Identification, Interpretation, and Evolution of Gene Regulatory Enhancer Landscapes

By

Mary Lauren Benton

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in

Biomedical Informatics
August 31, 2020
Nashville, Tennessee

Approved:

John A. Capra, Ph.D.

Emily Hodges, Ph.D.

Jacob J. Hughey, Ph.D.

Antonis Rokas, Ph.D.

Douglas M. Ruderfer, Ph.D.

To my family, for their infinite love and support.

Acknowledgements

This work would not have been possible without the support and guidance from my mentor, Dr. Tony Capra. I am grateful to Tony for the scientific discussions and numerous rounds of manuscript revisions, but more importantly for teaching me what it means to be a scientist. I also want to thank my committee and many collaborators from the past five years for providing thoughtful feedback throughout each of my projects. I want to give a specific thank you to Doug Ruderfer for the ten minutes of jokes at our Wednesday meetings and for asking the tough questions.

I also want to extend a special thank you to all the Capra Lab members, past and present, for many hours of scientific discussion, daily debugging sessions, and numerous rubber-duck and balloon related shenanigans. You kept me sane and kept me excited about science— no easy feat. To my lab twin, Laura Colbran, I'm so happy we got to sit within a few feet of each other every day for all of graduate school. I'd have to write a novel to thank you for everything, so I'll just say #MaryLaura4ever.

To my friends from my QCB beginnings—Andy Perreault, Corey Hayford, and Ian Setliff—I will be eternally grateful. Thanks for always being there for the many concerts, lunches, coffee walks, and trips to get smoked meats. You all reminded me that there is a life outside the lab, even it's still fun to talk science on the weekends. To Marcus Wild, thank you for being my constant in a weird 2020; drinks at the Devil's Advocate might be the best decision I ever made. And to my parents, you raised me to be the person I am today and have supported me through each new endeavor. Thank you for believing in me, giving me the confidence to pursue my dreams, and always answering the phone when I have questions about adulting.

Finally, I am grateful to the Biomedical Informatics training grants, T32LM012412 and T15LM007450, that financially supported this work.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction

Motivation

The improper regulation of gene expression plays an important role in the etiology of complex disease. The vast majority of loci associated with disease in genome-wide association studies (GWAS) are in non-protein-coding regions of the human genome, and many have been shown to contribute to disease risk by altering gene regulatory elements such as enhancers. In addition to single nucleotide variants, larger structural variants can also disrupt gene regulatory mechanisms and lead to disease by causing a loss of enhancer function or substantial changes to the regulatory architecture of the genome. Despite numerous examples of enhancers that are mutated in disease, predicting whether mutations in a given enhancer will influence phenotype is still a difficult task. This is partially because we still lack a comprehensive set of genome-wide enhancer annotations. Currently, there are many approaches to identify enhancer sequences on a genome-wide scale but no gold-standard validation set. This complicates the use of enhancers in biomedical research. Furthermore, strategies for interpreting enhancer mutation consider enhancers in isolation, despite evidence that redundancy in insect and mammalian enhancer landscapes buffers the phenotypic effects of enhancer loss on the expression of genes. We do not fully understand the effects of enhancer identification or enhancer landscapes on gene regulation and disease. This work highlights the key limitations in our understanding putative enhancer annotations, defines a integrative model of gene regulatory architecture, and quantifies the impact of genetic variation on enhancer landscapes.

Characteristics and Identification of Gene Regulatory Enhancers

Enhancers are traditionally defined as short genomic sequences that regulate the transcription of one or more target genes irrespective of distance or orientation from the target[1–3]. These serve as one component of the gene regulatory architecture of the cell, which also includes elements such as promoters and

insulators. Enhancers have previously been associated with a range of biochemical, functional, and sequence attributes that frequently serve as the primary definition of an enhancer element or are integrated into computational models. This section will highlight the genomic features, computational models, and experimental data that have previously been used to identify enhancers.

Sequence features, biochemical signatures, and functional attributes of enhancers

Enhancers regulate the expression of target genes through binding of specific transcription factors. As a result, active enhancers localize in regions of the genome not occupied by nucleosomes. Open chromatin is assayed using experimental techniques to identify features such as sensitivity to DNase I nuclease using DNase I sensitivity (DNase-seq)[4,5], nucleosome depletion using Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq)[6,7], or accessibility to transposase using transposase Tn5 mediated Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq)[8]. Each of these techniques relies on high-throughput sequencing of the resulting accessible DNA fragments that are then aligned to a reference genome. The alignment creates regions of enrichment for sequencing reads, or "peaks", where transcription factors are able to bind[4–8]. Regions of open chromatin are often used alone or in conjunction with other features as a signature of enhancer elements[9–12].

More directly, binding assays for known transcription factors (TFs)[13] and enhancer associated proteins, such as the histone acetyltransferase p300[14–16], have been used to identify enhancer elements[2]. An experimental protocol known as chromatin immunochromatin precipitation followed by sequencing (ChIP-seq) can successfully locate bound TFs genome-wide[17]. In ChIP-seq, protein-DNA complexes are cross-linked to preserve the interaction, then the DNA is broken up and the fragments of interest are precipitated using specific antibodies and sequenced. Much like the assays to detect open chromatin, ChIP-seq yields "peaks" of enrichment in genomic locations where the relevant marker was bound. ChIP-seq results in a map of genomic locations bound by the TF of interest at the time the assay was performed[17]. However, this approach can be prohibitively expensive and time-consuming when probing for multiple relevant factors. Furthermore, due to the high false-positive rate for enhancer prediction from

TF ChIP-seq, evidence of binding in a ChIP-seq assay is not sufficient to confirm enhancer activity[18]. False-positives may result from the inability to assay the combinatorial binding patterns required for enhancer activity and because it is difficult to rule our ChIP-seq peaks resulting from transient interactions[18–20].

Where TF binding has not been experimentally determined, *de novo* identification of TF binding motifs using computational methods or calculation of enrichment for computationally derived motifs and known motifs of interest can serve as a proxy[2,18,21]. However, while the presence of relevant TF binding motifs suggests potential enhancer function, it is not a guarantee of activity. Successful binding of TFs is often context-dependent and the contribution to enhancer activity may rely on successful interactions with other factors in the flanking sequence. A study in mouse adipocytes showed that the activity of enhancer sequences relied on the interactive binding of dozens of TFs in the surrounding sequence[19]. Furthermore, the presence of a motif does not necessitate binding[18]. Another recent study integrating TF binding motifs with ChIP-seq data in K562 cells reported an average of 430 unbound motifs to one bound motif, suggesting that binding is rare compared the number of predicted TF motifs[22].

Early studies of enhancers often focused on evolutionarily conserved regions of the genome[23–26]. Many enhancers, especially those involved in development, are conserved; therefore, sequence conservation is still used as a metric to rank and validate putative enhancers. It is important to note that many conserved sequences do not function as enhancers, and many active enhancers are not conserved. Indeed, recent work suggests that activity of individual enhancer elements evolves rapidly across mammalian species[27].

Combinations of specific post-translational modifications on the histone proteins of surrounding nucleosomes are also associated with enhancer activity. ChIP-seq is used to detect key histone modifications and establish putative enhancer locations, including monomethylation of lysine 4 on histone H3 (H3K4me1) and acetylation of lysine 27 on histone H3 (H3K27ac). These two biochemical modifications are thought to mark active enhancers, both separately and in combination[2,14,28]. The trimethlyation of lysine 4 on histone H3 (H3K4me3) is also frequently used to exclude H3K27ac regions

with suspected promoter activity from putative enhancer sets[2,29]. Recent studies have suggested other markers to identify enhancers or subtypes of enhancers, although these are less commonly invoked[30–33]. For example, where the repressive H3K27me3 mark coincides with H3K4me1 marked enhancer regions are often considered bivalent or poised enhancers, suggesting a mechanism for finer control of enhancer activation[30].

Finally, recent work describes a subset of enhancers that are transcribed into characteristic bi-directional enhancer RNAs (eRNAs)[34–39]. The FANTOM5 consortium used a technique called cap analysis of gene expression followed by sequencing (CAGE-seq) to quantify and map eRNAs across a broad set of tissues and cell lines[34]. These eRNA-predicted enhancers validate at a high rate (~70%) relative to those predicted by other genomic properties. However, CAGE likely misses potential enhancers because eRNAs are unstable and quickly degraded. Other approaches focus on sequencing nascent RNA transcripts to more efficiently detect unstable eRNAs. These include global run-on sequencing (GRO-seq)[40] and precision run-on sequencing (PRO-seq)[38,39], as well as GRO-cap[36], PRO-cap[38], and native elongating transcript-cap analysis of gene expression (NET-CAGE)[37] which include additional capture steps to improve specificity. For example, the NET-CAGE approach identified over 20,000 novel enhancer candidates in humans compared to CAGE-seq, which were broadly enriched for markers relevant to enhancer function[37]. Despite recent methodological improvements, there are limitations to the eRNA enhancer identification approach. The bi-directional transcription pattern is not exclusive to enhancers[41] nor are all functional enhancer elements bidirectionally transcribed[42,43]. Furthermore, the numerous molecular functions proposed for noncoding eRNAs are still not well understood[35,43] and will require continued experimental characterization to fully resolve.

Limitations of using biochemical and sequence attributes for enhancer identification

While informative, none of these attributes are comprehensive, exclusive to enhancers, or completely reliable indicators of enhancer activity. Because enhancers are context- and stimulus-dependent, it is difficult to assay a complete set of enhancer elements genome-wide. Approaches that rely on the

identification of eRNAs, TF binding, or histone modification signal only capture a snapshot of the regulatory architecture at the time of the assay in that particular biological context. Enhancers identified using eRNAs are a particularly limited set of regions; previous work has demonstrated that they fail to capture all of the sequences with validated activity in transgenic assays, suggesting that transcribed enhancers do not form the entire regulatory landscape[34]. While capturing a larger proportion of genomic sequence, enhancer identification based on biochemical and sequence features is prone to false-positive predictions of enhancer sequences because these are not exclusive to enhancer regions[33,44,45]. Instead, there is likely to be a spectrum of genetic elements, including promoters and insulators, that share similar attributes[46–48]. Indeed, recent experimental work blurs the distinction between enhancers and promoter elements. Enhancers across the genome can act as promoters, and a fraction of promoters can act as distal enhancers in some contexts[49,50]. The inability to defined a single 'histone code' to identify enhancers through combinations of histone modifications is another a notable example of this complexity[2,30,33,44]. H3K27ac marks both active promoters and enhancers, despite numerous studies using the single mark as the definition of an enhancer[29,51–55]. Another frequently cited enhancer mark, H3K4me1, has been discovered in regions without demonstrated enhancer activity[30,44]. Other papers highlight novel histone modifications correlated with enhancer states (H3K64ac, H3K122ac, H3K79me3, and H4K16ac), noting that these can mark active enhancer regions lacking the H3K27ac mark[31–33]. Low validation rates (20-33%) in previous experimental validations of putative enhancers using small-scale transgenic assays, suggest that definitions of enhancers based on combinations of histone modifications alone have low specificity[13,34]. Furthermore, the impact of bias and technical limitations of current functional genomics assays on enhancer identification is not completely understood[56].

