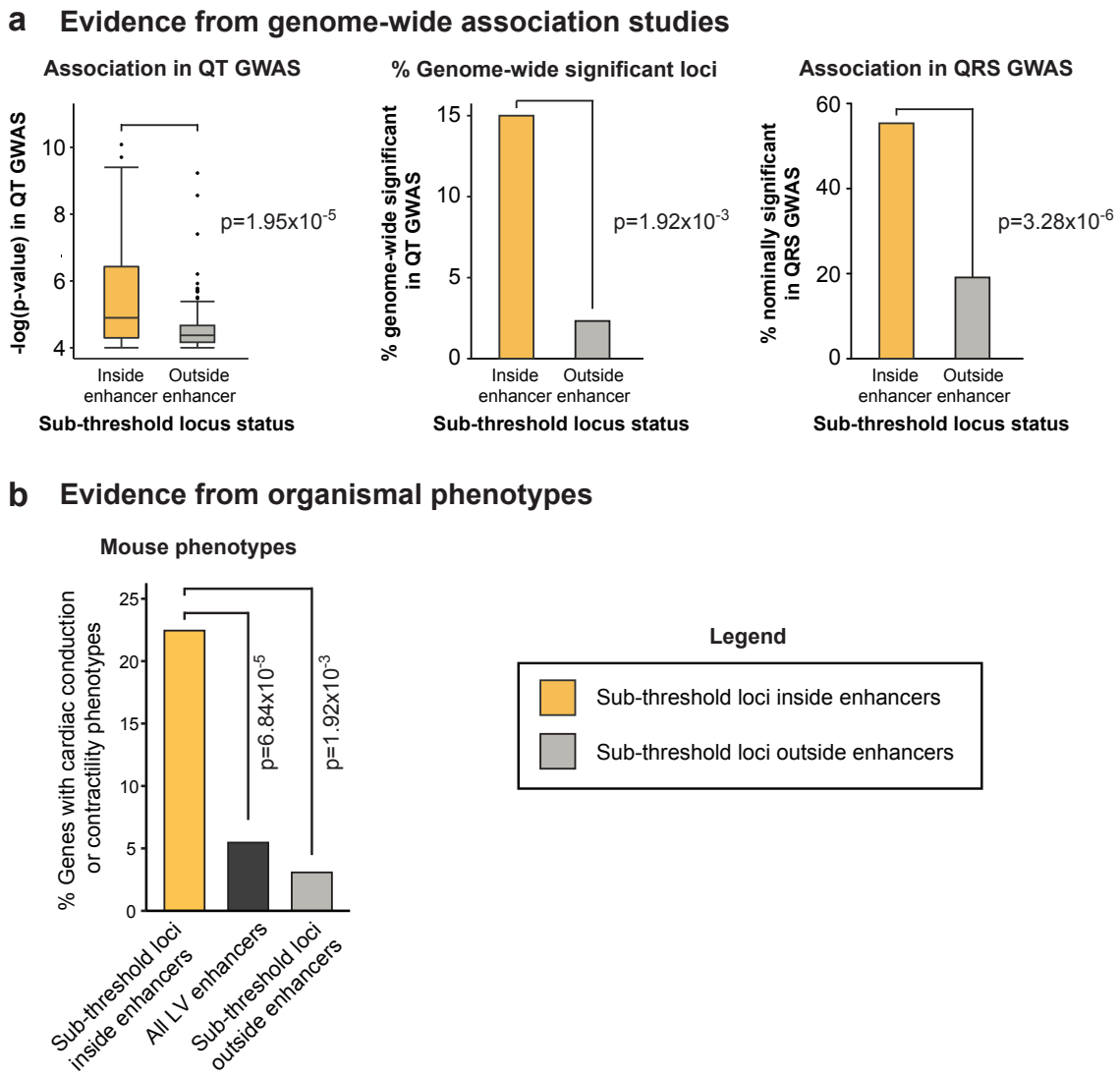


Figure 13: Epigenomic prioritization distinguishes biologically relevant sub-threshold loci
a, *Left,* Sub-threshold loci overlapping enhancers have significantly stronger association signals than loci outside enhancers in the QT interval GWAS. *Middle*, Loci overlapping enhancers have significantly more likely to be newly genome-wide significant in the held-out QT interval GWAS, than loci outside enhancers. *Right,* Sub-threshold loci overlapping enhancers are significantly more likely to be nominally significant ($p<0.05$) in QRS GWAS than sub-threshold loci not overlapping enhancers. **b,** Genetic perturbation of genes with predicted links to 60 enhancer-overlapping sub-threshold loci are significantly more likely to result cardiac conduction or contractility phenotypes than genes linked to all LV enhancers and genes nearby non-enhancer overlapping sub-threshold loci.

Enrichment of sub-threshold loci for non-cardiac human diseases in enhancer elements

We also assessed the broad applicability of using enhancer annotations to prioritize biologically promising sub-threshold variants. We therefore expanded our sub-threshold locus analyses to also two complex human diseases for which summary-level p-value information was available from a large cohort GWAS study: LDL cholesterol and Alzheimer's disease. Both LDL cholesterol and Alzheimer's disease show an enrichment pattern for sub-threshold SNPs in enhancer elements similar in that observed for QT interval sub-threshold SNPs (Fig. 14). Notably, this enrichment pattern persists after removal of genome-wide significant loci previously identified by the published LDL cholesterol and Alzheimer's disease studies. This indicates that sub-threshold loci associated with a wide range of complex human traits can be prioritized using epigenomic signatures, and that the results presented above for QT interval are not restricted to genetic architecture of cardiac electrophysiological traits.

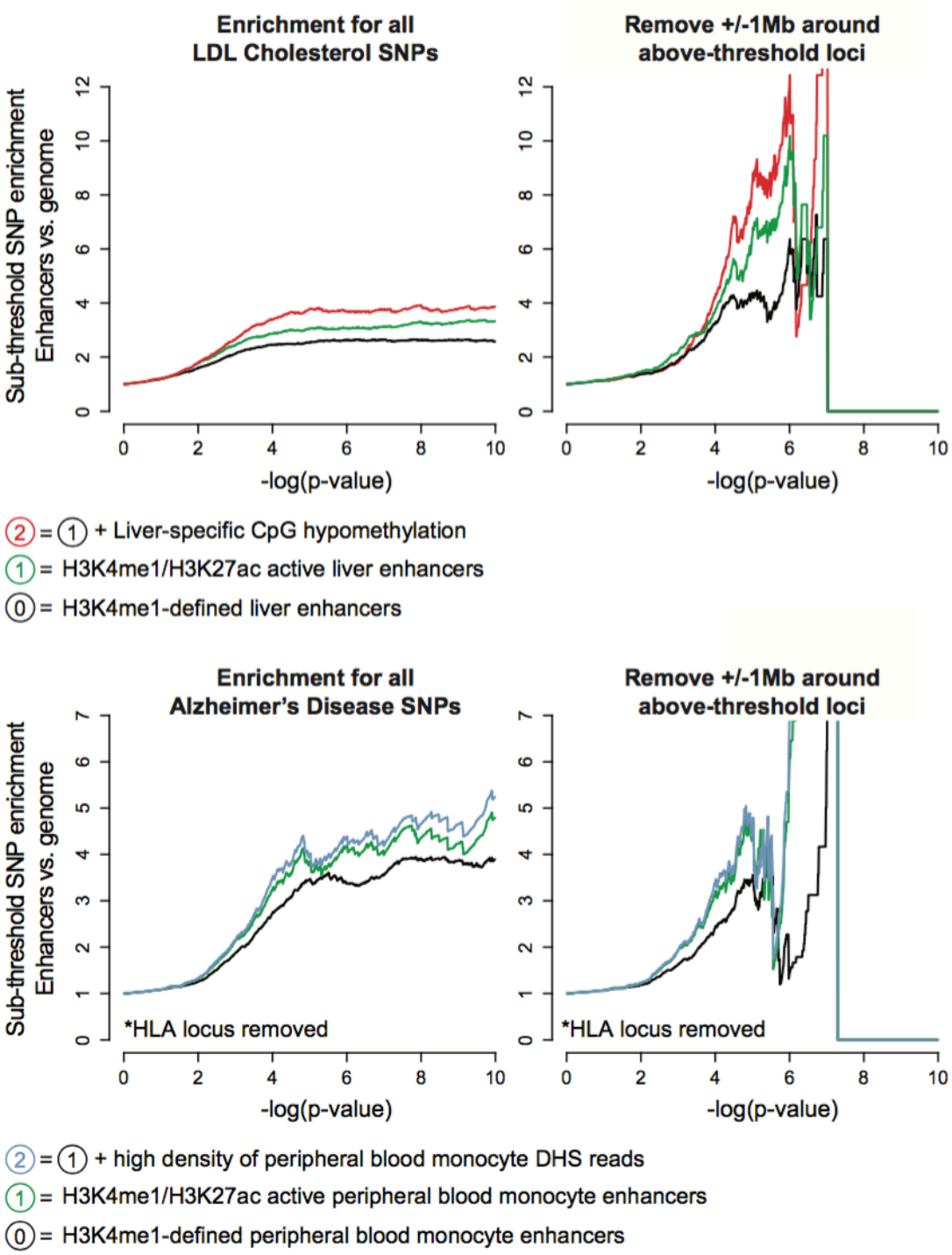


Figure 14: Enhancers harbor additional sub-threshold loci associated with Alzheimer's disease and LDL cholesterol. *Top,* Enhancer characteristics learned for QT/QRS loci (e.g. H3K27ac, CpG hypomethylation) are also effective for enrichment of LDL cholesterol-associated sub-threshold loci in human adult liver enhancers[25]. *Bottom,* Enrichment of Alzheimer's disease-associated sub-threshold SNPs in enhancers from peripheral blood monocytes. Tissue type was chosen using results by Gjoneska *et al.* (2015) computing enrichment of genome-wide significant Alzheimer's disease loci across enhancers from Roadmap Epigenomics tissues[31].

Sub-threshold locus at rs1743292/rs1772203 functionally disrupts enhancer activity

Only a very small number of above-threshold GWAS loci, including *SORT1* for LDL cholesterol, the *FTO*/*IRX3* locus for obesity, and the *SCN5A*/*SCN10A* locus for QRS duration, have been investigated in detail[32-37]. These studies all identified SNPs within non-coding regulatory elements that disrupt expression of a nearby gene that plays a critical role in controlling a human phenotype. In contrast, no sub-threshold locus has been experimentally studied or validated to date. We selected one locus on chromosome 6 where our results from Figures 10-13 suggest that sub-threshold SNPs disrupt enhancer activity and therefore expression of a gene involved in cardiac electrophysiology. We set out to investigate whether this locus can serve as an example for future investigations of other sub-threshold loci.

The sub-threshold locus on chromosome 6 contains 8 SNPs with reported p-values less than $1x10^{-4}$ and another 2 SNPs in LD that do not have calculated p-values. We focused on the 3 SNPs in this locus that overlap active LV enhancers: rs1743292 ($p=6.48x10^{-5}$) and rs112332323 (p-value not available) that both overlap a 3.6kb predicted enhancer, and rs1772203 ($p=5.87x10^{-5}$) that overlaps a 2.8kb predicted enhancer (Fig. 15a,b). We cloned fragments corresponding to both enhancers upstream of a minimal promoter driving the luciferase gene, and compared luciferase activity between constructs carrying either the major or minor haplotypes at each site (rs1743292 enhancer: Fig. 15c, rs1772203 enhancer: Fig. 15e, SNPs differing between cloned constructs listed at bottom of Fig. 15b). We observed that the activity of both enhancers is dependent on the sub-threshold haplotype: at the rs1743292 enhancer, the major haplotype has 45% greater activity ($p=2.99x10^{-12}$), while the minor haplotype is 28% more active in the rs1772203 enhancer ($p=1.79x10^{-5}$).

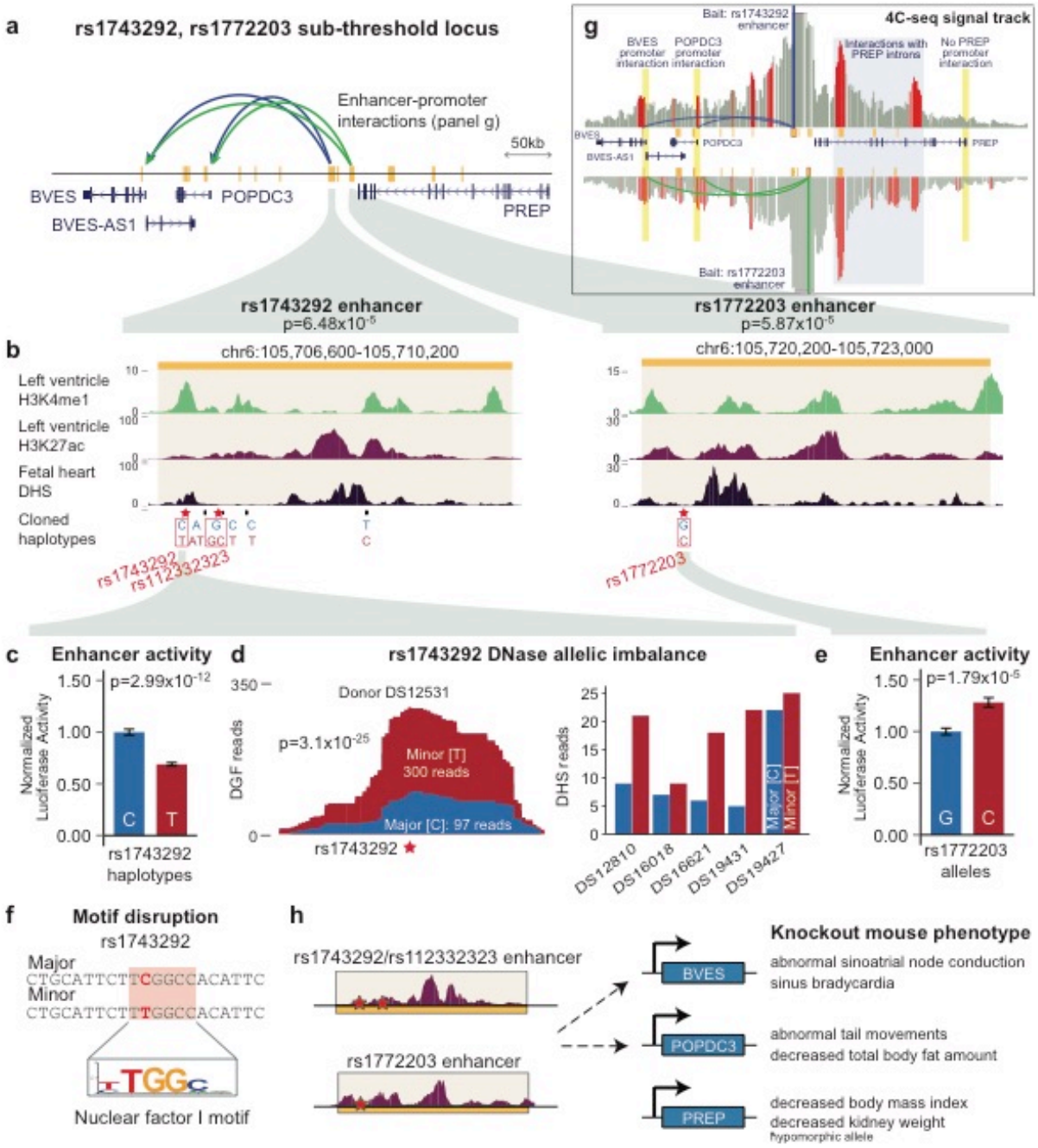


Figure 15: The rs1743292/rs1772203 sub-threshold locus disrupts activity of cardiac enhancers that interact with *BVES*, a gene important for cardiac electrophysiology

a, Overview of rs1743292/rs1772203 sub-threshold locus. Gold rectangles represent predicted active LV enhancers, blue and green lines represent enhancer promoter interactions from the rs1743292 and rs1772203 enhancers, respectively (see panel g). **b,** Detailed view of cardiac enhancers overlapping rs1743292 (left) and rs1772203 (right). **c,** rs1743292 haplotypes differing at 6 SNPs (listed at bottom of panel b) affect activity of cardiac enhancer in human iPSC-derived cardiomyocytes, n=24 per haplotype. Error bars represent standard error of the mean. **d,** *Left,* rs1743292 alters level of DNase I hypersensitivity in a heterozygous human fetal heart sample. *Right,* Allelic imbalance of DHS reads at rs1743292 observed for 5 of 5 human individuals. **e**,

rs1772203 allele affects activity of cardiac enhancer in human iPSC-derived cardiomyocytes, n=16 per allele. Error bars represent standard error of the mean. **f,** rs1743292 SNP overlaps a predicted nuclear factor I (NF-I) motif. **g,** 4C-seq analysis of the rs1743292 (blue) and rs1772203 (green) enhancers identifies enhancer-promoter interactions with nearby BVES, BVES-AS1 and POPDC3 genes, and additional enhancer-enhancer interactions within introns in PREP. **h,** Genetic perturbation of *Bves,* but not *Popdc3* or *Prep* leads to cardiac electrophysiological defects in mouse models.

In the fetal human heart, rs1743292 overlaps a strong DNase I hypersensitivity peak marking a local region of open chromatin signifying potential transcription factor binding (*DHS track*, Fig. 15b)[38]. Thus, to provide evidence that the rs1743292 locus alters enhancer activity in humans, we re-aligned the DHS sequencing reads from heterozygous human individuals in an allele-specific manner to assess the difference in the number of reads that map to either allele[3]. In fetal heart tissue from one individual sequenced to high depth, rs1743292 shows a significant allelic imbalance for DHS reads with 97 reads mapping to the major C allele and 300 reads mapping to the minor T allele (*left,* Fig. 15d, $p=3.1x10^{-25}$, binomial test). This trend is consistent in all five additional human individuals heterozygous at rs1743292 sequenced at lower depth (*right,* Fig. 15d), suggesting that rs1743292 can affect enhancer activity potentially through altering chromatin accessibility or transcription factor binding. Moreover, using motif analysis, we observed that rs1743292 alters a predicted binding site for the cardiac-expressed nuclear factor NF-I family (Fig. 15f), which contains a family member (NF-1a) that itself has been associated by GWAS with cardiac electrophysiology[39].

We used 4C-seq to identify genes that could be regulated by the rs1743292 or rs1772203 enhancers. We observed that both enhancers form interactions with promoters of the upstream popeye-domain containing (POPDC) family members *BVES/POPDC1* and *POPDC3,* and with predicted enhancers situated within introns of the downstream *PREP* gene (Fig. 15a,g). This suggests that both enhancers may contribute to regulating the gene expression of *BVES* and *POPDC3*, of note because the POPDC protein family of transmembrane proteins has recently reported roles in cardiac pacemaking[40,41].

We sought to investigate the roles of the three candidate target genes (*BVES*, *POPDC3*, *PREP*) of the rs1743292/rs1772203 locus in regulating myocardial repolarization. Consistent with the genetic association between this locus and QT interval length, we

found that mice homozygous for loss-of-function copies of *BVES* exhibit cardiac conduction and pacemaker defects (Fig. 15h)[30,40]. In contrast, *POPDC3* and *PREP* mouse loss-of-function models have no reported cardiac abnormalities, and instead show altered body fat, suggesting that this genetic locus alters QT interval length through the *BVES* gene[30].

Strengthening our evidence implicating *BVES* in QT interval, we observed that across 59 human tissues, *BVES* is most highly expressed in human left ventricle, whereas *POPDC3* has much lower expression in cardiac tissue than skeletal muscle, and *PREP* is constitutively expressed across a wide range of tissues (Fig. 16). We also used antisense morpholino oligonucleotides to knockdown transcripts from the *BVES, POPDC3* and *PREP* orthologs in zebrafish, observing that *bves* knockdown leads to a reproducible shortening of the zebrafish ventricular action potential duration (APD), the cellular correlate of the QT interval, ($p=0.002$ and 0.09 for two independent morpholino sequences), whereas there is no reproducible difference in ventricular APD following loss of *popdc3* or *prep* transcripts (Fig. 17). Collectively, these data from multiple organisms provide evidence that SNPs within the rs1743292/rs1772203 locus alter QT interval duration through disruption of *BVES* expression.

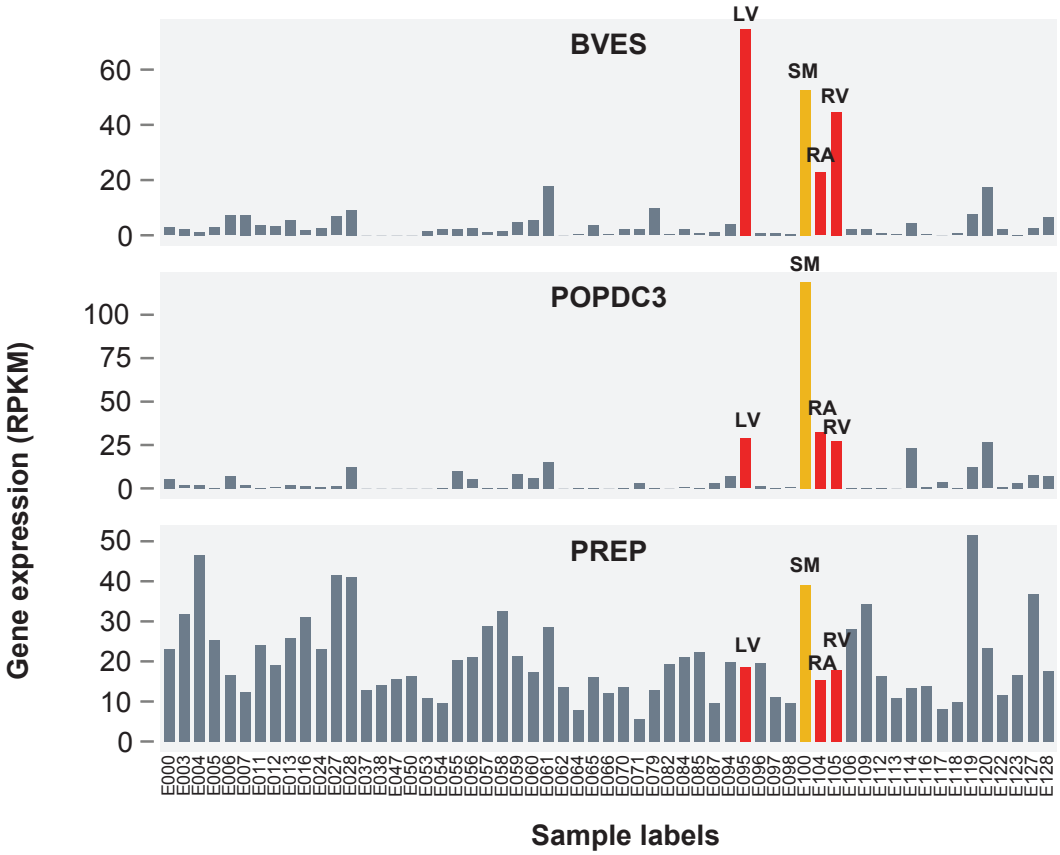


Figure 16: Expression patterns of BVES, POPDC3 and PREP across 59 human tissues. Red bars correspond to chambers of the adult human heart (LV = left ventricle, RV = right ventricle, RA = right atrium), orange corresponds to skeletal muscle (SM = skeletal psoas muscle). Sample labels (E##) correspond to labels assigned by the Roadmap Epigenomics Consortium and available on the Roadmap Epigenomics website[9].

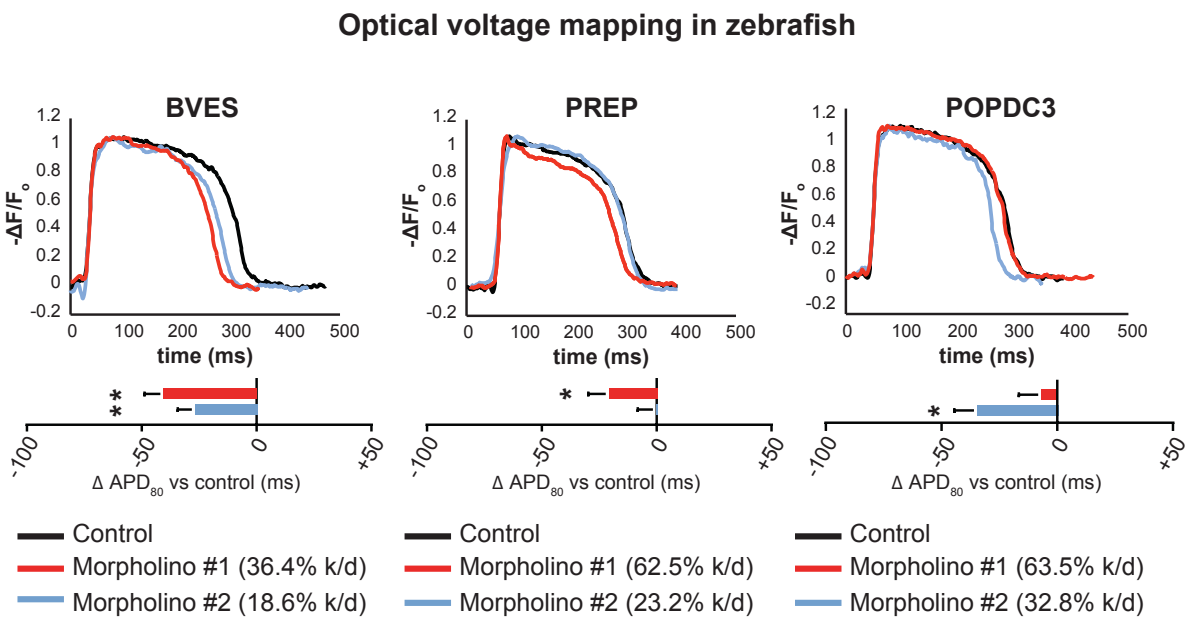


Figure 17: Knockdown of *bves* in zebrafish leads to ventricular repolarization defects. Effect of gene knockdown using two independent morpholino sequences (red, blue) on ventricular action potential duration compared against control scrambled morpholino (black). *Top*: Sample optical voltage mapping traces from one matched morpholino and control knockdown pair. *Bottom*: Differences in APD_{80} between control and antisense morpholino oligonucleotide-mediated knockdown zebrafish. * corresponds to $p<0.05$ from unpaired two-tailed Student's t-test, n=19 for control scrambled morpholinos, n=20 for each morpholino targeting *BVES, POPDC3, or PREP* transcripts, error bars represent standard error of the mean.

These results provide evidence that cardiac enhancers can be used to identify novel sub-threshold loci and genes associated with cardiac traits. As demonstrated with the luciferase enhancer reporter assays, and specifically the rs1743292/rs1772203 locus, sub-threshold loci harbor SNPs that affect enhancer activity and regulate genes involved in QT interval. In the current QT interval GWAS, rs1743292 had an effective sample size of 68,900 individuals with 12.76% power to detect the locus at genome-wide significance. To detect rs1743292 at genome-wide significance with 80% power would require a GWAS cohort of 146,700 individuals. Thus, our study demonstrates that genome-wide enhancer maps are a powerful tool for identifying sub-threshold loci with *bona fide* roles in human cardiovascular physiology that would have remained otherwise unrecognized from existing GWAS cohorts.

Discussion

A major limitation in the human genetics field is the inability to ascribe function to the vast majority of non-coding SNPs associated with complex human traits. Using enhancer annotations from hundreds of cell types and tissues, we find ~50% of QT/QRS GWAS loci overlap enhancers, and that these enhancers share common characteristics, including H3K27ac marks, CpG hypomethylation, and greater evolutionary conservation. The high density of common variation we observed in non-coding enhancers may be due to weaker evolutionary selection against the subtle phenotypes that arise from disruption of transcriptional regulatory units compared to the more severe disruption of protein-coding sequences commonly observed in rare Mendelian diseases.

Studies of genetic heritability have indicated that many additional loci lie below the genome-wide significance threshold[26]. Our study contributes fundamental insights to overcoming the difficult problem of discovering the biologically relevant sub-threshold genetic signals that are orders of magnitude weaker than discovered by traditional GWAS. Three prior studies have observed the general enrichment of either sub-threshold SNPs or SNPs that explain a disproportionately high amount of heritability in cell type-specific regulatory elements[3,42,43]. However, our study is unique in demonstrating the advantage of combining different epigenomic features to produce greater enrichments of sub-threshold loci. Critically, no previous study to our knowledge has implicated any specific sub-threshold locus in any complex human trait, while we establish that 13 of the 18 sub-threshold loci tested in this study are capable of altering enhancer activity. We also leverage GWAS summary statistics and genetic perturbations in mouse to demonstrate that epigenetic marks can discriminate true positive sub-threshold signals from noise, a key problem that, until now, has prevented the study of these loci. Finally, we perform an in-depth molecular dissection of the rs1743292/rs1772203 sub-threshold locus and implicate the popeye-domain containing family of transmembrane proteins in regulating myocardial repolarization. The study of above-threshold GWAS loci is generating more biological insights on new causal genes contributing to human disease, however there

remains a wealth of untapped signals in the sub-threshold region. The work presented represents a first step towards deciphering this signal and opens the door for the discovery of greater numbers of disease loci, genes, and pathways.

Our study focused on QT interval and QRS duration due to their clear tissue of origin and a wealth of existing GWAS data, however we believe our approach could generalize to any well-powered GWAS on any trait. To this end, we chose two recently published, well-powered GWASs that relate to human diseases affecting large segments of the population: LDL cholesterol levels and Alzheimer's disease. For both traits, we observed the enrichment of SNPs well into the sub-threshold significance range, that the enrichment signature persists following removal of all above-threshold loci, and that functional features that improve enrichment of QT-associated sub-threshold loci are also effective when applied to sub-threshold loci associated with LDL cholesterol and Alzheimer's disease (Fig 14). These results suggest that epigenomics can be applied more broadly to identify new loci with sub-threshold statistical significance from GWAS of many complex human diseases. One important future extension of this work would be to build a formal machine learning classifier that can be first trained on above-threshold GWAS loci before being applied to quantitatively rank sub-threshold loci by predicted biological relevance.

Finally, investigating the differences between above-threshold and sub-threshold loci to elucidate the factors that drive loci to different degrees of association with a trait will be an important area of future investigation. Many reported genome-wide significant loci have been discovered by GWAS despite low power, likely due to the existence of many other variants of similar effect that go undetected, termed the "winner's curse", and thus this difference could be driven in part by random chance. However, we also hypothesize that sub-threshold loci with weaker effect sizes may act in different pathways from loci with stronger effect sizes, and that sub-threshold variants could have weaker effects on gene expression.

In summary, our results provide a critical roadmap for the systematic analysis and re-analysis of genome-wide association studies to prioritize novel biologically relevant loci with weak association signals. As demonstrated with the rs1743292/rs1772203 locus, these loci would otherwise require substantially greater cohort sizes to reach statistical significance. Thus, we expect that this approach can be exploited to broadly improve the understanding of the biological pathways that contribute to complex human traits and disease.

Acknowledgements

We are thankful to E. Gjoneska for advice on mammalian luciferase assays, P. Kheradpour for advice on TF motif analysis, G. Quon and J. Wamstad for advice on epigenomics analyses and D. Altshuler, L. Ward and V. Agarwala for advice on human genetics analyses. We are grateful to the members of the Boyer and Kellis laboratories for helpful suggestions and discussions. This work was supported by the CIHR Doctoral Foreign Research Award to X.W., NIH grant (R01HL109004) to D.J.M., NIH grant (R01HG004037) to M.K., and the NHLBI Bench to Bassinet Program (U01HL098179) to L.A.B.