As implied by the limitations of the genomic attributes correlated with enhancer activity, the biological characteristics of the gene regulatory architecture complicate enhancer identification. First, the lack of enhancer activity in one cell type or cellular condition is not sufficient to rule out potential regulatory activity[28,57]. A number of case studies describe enhancer regions with activity in only one tissue or developmental time point[28,57]. Contrary to the switch-like model of enhancer activity, enhancers

have been shown to occupy multiple intermediate states between inactivity and activity referred to as 'poised', 'primed', or 'latent' enhancers[2,30,45,58]. Genetic variation between individuals has been shown to alter epigenetic modifications and enhancer activity which may confound the generalizability of identification approaches built using these features[59,60]. However, the proportion of epigenetic modifications that are variable across individuals is estimated to be small on (1–15%)[61].

Computational approaches for enhancer identification

Due to the large number of biological features indicative of enhancer activity combined with advances in machine learning techniques, many computational approaches have been developed to identify enhancers[10,11,62–76]. These can be stratified into two groups: (1) unsupervised or semi-supervised approaches that segment the genome into functional states, and (2) supervised approaches that classify sequences as either an "enhancer" or "non-enhancer" based on labeled training data.

Unsupervised approaches for enhancer identification often rely on chromatin segmentation approaches such as the popular ChromHMM, Segway, EpiCseg, or GenoSTAN[72,74,75]. These models use hidden Markov models or dynamic Bayesian networks to assign each genomic region to a specific functional state based on the combination of input features at that location. While the number of functional states much be specified in advance, these models can integrate a wide range of biochemical features to determine the segmentation. Following the segmentation, human experts can assign states to biological annotations such as "enhancer" or "promoter" using the level of enrichment for input markers in each state[69]. Recently, however, additional machine learning models have been proposed to automate this process[76]. Unsupervised approaches do not require labeled training data to make predictions, which is useful given the small fraction of validated enhancer sequences and the challenges involved in defining appropriate positive and negative training sets[72]. Furthermore, these methods can be leveraged to simultaneously predict multiple types of functional annotations, including different subtypes of enhancers[62,63].

Supervised classification algorithms used sets of known enhancer and non-enhancer regions to learn the features that distinguish enhancers from other functional regions and genomic background. These can range from simple rule-based intersections of markers associated with enhancers to more complex machine learning frameworks. The simple, but widely adopted, methods use intersections of histone modifications and other genomic annotations to identify enhancers[12,14,28–30,44,77–81]. For example, the co-occurrence of H3K4me1 and H3K27ac, or the presence of H3K27ac without H3K4me3 are often labeled as putative enhancers[14,28,29,44,80,81]. Since enhancers are typically defined as distal regulatory regions that do not contain protein-coding sequence, these simple intersections can be filtered to exclude exonic sequences and regions close to a transcription start site (TSS)[10–12,77,82]. However, enhancers have been reported in both coding sequences and intronic regions nearby genes which will be incorrectly classified by these rule-based approaches[2,83]. More recently, the application of supervised machine learning frameworks for enhancer prediction have become popular. These classification models are trained on similar input data as the unsupervised or rule-based approaches—histone modifications, regions of open chromatin, transcription factor binding motifs, and other genomic annotations—but include explicit positive and negative labels on the training data[10,11,64–68,70,71]. While these methods use a range of underlying statistical models, each one learns higher order patterns that allow the model to best classify the sequence or region by enhancer status. As the amount of available training data and computational resources continue to grow, a new wave of deep learning models for enhancer prediction suggest that increasingly complex modeling approached may improve our ability to efficiently and accurately identify enhancers *in silico*[84–89].

Limitations of computational approaches for enhancer identification

Although computational enhancer prediction approaches are widely used, it remains difficult to accurately quantify their performance. Due to the lack of a gold standard enhancer set, predictions are validated through both low- and high-throughput transgenic reporter assays or overlap with other attributes correlated with enhancer activity. However, as discussed previously, attributes such as evolutionary

conservation, proximity to genes, or the presence of trait-associated genetic variation are not comprehensive or conclusive evidence of enhancer function. Biases in the validation attributes themselves may prevent an accurate assessment of model performance. In addition, interpreting the statistics that underlie the computational enhancer predictions is nontrivial. Despite advances in model interpretation[90], machine learning algorithms largely remain 'black boxes' making it difficult to obtain generalizable biological and mechanistic insights into the regulatory architecture from even well-performing models[91].

Experimental approaches for identification and validation of enhancer function

For many years, experimental identification and validation of enhancers was limited to low-throughput assays in cell lines and transgenic embryos. In these transgenic reporter assays, the putative enhancer sequence is incorporated into a bacterial plasmid upstream of a minimal promoter and reporter gene[2,92]. If the sequence activity is able to drive expression of the reporter gene, the sequences is considered an enhancer. While informative and still commonly used, transgenic reporter assays are time consuming and require all sequences are known in advance[2]. However, the recent development of sophisticated high-throughput methods allows for experimental validation of enhancer function on a much larger scale.

Massively parallel reporter assays (MPRAs), are promising approaches to generate genome-wide enhancer maps with demonstrated activity in a given cellular context[93]. In MPRAs, reporter constructs containing the putative enhancer sequences include unique DNA barcodes to create libraries with thousands of barcode-labeled plasmids. These constructs can all be tested in a single experiment, validating the activity of thousands of sequences at one time. Recent studies have used MPRA protocols extensively to assess the regulatory potential of a large number of candidate regions, to characterize the dynamics of enhancer evolution[94], and to quantify the impact of genetic variation on enhancer activity[62,95–103]. Although computationally predicted enhancers validate at a relatively low rate in MPRAs (~26%), the results do corroborate some previously described enhancer attributes[98]. Active sequences in MPRAs are enriched for regulatory elements in DHSs and evolutionary conservation, and co-occur with the expected histone modifications and relevant TF binding motifs[47,93,96].

Self-transcribing active regulatory region sequencing (STARR-seq), an MPRA variation originally described in Drosophila, has also recently been applied to human enhancer sequences[104–106]. STARR-seq allows the activity and strength of the enhancer to be directly quantified by incorporating the enhancer sequences downstream of the minimal promoter. Thus, the enhancer sequence itself transcribed when the enhancer is active and eliminates the need for additional barcodes[104]. An additional benefit of this protocol is that it allows for scans for enhancer activity based on libraries of randomly fragmented DNA sequences. This 'shotgun' method sidesteps the need to synthesize libraries of known candidate sequences and instead can assay enhancer activity of short sequences genome-wide[102,104,106].

Limitations of experimental approaches for enhancer identification

Despite the recent technical advances in experimental enhancer identification, there remain limitations that preclude the adoption of a gold standard enhancer identification method. MPRAs are a vast improvement on traditional reporter assays because they are high-throughput; however, the enhancer sequences identified still represent only a subset of active enhancers in a given cell type[93]. Neither transgenic assays nor MPRAs can completely account for the cell-type and stimulus-dependent nature of enhancers, causing inactive sequences to be more difficult to interpret. Additionally, MPRAs suffer from high library complexities, restrictions on the length of the sequence that is able to be assayed, and remove the enhancer sequence from its endogenous context[3,93,100]. The lack of genomic context in MPRAs may alter an enhancer's ability to drive gene expression and bias quantifications of activity[100]. Comparisons of a traditional episomal MPRA with that of a novel lentivirus MPRA (lenti-MPRA) that integrates the putative enhancer sequence into the genome demonstrated that integrated MPRAs were more reproducible[100]. Although the results from the two types of MPRAs were highly correlated with each other, the lenti-MPRA was more highly correlated with other relevant genomic annotations. Another MPRA variant, the parallel targeting of chromosome positions by MPRA (patchMPRA) highlighted the relevance of chromatin structure on the level activity of regulatory sequences[107]. This suggests that both noise and a lack of genomic context in episomal assays may obscure or alter relevant signals, and that it is

crucial to consider enhancer sequences in their endogenous context to fully explain their activity[100,107]. Furthermore, other recent studies describe promoter-dependent enhancer activity, transcription originating in unintended locations along the reporter constructs, and inflammatory responses induced by plasmid transfection, all of which can confound readouts of enhancer activity[105,108]. In order to reliably identify and validate enhancers, these caveats must be understood and addressed.

3D Chromatin Structure and Interaction

Experimental approaches to measure chromatin conformation

Enhancers are thought to interact with their target promoters through loops in the three-dimensional structure of chromatin. Understanding where these loops form and the genomic landscape at each of these interaction points, is important for the functional characterization of gene regulatory architecture. Chromatin conformation assays generate long-range interaction maps of the genome, imply enhancer-promoter interactions, and designate three-dimensional compartments with localized regulatory activity[109,110]. Chromatin conformation capture (3C), circular chromatin conformation capture (4C), chromosome conformation capture carbon copy (5C), and Hi-C all generate contact maps through formaldehyde cross-linking interacting DNA segments and sequencing[2,109]. These methods differ in the number of interactions they are able to detect; 3C, 4C, and 5C probe connections with pre-specified genomic loci or within specific regions, Hi-C generates all-to-all contact maps of genome-wide interactions[111].

Where specific genomic loci are of interest, such as interactions with promoters or single transcription factors, other targeted assays such as ChIA-PET, HiChIP, and promoter-capture Hi-C can be used[112–114]. ChIA-PET combines ChIP-seq with a proximity ligation step to precipitated interaction genomics regions that are cross-linked to the protein or histone modification of interest[112]. The ChIP step can be targeted to elucidate enhancer-promoter contacts or other interactions associated with specific protein complexes. HiChIP improves the ChIA-PET protocol by performing ChIP-seq on a Hi-C

interaction library, allowing for more efficient capture of loci interacting with a protein of interest[114]. In promoter-capture Hi-C, known promoter sequences are used to pull down fragments of interest from Hi-C libraries and sequenced[113]. The result is a set of interacting regions enriched for promoter sequences that increases the ability to define enhancer-promoter interactions. Further experimental characterization showed differences in distal regions interacting active and inactive genes, providing evidence that promoter-capture Hi-C accurately links enhancers and their targets[115].

Topologically associating domains and their boundaries

While mapping the 3D chromatin architecture, researchers discovered that the genome contained many approximately 1Mb regions that were enriched for chromatin interactions within the region. These same 1 Mb regions were depleted for interactions with external loci. Referred to as topologically associating domains, or TADs, these regions are defined from properties of the contact maps derived from the chromatin conformation assays. Thought to act as 'regulatory domains', TADs restrict most interactions between enhancers and genes to loci within the same TAD[110,116,117]. TADs are largely conserved across developmental time points, cell types, and even species[117–119], although the degree of conservation is the subject of some debate[120]. Other recent work underscores their importance by demonstrating that variants altering the boundaries or structure of TADs is associated with cancers and severe developmental and neurological disorders[121–124]. Disruptions to TADs can change the existing enhancer-promoter contacts, leading to 'enhancer hijacking' and the subsequent mis-expression of genes[121,123,124].

The boundaries of TADs often contain clusters of CCCTC-binding factor (CTCF) motifs, many of which show evidence of evolutionary constraint[119,125]. The current loop extrusion model of TAD formation proposes that DNA slides through the ring-shaped cohesin complex until it is stopped by CTCF bound to convergently oriented motifs, creating a DNA loop[118,126–128]. These regulatory loops are thought to facilitate enhancer-promoter contacts and insulate promoters from ectopic enhancer activity[117,129]. Recent work suggests that direction-specific CTCF binding helps control enhancer-promoter contacts by insulating certain genes from enhancer activity and promoting loops that create distal interactions[130].

Clustered CTCF binding sites in a region have also been linked to the strength of the TAD boundary and evolutionary conservation, further highlighting the importance of the factor in TAD maintenance and regulatory insulation[110,125,131,132]. Within TADs, a similar process creates smaller interaction loops, sometimes referred to as sub-TADs, that can bring regulatory elements in contact with other regulatory elements or transcription start sites in order to maintain gene expression[118,133,134]. The disruption of CTCF binding sites in TAD domains is associated with changes to the TAD architecture and some of the severe phenotypes discussed previously[121,123,124,135].

Limitations of current approaches

While informative about genomic state, current approaches to map 3D genomic interactions suffer from a few key limitations. Chromatin interaction data published from studies across diverse sets of human tissues have variable read depth across tissues and relatively low resolution[136]. Low read depth, and a decreased ability to detect short (< 10 kb) interactions, results in downstream biases that lead to difficulty determining significant interactions and limit the ability to compare across tissues[109,137,138]. Poor resolution can especially impair enhancer-promoter mapping since the interaction anchors may encompass many annotations or be limited to a subset of cells and difficult to distinguish[137,138]. Furthermore, because enhancers are both cell-type and context dependent, the interactions between regulatory elements and their target genes may be similarly dependent[139,140]. To date, few biological contexts have been assayed using high-resolution approaches, so it can be difficult to generalize results across different cell types. Computational models trained on experimental results and genomic sequence information may help to bridge this gap by predicting chromatin interactions in novel biological contexts[141,142]. Despite claims that CTCF and TADs are conserved across species and cell types, there are distinct and potentially impactful differences that must be considered[139,140,143,144]. Although TADs and CTCF-mediated boundaries are crucial for maintaining appropriate enhancer-promoter contacts in specific cases, CTCF binding and TAD formation may actually be quite dynamic[145]. Variation in TAD structures or boundaries does not always result in a large effect on gene expression, suggesting that their role in constraining interactions may be

limited to a subset of elements and that the loss of a TAD boundary is not sufficient to induce new interactions[132,146].

Models of Gene Regulatory Architecture

Due to the fact that enhancers regulate the expression of genes over large distances and with varying orientation to the target promoters, characterizing the effective regulatory landscape of a gene remains challenging[2]. Previous work has shown that multiple enhancers may regulate a single target gene, with different levels of cooperativity and redundancy[12,147–149]. Although long established in Drosophila[150–152], new evidence suggests that mammalian genomes also contain redundancy in the enhancer landscape of genes[147,153,154]. This may help to maintain stable gene expression levels, both across evolutionary time and within a single species, and can provide robustness to genetic variation[147,154]. Our understanding of such phenomena requires appropriate models of gene regulatory architecture.

Approaches to link enhancers to target genes

Linking enhancers to their target genes is still an open area of research. The simplest approach involves using the nearest gene or other proximity-based rules, such as those used by GREAT[155]. The default GREAT approach creates a 'basal regulatory domain' around each TSS, that is then extended upstream and downstream until it reaches the basal regulatory domain of another gene or reaches 1 Mb. This reduces the number of overlapping regulatory domains in gene-dense regions. However, from the early case studies of enhancers, we know that they often do not interact with the nearest gene[113,156,157] and can target genes that are further than 1 Mb away in linear sequence[158,159]. Despite these limitations and high potential for false positives, the proximity-based approach can be applied uniformly for all enhancers sets and is still used in current research[160].

Variants associated with changes in gene expression across individuals have previously been used to inform enhancer-gene linkages[161]. Consortia such as the Genotype-Tissue Expression (GTEx) project

have conducted large-scale studies of expression quantitative trait loci (eQTL) across dozens of human tissues and cell lines[162]. Where eQTL overlap with regulatory annotations, such as enhancers, can be used to infer the element's target gene[78,163]. However, this view is complicated by limitations of eQTL mapping and potential redundancy in the enhancer landscape of a gene[161,164]. Machine learning approaches that integrate eQTL with other genomic features may improve target-prediction performance in the context of non-coding variant interpretation. For example, the Inference of Connected eQTLs (ICE)[165] algorithm trained a gradient-boosted decision tree classifier on known GTEx eQTL to predict the target genes of non-coding variants, many of which may fall inside enhancers.

More directly, data from chromatin conformation assays can be used to locate regions of the genome that are in close physical proximity. The interacting loci can be annotated with other enhancer and promoter regions to infer enhancer-gene linkage in a given cell type[161,166] or interactions with enhancer and promoter-associated features can be probed directly using variants of Hi-C[113,115,167]. However, recent work claims that models using chromatin conformation alone perform poorly compared to CRISPR-validated enhancer target maps. They suggest a combinatorial approach, called the activity-by-contact (ABC) model, that weights enhancer-gene links by the inferred strength of the enhancer from biochemical markers and the frequency of chromatin contact[168].

An alternative approach integrates histone modification ChIP-seq with gene expression data from RNA-seq. The correlation between the strength of the functional genomics peak and the level of gene expression is considered evidence of a link between the putative regulatory element and the gene. To improve performance, statistical or machine learning models have been trained on the correlations and information about the local chromatin conformation in order to predict the enhancer-gene links[51,169,170]. Some models (TargetFinder[171], EAGLE[172], 3DPredictor[173], PEP-motif[174], EP2vec[175], CT-FOCS[176]) use combinations of functional genomics, genomic window, and sequence features to make predictions about enhancer-promoter interactions without explicit gene expression correlations. However, predictions made by all of these methods are limited by the availability of training and validation data across tissues. Experimentally derived chromatin interactions from Hi-C or similar approaches serve as the *de facto* gold

standard for validation but suffer from their own limitations[137] and the accuracy of computational models has been questioned[177,178]. Recent work to benchmark many common computational predictors will improve our ability to choose the appropriate and accurate enhancer-gene linkages, although further experimental work is required to fully validate these predictions[179].

Evidence for redundancy in gene regulatory architecture

Current models of transcriptional regulation suggest that genes are often regulated by multiple enhancer elements[12,147–149]. This has been well established in Drosophila, where redundant "shadow enhancers" provide overlapping regulatory functions and robustness to genetic variation[150–152]. Shadow enhancers are thought to be pervasive in the Drosophila genome, occurring near the majority of genes, although their exact functional mechanisms are not completely known. In some cases, the enhancers may act redundantly, where in others they also serve to fine-tune expression across specific conditions or developmental stages[150]. A more complex role for shadow enhancers is supported by constraint at these loci and the recent work suggesting showing positional effects of shadow enhancer function[150,180].

Case studies of specific enhancer clusters have shown that there is some level of redundancy in the regulation of gene expression in mammalian species as well[147,153,181,182]. Similar to shadow enhancers, groups of enhancers in mammalian species may act additively, synergistically, or redundantly. The idea of enhancer cooperation speaks to lines of earlier work describing super-enhancers[12,148,149,183], and more recently, *cis*-regulatory domains[184]. These regulatory clusters or domains are known to contribute to the redundant or cooperative regulation of target genes[12,148,182–184]; the enhancer landscape is largely an extension of this idea. Recent work on the contribution of enhancer landscapes to the maintenance of stable gene expression levels has demonstrated that the number of regulatory elements is related to gene expression stability[154]. Furthermore, deletions of individual enhancer elements often do not result in changes to gene expression or organismal phenotype, supporting the idea that enhancer redundancy is widespread[147,182,185]. Genes that require more precise levels of expression, such as dosage-sensitive and

developmental genes, have also been associated with a larger amount of enhancer sequence, suggesting that some genes depend on more complex enhancer landscapes[164].

Interpretation of Non-coding Genetic Variation

The regulatory genome plays a large role in all essential cellular and evolutionary processes, from organismal development to speciation[2,186–188]. Not surprisingly, genetic variants that alter the function of gene regulatory elements can have profound phenotypic impacts[124,189]. Many types of functional alteration have been implicated in the development of disease phenotypes, including the deletion of individual regulatory elements and changes to the chromatin architecture inducing ectopic enhancer interactions. However, the effects of disrupted enhancer function are difficult to fully characterize. Studying the effects of enhancer variation in different genomic contexts will improve our understanding of the mechanistic and phenotypic effects of gene regulation and allow for more accurate prioritization of uncharacterized non-coding variants.

Genetic variants can disrupt proper gene regulation

Genetic variants that disrupt gene regulatory regions contribute to the architecture of complex disease[2]. Broadly, regulatory elements are enriched for overlap with disease-associated variants, including those identified in genome wide association studies[189–192]. Thus, many variant prioritization methods consider the impact of single nucleotide variants (SNVs) on features of gene regulatory elements, such as alteration of important transcription factor binding motifs[189,193–196] or disruption of enhancer cooperation[197], to predict the effects of a given variant.

Structural variants (SVs) can also disrupt the gene regulatory architecture and lead to a gain or loss of enhancer function[121,124,198]. These include large deletions and duplications of sequences as well rearrangements like inversions and translocations. SVs often affect entire chromatin domains and may generate new enhancer-gene contacts, often referred to as ‘enhancer adoption’ or ‘enhancer hijacking’[121,199–203]. This outcome may be even more deleterious than the disruption of individual

regulatory elements, and SVs at TAD boundaries show evidence of purifying selection[204]. The effects of SVs on gene expression have previously been studied at specific loci or in single disease contexts, providing a range of mechanisms by which regulatory variation can result in disease[121,124,200,201,205–208].