Author Contributions

X.W., M.K. and L.A.B. designed the project; X.W. performed human genetics and epigenomics analyses; C.N.-C. provided association results and analyses for sub-threshold QT interval loci; P.H. provided association results for sub-threshold QRS GWAS loci; P.H.L.K., E.W., X.W., G.R. and W.L. performed 4C-seq experiments and analysis, X.W., G.R., V.S. and E.B. constructed luciferase reporter constructs and performed human cardiomyocyte luciferase assays. N.R.T., R.M. X.-X.N., J.Y., J.L.-M., E.V.D. P.T.E. and D.J.M. performed zebrafish morpholino experiments; X.W., M.K. and L.A.B. wrote the manuscript with input from all authors.

Author Information

High-throughput RNA-sequencing data for *in vitro* differentiated human cardiomyocyte cells and 4C-seq data for 10 enhancers are deposited at the Gene Expression Omnibus under accession number GSE53567. The authors declare no competing financial interests.

Methods

Identifying GWAS loci associated with cardiac traits

We compiled a list of all SNPs associated with electrocardiographic QT interval (reflecting myocardial repolarization) or QRS duration (reflecting cardiac conduction) from the NHGRI GWAS catalog of published GWAS (accessed on July 09, 2013), and removed loci identified from studies with small sample sizes (<5000 individuals). As the GWAS catalog reports SNPs with $p<1x10^{-6}$, we performed a sensitivity analysis using only loci with $p<5x10^{-8}$ to demonstrate that two different cut-offs does not meaningfully affect enrichment results for left ventricle. We used genotype data from the 1000 Genomes project to identify all SNPs in LD ($r^2>0.8$, CEU population) with the lead SNPs. For cases where two lead SNPs were in LD with each other (i.e. different studies reported different SNPs from the same haplotype block), we merged the resulting loci. To avoid over-counting, if the sets of LD SNPs from two independent lead SNPs overlapped, we randomly assigned each of the shared LD SNPs to only one of the two lead SNPs.

RNA-seq data and enhancer annotations

Epigenome Roadmap datasets. Processed RNA-seq data for 59 human tissues and enhancer annotations (for 127 H3K4me1-defined and 88 "strong" H3K4me1/H3K27ac-defined enhancer sets) were downloaded from the Roadmap Epigenomics Project[9]. Initial analyses across all 127 tissues were performed on cardiac enhancers defined by ChromHMM by the Roadmap Epigenomics Project using five chromatin modifications including H3K4me1 but not H3K27ac (15-state model). "Strong" cardiac enhancers, available for a subset of 88 tissues, were defined by ChromHMM by the Roadmap Epigenomics Project using six chromatin modifications including both H3K4me1 and H3K27ac (18-state model).

Human differentiated cardiomyocyte RNA-seq dataset. hESCs were differentiated to cardiomyocytes as previously described [44] and were obtained from David Elliott at Monash University. RNA was extracted using TRIzol reagent according to the

manufacturer's instructions. 10μg RNA was used for library construction according to Illumina RNA-seq library kit with minor modifications. Briefly, mRNA was isolated using Dynabeads mRNA Purification Kit (Invitrogen, Catalog #61006) followed by fragmentation and ethanol precipitation. First and second strand synthesis were performed followed by end repair, A-tailing, paired end adaptor ligation and size selection on a Beckman Coulter SPRI TE nucleic acid extractor. 200-400 bp dsDNA was enriched by 15 cycles of PCR with Phusion High-Fidelity DNA Polymerase (NEB, Catalog #M0530) followed by gel purification of 250 bp fragments from the amplified material. Amplified libraries were sequenced on an Illumina GAIIx sequencer. Reads were mapped against the hg19 version of the human genome using RSEM v. 1.2.3 and bowtie v. 0.12.7 using flags "rsem-calculate-expression --phred64-quals -p 4 --output-genome-bam --calc-ci --paired-end --bowtie-chunkmbs 1024, without in-silico polyA addition to the transcripts.

Enrichment of genomic features in QT/QRS loci

We used genomic features annotated by combinations of histone modifications (e.g. enhancers and promoters using ChromHMM by the Roadmap Epigenomics Project) or by GENCODE (e.g. protein-coding exons). Previous studies have compared the number of GWAS SNPs overlapping a feature against the number expected for a randomly chosen region of similar size[3,45]. However, this approach does not control for biases associated with the location of GWAS SNPs. We controlled for these biases by following the Variant Set Enrichment approach where we generate a background distribution for genomic feature enrichment in loci around sets of 112 randomly sampled control lead SNPs[2]. We chose control lead SNPs from a genome-wide genotyping array (Affymetrix 660W) matched for size of the LD block (+/- 5 SNPs), minor allele frequency of the lead SNP (+/- 0.1), distance to the nearest gene (+/- 25kb if outside gene), and number of nearby genes within a +/- 500kb interval (+/- 3 genes). We also considered differences in local GC content (+/-25nt) but did not observe a strong difference between GWAS and control lead SNPs (p=0.06). To calculate enrichment of genomic regions in GWAS loci, we compared the number of GWAS loci that overlapped an enhancer to 100,000 sets of

equally sized randomly sampled control lead SNPs. The 112 GWAS SNPs compiled from the NHGRI GWAS catalog includes 57 loci with p-values between $1x10^{-6}$ and $5x10^{-8}$ that have a higher false positive rate. In a sensitivity analysis, we examined the subset of 55 loci that met the more stringent $5x10^{-8}$ statistical threshold and found that sets of cardiac enhancers (specifically fetal heart and adult left ventricle) were also most highly enriched in these loci compared to the 123 non-cardiac tissues.

Comparing differences between QT/QRS-associated LV enhancers and all LV enhancers

H3K27ac, DNase I Hypersensitivity and CAGE-seq read enrichment: To score the presence of epigenomic marks in enhancers, we averaged the wig signal tracks over every enhancer with the UCSC bigWigAverageOverBed tool. Fold difference in signals between QT/QRS enhancers and all LV enhancers were calculated by comparing the median signal values of the two groups. P-values were calculated using the Mann-Whitney U test. *Activity in other cardiac and non-cardiac tissues:* Overlap with enhancers in other tissues was calculated using the intersectBed function in the BEDTools suite[46]. *CpG hypomethylation and hypermethylation:* Whole-genome bisulfite sequencing data for 37 human tissues, including the left ventricle, was obtained from the Roadmap Epigenomics Project[9]. We identified LV-specific hypo and hypermethylated CpGs as those that differed in percent methylation with the mean of 34 non-cardiac tissues by both (i) 2 standard deviations and (ii) at least a difference in absolute percent methylation of 15 percent. *Evolutionary Conservation:* We calculated evolutionary conservation of enhancers using the methodology outlined by Nord *et al.* (2013)[47]. Briefly, we first identified the 100bp region of each enhancer with greatest average evolutionary conservation across primates (primate subset of 46-way phyloP conservation track obtained from UCSC). To quantify differences in evolutionary conservation of GWAS enhancers against all LV enhancers, we randomly selected 1000 size-matched sets of LV enhancers (size within +/-1kb of corresponding QT/QRS enhancer), as the 100bp segment of greatest conservation in longer enhancers is statistically more likely to have greater conservation than a shorter segment.

Comparing differences in TF motif disruption

We obtained TF motif instances in the human genome (hg19) for 651 human motifs from the ENCODE project[48], and filtered these to only consider 287 motifs that correspond to TFs expressed in the left ventricle (>1 RPKM by RNA-seq). We quantified the number of QT/QRS loci containing a SNP that disrupted an enhancer motif corresponding to an expressed TF in the left ventricle, and compared this against randomly sampled sets of control loci matched for MAF, LD block size, distance to the nearest gene and presence on the Affymetrix 660W genotyping array.

Enrichment of QT SNPs below genome-wide significance in enhancers

Enrichment analysis. We used a sliding -log(p-value) threshold from 0 to 10 with steps of 0.1. At each cut-off, we computed the proportion of SNPs in enhancers with p-values more significant than the cut-off (foreground) against the proportion of SNPs in the whole genome. *Grouping SNPs in LD*. For each pair of SNPs, if the two SNPs are in LD (r^2>0.2, CEU population from 1000 Genomes project) we remove the SNP with the weaker p-value.

Enrichment of LDL cholesterol and Alzheimer's disease-associated sub-threshold loci in enhancers

Summary GWAS data for LDL cholesterol was obtained from Willer *et al.* (2013)[25], and summary GWAS data for Alzheimer's disease (AD) was obtained from Lambert *et al.* (2013)[49]. Enrichment analyses were performed as described above for QT interval. For enrichment of Alzheimer's disease-associated SNPs, the region encompassing the HLA locus was excluded (chr6:24,182,924-34,537,546 in hg19), as this region contained approximately 25% of all low p-value SNPs ($p<1x10^{-5}$) in the genome therefore and could skew enrichment results.

The liver tissue was chosen for LDL cholesterol enrichment based on biological relevance. Tissue choice for AD SNPs was made using genome-wide enrichment analyses performed

by Gjoneska *et al.* 2015[31]. For this analysis, we chose the second-most enriched tissue from Gjoneska *et al.* (peripheral blood monocytes, with most significant p-value) instead of the most enriched tissue (peripheral blood mononuclear cells, PBMCs, with second-most significant p-value) because the enrichment of AD SNPs in PBMC enhancers was substantially weaker than peripheral blood monocytes following removal of SNPs within the HLA locus. For AD GWAS, removal of SNPs within +/- 1Mb of above-threshold loci was performed using 13 above-threshold loci with $p<5x10^{-8}$ (Stage 1 analysis) listed in Table 2 of Lambert *et al.* For LDL cholesterol analyses, we first attempted to remove all SNPs within +/- 1Mb of above-threshold loci reported in Supplemental Files 2 & 3 of Willer *et al.*, however many SNPs with $p<5x10^{-8}$ remained. Therefore, we performed LD pruning ($r^2>0.2$ from CEU population) on summary-level p-value data from Willer *et al.* to define above-threshold loci and then removed 68 unique genomic intervals from the analysis. Enhancer functional characteristics applied to the enhancer sets were chosen based on the availability of additional data for the chosen tissue. DNase I hypersensitivity data not available for human liver, and genome-wide CpG methylation data was not available for peripheral blood monocytes.

Comparison of QT sub-threshold loci in QRS GWAS data

To assess whether QT sub-threshold loci overlapping enhancers are more likely to represent true biological signals, we queried the p-values of these loci in a related GWAS of QRS duration. In total, the QT GWAS we used to identify the sub-threshold loci consisted of 76,061 individuals, while the QRS GWAS queried consisted of 60,255 individuals. We compared the total sizes of each cohort used in the two studies and calculated that a minimum total of 46,452 individuals must be different between the two studies. Specifically, there are at least 31,129 individuals present in QT GWAS that are not present in the QRS GWAS, and at least 15,323 individuals present in the QRS GWAS that are not present in the QT GWAS.

We used summary-level p-value data from the QRS GWAS testing four clinically applied QRS traits: Sokolow-Lyon, Cornell, 12-lead-voltage duration products, and QRS duration.

For each SNP, the assigned p-value represented the minimum p-value across these four traits. For each sub-threshold locus, we identified all SNPs in strong LD (r^2>0.8, CEU population from 1000 Genomes project), and assigned the p-value as the minimum of all p-values for LD SNPs in the QRS GWAS data.

Identifying candidate genes near sub-threshold loci using activity correlation across human tissues

From the Roadmap Epigenomics Project, we were able to obtain matching "strong" enhancer annotations and RNA-seq data for 59 of the 127 tissues, including LV. For each LV enhancer, we considered all genes with expression ≥1 RPKM in LV and *in vitro* differentiated human cardiomyocytes and distance within +/-500kb as potential targets. We then split the RNA-seq data for the 59 tissues into two groups, depending on whether the enhancer is present or absent in each tissue, and applied a one-sided Mann Whitney U test to ask whether each potential target gene showed significantly greater expression in tissues where the enhancer was active. Genes differentially expressed between tissues with active and inactive enhancers ($p<0.05$) were considered computationally-determined potential target genes. For determining targets of sub-threshold enhancers, we first filtered our set of sub-threshold enhancers to remove those unlikely to be associated with QT interval. To do this, we excluded sub-threshold SNPs if the -log(p-value) was lower than 80% of the -log(p-value) of the most statistically significant SNP in LD (r^2>0.2), as these are unlikely to be causal.

Cardiac phenotypes for genes with mutations in mouse

For sub-threshold loci overlapping enhancers, and the set of all active LV enhancers, we identified nearby genes using the enhancer-gene linking method described above. This methodology was not applicable to the 129 sub-threshold loci that do not overlap enhancers, and therefore we identified the two nearest genes within 1Mb using GREAT

v2.0.2 and selected only genes with expression in adult human left ventricle data (>1 RPKM). Mouse orthologs of human genes were identified using the Ensembl Genes 79 database through BioMart, and all queries of the MGI mouse phenotypes database were made between April 26, 2015 and May 6, 2015. We used three search terms relevant to QT interval: "ventricle muscle contractility", "cardiac contractility" and "conduction" (excluding non-cardiac conduction terms).

Quantifying allelic imbalance at SNPs

We used DNase I hypersensitivity and digital genomic footprinting data from the ENCODE[48] and Roadmap Epigenomics Projects[8] because samples were sequenced to a greater depth than the chromatin modification ChIP-seq data, and there were data available from more individuals. To quantify allelic imbalance, we mapped DHS/DGF reads to a version of the human genome (hg19) downloaded from the UCSC genome browser with all SNPs (dbSNP141) masked by ambiguous nucleotides (N's) using Bowtie2 (v2.2.0, flags: -N 1, --sensitive, --end-to-end, --no-unal). As genotypes were not available, we considered a sample heterozygous at a particular SNP if reads from the hg19-defined reference and alternate alleles each mapped to 3 or more unique positions. Using this methodology, we observed the median difference in reads mapping to the reference versus alternate alleles to be 0. In total, reads mapped to the reference allele more often than alternate at 6537 of 13826 heterozygous SNPs, and vice versa at 5884 of 13826 heterozygous SNPs, with equal numbers of reads mapping to both alleles at the remaining 1405 SNPs. To quantify statistical significance of allelic imbalance at SNPs, we followed Maurano *et al.* (2012) and considered only SNPs with more than 21 reads. We performed a binomial test under the null hypothesis where reads map to both alleles at equal frequency, followed by Benjamini-Hochberg multiple testing correction across all heterozygous enhancer-overlapping SNPs.

4C-Seq Methods

Human iPSC-derived cardiomyocytes (iCMs) (Cellular Dynamics, Catalog #: CMC-100-010-001) were thawed according to manufacturer's instructions and diluted to a final

plating density of $0.2x10^6$ cells per mL with plating medium (Cellular Dynamics, Catalog#: CMM-100-110-001). After 7 days in culture, iCMs were homogenized using a douncer, cross-linked and further processed as 4C template using DpnII as the first restriction enzyme and Csp6I as the second enzyme following the procedure outlined in van de Werken *et al.* (2012)[50]. The median spacing between GATC fragments (recognized by DpnII) in the hg19 human genome is 264 nt. Sequencing of the 4C-Seq library was performed on an Illumina HiSeq 2000, and sequencing reads were aligned to a reduced genome consisting of sequences that flank DpnII restriction sites. Primer sequences used for sequencing the 4C-seq library are listed in Supplementary File 3. The human genome (hg19) was used as reference genome for mapping 4C sequence captures. Non-unique sequences that flank a restriction site were removed from the analysis.

To map 4C-seq reads to the genome, we first binned reads according to the reading primers used in each lane. We allow a single mismatch in the reading primer that overlaps the primary restriction cut site (DpnII). The binned sequences were mapped to an *in silico* library of potential fragment ends generated based on the restriction enzymes used for the 4C template preparation. We did not allow any mismatch in the fragment-end, and for analysis we focused on the unique fragends only (excluding repetitive fragment ends). As biases from sequencing yield or restriction cutting may be introduced by 4C-seq, we computed 4C-seq coverage in a genomic region by averaging mapped reads in running windows of 21 4C-seq fragment-ends. For peak-calling in a single 4C experiment, we perform explicit background modeling of the up- and downstream genomic regions independently. We assume that in a completely unstructured chromatin fiber the contact probability monotonically decreases as a function of the distance to the viewpoint. We model this by performing monotonic regression of the 4C signal as a function of the distance to the viewpoint. For this we use the R package isotone, which implements the monotonic regression[51]. We then compare the observed 4C signal to the predicted value from the background model and call the extremes that reach a significance threshold as peaks. For a given threshold q and a distribution F of residuals

from the background model, every observation greater than Q3(F)+q*IQR(F), where Q3 is the third quartile of F and IRF(F) the inter-quartile range, is considered significant.