Strategies for non-coding variant interpretation using regulatory annotations

Predicting the effect of variation in non-protein-coding regions on phenotypes, especially through changes to gene regulation, is a challenging but essential task. There are many mechanisms by which genetic variants can lead to disease[209], thus variant interpretation and variant effect prediction take many forms. Of particular interest are computational approaches that have been developed to predict the pathogenicity of non-coding variants or prioritize the most relevant set of candidates for experimental follow-up[210–215]. These incorporate a range of genome-wide annotations, including enhancer-associated histone modifications, DNA sequence, transcription factor binding profiles, and chromatin accessibility into both supervised and unsupervised machine learning algorithms to predict the effects of non-coding variation on gene expression or functional genomics marks[214,216–220]. Although many of these approaches do not explicitly model variant effects in terms of changes to enhancer function, the underlying assumptions remain that non-coding variation influences gene regulatory function and that this function can be captured by markers frequently associated with enhancer activity. Future work incorporating additional information about the broader chromatin context[185,197,221], training with data from large-scale CRISPR and MPRA enhancer screens[212], and employing more complex computational models[216,219,222] will continue to improve our ability to predict the effect of non-coding variants and causal mechanisms.

Experimental assessment and validation of non-coding variant effects remains the gold standard; however, determining the causal mechanisms underlying a disease-associating can be expensive and time consuming. Recent advances in MPRA technology have allowed for large saturation mutagenesis assays to quantify the impact of all possible variants in a regulatory sequence[223]. CRISPR-based assays will also be valuable to assess the impact of regulatory alterations *in vivo*, especially those that influence enhancer-gene targeting causing ectopic gene expression[224]. As these techniques become more sophisticated, they

can generate increasing amounts of functionally relevant data for future algorithm development. Continued additions to manually curated databases, such as DiseaseEnhancer, will also provide a record of causal variants with experimentally validated impacts on enhancer sequences[225].

Limitations of current non-coding variant interpretation approaches

Despite considerable progress, interpreting the effect of non-coding variants on gene expression and downstream phenotypes remains challenging. Due to the tissue and context-specific nature of gene regulatory elements such as enhancers and the inherently dynamic process of gene regulation, our ability to accurately train computational models often lags behind our current understanding of the regulatory process. Supervised machine learning models in particular suffer from our limited ability to create true positive and negative training sets. Furthermore, there are many examples of non-coding variants that have large impacts on enhancer function or chromatin architecture, yet no detectable effects on gene expression or phenotype[132,146,147]. Redundancy in the gene regulatory landscape of a gene is not well understood and typically not included in variant effect prediction. Genetic variants rarely occur in isolation and studies of single variants on individual regulatory elements can miss important joint effects[198,206]. Finally, previous work has largely focused on the potential regulatory effects of SNVs, considering the effects of SVs on gene regulatory architecture in small-scale studies of individual variants. Expanding the functional annotation of all SVs to include regulatory effects will fill an important gap. Considering the impact of regulatory variation in combination, and across multiple scales, will advance our knowledge of the interactions between variants and lead to more complete interpretation and prediction of genome-wide effects.

Chapters

Non-coding regulatory regions are crucial for the maintenance of proper transcriptional programs in the cell. Accurately identifying the elements, such as enhancers, involved in gene regulation and how they maintain gene expression levels in different biological contexts is a crucial step towards understanding the

impact of genetic variants that alter gene regulatory architecture. While many enhancer identification approaches are in common use, we hypothesized that enhancers predicted by different approaches would differ significantly in their genomic and functional attributes. Chapter II provides a comprehensive quantification of the genomic, evolutionary, and functional differences between enhancer sets identified by different strategies. Furthermore, we conclude that our ability to generate high-confidence and biologically relevant sets of enhancers by focusing on enhancers identified by multiple strategies is limited. Combinations of available enhancer sets are not more likely to overlap markers of functional relevance, and machine learning models fail to distinguish between unique and reproducible enhancer sequences. Chapter III uses a novel cohort with genome and RNA sequencing to functionally annotate structural variants (SVs) impacting putative gene regulatory sequences. We demonstrate that SVs disrupting regulatory elements have a substantial impact on gene expression and have likely been selected against. In Chapter IV we develop a framework to define enhancer landscapes—groups of enhancers putatively regulating the same target gene—in multiple human tissues and quantify their influence on gene regulation. We observe that differences in gene function and constraint on gene expression are reflected in the features of their enhancer landscapes, including the number and tissue-specificity of associated enhancers. Ultimately, our results highlight the importance of accurate identification of gene regulatory elements and models of enhancer activity that consider the broader genomic context to our understanding of gene expression dynamics and ability to interpret regulatory genetic variation.

CHAPTER II

Evaluating the Genomic and Functional Differences Between Genome-wide Enhancer Sets[1]

Introduction

Accurately identifying gene-regulatory enhancers remains a challenging task. This is mainly due to the large number of genome-wide approaches currently in use and the lack of a comprehensive gold standard. In practice, most studies consider enhancers defined by a single approach in downstream analyses. These are based on subsets of experimental features correlated with enhancer activity and complex computational techniques, making it difficult to compare results across studies. This chapter evaluates the genomic similarities between enhancer sets identified by representative strategies in four biological contexts. We then quantify the differences between the functional attributes of enhancer sets, including enrichment for transcription factor binding motifs and overlap with experimentally validated enhancer sequences or trait-associated genetic variation. Finally, we explore strategies to combine and prioritize enhancer sets to generate reliable maps of genome-wide enhancer activity.

We find that enhancers identified by different strategies have significant dissimilarity in their genomic, evolutionary, and functional characteristics within each context. This disagreement is sufficient to influence the interpretation of candidate SNPs from GWAS studies, and to lead to disparate conclusions about disease mechanisms. Most regions identified by enhancers are supported by only one method, and we find limited evidence that regions identified by multiple methods are better candidates than those identified by a single method. As a result, we cannot recommend the use of any single enhancer identification strategy in all situations. This chapter highlights the inherent complexity of

---

[1] Adapted from Benton, M.L., Talipineni, S.C., Kostka, D. *et al.* Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. *BMC Genomics* **20,** 511 (2019). https://doi.org/10.1186/s12864-019-5779-x

enhancer biology and identifies an important challenge to mapping the genetic architecture of complex disease. In order to enable robust and reproducible results, we must foster a deeper appreciation for the dynamic and complex nature of gene regulatory elements.

Methods

Enhancer identification methods

Here, we summarize how we defined human enhancer sets across four biological contexts. All analyses were performed in the context of the GRCh37/hg19 build of the human genome. We used TSS definitions from Ensembl v75 (GRCh37.p13).

We downloaded broad peak ChIP-seq data for three histone modifications, H3K27ac, H3K4me1, and H3K4me3 from the ENCODE project[9] for two cell lines, K562 and Gm12878, and from the Roadmap Epigenomics Consortium[191] for two primary tissues, liver and heart. The ENCODE broad peaks were generated by pooling data from two isogenic replicates. The Roadmap Epigenomics broad peaks were also generated with data from two replicates. We chose broad peak files because we expect histone modifications to flank active enhancer regions, and the broad peaks represent wide regions of relative enrichment that are likely to encompass the functionally relevant sequences. See below for details on consistent peak calling. We downloaded the "enhancer-like" annotations from ENCODE (version 3.0); these combine DHSs and H3K27ac ChIP-seq peaks using an unsupervised machine learning model. We retrieved ChromHMM enhancer predictions[74] for the K562 and Gm12878 cell lines from the models trained on ENCODE data[69]. We downloaded ChromHMM predictions for liver and heart tissues from the 15-state segmentation performed by the Roadmap Epigenomics Consortium. For all ChromHMM sets, we combined the weak and strong enhancer states. We considered two enhancer sets for K562 and Gm12878 based on supervised machine learning techniques—one described in Yip et al. 2012[10], and the other in Ho et al. 2014[11]. Yip12 predicted 'binding active regions' (BARs) using DNA accessibility and histone modification data; the positive set contained BARs overlapping a 'transcription-related factor' (TRF), and

the negative set contained BARs with no TRF peaks. The predicted regions were filtered using other genomic characteristics to determine the final set of enhancers[10]. Ho14 used H3K4me1 and H3K4me3 ChIP-seq peaks in conjunction with DHSs and p300 binding sites to predict regions with regulatory activity both distal and proximal to TSSs. The distal regulatory elements make up their published enhancer set[11]. For K562 and Gm12878 we also downloaded p300 ChIP-seq peaks from ENCODE[9]. We downloaded enhancer regions predicted by the FANTOM consortium for the four sample types analyzed[34]. We also downloaded enhancer predictions in liver from Villar et al. 2015[27]. We downloaded regions of nascent bidirectional transcription from GRO-cap data generated for the two cell lines, K562 and Gm12878[36]. The transcribed regions on matched positive and negative strands were merged into a single annotation.

To represent enhancer identification strategies in common use, we created two additional enhancer sets for this study using histone modification ChIP-seq peaks and DNase-seq peaks downloaded from ENCODE and Roadmap Epigenomics. The H3K27acPlusH3K4me1 track is a combination of H3K27ac and H3K4me1 ChIP-seq peak files[28,30,44]. If both types of peaks were present (i.e., the regions overlap by at least 50% of the length of one of the regions) the intersection was classified as an enhancer. Similarly, to create the H3K27acMinusH3K4me3 set for each context, we intersected H3K27ac and H3K4me3 ChIP-seq peak files and kept regions where H3K27ac regions did not overlap a H3K4me3 peak by at least 50% of their length. We derived the combination of H3K27ac and H3K4me3 and the 50% overlap criterion from previous studies[14,27,44]. The DNasePlusHistone track is based on the pipeline described in Hay et al. 2016[12]. It combines H3K4me1, H3K4me3, DNaseI hypersensitive sites (DHSs), and transcription start site (TSS) locations. We filtered a set of DHSs, as defined by DNase-seq, for regions with an H3K4me3 / H3K4me1 ratio less than 1, removed regions within 250 bp of a TSS, and called the remaining regions enhancers.

For all enhancer sets, we excluded elements overlapping ENCODE blacklist regions and gaps in the genome assembly[226]. Additionally, due to the presence of extremely long regions in some enhancer sets, likely caused by technical artifacts, we removed any regions more than three standard deviations

above or below the mean length of the dataset. This filtering process removed relatively few annotations (Provided in Benton *et al*.[192] Table S11).

When considering the agreement between biological replicates for K562, Gm12878, and liver H3K27ac ChIP-seq data, we downloaded the FASTQ files from ENCODE[9] and Villar *et al.* 2015[27], respectively, aligning each to GRCh37.p13 using BWA[227] (v.0.7.15, default options). We called peaks of broad enrichment using MACS[228] (v.1.4.2, default options). We processed each of the replicate peak files using the same pipeline as the published peak files.