Generating enhancer reporter constructs

Sub-threshold loci were considered candidates for testing by the luciferase reporter assay if the sub-threshold SNP overlapping the active LV enhancer either (i) overlaps a fetal heart DNase I hypersensitivity site, or (ii) is an eQTL in the left ventricle (i.e. the SNP genotype is associated with differential expression of a nearby gene). We generated allele-specific enhancer constructs using two strategies outlined below: (i) PCR from genotyped heterozygous individuals, or (ii) direct synthesis of enhancer fragments. (i) *Enhancer cloning from heterozygous individuals:* We designed primer sequences to clone the entire predicted enhancer sequence defined by ChromHMM, and appended a 5'CACC sequence to forward primers to permit directional TOPO cloning. We designed primer sequences to clone fragments of up to 3kb. For enhancers annotated as larger than 3kb, we either selected a 3kb fragment centered at the region of greatest histone modification density (H3K4me1, H3K27ac), or generated multiple fragments spanning the enhancer. Primer sequences and samples for human genomic DNA (Coriell Cell Repositories) are listed in Supplementary File 3. We PCR amplified enhancers from human genomic DNA using Q5 High-Fidelity DNA Polymerase (NEB, Catalog # M0491S) and purified fragments corresponding to the correct length using a QIAquick Gel Extraction Kit (Qiagen, Catalog # 28706). (ii) *Direct synthesis of enhancer fragments:* Enhancer fragments up to 1kb in size were chosen so that the fragment covers both the sub-threshold SNP as well as peak within the DNase I hypersensitivity signal, and a 5'CACC sequence was appended to permit directional TOPO cloning. Fragments were synthesized using the gBlocks Gene Fragments service from Integrated DNA Technologies (sequences are listed in Supplemental File 3). Enhancer fragments from both methods were cloned into Gateway-compatible entry vectors using a pENTR/D-TOPO Cloning Kit (Life Technologies, Catalog # K2400) and transformed into TOP10 *E. coli* bacteria following manufacturers guidelines. We used Sanger sequencing to verify that purified entry vectors carried enhancers with the correct insertion orientation and no mutations beyond

the expected polymorphisms. Entry vectors were then Gateway-cloned using LR Clonase II Plus (Life Technologies, Catalog # 12538-120) into a Gateway-converted pGL4.23 destination vector (Promega, Catalog # E8411) for luciferase assays in human cell lines [52]. We used Sanger sequencing to confirm a second time the correct enhancer orientation and sequence inside the destination vectors.

Human cardiomyocyte luciferase assays

Human iCMs (Cellular Dynamics, Catalog #: CMc-100-010-001) were thawed according to manufacturer's instructions and diluted to a final plating density of $0.2x10^6$ cells per mL with plating medium (Cellular Dynamics, Catalog#: CMM-100-110-001). 96-well tissue culture treated plates were coated with 0.1mL of 0.1%(w/v) gelatin per well and incubated at 37°C for at least two hours. The gelatin solution was aspirated off and wells rinsed with 100uL of PBS, aspirated, and let sit in the tissue culture hood. Using a multichannel pipette, 100ul of cells were seeded per well to obtain a target density of $20x10^3$ iCMs. The plates were kept on a flat bench at room temperature for 10-15 minutes to allow for cells to settle down uniformly, followed by incubation at a tissue culture incubator set at 37°C and 7% CO_2. 48 hours post-seeding, the iCM plating medium was replaced with 100uL of Maintenance Medium (Cellular Dynamics, Catalog #:CMM-100-120-001). The Maintenance Medium was replaced every other day.

3-4 days post-plating, iCMs began beating spontaneously and 7 days post-plating, they formed electrically connected syncytial layers that beat simultaneously. At this stage, the cells were transfected with the appropriate Luciferase reporter constructs and controls for downstream analyses. Media was replaced an hour before transfections. For each well, 95ng of enhancer firefly Luciferase reporter (cloned into pGL4.23, Promega) and 5ng of Renilla Luciferase transfection control vector (pGL4.73, Promega) was mixed with 10ul of OPTIMEM (Life Technologies, Catalog #:51985-034). 0.2ul of Viafect transfection reagent was added to the DNA/OPTIMEM mixture. After mixing, the transfection cocktail was incubated at room temperature for 5 min and 10ul dispensed into the well with iCMs and plates transferred to 37°C. Media was changed 24 hours after transfection.

8 independent wells of iCMs were transfected per construct to account for variability in plating and transfection efficiencies. A mammalian expression vector, pEF-GFP (Addgene, Plasmid 11154), was used to visually monitor transfection efficiency. At least 65-70% of the population of iCMs expressed GFP 24 hours post-transfection.

Luciferase activity was measured 24hr after transfections using the Dual-Luciferase Reporter Assay System (Promega, Catalog#:E1980). After aspirating media, cells were rinsed with PBS once, and lysed with 20uL of 1X Passive lysis buffer in the Luciferase assay kit. 15 minutes after gentle shaking on an orbital shaker and complete lysis, the plate was stored at -80°C until further processing. Samples were prepared and luminescence measured according to Manufacturer's Assay protocol for 96-well plates using the Varioskan Flash Multimode Reader (Thermo Scientific).

Data Analysis. For all transfection wells, luminescence values of a blank non-transfection control were subtracted from all measured activity values. Firefly luciferase activity was then normalized to Renilla luciferase activity to control for transfection efficiency in each well. As luciferase reporter assay reagents decrease in activity during regular storage, the reference and alternate alleles of each reporter construct were spotted on the same 96-well plates to control for plate-to-plate variability in reagent activity. For each enhancer, we merged readings from multiple days by normalizing the activity of reporters to the reference allele. Each reporter construct was transfected into wells of at least two separate 96-well plates and readings for all wells were merged. Wells where Renilla luciferase activity (transfection control) was substantially lower (>90%) than neighboring wells were excluded from analyses. Statistical significance was determined by unpaired Student's t-test assuming equal variance. Minimum sample size of n=8 per enhancer construct was chosen to achieve 95% power for effect size (Cohen's d) of 2 (0.2 difference in activity between haplotypes with standard deviation of 0.1 normalized luciferase activity units) at p=0.05.

Zebrafish antisense morpholino oligonucleotide-mediated knockdown and optical voltage mapping

Zebrafish (TuAB strain) were cared for according to standard techniques. All animal experiments were approved by the Partners Subcommittee on Research Animal Care (SRAC) and were conducted in compliance with the regulations published in the US National Institute of Health *Guide for the Care and Use of Laboratory Animals.* At the single cell stage, fertilized oocytes were injected with standardized concentrations and volumes of antisense morpholino oligonucleotides (5'CAATAGATGGCGCTGTGTACCTGTC3' and 5'AGAGCAGCCTGAAAGACAATAAAGA3' for *bves*, 5'GGTTAATCCACTCACCTGCCTGAAA3' and 5'CCGTCACTCGTATCCTGTTTTAGTG3' for *popdc3,* XXX and YYY for *prep,* 5'GTTCAATTGTTTCTCACCTGCCAGA3' and 5'CTAATCCTGTGAAAGCAGAAGATCC3' for *popdc2*) dissolved in Danieau's solution (58 mM NaCl, 0.7mM KCl, 0.4 mM MgSO4, 0.6 mM Ca(NO3)2, 5.0 mM HEPES pH 7.6). Controls were injected with an equivalent dose of non-targeting morpholino of equal length but differing nucleotide composition (5'ATCCTCTTGAGGCGAACAAAGAGTC3'). RNA was harvested at 72 hours using Trizol (Life Technologies) according to the manufacturer's instructions, cDNA synthesized by iScript reverse transcriptase (Bio-Rad) and semi-quantitative PCR was used to assess relative percentage of gene knockdown. All studies of morpholino efficacy are a result of samples obtained from three independent injections. For evaluation of ventricular action potential duration, embryo hearts were microdissected at 72 hours of development and stained with di-8-ANEPPS (Invitrogen). Cardiac contraction was arrested with 15uM blebbistatin (Sigma-Aldrich). Hearts were then field paced at 2Hz and imaged at 1000 frames per second as previously described[40]. Analysis of action potential durations was performed using an in-house developed MatLab program. The action potential duration at 80% repolarization was utilized for all analyses. A minimum *n* of 9 embryos was required for all ventricular action potential studies, based on power

calculations for effect size (Cohen's d) of 1.5 at p=0.05. No animals were excluded from analyses unless ventricular depolarization could not be induced at 120 paces per minute. No randomization of samples or blinding of investigators was utilized during these protocols. Statistical comparisons were performed using one-way ANOVA with Fisher's Least Significant Difference testing with all comparisons being to clutchmate controls. All distributions were normal, and variances between control and experimental groups were not statistically significant.

References

1. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).

2. Cowper-Sal lari, R. *et al.* Breast cancer risk–associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature Genetics* **44,** 1191–1198 (2012).

3. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337,** 1190–1195 (2012).

4. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics* **45,** 124–130 (2013).

5. Arking, D. E. *et al.* Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nature Genetics* **46,** 826–836 (2014).

6. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106,** 9362–9367 (2009).

7. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456,** 18–21 (2008).

8. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322,** 881–888 (2008).

9. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes.

Nature **518,** 317–330 (2015).

10. Rabkin, S. W., Mathewson, F. L. & Tate, R. B. The electrocardiogram in apparently healthy men and the risk of sudden death. *Br Heart J* **47,** 546–552 (1982).

11. Heist, E. K. & Ruskin, J. N. Drug-Induced Arrhythmia. *Circulation* **122,** 1426–1435 (2010).

12. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

13. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).

14. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).

15. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences* **107,** 21931–21936 (2010).

16. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–283 (2010).

17. Hon, G. C. *et al.* Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. *Nature Genetics* **45,** 1198–1206 (2013).

18. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480,** 490–495 (2011).

19. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–461 (2014).

20. Hamer, D. & Sirota, L. Beware the chopsticks gene. *Molecular psychiatry* **5,** 11–13 (2000).

21. van der Harst, P. *et al.* 52 Genetic Loci Influencing Myocardial Mass. *J. Am. Coll. Cardiol.* **68,** 1435–1448 (2016).

22. Gieger, C. *et al.* New gene functions in megakaryopoiesis and platelet formation. *Nature* **480,** 201–208 (2011).

23. Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nature Genetics* **43,** 246–252 (2011).

24. Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nature Genetics* **41,** 703–707 (2009).

25. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45,** 1274–1283 (2013).

26. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genetics* **43,** 519–525 (2011).

27. van de Werken, H. J. G. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods* **9,** 969–972 (2012).

28. van der Harst, P. Fifty-two genetic loci influencing human electrically active myocardial mass. *In revision*

29. Fullwood, M. J. *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462,** 58–64 (2009).

30. Bello, S. M., Smith, C. L. & Eppig, J. T. Allele, phenotype and disease data at Mouse Genome Informatics: improving access and analysis. *Mamm. Genome* **26,** 285–294 (2015).

31. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518,** 365–369 (2015).

32. van den Boogaard, M. *et al.* Genetic variation in T-box binding element functionally affects SCN5A/SCN10A enhancer. *J. Clin. Invest.* **122,** 2519–2530 (2012).

33. van den Boogaard, M. *et al.* A common genetic variant within SCN10A modulates cardiac SCN5A expression. *J. Clin. Invest.* **124,** 1844–1852 (2014).

34. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466,** 714–719 (2000).

35. Arnolds, D. E. *et al.* TBX5 drives Scn5a expression to regulate cardiac conduction system function. *J. Clin. Invest.* **122,** 2509–2518 (2012).

36. Claussnitzer, M. *et al.* FTOObesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med* **373,** 895–907 (2015).

37. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507,** 371–375 (2014).

38. Neph, S. *et al.* Circuitry and Dynamics of Human Transcription Factor Regulatory

Networks. *Cell* **150,** 1274–1286 (2012).

39. Ritchie, M. D. *et al.* Genome- and phenome-wide analyses of cardiac conduction identifies markers of arrhythmia risk. *Circulation* **127,** 1377–1385 (2013).

40. Froese, A. *et al.* Popeye domain containing proteins are essential for stress-mediated modulation of cardiac pacemaking in mice. *J. Clin. Invest.* **122,** 1119–1130 (2012).

41. Kirchmaier, B. C. *et al.* The Popeye domain containing 2 (popdc2) gene in zebrafish is required for heart and skeletal muscle development. *Dev. Biol.* **363,** 438–450 (2012).

42. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *American Journal of Human Genetics* **95,** 535–552 (2014).

43. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47,** 1228–1235 (2015).

44. Elliott, D. A. *et al.* NKX2-5eGFP/w hESCs for isolation of human cardiac progenitors and cardiomyocytes. *Nature Methods* **8,** 1037–1040 (2011).

45. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).

46. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

47. Nord, A. S. *et al.* Rapid and Pervasive Changes in Genome-wide Enhancer Usage during Mammalian Development. *Cell* **155,** 1521–1531 (2013).

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

49. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics* **45,** 1452–1458 (2013).

50. van de Werken, H. J. G. *et al.* Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nature Methods* **9,** 969–972 (2012).

51. Mair, P., Hornik, K. & de Leeuw, J. Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. *Journal of statistical software* **32,** 1–24 (2009).

52. Fisher, S. *et al.* Evaluating the biological relevance of putative enhancers using Tol2 transposon-mediated transgenesis in zebrafish. *Nat Protoc* **1,** 1297–1305 (2006).

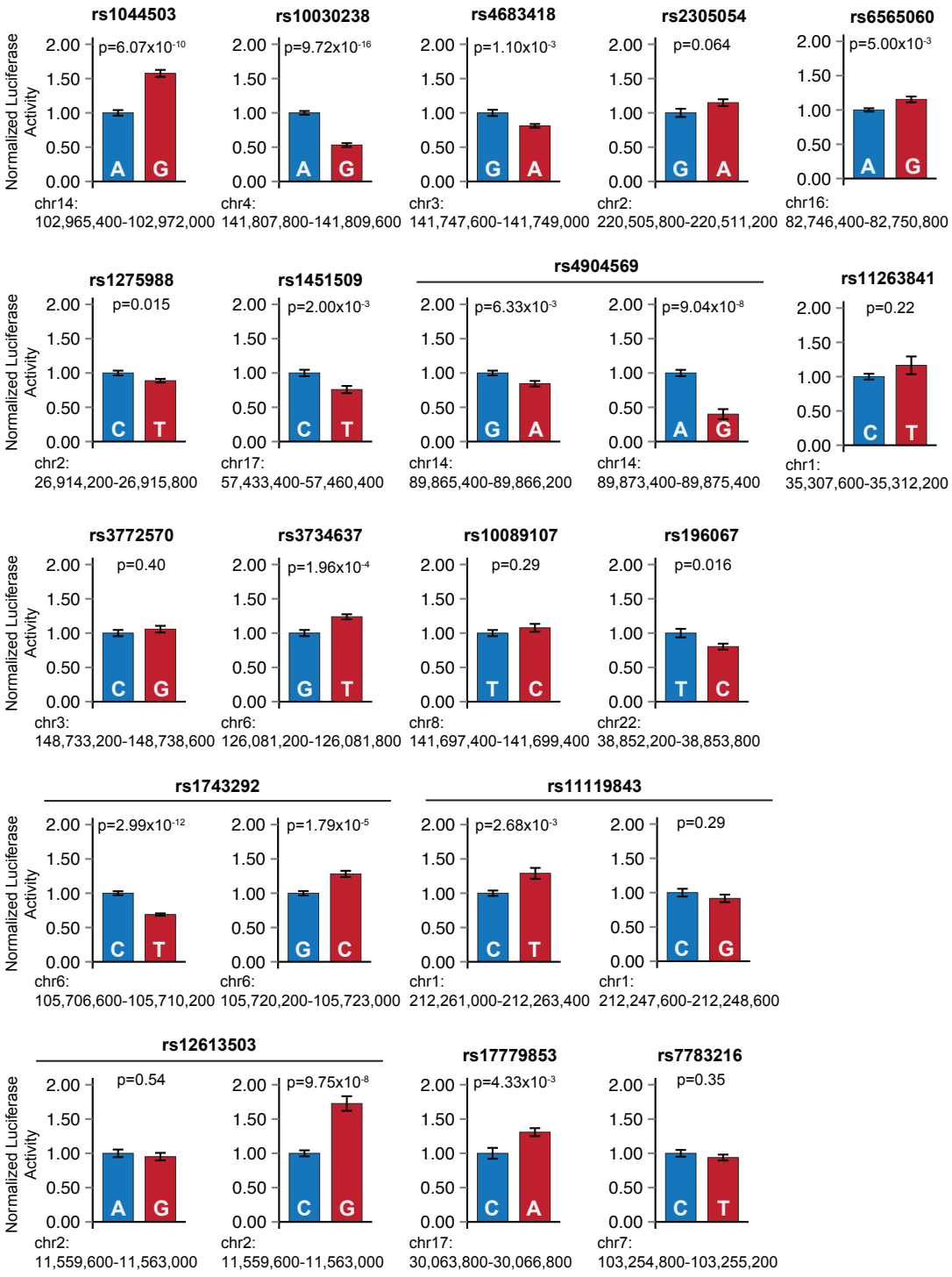


Figure S1: Sub-threshold SNP alleles affect enhancer activity. For each sub-threshold locus, enhancers carrying one of two haplotypes were cloned upstream of a minimal promoter and firefly luciferase reporter gene. *Blue*: enhancer carrying reference allele; *red*: enhancer carrying alternate allele. Error bars represent standard error of the mean.

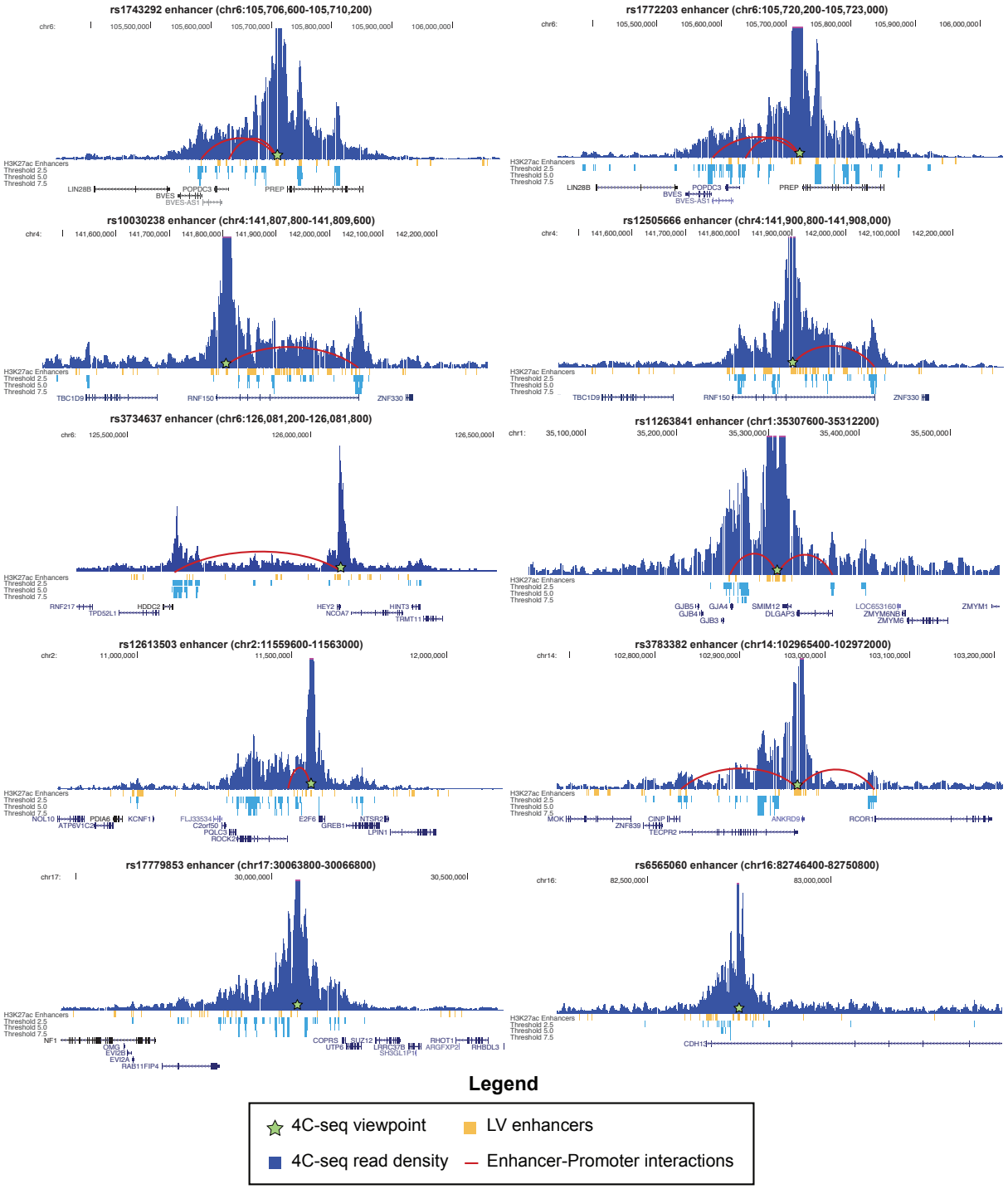


Figure S2: 4C-seq interactions with 10 enhancers in 8 sub-threshold loci. Height of blue bars represents interaction strength with 4C viewpoint. Red curves indicate enhancer-promoter interactions called within an annotated GENCODE promoter (up to 2.5kb upstream of TSS) at a threshold of 5.0.

Chapter 3

Genome-wide mapping of transcriptional regulatory activity in human cells

Abstract

Genome-wide profiling of epigenomic marks has allowed for the rapid prediction of tens of thousands of transcriptional regulatory elements, most notably distal enhancers that would otherwise have remained hidden within non-coding DNA. Many epigenomically-predicted enhancers, however, fail to promote transcription in an experimental context. Moreover, these predicted enhancers are on average 1kb in length, making it difficult to identify specific segments of DNA within each enhancer that are critical for driving transcriptional activity and would lead to deleterious consequences when disrupted by genetic variants. Here, we describe a high-throughput experimental assay, ATAC-STARR, that quantifies the transcriptional regulatory activity of millions of DNA fragments. DNA fragments used in ATAC-STARR are preferentially extracted from regions of open chromatin, including promoters and distal enhancer elements, and fragments tested can be up to 1kb in length, five times longer than sequences used in other high-throughput reporter assays. In a pilot study, we used ATAC-STARR to test fragments with sizes between 25-500nt and we identified 537 significantly up-regulated fragments in 382 unique loci in the GM12878 lymphoblastoid cell line at FDR<0.05. Active regulatory fragments are more likely to be larger in size, enriched for activating histone modifications including H3K27ac and for motifs and binding of immune transcription factors. Our results suggest that ATAC-STARR can be a general strategy for experimentally assaying the DNA regulatory landscape of different cell types.

Introduction

Transcriptional enhancer elements were originally characterized in the 1980's as non-coding DNA elements with the ability to promote gene expression at a distance from a gene promoter[1,2]. Transcriptional enhancers are regions of the genome that bind transcription factors and act to either promote or repress the transcription of nearby genes. Unlike protein coding genes, which can be identified with high sensitivity from both open reading frame scanning using the codon table and evolutionary conservation, enhancers do not have any highly predictive sequence patterns and are not highly conserved under existing metrics of evolutionary conservation. The combination of histone modifications, such as the presence of activating H3K4me1 and H3K27ac and absence of repressive H3K27me3, are typically used to predict the positions of enhancers through scanning of the genome with unsupervised machine learning models[3-5]. However, many putative enhancers predicted only from their epigenomic contexts do not experimentally drive reporter gene transcription, and prediction of enhancers using commonly profiled histone modifications may miss enhancers that co-localize with a different set of histone modifications[6,7]. Both of these concerns necessitate the development of high-throughput assays to experimentally test the activity of enhancers genome-wide.

High-throughput enhancer reporter assays, first developed in 2012, use *in vitro* oligonucleotide synthesis to generate and clone tens of thousands of distinct DNA sequences into a common enhancer reporter vector. To date, these assays such as the Massively Parallel Reporter Assay (MPRA) and Cis-Regulatory Element analysis by high-throughput sequencing (CRE-seq) have been successfully used to test the activity of thousands putative transcriptional enhancer elements[8-11]. However, technical limitations of large-scale *in vitro* oligonucleotide synthesis restrict input DNA fragments to a maximum of 230nt in length. Moreover, only a limited number of DNA sequences (up to $2x10^5$) can be synthesized per array, and sequences are often redundantly synthesized to compensate for a nucleotide synthesis error rate of 1/200.

Here, we perform a genome-wide screen of transcriptional regulatory activity using enhancer constructs of comparable size scale to those individually tested in low-throughput studies. To avoid the limitations of *in vitro* oligonucleotide synthesis, we developed a modified version of STARR-seq (Self-Transcribing Active Regulatory Region sequencing), termed ATAC-STARR, where hundreds of thousands of fragments of human genomic DNA generated by Tn5 transposition using ATAC-seq (Assay for Transposase-Accessible Chromatin with high throughput sequencing)[12] and enriched at sites of open chromatin and higher transcriptional regulatory activity are cloned into an enhancer reporter construct. In a pilot experiment, we identify 537 up-regulated fragments in 382 genomic loci that have marks of activating regulatory regions, suggesting ATAC-STARR is capable of identifying regulatory DNA sequences in an unbiased manner.

Results

Experimental design and cloning strategy

ATAC-STARR combines the selective fragmentation of genomic DNA at regions of open chromatin from ATAC-seq with high-throughput cloning of fragments into an enhancer reporter vector. Fragments are cloned into the 3'UTR of a reporter gene and drive self-transcription so that putative enhancer segments can be identified and quantified by high-throughput RNA sequencing to produce a quantitative readout of enhancer activity (Fig. 1). We first set out to use ATAC-STARR to investigate how fragment size affects activity of putative regulatory fragments by generating a library with fragment sizes that range from 25 to 500nt. We modified the existing ATAC-seq methodology so that after Tn5 transposition and a first round of PCR amplification, DNA fragments are run an agarose gel for size selection. As ATAC-seq libraries are commonly contaminated with mtDNA, we treated gel-extracted fragments with a CRISPR-Cas9 library of anti-mtDNA gRNAs to selectively cut fragments originating from the mitochondrial genome. We then performed a second round of PCR to selectively amplify non-digested fragments and to ensure that all remaining fragments are represented in the final library at high copy number. PCRs were performed using primers carrying random i7 barcode sequences to increase the number of potential technical replicates per unique enhancer fragment tested and to estimate variance of enhancer activity. Fragments were then cloned into a linearized STARR-seq backbone and transfected into GM12878 lymphoblastoid cells (Fig. 1, Methods). Overall our library construction method can be completed in 2-3 days and requires 10^4-10^5 cells as input starting material.

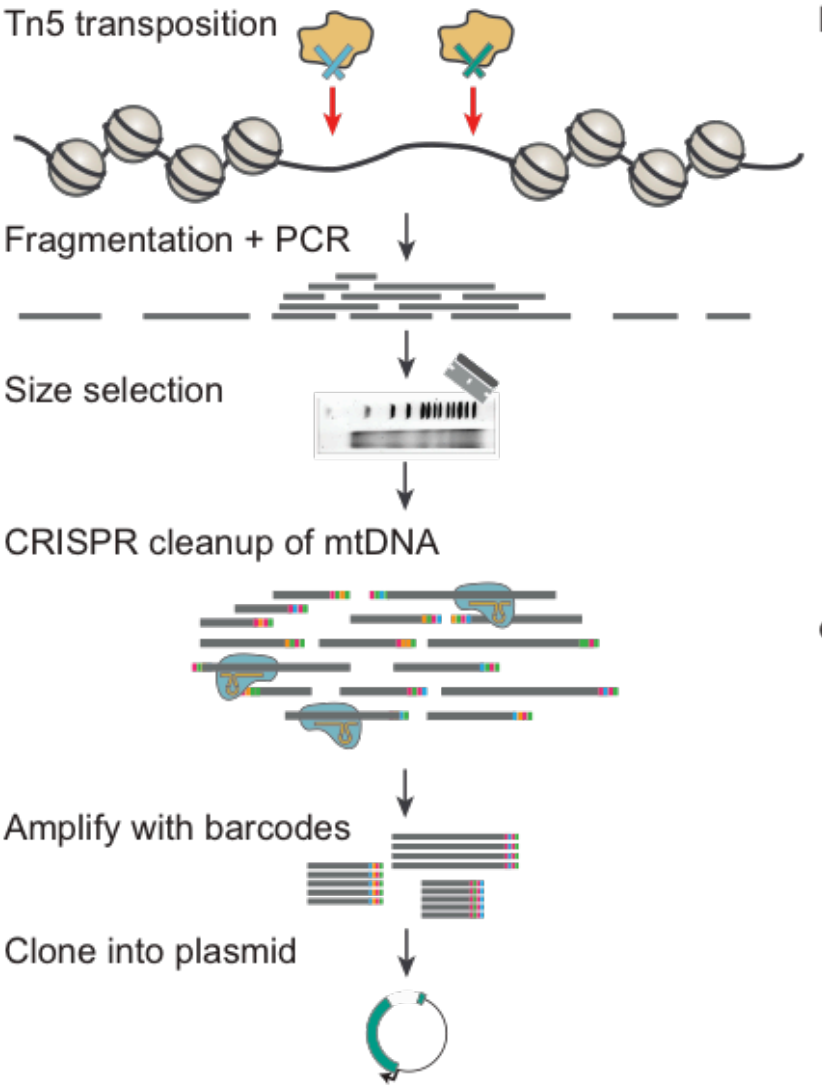


Figure 1: Overview of ATAC-STARR library construction method. The Tn5 transposase is first used to preferentially fragment genomic DNA at regions of open chromatin. Fragments are then size-selected on an agarose gel and mtDNA contamination is removed by selective CRISPR-Cas9 degradation. Fragment library is amplified by PCR and cloned into a enhancer reporter vector.

We sequenced the ATAC-STARR plasmid library to assess the library complexity and similarity with existing assays for open chromatin. We observed that qualitatively genome-wide signal tracks generated from ATAC-STARR fragments are highly similar to those generated by ATAC-seq and DNase I hypersensitivity mapping, and on a genome-wide basis the majority of open chromatin regions as assessed by DNase I hypersensitivity are captured as peaks in the ATAC-STARR library, suggesting that the complexity of a

traditional ATAC-seq and DNase I hypersensitivity experiment can be captured and cloned for use in ATAC-STARR (Fig. 2).