Enhancer attribute data

We downloaded evolutionarily conserved regions defined by PhastCons, a two-state hidden Markov model that defines conserved elements from multiple sequence alignments[229]. We concatenated primate and vertebrate PhastCons elements defined over the UCSC alignment of 45 vertebrates with humans into a single set of conserved genomic regions. We downloaded the full list of 20,458 unique GWAS SNPs from the GWAS Catalog (v1.0, downloaded 08-10-2016)[230]. We also manually curated the set of GWAS SNPs into subsets associated with phenotypes relevant to liver or heart for context-specific analyses (Table S4 in Benton *et al*.[192]). We downloaded all GTEx eQTL from the GTEx Portal (v6p, downloaded 09-07-2016)[231]. We concatenated the data from all 44 represented tissues and ran the enrichment analysis on unique eQTL, filtering at four increasingly strict significance thresholds: 10^{-6}, 10^{-10}, 10^{-20}, and 10^{-35}. We present the results from the p-value threshold of 10^{-10}, although the choice of threshold did not qualitatively alter the results. We also performed separate context-specific analyses on liver and heart specific eQTL from GTEx ($p < 10^{-10}$). To identify other variants tagged by the GWAS SNPs and eQTL, we expanded each set to include SNPs in high LD ($r^2 > 0.9$) in individuals of European ancestry from the 1000 Genomes Project, phase 3 [232].

Experimentally validated enhancer sequences with activity in the heart and all negative enhancer sequences were downloaded from the VISTA enhancer browser (downloaded 11-16-2017)[92]. We also downloaded sequences and Sharpr-MPRA activity levels for 15,720 putative enhancer regions tested for

regulatory activity in K562 cells using a massively parallel reporter assay (MPRA)[62]. The Sharpr-MPRA algorithm infers a regulatory score for each base pair in a region using a probabilistic model, with positive scores indicating activating regulatory regions and negative scores indicating repressive regions. Following Ernst et al., we summarized the overall regulatory activity of a given enhancer region as the activity value with the maximum absolute value and classified the enhancer regions into activating (n = 5,373) and repressive (n = 10,347) based on the score's sign[62].

Genomic region overlap and similarity

To quantify genomic similarity, we calculated the base pair overlap between two sets of genomic regions, *A* and *B*, by dividing the number of overlapping base pairs in *A* and *B* by the total number of base pairs in *B*. We also performed this calculation on element-wise level, by counting the number of genomic regions in *B* overlapping regions in *A* by at least 1 bp, and dividing by the number of genomic regions in *B*. We performed both calculations for each pairwise combination of enhancer sets. All overlaps were computed using programs from the BEDtools v2.23.0 suite[233].

We also evaluated the similarity between pairs of genomic region sets using the Jaccard similarity index. The Jaccard index is defined as the cardinality of the intersection of two sets divided by cardinality of the union. In our analyses, we calculated the Jaccard index at the base pair level. We also computed the relative Jaccard similarity as the observed Jaccard similarity divided by the maximum possible Jaccard similarity for the given sets of genomic regions, i.e., the number of bases in the smaller set divided by the number of bases in the union of the two sets. To visualize overlaps, we plotted heatmaps for pairs of methods using ggplot2 in R[234].

Genomic region overlap enrichment analysis

To evaluate whether the observed base pair overlap between pairs of enhancer sets is significantly greater than would be expected by chance, we used a permutation-based approach. We calculated an empirical *p*-value for an observed amount of overlap based on the distribution of overlaps expected under a null

model of random placement of length-matched regions throughout the genome. We used the following protocol: let *A* and *B* denote two sets of genomic regions; count the number of bp in *A* that overlap *B*; generate 1,000 random sets of regions that maintain the length distribution of *B*, excluding ENCODE blacklist regions and assembly gaps; count the number of bp in *A* that overlap regions in each of the random sets; compare the observed bp overlap count with the overlap counts from each iteration of the simulation and compute a two-sided empirical *p*-value. We used the same framework to evaluate element-wise comparisons by counting the number of regions in *A* that overlap *B* rather than the bps. This approach was performed using custom Python scripts and the Genomic Association Tester (GAT)[235]. We note that this measure of overlap significance is not symmetric, and accordingly we confirmed results of our element-wise results for both orderings of the pairs of enhancer sets.

Enhancer conservation, GWAS catalog SNP, and GTEx eQTL enrichment

In addition to comparing the overlap between pairs of enhancer sets, we also computed enrichment for overlap of evolutionarily conserved regions, GWAS SNPs, and GTEx eQTL with each of the enhancer sets. For conserved elements, we proceeded as described above for comparisons between pairs of enhancer sets, but considered the conserved elements as set *A* and the enhancers as set *B*. For GWAS tag SNPs, we considered each variant as a region in set *A* and the enhancer regions as set *B*. We used the same approach for testing all variants in LD ($r^2 > 0.9$) with GWAS tag SNPs and for testing enrichment for liver- and heart-specific GWAS tag SNP sets. We also tested for enrichment using only the variant with the maximum number of enhancer set overlaps for each GWAS SNP's LD block. In this analysis, *A* was the set of variants with maximum enhancer set overlap for each LD block and *B* was the set of enhancers. Enrichments were computed for the eQTL SNP sets using the same strategy as described for GWAS SNPs.

Enhancer set Gene Ontology annotation and similarity

We used GREAT to find Gene Ontology (GO) annotations enriched among genes nearby the enhancer sets. GREAT assigns each input region to regulatory domains of genes and uses both a binomial and a hypergeometric test to discover significant associations between regions and associated genes' GO annotation terms[155]. Due to the large number of reported regions in each enhancer set, we considered significance based only on the binomial test with the Bonferroni multiple testing correction (<0.05). We downloaded up to 1,000 significant terms for each enhancer set from the Molecular Function (MF) and Biological Process (BP) GO ontologies. We calculated the similarity between lists of GO terms using the GOSemSim package in R[236]. GOSemSim uses sematic similarity metric that accounts for the hierarchical organization and relatedness of GO terms when calculating the similarity score[237]. For each pair of enhancer sets, we calculated the similarity between their associated GO terms, and converted the resulting similarity matrix into a dissimilarity matrix. We also calculated the number of shared GO terms between pairs of methods and manually compared the top ten significant terms for each enhancer set.

Since enhancers often target genes over long distances, we also considered target predictions generated by the JEME algorithm to assign enhancers to potential target genes in each context[169]. JEME is a two-step process that considers the superset of all enhancers across contexts as well as context-specific biomarkers to make its predictions. By intersecting each enhancer set with the corresponding enhancer-target maps from JEME, we created a set of putatively regulated genes for each method in a given context. We performed GO enrichment analyses on the gene sets using the online tool WebGestalt[238]. We downloaded the top 1,000 significant terms ($p < 0.05$ after Bonferroni correction) for each enhancer set from the BP and MF GO ontologies and calculated the pairwise similarity between lists of GO terms using the same semantic similarity metric as above.

Clustering of enhancer sets by enriched transcription factor binding motifs

We used AME from the MEME suite to quantify enrichment for known motifs from the HOCOMOCO (v11) core database in each enhancer set[239,240]. Enrichment was calculated based on a comparison to

background sequences generated by randomly shuffling the enhancer sequences while maintaining their dinucleotide frequency distribution. We used the default E-value threshold of 10 to define significant enrichment.

We calculated the similarity between sets of enriched motifs using the Jaccard index. Because many of the enhancer sets were enriched for a high proportion of motifs from the HOCOMOCO database we also generated an expected Jaccard similarity index using size-matched lists of random motifs. This served as a baseline level of similarity expected from an enriched motif list of that size. We clustered the enhancer sets based on their Jaccard similarities using multidimensional scaling (MDS).

Genomic and functional clustering of enhancer sets

To identify groups of similar enhancers in genomic and functional space, we performed hierarchical clustering on the enhancer sets. For genomic similarity, we converted the pairwise Jaccard similarity to a dissimilarity score by subtracting it from 1 and then clustered the enhancer sets based on these values. For functional similarity, we clustered the lists of GO terms returned by GREAT for each enhancer set or sets of enriched TF binding motifs from AME. We calculated similarity of GO terms using the GoSemSim package in R and converted it to dissimilarity by subtracting the similarity score from 1. For both the genomic and GO similarity, we used agglomerative hierarchical clustering in R function with the default complete linkage method to iteratively combine clusters[241]. We visualized the cluster results as dendrograms using ggplot2 and dendextend[234,242]. We performed multidimensional scaling (MDS) on the Jaccard, GO term, and TF motif dissimilarity matrices using default options in R[241].

Combinatorial analysis of enhancer sets and enrichment for functional signals

We stratified genomic regions by the number of enhancer identification strategies that annotate them in order to determine whether regions predicted to be enhancers by more methods show greater enrichment for three signals of function—evolutionarily conserved base pairs, GWAS SNPs, or GTEx eQTL—compared to regions with less support. We divided all regions predicted by any enhancer identification

method in a given context into bins based on the number of methods that predicted it. Some enhancer regions had varying prediction coverage and were split across multiple bins. While infrequent (<3% of regions), we removed all regions less than 10 bp in length since these are unlikely to function as independent enhancers. For each enhancer support bin, from 1 to the number of prediction methods, we calculated the enrichment for overlap with each functional signal using the permutation framework described above. We considered three different proxy sets: evolutionarily conserved base pairs as defined by PhastCons elements, GWAS SNPs, and GTEx eQTL. In each enrichment analysis, the functional signal regions were set *A* and the enhancer regions with a given level of support were set *B*. We report the average enrichment for each enhancer support bin with the empirical 95% confidence intervals.

For enhancer sets with quantitative enhancer-level scores available across contexts, we ranked each enhancer by its score, and then analyzed whether regions that have higher scores are more likely to be predicted by other identification methods. We calculated the rank using the ChIP-seq or CAGE-seq signal scores for a subset of methods (H3K27acPlusH3K4me1, H3K27acMinusH3K4me3, DNasePlusHistone, FANTOM), and the machine learning derived score for EncodeEnhancerlike regions. Within each set, we sorted the enhancer regions by score and assigned ranks starting at 1 for the top-scoring region. We then partitioned the enhancer regions in each set by the number of other enhancer sets that overlap at least one base pair in that region. To compare the most confident enhancer predictions, we subset each of the ranked methods into the top 100, 500, or 800 top enhancers. We used these subsets to calculate the level of enrichment for overlap with GWAS Catalog SNPs, GTEx eQTL, and evolutionary conservation based on the number of methods that agree on each annotated region.

Machine learning prediction of shared enhancers from short sequence motifs

We trained a support vector machine (SVM) classifier to distinguish between enhancers that a reproducible across methods versus those that are unique to a single method. The feature set for the SVM classifier was the k-mer spectrum (k = 6) for enhancer sequences in each category. The positives were enhancer sequences in genomic regions identified by at least two methods and the negatives were

enhancer sequences identified by only one method. We applied our model using ten-fold cross validation and calculated the accuracy using the area under the receiver operating characteristic curve.