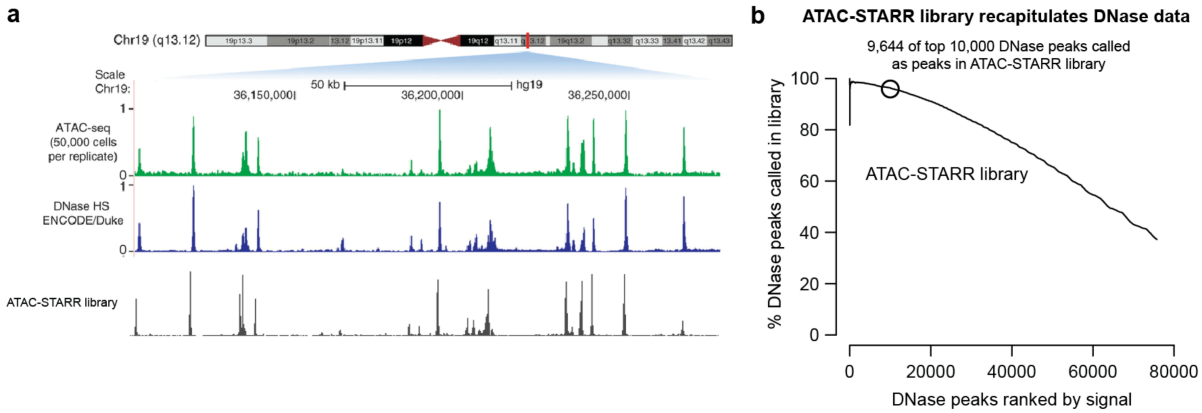


Figure 2: DNA fragments cloned into ATAC-STARR library recapitulate DNase I hypersensitivity and ATAC-seq library complexity and patterns. a. Top portion adapted from Buenrostro *et al.* (2013)[12]. ATAC-STARR library captures peaks detected by ATAC-seq and DNase I HS assays. b. Majority of top DNase I hypersensitivity peaks in GM12878 cell line are captured by fragments in ATAC-STARR plasmid library.

As ATAC-STARR fragments are cloned into the 3'UTR of a reporter gene, they drive self-transcription and the proportion of fragments observed in the transcribed RNA to input DNA represents a quantitative readout of enhancer activity. We therefore compared the abundances of fragments between our input control (non-transfected plasmid) and reporter gene RNA to determine the overall ability of ATAC-STARR to detect enhancer activity. As expected, we observed that biological replicates of plasmid and RNA samples are more similar to themselves than each other. When we compared RNA reads to the input plasmid, we observed a large abundance of fragments that were over-represented in RNA compared to plasmid, suggesting that many fragments are capable of driving reporter gene expression (Fig. 3).

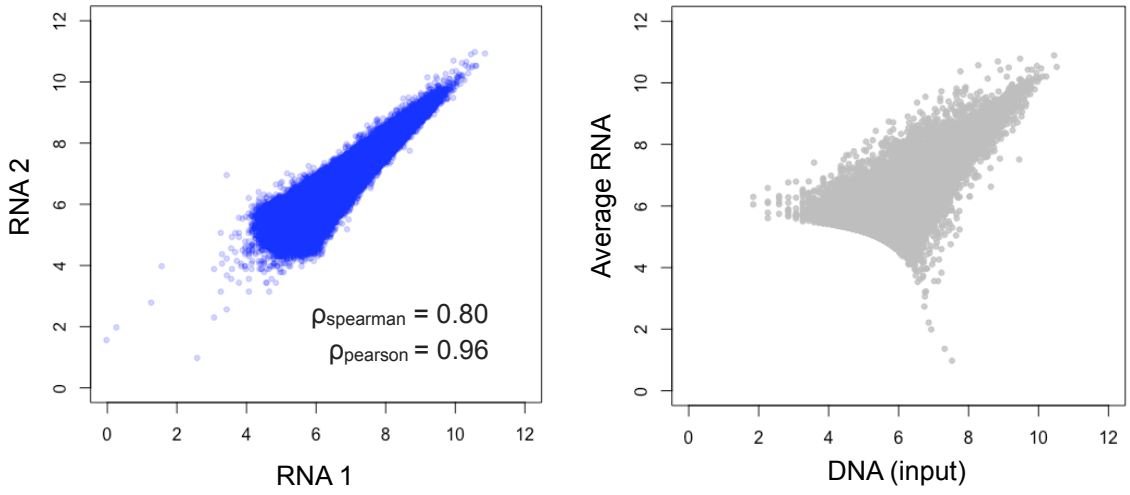


Figure 3: Correlation in ATAC-STARR fragment expression levels RNA and DNA. *Left,* Scatterplot of ATAC-STARR fragment abundance in two RNA biological replicates. *Right,* Scatterplot of ATAC-STARR fragment abundance between RNA and input plasmid DNA. Expression values are log2 transformed and only fragments with sufficient (>100 reads) representation are shown.

Identification of fragments with regulatory potential and size-dependent effects on activity

To identify fragments that drive reporter gene expression we compared the abundance of transcribed RNA reads to input plasmid reads using DESeq. At an FDR cut-off of 0.05, identified 537 fragments that were up-regulated in the collected RNA compared to input DNA. Among these up-regulated fragments, we observed a striking relationship between fragment length and regulatory activity, where longer fragments drove proportionally greater regulatory activity (Fig. 4). As ATAC-seq fragment length is related to the class of genomic element that generated the fragment (e.g. enhancers are enriched in fragments of length 50-150nt and 250-350nt and depleted between 150-250nt)[12], we adjusted for proportion of reads mapping to predicted enhancers and promoters in the genome and observed that the fragment length to activity relationship persists. Notably, most existing high-throughput enhancer activity studies use *in vitro* synthesized fragments with predicted enhancer lengths between 130nt and 200nt. We observed that fragments between 400 and 500nt in length are between 6-14 times more likely to drive enhancer

activity than those between 100 and 200nt, demonstrating the clear benefit of a STARR-seq approach that does not rely on *in vitro* synthesis for testing enhancer activity genome-wide (Fig. 4).

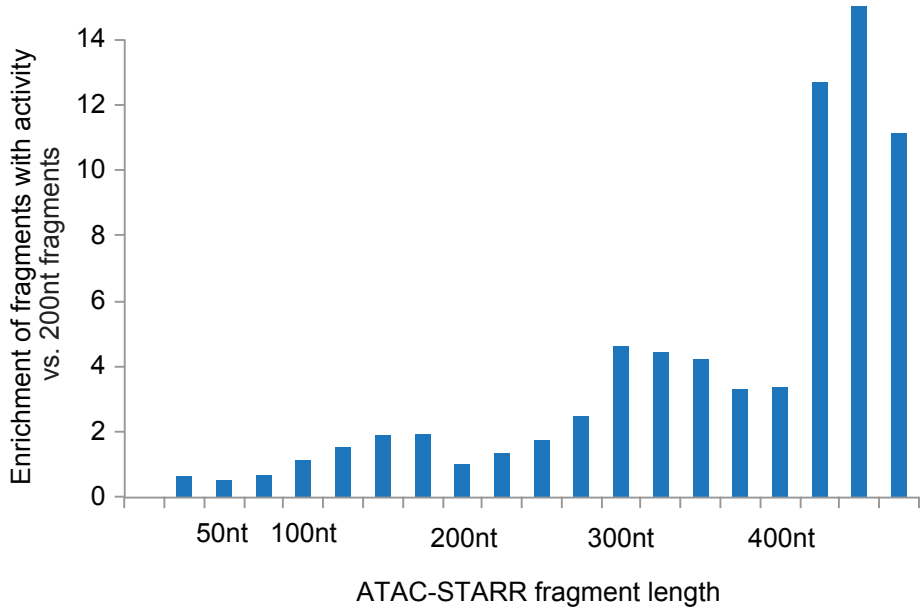


Figure 4: Activity of ATAC-STARR fragments split by fragment length. Activity calculated as the proportion of fragments in a given size range called as up-regulated (FDR<0.05), relative to fragments of length 200nt.

Transcriptionally active regions identified by ATAC-STARR are enriched for activating epigenomic marks

We surveyed the ATAC-STARR fragments that drive enhancer activity to assess whether they share common genetic or epigenomic characteristics. We first considered overlap of ATAC-STARR fragments with individual chromatin states - recurring combinations of histone modifications overrepresented in different classes of genomic elements (e.g. promoters, enhancers). We first considered the entire set of ATAC-STARR fragments in the input library. Consistent with previous ATAC-seq studies, these fragments are 19.82 and 5.98-fold more likely to be derived from promoters and active enhancers, respectively, compared to the entire genome (Fig. 5)[12]. We next considered the set of 537 activating ATAC-STARR fragments at FDR<0.05. We observed a modest enrichment of promoter fragments (1.89-fold enrichment vs. input), suggesting that some genomic

promoter elements are capable of driving reporter gene expression in the ATAC-STARR system. Notably, up-regulated fragments were substantially more enriched for enhancer elements (4.16-fold enrichment vs. input), consistent with previous reports that the STARR-seq assay is capable of detecting transcriptional enhancers[13] (Fig. 5). Finally, we noticed that fragments mapping to the "TSS Flanking Upstream" chromatin state were also enriched in up-regulated fragments, compared to the input library. The "TSS Flanking Upstream" chromatin state is characterized by presence of the activating promoter and enhancer histone modifications H3K4me1, H3K4me3, and H3K27ac, and shows enrichment between 400nt and 1kb upstream of annotated transcription start sites. However, the majority (61%) of these chromatin states are more than 5kb away from the nearest transcription start site, suggesting that some genomic regions with "TSS Flanking Upstream" annotations may also act as distal enhancers.

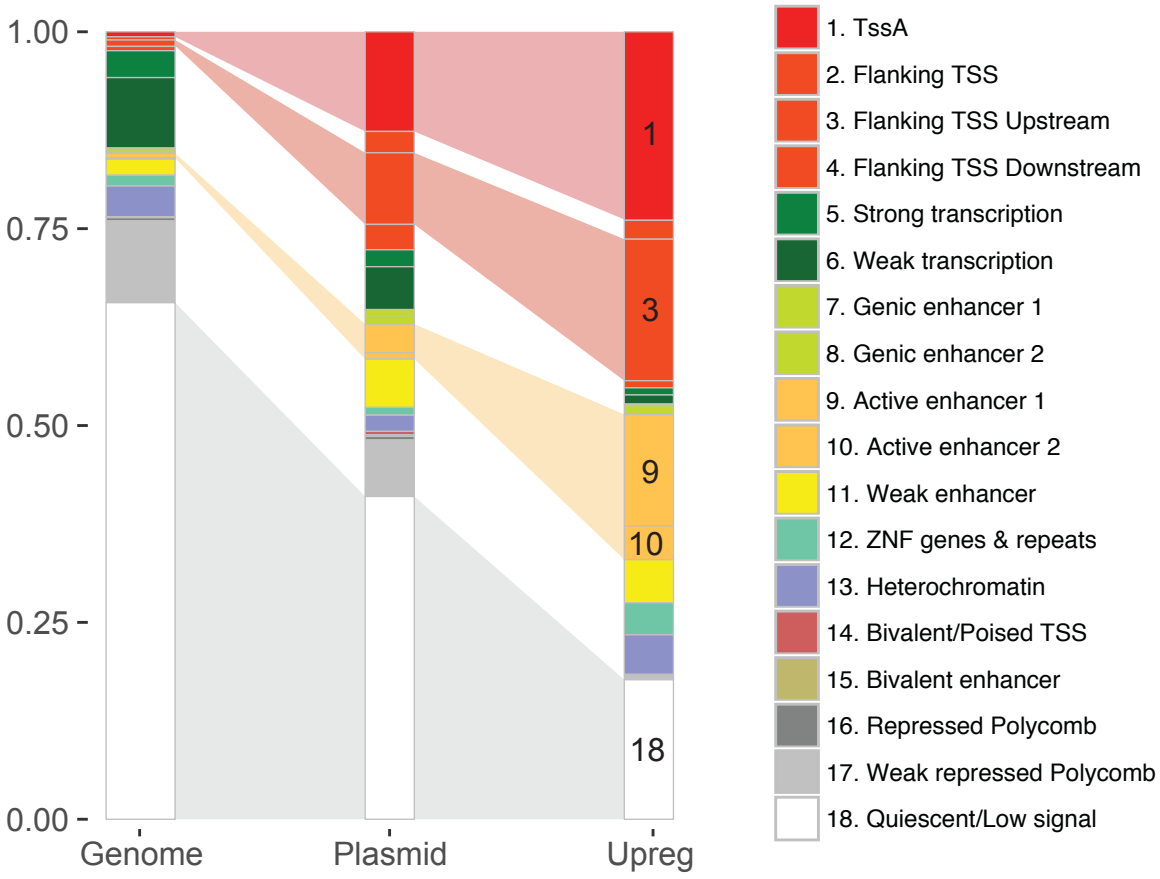


Figure 5: Up-regulated ATAC-STARR fragments are enriched in active regulatory regions. Proportion of different GM12878 chromatin states in the whole genome (left), input plasmid library (middle) and 537 up-regulated fragments (right). Genome annotations were obtained from the Roadmap Epigenomics Project using the 18-state model predicted with 7 histone modifications, and bar sizes correspond to percentage of nucleotide overlap [4]. Numbers in the right column mark representative chromatin states that are enriched in up-regulated fragments (states #1, 3, 9 and 10) and depleted (state #18).

We also explored the enrichment of individual histone modifications in the set of up-regulated ATAC-STARR fragments (Fig. 6). From the ENCODE project, we identified nine histone modifications or variants with ChIP-seq data available in the GM12878 cell line[14]. Consistent with the observation that promoter-derived fragments are modestly enriched in the up-regulated fragment set, enhancer-derived fragments are highly enriched in the up-regulated set and repressive fragments are under-enriched (Fig. 5), we observe that histone modifications enriched at enhancer elements are most enriched in

the up-regulated fragments, while H3K27me3, which marks repressive chromatin states is depleted in up-regulated fragments compared to the input library (Fig. 6).

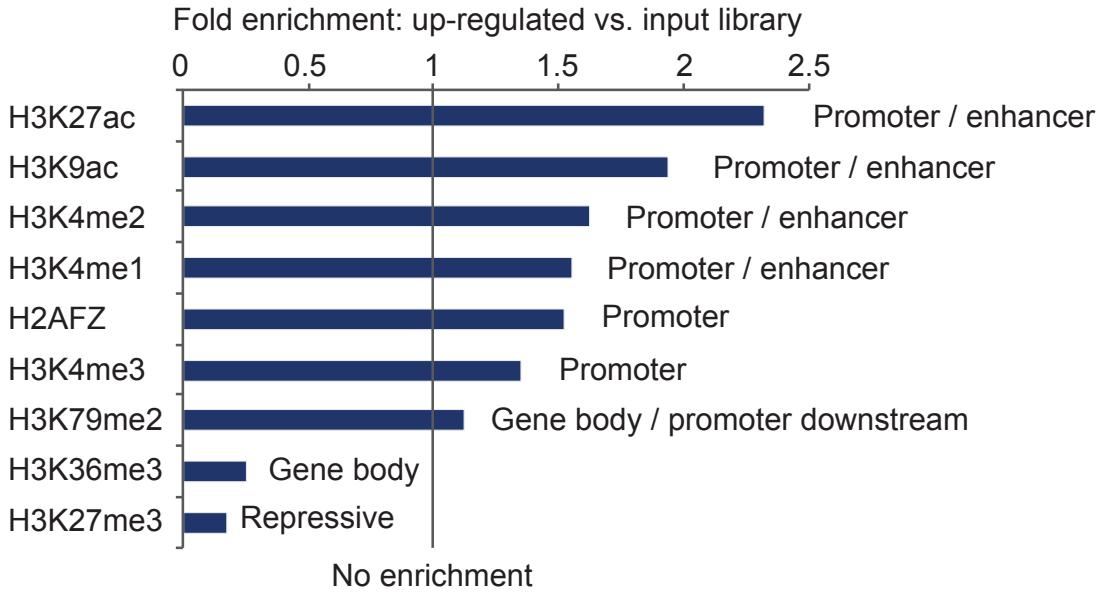


Figure 6: Enrichment of histone modifications and histone variants in up-regulated ATAC-STARR fragments vs. input library. Histone modification positions calculated using high-confidence ChIP-seq peaks identified from two biological replicates by the ENCODE project[14].

We hypothesized that the set of 537 activating ATAC-STARR fragments drive transcriptional activity through the binding of specific transcription factors. We therefore calculated the overrepresentation of transcription factor motifs in up-regulated ATAC-STARR fragments using 651 experimentally determined transcription factor motifs identified by the ENCODE project[14,15]. At FDR<0.05, we find that seven distinct motif families are enriched up-regulated ATAC-STARR fragments (Fig. 7). Notably, many of these motifs correspond to transcription factors expressed in the GM12878 cell line, including the IRF family of transcription factors, and to transcription factors that bind specifically in GM12878 compared to other human cell lines, including NFKB1 and RELA

[16,17]. Taken together, the enrichment of active chromatin states, activating epigenetic modifications, and transcription factor motifs indicate that ATAC-STARR is capable of identifying bona-fide transcriptional regulatory elements in the genome.

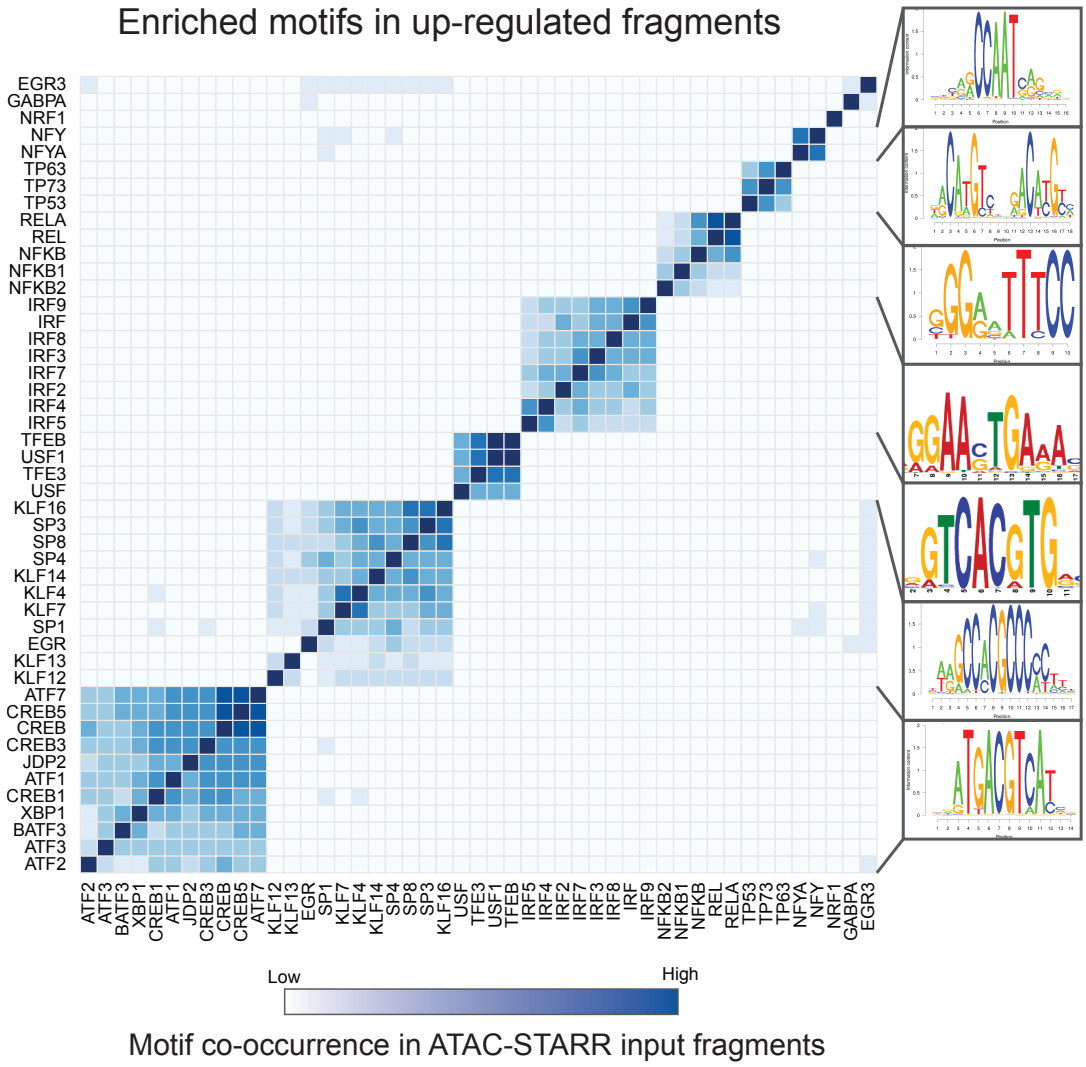


Figure 7: Enrichment of transcription factor motifs in up-regulated ATAC-STARR fragments. Motifs were identified using an FDR<0.05 threshold among 651 motifs identified by the ENCODE project[14,15]. As motifs can share strong sequence similarity, we used hierarchical clustering to group motifs together based upon their co-occurance in ATAC-STARR fragments. Darker cells correspond to greater co-occurrence between two motifs in the ATAC-STARR input library. Motif logos right of the matrix correspond to representative logos for each of the seven clusters.

High-resolution mapping of regulatory activity using ATAC-STARR

As the Tn5 transposase used to fragment DNA for the ATAC-STARR library preparation randomly inserts itself into the genome, we expect that for regions of the genome highly amenable to Tn5 digestion there will be many fragments generated with slight offsets in start and end site positions. We decided to leverage regulatory activity of these slightly offset DNA fragments to identify short sequences that are important for driving enhancer activity. As proof-of-principle, we focused on one region on chromosome 6 with a high density of overlapping ATAC-STARR fragments that show significant positive regulatory activity (Fig. 8). We observed a distinct peak in regulatory activity overlapping an 8-nucleotide POU2F2 motif and region of evolutionary conservation. From mining data from the ENCODE project, we also observed binding of POU2F2 at this motif through ChIP-seq in GM12878 and GM12891 cells[14].

Given the combination of a regulatory activity peak, evolutionary conservation and experimentally-observed POU2F2 binding at this site, we hypothesized that transcription factor binding at this motif would be sufficient to drive regulatory activity. Indeed, when we expanded the regulatory activity signals at this site to consider every fragment individually, we observed that 30 of 31 fragments overlapping the complete POU2F2 motif drove reporter gene transcription (Fig. 8). In contrast, fragments in the same general genomic region that did not overlap the POU2F2 motif did not show any robust ability for driving regulatory activity. Taken together, these results indicate that ATAC-STARR is capable of identifying DNA elements in a genomic locus critical to transcriptional regulation.

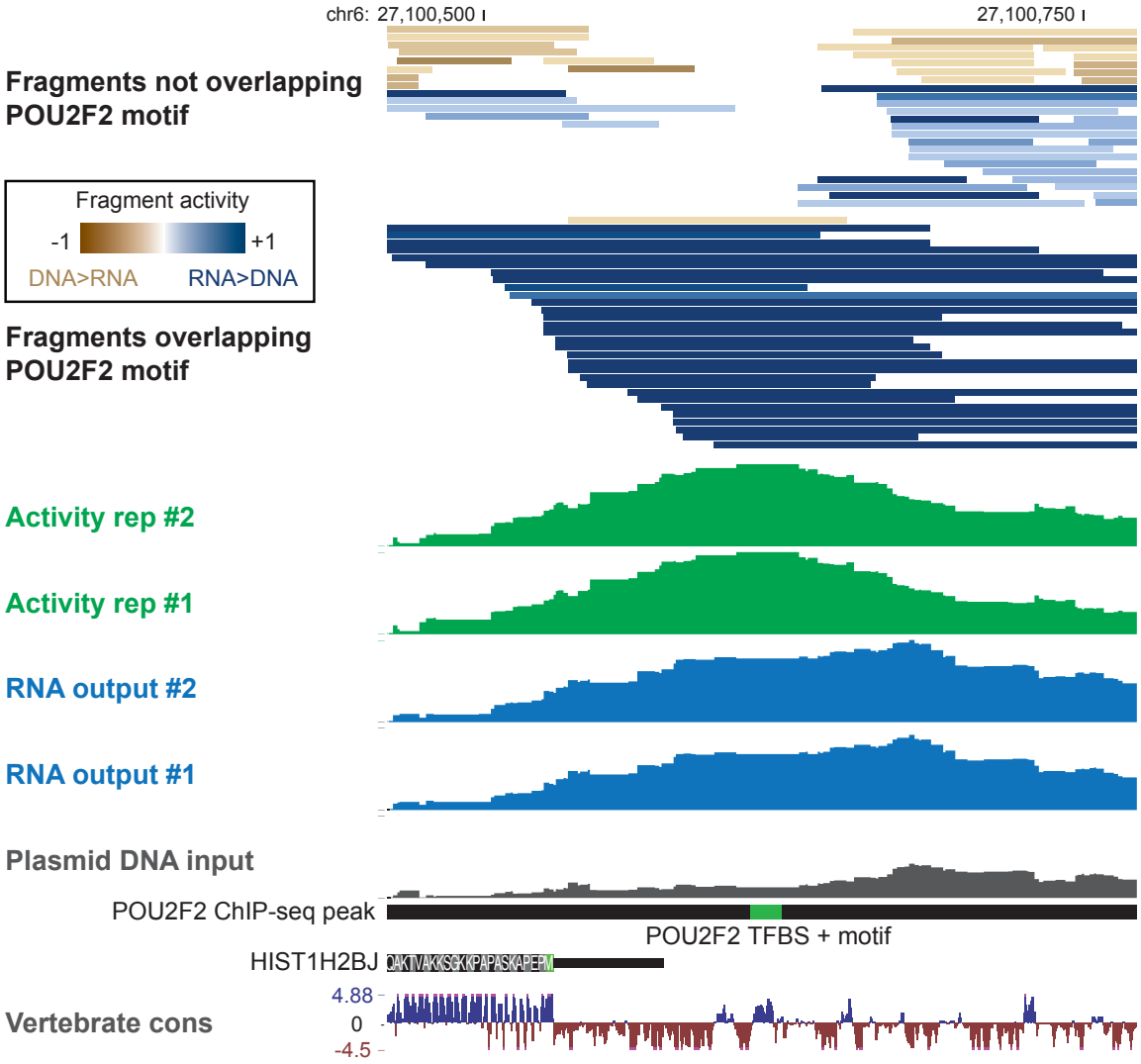


Figure 8: High-resolution mapping of regulatory activity. A locus with high density of up-regulated ATAC-STARR fragments was selected as a proof-of-concept to assess whether ATAC-STARR is capable of identifying short segments of DNA that drive regulatory activity. RNA outputs #1 and #2 correspond to two biological replicates for ATAC-STARR library transfection, and Activity rep #1 and #2 are calculated by (RNA-DNA)/DNA. Green box in POU2F2 ChIP-seq peak corresponds to canonical 8nt POU2F2 motif.

Discussion

Here, we present a high-throughput experimental assay, ATAC-STARR, to test the transcriptional regulatory activity for hundreds of thousands of DNA regions preferentially located in regions of open chromatin. We initially test ATAC-STARR in a pilot study using a library of sequences from the GM12878 lymphoblastoid cell line ranging from 25-500nt in length. We find that larger sequences in the 400-500nt range are substantially more likely to show activity in this experimental context, and that up-regulated ATAC-STARR fragments are significantly more likely to contain motifs for immune transcription factors and to be marked by activating histone modifications at their endogenous locus. Finally, we use the dense tiling of ATAC-STARR fragments to perform a high-resolution mapping of regulatory activity upstream of a promoter element to identify an 8nt evolutionarily conserved POU2F2 motif critical to regulatory activity.

While we performed our pilot study in the GM12878 cell line, the ATAC-STARR methodology can be readily applied to study the transcriptional regulatory architecture of any cell line. For cell lines with poor transfection efficiencies, a non-integrating lentiviral infection method can be used instead of transfection, as both approaches have shown highly similar results in other high-throughput reporter assays[18]. Moreover, while not explored in this study, ATAC-STARR can be used to identify genetic variants such as single nucleotide polymorphisms (SNPs) that alter regulatory element activity by mapping reads in an allele-specific manner[19,20]. This application of ATAC-STARR relies on the presence of different alleles for a given genetic variant being present in the input genomic DNA. The use of a previously genotyped cell line, such as GM12878, is preferable for such an application to more easily quantify allelic imbalance at genetic variants. As no human individual exists who is heterozygous at every common genetic variant, future studies can pool multiple cell lines together to generate an ATAC-STARR library heterozygous at more genetically important loci. This problem can be addressed if genomic DNA from a genetically heterogeneous pool of individuals is used. For example, to build an ATAC-STARR library heterozygous at many disease-associated SNPs for

autoimmune diseases, one can generate the library using pooled lymphoblastoid cells grown from individuals from distinct human populations first computationally chosen to maximize genetic diversity over the most relevant set of variants.

ATAC-STARR has several major advantages compared to other high-throughput reporter assay technologies such as the massively parallel reporter assay. First, *in vitro* synthesis fragments are limited to 230nt in length, while ATAC-STARR test DNA fragments up to 1kb in length, a limitation imposed by the cluster generation step of Illumina high-throughput sequencing platforms. Second, ATAC-STARR libraries can be an order of magnitude more complex than MPRA libraries, which are limited to 200,000 spots per synthesis run[8-10]. Finally, library construction in ATAC-STARR is an order of magnitude less expensive than MPRA, as ATAC-STARR bypasses a costly *in vitro* synthesis step.

One major limitation of ATAC-STARR is the requirement for using genomic DNA, while technologies involving *in vitro* synthesis can readily introduce changes to DNA not observed in the human population to better fine-map regulatory sub-regions of enhancers. However, the input fragment library in ATAC-STARR can also be modified to introduce non-existing mutations through error-prone PCR or introduction of mutagens during fragment amplification. Finally, ATAC-STARR can be coupled with a fragment capture technology to selectively test a subset of enhancers or promoters at higher resolution while retaining the advantages of having larger fragment sizes.

In summary, we present a pilot study for ATAC-STARR, a high-throughput method to assay the regulatory activity of tens of thousands of open chromatin regions located genome-wide. As ATAC-STARR can be readily applied to any human cell type, we envision this approach or similar technologies being used to quantify the transcriptional regulatory landscape of DNA sequences for a variety of human tissues.

Methods

ATAC-STARR library construction

We performed 16 ATAC-seq reactions on 50,000 GM12878 cells each using a modified protocol based upon Buenrostro *et al.*[12]. We performed cell collection, lysis, and Tn5 digestion as described by Buenrostro *et al.*, Tn5-fragmented DNA was cleaned up using a MinElute PCR purification kit (Qiagen #28004, four reactions per column eluted in 20uL EB buffer) and the resulting 80uL of eluate was split into 16 reactions of PCR. PCR was performed using custom primers (F: 5'-TAGAGCATGCACCGGCAAGCAGAAGACGGCATACGAGATNNNNNNCGGTCTCGTGGGCTCGGAGATGT-3', R: 5'-GGCCGAATTCGTCGATCGTCGGCAGCGTCAGATGTG-3') and NEBNext Ultra II Q5 DNA polymerase master mix (NEB #M0544L). PCR reactions were pooled and cleaned up with a Qiagen MinElute PCR purification kit (two PCR reactions per column eluted in 20uL EB buffer) and run on a 1% agarose gel stained with SYBR Gold (Thermo Fisher #S11494). Size selection of ATAC-seq fragments was performed by gel excision using a razor blade and DNA from gel slabs was purified using a MinElute Gel Extraction kit (Qiagen #28604) and eluted in 12 uL of warm buffer EB. The resulting size-selected ATAC-seq fragment library was treated with an anti-mitochondrial DNA CRISPR/Cas9 library following the protocol outlined in Montefiori *et al.* using 10X excess of Cas9 protein, and inactivated with 1uL of Proteinase K for 30min at 37C. We cleaned up the reaction with a Qiagen MinElute PCR purification kit and split into 16 PCR reactions for a second round of PCR using the same conditions and primers as described above to increase copy number of lowly represented DNA fragments. PCR products were cleaned up using two rounds of AMPure bead selection (0.7X for large molecular weight library, 0.8X for smaller library) to remove small fragments, eluted in 40uL of water and quantified using Qubit dsDNA HS Assay kit (Thermo Fisher #Q32854).