Results

A panel of enhancer identification strategies across four biological contexts

To evaluate the variation in enhancer sets generated by different enhancer identification strategies, we developed a consistent computational pipeline to compare enhancer sets genome-wide. Our approach is based on publicly available data, and we applied it to a representative set of methods in four common cell types and tissues (biological contexts): K562, Gm12878, liver, and heart cells (Figure 1). Given the large number of enhancer identification strategies that have been proposed[2,71], it is not possible to compare them all; so for each context, we consider methods that represent the diversity of experimental and computational approaches in common use.

For all contexts, we consider two enhancer sets derived solely from chromatin immunoprecipitation followed by sequencing (ChIP-seq) for histone modifications informative about enhancer activity from the ENCODE Project[9]. The "H3K27acPlusH3K4me1" set includes all H3K27ac ChIP-seq peaks that also overlap an H3K4me1 peak, and the "H3K27acMinusH3K4me3" set contains H3K27ac peaks that do not overlap an H3K4me3 peak[27,44,80]. We used broad peak files processed using a consistent custom pipeline and quality control criteria by ENCODE. In liver only, we consider an additional set of enhancers identified using the H3K27acMinusH3K4me3 definition on different samples ("Villar15")[27]. We also consider a method that incorporates DNase I hypersensitive sites (DHSs) with histone modifications to generate the "DNasePlusHistone" enhancer set, which is composed of DHSs where the ratio of H3K4me1 to H3K4me3 is less than one[12]. For the two cell lines we also include ChIP-seq peaks for the transcription cofactor p300, "p300", that is known to be associated with active enhancers[9,15,44]. Since transcriptional signatures are increasingly used to identify enhancers, we consider "FANTOM" enhancers identified from bidirectionally transcribed eRNA detected via cap analysis of

gene expression (CAGE) by the FANTOM5 Project[34,243,244]. For K562 and Gm12878 we include a set of transcribed regions defined by nascent bidirectional transcription in a modification of the global run-on sequencing (GRO-seq) as "GRO-cap"[36]. Finally, we also include several methods that combine machine learning with functional genomics data, such as the ENCODE consortium's "EncodeEnhancerlike" made by combining DHSs and H3K27ac peaks using an unsupervised ranking method and the "ChromHMM" predictions generated by a hidden Markov model trained on ChIP-seq data from eight histone modifications, CTCF, and RNA Pol II[69,245–247]. For the K562 and Gm12878 cell lines we include enhancer predictions made by two supervised machine learning methods trained to identify enhancers based on ChIP-seq data in conjunction with other functional genomic features. We will refer to these sets as "Yip12" and "Ho14"[10,11]. An overview of the data and computational approaches used by each method is given in Figure 1and full details are available in the Methods.

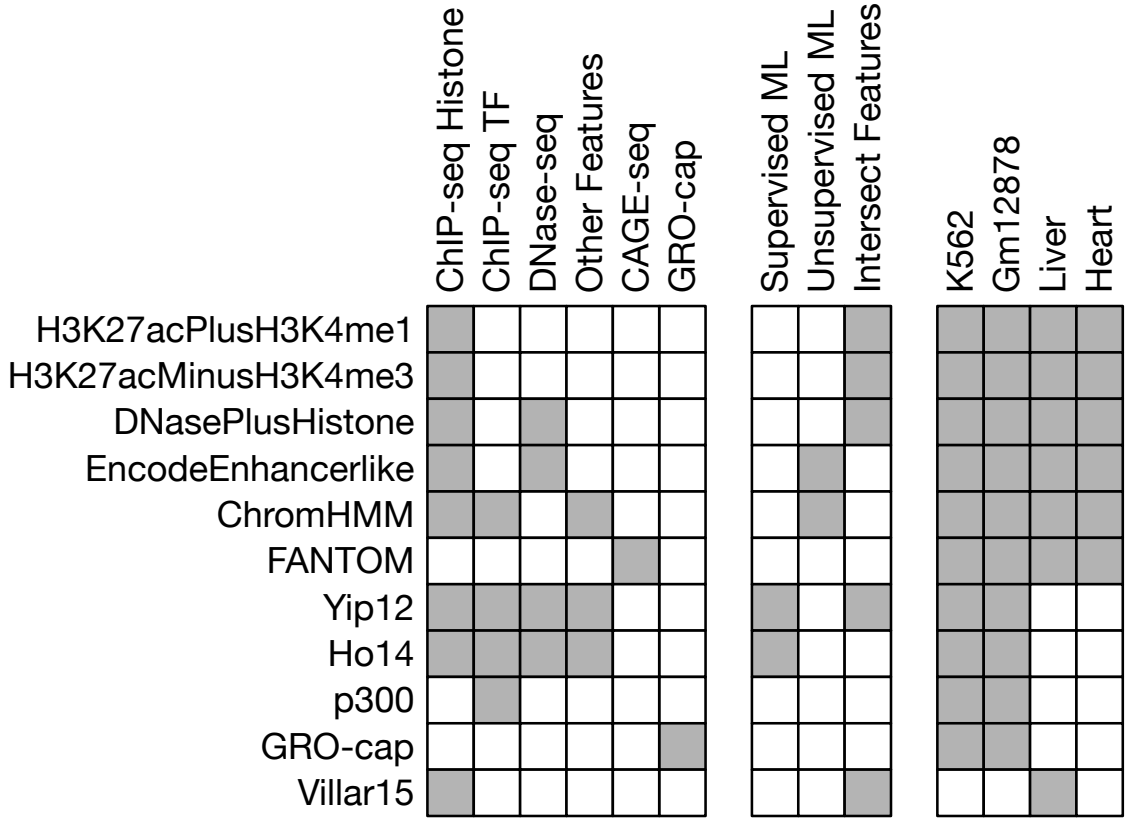

Figure 1. Eleven diverse enhancer identification strategies were evaluated across four contexts.

Each row summarizes the data sources, analytical approaches, and contexts for the eleven enhancer identification strategies we considered. The leftmost columns of the schematic represent the experimental assays and sources of the data used by each identification strategy. The middle columns describe the computational processing (if any) performed on the raw data (ML: machine learning). The rightmost columns give the contexts in which the sets were available. Table 1 gives the number, length, and genomic coverage of each enhancer set.

Genomic coverage of different enhancer sets varies by several orders of magnitude

Enhancer regions identified in the same context by different methods differ drastically in the number of enhancers identified, their genomic locations, their lengths, and their coverage of the genome (Table 1, Figure 2; Fig S1 in Benton *et al*.[192]). As noted above, we expected to find variation between enhancer sets in these attributes. Nevertheless, the magnitude of differences we observed is striking. For each attribute we considered, enhancer sets differ by several orders of magnitude (Table 1, Figure 2). For instance, FANTOM identifies 326 kilobases (kb) of sequence with liver enhancer activity, EncodeEnhancerlike identifies 89 megabases (Mb), and H3K27acMinusH3K4me3 identifies almost 138 megabases (Mb).

Table 1. Summary of all enhancer sets analyzed in this study.

Context	Enhancer Set	Number of Base Pairs (kb)	Number of Enhancers	Median Length	Genome Coverage
K562	H3K27acPlusH3K4me1	22,113	6,642	1,903	0.0078
	H3K27acMinusH3K4me3	34,072	19,698	525	0.0120
	DNasePlusHistone	6,620	13,402	431	0.0023
	ChromHMM	96,545	100,837	600	0.0339
	EncodeEnhancerlike	39,961	36,008	878	0.0140
	Ho14	29,027	35,769	556	0.0102
	Yip12	5,389	13,303	342	0.0019
	p300	7,939	26,463	316	0.0028
	GRO-cap	3,905	23,825	160	0.0014
	FANTOM	390	1,084	344	0.0001
Gm12878	H3K27acPlusH3K4me1	28,355	8,019	2,749	0.0099

	H3K27acMinusH3K4me3	20,868	11,238	701	0.0073
	DNasePlusHistone	9,286	19,815	386	0.0033
	ChromHMM	73,929	69,314	800	0.0259
	EncodeEnhancerlike	50,224	38,872	1,018	0.0176
	Ho14	41,543	39,550	674	0.0146
	Yip12	5,389	13,303	342	0.0019
	p300	6,480	17,532	360	0.0023
	GRO-cap	3,646	21,308	160	0.0013
	FANTOM	1,025	2,826	343	0.0004
Liver	H3K27acPlusH3K4me1	87,576	37,644	1,831	0.0307
	H3K27acMinusH3K4me3	137,874	77,014	1,096	0.0484
	DNasePlusHistone	51,292	170,212	152	0.0180
	ChromHMM	108,375	101,260	800	0.0380
	EncodeEnhancerlike	89,129	37,426	1,849	0.0313
	FANTOM	326	869	347	0.0001
	Villar15	86,139	27,725	2,545	0.0302
Heart	H3K27acPlusH3K4me1	59,892	42,910	1,102	0.0210
	H3K27acMinusH3K4me3	157,468	141,162	684	0.0553
	DNasePlusHistone	33,224	103,898	168	0.0117
	ChromHMM	93,067	113,092	600	0.0327
	EncodeEnhancerlike	186,866	47,235	2,872	0.0656
	FANTOM	611	1,720	335	0.0002