The pSTARR-seq_human plasmid used for generating the plasmid library was a gift from Alexander Stark (Addgene plasmid #71509). The linear backbone used for the subsequent

cloning steps was generated by digesting 4ug of circular pSTARR-seq_human for 4-6 hours with AgeI and SalI restriction enzymes (NEB #R3552S and R3138S), followed by gel excision and purification of a linear 3.5kb fragment corresponding to the human STARR-seq plasmid backbone. For each library, we performed 20 individual InFusion HD cloning reactions (Takara Bio #638911) using a 3.5:1 molar ratio of insert to vector backbone, following manufacturer's instructions. Reactions were collected and cleaned up using the Qiagen MinElute Enzymatic Reaction cleanup kit and transformed into twenty 20uL aliquots of MegaX DH10B T1R electrocompetent bacteria using the [electroporation conditions]. Recovered bacteria were grown in 2L of pre-warmed luria broth and 100ug/mL of carbenicillin, and serial dilutions were plated to estimate the number of clones in the library. Plasmids were collected from bacteria after growing overnight at 37C using the Plasmid Plus MegaPrep kit following manufacturer's instructions. Plasmid concentration was quantified using Nanodrop One (Thermo Scientific) and diluted to a 3ug/uL concentration for subsequent transfections. To ensure fragment library quality and diversity, a small aliquot of the fragment library was amplified by PCR using P5 and P7 primers and run on an Illumina MiSeq sequencer using the 50-cycle kit as per manufacturer's instructions.

Cell culture and transfections

GM12878 cells were obtained from the Coriell biorepository and grown in RPMI 1640 Medium with GlutaMAX Supplement (Thermo Fisher #61870127) with 15% fetal bovine serum (Sigma Aldrich #F2442) and 1% pen/strep at a density of between $2x10^5$ and $1x10^6$ cells/mL with regular media changes every 2-3 days. Approximately 24 hours before transfection GM12878 cells were split to a density of $4x10^5$ cells/mL to ensure the presence of actively dividing cells for increased transfection efficiency. For transfection, cells were collected for 5 min at 300g, washed once with pre-warmed PBS, and collected again for 5 min at 300g. PBS was aspirated and cell pellets were re-suspended in Resuspension Buffer R (Thermo Fisher #MPK10096) at a concentration of 7.5 million cells per 100uL. DNA was added to cells at a concentration of 5ug of plasmid per 1 million cells. In total, we transfected between 140-150 million cells per biological replicate

using the 100uL tips from the Neon Transfection System using the following conditions: 1200V with 3 pulses of 20ms. Transfected cells were immediately recovered in pre-warmed GM12878 media without antibiotic.

RNA isolation and cDNA generation

Cells were collected 24 hours post-transfection, washed twice in chilled PBS (spun for 5 min at 300g) and RNA was collected using the Qiagen RNEasy Maxi kit (Qiagen #75162) following manufacturer's instructions and performing the optional on-column DNase treatment step (Qiagen #79254). Poly A+ RNA was extracted from total RNA using the Oligotex mRNA Midi kit (Qiagen #70042), and any remaining DNA was digested with a second DNase treatment step using Turbo DNase (Thermo Fisher #AM2238) following manufacturer's instructions. Treated mRNA was cleaned up and concentrated using the Qiagen RNEasy MinElute Cleanup kit (Qiagen #74204). We generated cDNA from mRNA using Superscript III reverse transcriptase (Thermo Fisher #18080085) with a gene-specific RT primer located in the 3'UTR of the sgGFP reporter gene downstream from the inserted fragments (5'-CAAACTCATCAATGTATCTTATCATG-3'). Reverse transcription was performed following manufacturer's recommendations except with 2ug of poly A+ mRNA and 1uL of 12.5uM primer per 20uL reaction, and extension was performed for 60 minutes at 50C. Reactions were cleaned up using a MinElute PCR purification kit (Qiagen #28106, two reactions per column) and eluted in 15uL of pre-warmed buffer EB.

Library construction and high-throughput sequencing

We tested the number of cycles needed to amplify single-stranded cDNA by performing a 10uL qPCR reaction with a 2uL aliquot of cDNA, library preparation primers (F: 5'-AATGATACGGCGACCACCGAGATCTACAC[TAGATCGC]TCGTCGGCAGCGTC-3', R': 5'-CAAGCAGAAGACGGCATACGAGAT-3') and NEB Ultra II Q5 polymerase with 1X SYBR green. We also diluted plasmid DNA to a comparable concentration as ssDNA before amplification for final library preparation. We performed PCR in 50uL reactions using 10uL aliquots of cDNA so that the concentration of DNA and primer in

the larger PCR reaction was equivalent to the qPCR reaction. PCR cycle number was chosen to be the cycle before the cycle threshold was reached in qPCR (8 cycles). PCR reactions were cleaned up using Qiagen's MinElute PCR Purification kit. Each library batch (five transfected biological replicates, five plasmid controls) were sequenced on two flowcells on a NextSeq 500 machine using the 75-cycle kit as per manufacturer's instructions for 2x37 PE reads with 2x8nt barcodes.

Read mapping, data processing and enrichment analyses

Reads were split by barcode and aligned to the human genome (hg19) using bowtie2 v2.2.9. Alignment files were filtered to (i) remove reads mapping to chrM, (ii) select reads passing the -q 30 filter in samtools, and (iii) remove reads aligning to the ENCODE hg19 blacklist regions. Expression of each unique DNA fragment was calculated for both RNA and plasmid samples, and differentially expressed fragments were identified using DESeq2 at FDR<0.05.

Enrichment of up-regulated ATAC-STARR fragments in different genomic regions was calculated using BEDTools2 v2.19.0. BED files of motif annotations, histone modification ChIP-seq were downloaded from the ENCODE project website[14], and 18-state ChromHMM predictions for GM12878 were downloaded from the Roadmap Epigenomics Consortium website[4].

Acknowledgements

We are thankful to R. Tewhey for advice on library construction, RNA isolation, and library construction, E. Brown for advice on transfections, V. Subramanian for advice on library construction, and D. Sobreira for general experimental advice.

Author Contributions

X.W., M.C. and M.K. designed the project; X.W. performed ATAC-STARR library construction, transfection, RNA isolation and sequencing library preparation steps and all computational steps; S.M.G. performed CRISPR/Cas9 cleanup of mtDNA contamination.

References

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

2. O'Kane, C. J. & Gehring, W. J. Detection in situ of genomic regulatory elements in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.* **84,** 9123–9127 (1987).

3. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28,** 817–825 (2010).

4. Consortium, R. E. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).

5. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).

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

7. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457,** 854–858 (2009).

8. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell* **165,** 1519–1529 (2016).

9. Grossman, S. R. *et al.* Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proceedings of the National Academy of Sciences* **114,** E1291–E1300 (2017).

10. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* **30,** 271–277 (2012).

11. Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C. & Cohen, B. A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proceedings of the National Academy of Sciences* **109,** 19498–19503 (2012).

12. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods* **10,** 1213–1218 (2013).

13. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339,** 1074–1077 (2013).

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

15. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* (2013). doi:10.1093/nar/gkt1249

16. Xu, D., Zhao, L., Del Valle, L., Miklossy, J. & Zhang, L. Interferon regulatory factor 4 is involved in Epstein-Barr virus-mediated transformation of human B lymphocytes. *J. Virol.* **82,** 6251–6258 (2008).

17. Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523,** 486–490 (2015).

18. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity.

19. Wang, X. *et al.* Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. *eLife Sciences* **5,** e10557 (2016).

20. Kumasaka, N., Knights, A. J. & Gaffney, D. J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature Genetics* **48,** 206–213 (2016).

Chapter 4: Future directions

Identification of therapeutic targets using GWAS

Variants with weak effect sizes

Understanding the genetic basis of human disease is critical for understanding human biology and is being embraced as increasingly imperative for the development of new therapeutics. The discovery that most variants associated with genetically complex human traits have weak effect sizes and are preferentially located in non-coding regions has made the study of distal regulatory elements critical for interpreting the genetics of complex disease. When taken alone, the weak observed effect sizes of genetic variants identified by GWAS make a poor case for using GWAS as a starting point for therapeutics development. However, there are increasing numbers of examples where the genes identified by GWAS of a complex trait converge with those causal for the Mendelian forms of the trait. For example, common variation nearby the cardiac ion channel genes KCNQ1, KCNH2, SCN5A, KCNE1, and KCNJ2 is associated with QT interval length by GWAS, and rare coding variants in these genes are causal for long QT syndrome[1]. Similarly, common and rare variation converge for Proprotein convertase subtilisin/kexin type 9 (PCSK9) in LDL cholesterol GWAS and familial hyper- and hypocholesterolemia, and therapeutic inhibition of PCSK9 is effective for lowering LDL levels in patients[2-4]. The convergence in genes implicated by GWAS and family-based studies suggest that the two approaches capture genetic variants on different ends of an allelic series, and that stronger perturbation of GWAS target genes with weak effect sizes can have biologically meaningful phenotypic consequences. Our investigation of the *Bves* sub-threshold locus in Chapter 2 supports this view - mouse genetic knockouts and zebrafish morpholino knockdowns of *Bves* give rise to an organismal electrophysiological phenotype while the sub-threshold SNP independently has a modest odds ratio [5].

Causal gene identification from a pathogenic non-coding variant

As discussed in Chapter 2, the systematic identification of gene targets for non-coding regulatory elements remains an unsolved problem. Physical mapping of chromatin interactions has revealed that transcriptional enhancers regulate gene targets up to 1 megabase away, complicating the identification of regulatory interactions[6]. Many

techniques have been used to identify potential gene targets of enhancers. Computational solutions to this problem have been proposed, including the selection of targets by genomic proximity or the identification of enhancer-gene pairs that show correlated activity patterns across different human tissues[7]. However, these computational predictions have high false positive and false negative rates, complicating efforts to use the results as a starting point for extensive follow-ups on predicted gene targets. Genetic methods for assigning gene targets involve identifying expression Quantitative Trait Loci (eQTLs), where the expression of a gene consistently varies between with individuals differing in genotype at a nearby genetic variant[8]. While this approach in theory has the capability to identify all genes affected by a particular genetic locus, eQTL studies to date have been underpowered (~200-300 individuals) and lack resolution within a genetic locus composed of multiple variants in high linkage disequilibrium[8]. Many genetic loci implicated by GWAS currently do not have eQTL gene targets, and targets implicated by eQTLs are not necessarily the causal genes at a GWAS locus, limiting the effectiveness of eQTL studies for gene target identification.

Experimental approaches may represent the clearest options for linking enhancers to nearby genes. Chromatin conformation capture technologies offer the ability to identify physical interactions between regions of the genome, and can be performed on a global level using Hi-C and ChIA-PET methods[6,9,10]. However, the presence of a physical interaction between a predicted enhancer and a promoter does not necessarily imply a causal regulatory relationship between these two elements. Moving forward, the most conclusive evidence of regulation may come from perturbations of predicted enhancer elements through genome editing, followed by measurement of changes in expression of nearby genes. While the eQTL approach for gene target discovery is confounded by the presence of multiple genetic variants in a haplotype block, genome editing using for example CRISPR/Cas9 allows for a precise modification of the sequence at a genomic region. Currently genome modification using CRISPR/Cas9 cannot be applied in a high-throughput method to many non-coding regulatory regions while still tracking

expression of nearby genes, however single-cell experimental technologies may make this a possibility in the future.

Expanding the collection of epigenomic maps

Another important insight from the work presented in this thesis concerns the heterogeneity of enhancer elements active in a given tissue. Current studies integrating epigenomics and human genetics use either the entire set of ChromHMM-predicted enhancer elements from a cell type, or the set of all peaks for an individual histone modification to predict the influence of common genetic variation on transcriptional regulation[11-13]. Our results in Chapter 2, however, suggest that there may be heterogeneous subgroups of enhancer elements in a given tissue, and that accounting for this variability can improve enrichment and functional interpretation of GWAS loci. Moving forward, it will be important to generate more detailed enhancer annotations for each human tissue that stratify predicted enhancers by functional annotations. Moreover, mapping the activity of enhancer subgroups after exposure to different environmental stimuli can also reveal important functional mechanisms for disease variants. For example, recent studies have identified SNPs that influence gene expression only in certain environmental contexts[14,15], suggesting that causal variants for GWAS may be missed by epigenomic variant interpretation approaches if the enhancer profile is generated from cells in the wrong environmental condition. To generate more comprehensive maps of enhancer elements, it may be valuable to perform epigenomic mapping of primary human tissues from phenotypically heterogeneous individuals at different stages disease progression.

Single cell epigenomic profiling technologies offer a promising option for mapping disease-relevant cell types moving forward[16]. The pathogenesis of some diseases may be driven in part by cell types that exist at low frequencies within a tissue. Unique enhancer elements present in these cell types will often be missed by conventional whole-tissue epigenomic mapping approaches, as epigenetic signatures arising from more prevalent cells will conceal signals from these lowly represented enhancers. Clustering of single cell

epigenomic data can reveal these rare sub-populations, which can then be analyzed and studied separately without worry of the signal being drowned out by more prevalent cell types.

High-throughput interrogation of transcriptional regulatory elements

Finally, one major next step for genome-wide assays of transcriptional regulatory activity will be the perturbation of regulatory DNA in the native chromatin context of the human genome. Recent studies have performed high-density endogenous editing of specific loci in the genome, identifying variants associated with cell viability or transcription of an integrated reporter gene[17,18]. Moving forward, we expect similar approaches to be applied for identifying non-coding DNA critical in a variety of human diseases by using the CRISPR-Cas9 system to systematically edit every enhancer or motif for important transcription factors, and identify cells displaying pre-determined disease phenotypes. Assimilation of genome-wide nucleotide-resolution maps of regulatory element activity in a variety of environmental conditions with maps of genetic variation associated with human disease will form the basis of the future studies of human genetics and lay the groundwork for a comprehensive understanding of every disease association in the genome.

References

1. Arking, D. E. *et al*. Genetic association study of QT interval highlights role for calcium signaling pathways in myocardial repolarization. *Nature Genetics* **46,** 826–836 (2014).

2. Global Lipids Genetics Consortium *et al*. Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45,** 1274–1283 (2013).

3. Cohen, J. C., Boerwinkle, E., Mosley, T. H., Jr & Hobbs, H. H. Sequence Variations in PCSK9,Low LDL, and Protection against Coronary Heart Disease. *N Engl J Med* **354,** 1264–1272 (2006).

4. Abifadel, M. *et al*. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. *Nature Genetics* **34,** 154–156 (2003).

5. Wang, X. *et al*. Discovery and validation of sub-threshold genome-wide association study loci using epigenomic signatures. *eLife Sciences* **5,** e10557 (2016).

6. Li, G. *et al*. Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell* **148,** 84–98 (2012).

7. McLean, C. Y. *et al*. GREAT improves functional interpretation of cis-regulatory regions. *Nat Biotechnol* **28,** 495–501 (2010).

8. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).

9. de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev*. **26,** 11–24 (2012).

10. Lieberman-Aiden, E. *et al*. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326,** 289–293 (2009).

11. Trynka, G. *et al*. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics* **45,** 124–130 (2013).

12. Gusev, A. *et al*. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *American Journal of Human Genetics* **95,** 535–552 (2014).

13. Finucane, H. K. *et al*. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* **47,** 1228–1235 (2015).

14. Zhernakova, D. V. *et al*. Identification of context-dependent expression quantitative trait loci in whole blood. *Nature Genetics* **49,** 139–145 (2016).

15. Moyerbrailean, G. A. *et al*. High-throughput allele-specific expression across 250 environmental conditions. *Genome Res*. **26,** 1627–1638 (2016).

16. Buenrostro, J. D. *et al*. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523,** 486–490 (2015).

17. Sanjana, N. E. *et al*. High-resolution interrogation of functional elements in the noncoding genome. *Science* **353,** 1545–1549 (2016).

18. Rajagopal, N. *et al*. High-throughput mapping of regulatory DNA. *Nat Biotechnol* **34,** 167–174 (2016).