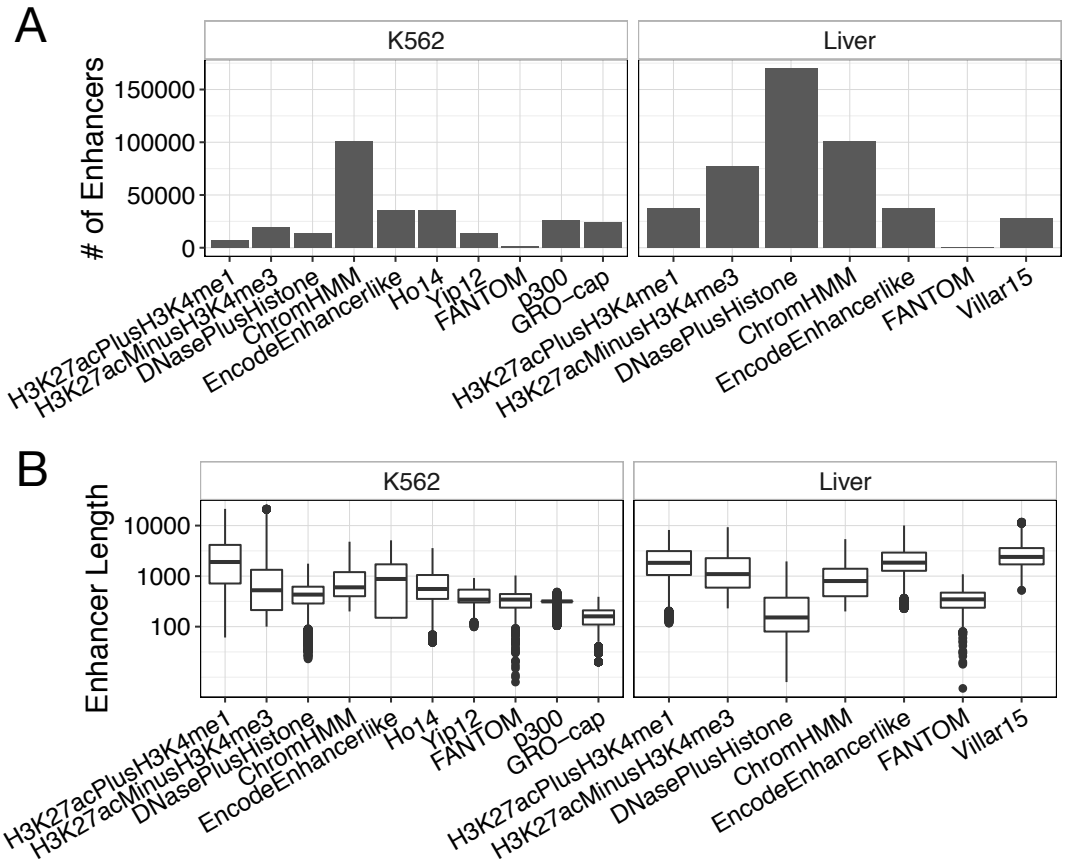

Figure 2. Enhancer identification methods vary in the number and length of predicted enhancers.
(A) The number of K562 and liver enhancers identified by each method varies over two orders of magnitude. There is considerable variation even among methods defined based on similar input data, e.g., histone modifications. (B) The length of K562 and liver enhancers identified by different methods shows similar variation. Enhancer lengths are plotted on a $\log_{10}$ scale on the y-axis. Data for other contexts are available in Table 1 and Fig S1 (provided in Benton *et al.*[192]).

In addition, methods based on similar approaches often differ substantially due to technical factors; e.g., Villar15, which uses the same enhancer definition as H3K27acMinusH3K4me3, only annotates 86.1 Mb with enhancer function in liver. Enhancer sets also vary in their relative distance to other functional genomic features, such as transcription start sites (TSSs). For example, in liver, the average distance to the nearest TSS ranges from 14 kb for EncodeEnhancerlike to 64 kb for DNasePlusHistone (Table S1 in Benton *et al.*[192]). Overall, as expected, methods based on histone modifications tend to identify larger numbers of longer enhancers compared with CAGE data, and

machine learning methods are variable. However, these differences span orders of magnitude. We highlight these trends in liver, but they are similar in other contexts (Table 1, Figure 2; Fig S1 in Benton *et al.*[192]).

Enhancer sets overlap more than expected by chance but have low genomic similarity

Given the diversity of the enhancer sets identified by different methods, we evaluated the extent of both bp and element-wise overlap between them. All pairs of enhancer sets overlap more than expected if they were randomly distributed across the genome (Figure 3A,B ~5–100x, $p < 0.001$ for all pairs, permutation test). As expected due to the greater cellular heterogeneity and genetic variation of tissue samples vs. cell lines, enhancer sets identified by different methods in the same cell line have more significant overlap than enhancer sets identified in tissues (Figure 3B).

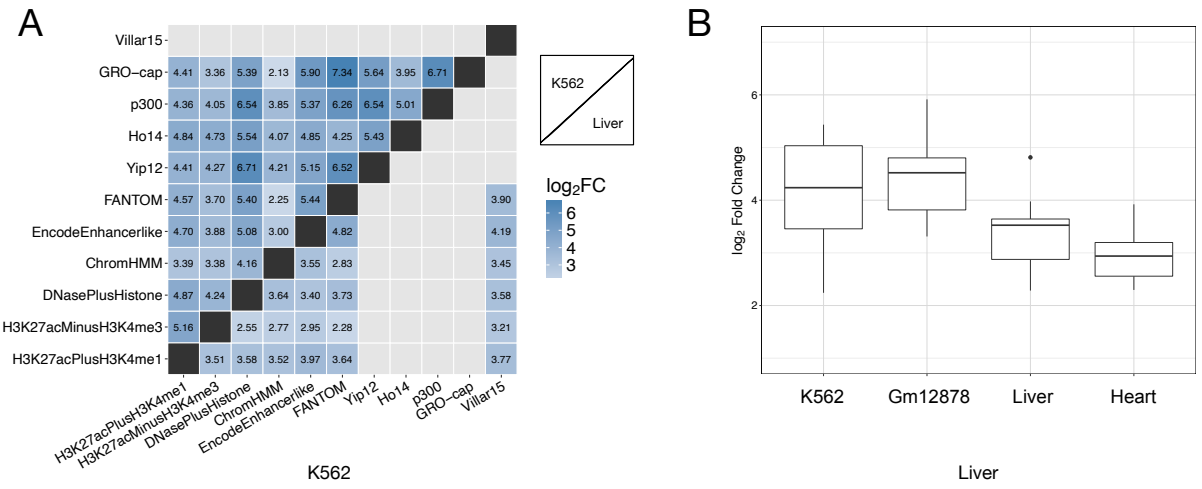

Figure 3. Enhancer sets have low genomic overlap.
(A) Pairwise bp enrichment values ($\log_2$ fold change) for overlap between each K562 (upper triangle) or liver (lower triangle) enhancer set, compared to the expected overlap between randomly distributed, length-matched regions. (B) The $\log_2$ enrichment for bp overlap compared to a random genomic distribution for each pair of enhancer sets within each context. Only contexts with annotations across all biological contexts are included. The fold changes across annotations for the primary tissues—liver and heart—are significantly lower than the cell lines—K562 and Gm12878 ($p = 6.88E-11$ Kruskal-Wallis test, followed by Dunn's test with Bonferroni correction for pairwise comparisons between contexts). The patterns are similar for element-wise comparisons (Fig S3 provided in Benton *et al.*[192]).

However, most (54%) predicted enhancers are "singletons" that are annotated by only a single enhancer identification strategy. Furthermore, the magnitude of overlap between enhancer sets is typically low: less than 50% for nearly all pairs of methods (median 17% bp overlap for K562 and 30% for liver; Figure 4A,B; Fig S2 and Table S2 in Benton *et al.*[192]). Furthermore, the largest overlaps are in comparisons including one enhancer set with high genome coverage or in comparisons of sets that were identified based on similar data. These patterns were similar when evaluating overlap on an element-wise basis (median element-wise overlap: 18%–34%; Figs S3-S4 and Table S2 in Benton *et al.*[192]).

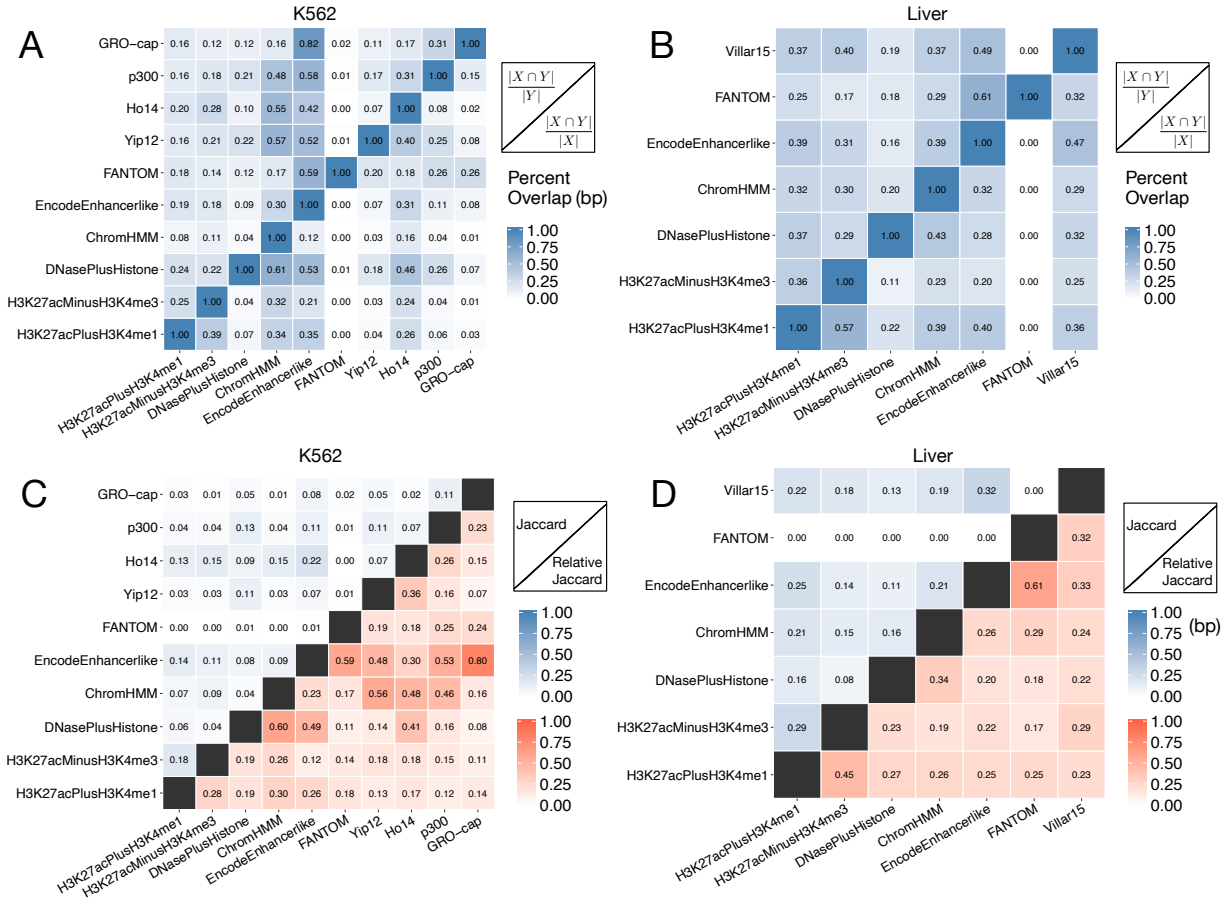

Figure 4. Enhancer sets have low genomic similarity.
The percent base pair (bp) overlap between all pairs of (A) K562 enhancer sets and (B) liver enhancer sets. Percent overlap for each pair was calculated by dividing the number of shared bp between the two sets by the total number of base pairs of the set on the y-axis. The highest overlap is observed for pairs based on similar input, e.g., machine learning models trained on the same functional genomics data, or comparisons with large sets, e.g. ChromHMM. Comparisons between biological replicates average 76% overlap. (C, D) The Jaccard similarity between all pairs of (C) K562 or (D) liver enhancer sets. The upper triangle gives the Jaccard similarity, and the lower triangle gives the

relative Jaccard similarity in which the observed similarity is divided by the maximum possible similarity for the pair of sets.

To further quantify overlap, we calculated the Jaccard similarity index—the number of shared bp between two enhancer sets divided by the number of bp in their union—for each pair of methods. Overall, the Jaccard similarities are also low for all contexts, with an average of 0.07 for K562 and 0.13 for liver and all pairwise comparisons below 0.35 (Figure 4C,D, Fig S2 in Benton *et al*.[192], upper triangle). Since the Jaccard similarity is sensitive to differences in set size, we also computed a "relative" Jaccard similarity by dividing the observed value by the maximum value possible given the set sizes. The relative similarities were also consistently low (Figure 4C,D, Fig S2 in Benton *et al*.[192], lower triangle). To assess the influence of biological variation on the observed overlaps, we compared the overlap of replicates from H3K27ac ChIP-seq data in K562, Gm12878, and liver generated by the same laboratory and processed by the same peak calling pipeline. H3K27ac ChIP-seq data are used in the definition of most of the enhancer sets considered here, so high variability in this data would likely impact many of the predictions. We expected the replicates to have high overlap and serve as an "upper bound" on similarity in practical applications. On average, the replicates overlap 76% at the bp level (with a range of 54–88%) and 84% element-wise (with a range of 51-89%). The only value less than 66% comes from a single K562 comparison. Thus, while there is variation, the amount of overlap observed between enhancers identified by different methods almost always falls far below the variation between ChIP-seq replicates.

Enhancer sets have different levels of evolutionary conservation

Enhancers identified by different methods also differ in their levels of evolutionary constraint. Using primate and vertebrate evolutionarily conserved elements defined by PhastCons[229], we calculated the enrichment for overlap with conserved elements for each enhancer set. All enhancer sets have more regions that overlap with conserved elements than expected from length-matched regions drawn at random from the genome. However, enhancers identified by some methods are more likely to be

conserved than others (Figure 5). Across each context, the histone-mark-based, ChromHMM, Villar15, and Ho14 enhancer sets are approximately 1.3x to 1.8x enriched for overlap with conserved elements. Adding DNaseI hypersensitivity data, as in the DNasePlusHistone and EncodeEnhancerlike sets, increases the level of enrichment slightly compared to solely histone-derived enhancers (1.9x–2.3x). In contrast, the FANTOM, Yip12, and p300 enhancers are nearly twice as enriched for conserved regions as the histone-based sets (2.7x, 3.3x, and 2.9x, respectively). GRO-cap enhancers in K562 and Gm12878 are the most enriched for overlap with conserved elements (4.7-4.8x). Evolutionary conservation was considered in the definition of the Yip12 set, but not directly in the FANTOM, p300, or GRO-cap sets. Here we considered element-wise enrichment for the number of enhancer regions overlapping conserved elements; enrichment trends are similar when we consider the number of conserved base pairs overlapped by each enhancer set (Fig S5 in Benton *et al.*[192]).

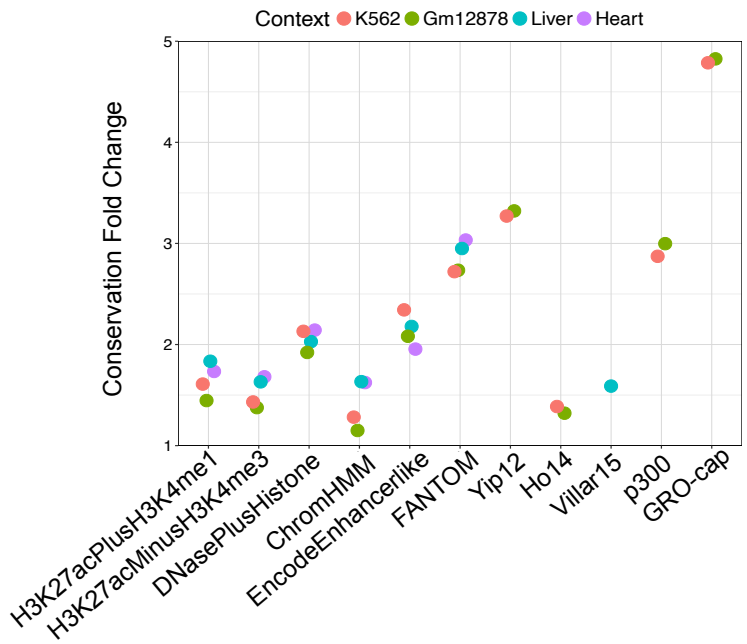

Figure 5. Enhancer sets vary in their degree of evolutionary conservation.
Each point represents the enrichment (fold change compared to randomly shuffled regions) for overlap between a conserved element (combined primate and vertebrate PhastCons) and each enhancer set. Methods based on transcriptional assays and TF binding profiles (GRO-cap, FANTOM, p300, and Yip12) are the most enriched for conserved elements, while sets based on histone modification data alone are among the least enriched.

Identification strategies highlight different subsets of experimentally validated enhancers

Though we lack unbiased genome-wide gold standard sets of enhancers, nearly two thousand human sequences have been tested for enhancer activity *in vivo* in transgenic mice at E11.5 by VISTA[92] and thousands more have been tested in cell lines via MPRAs. Strong ascertainment biases in how regions were selected for testing in these assays prevent their use as a gold standard, but they do provide an opportunity to examine overlap between validated and predicted enhancers. We evaluated the overlap and enrichment of each heart enhancer set with 1,837 regions tested for enhancer activity in the developing heart by VISTA and for each annotated K562 enhancer with 15,720 regions tested in K562 cells by Sharpr-MPRA[62]. All heart enhancer sets are significantly enriched for overlap with the 126 VISTA heart positives (Fig S6 and Table S3 in Benton *et al.*[192]; $p < 0.001$ for all), and each set is at least ~3x more likely to overlap validated enhancers than expected if it was randomly distributed across the genome. However, the heart enhancer sets are also significantly enriched for overlap with VISTA negatives ($p <= 0.004$). This is not surprising as the regions tested by VISTA were largely selected based on having evidence of enhancer activity, and they may have enhancer activity in other contexts not tested by VISTA, including adult heart. However, there is substantial disagreement among the enhancer sets about the status of the VISTA heart enhancers; 16% ($n = 20$) of validated heart enhancers are not predicted to have enhancer activity by any method, and 17% (22) are only predicted by one method (Fig S6 in Benton *et al.*[192]). Similarly, all of the enhancer sets in K562 are significantly enriched for overlap with both activating and repressive regions characterized by Sharpr-MPRA (Fig S7 in Benton *et al.*[192]; $p < 0.001$ for all). There is little variation between the methods in terms of overall enrichment, with most having ~2x relative enrichment for activating regions. Nearly half of the activating regions in the MPRA (47%; 2,508 / 5,373) were not identified by any of the enhancer sets, and 30% of activating regions overlapping a predicted enhancer are unique to a single set (Fig S7 in Benton *et al.*[192]; 891 / 2,747). Thus, comparisons with validated enhancers from both VISTA and MPRAs suggest that different strategies identify different subsets of active regulatory regions in the same context, and that all strategies miss a sizable portion of functional enhancer sequences. However, we again caution against interpreting the relative performance

of different enhancer identification strategies on these data, since there are strong ascertainment biases in how regions were selected for testing. For example, ChromHMM enhancer predictions and DNase I hypersensitivity data were used to select the regions tested by Sharpr-MPRA.

Interpretation of GWAS hits and eQTL is contingent on the enhancer identification strategy used

Functional genetic variants—in particular mutations associated with complex disease—are enriched in gene regulatory regions. Thus, genome-wide enhancer sets are commonly used to interpret the potential function of genetic variants observed in GWAS and sequencing studies. To illustrate this situation, Figure 6A shows a 9kb region at the human chromosome 1p13 locus containing the noncoding region between CELSR2 and PSRC1 associated with both low-density lipoprotein (LDL) cholesterol levels and myocardial infarction (MI) in GWAS[248]. It gives the locations of variants in high LD with the tag SNP, rs12749374, and includes regions identified as liver enhancers by the representative methods analyzed in this study. A comprehensive series of studies showed that the minor allele of rs12740374 creates a C/EBP binding site, causing increased expression of SORT1 specifically in liver and leading to the association with both LDL cholesterol and increased risk of MI[248]. In this example, the region containing the casual SNP is predicted to be an enhancer by four of the seven methods (Figure 6A). This represents a case in which the available enhancer data help to highlight the causal locus.

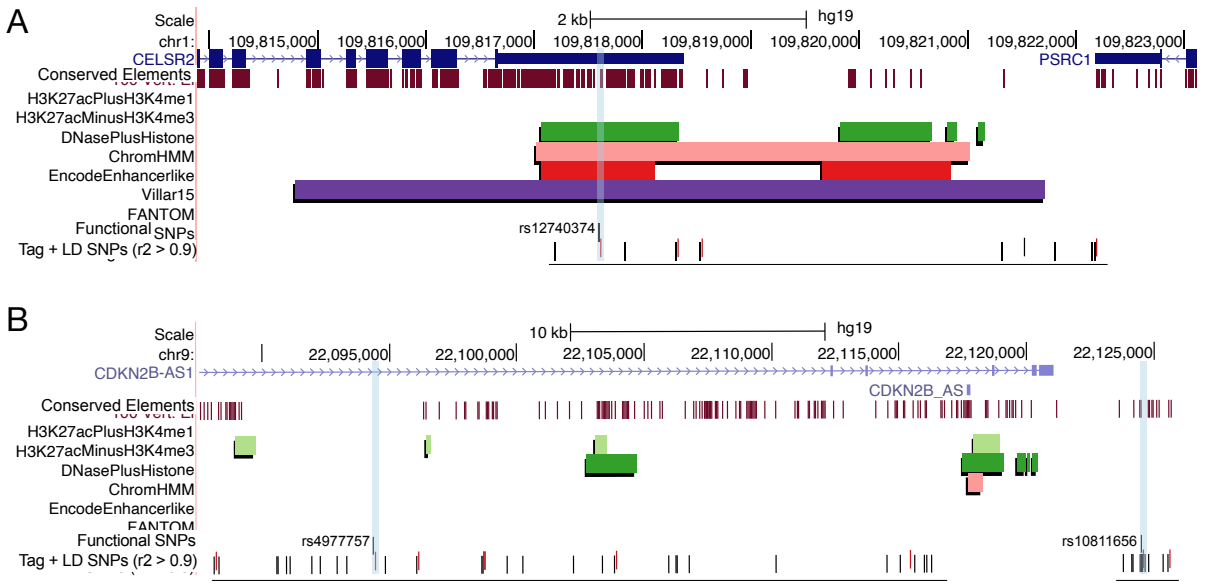

Figure 6. Causal GWAS variants overlap with different enhancer sets.
(A) The 9 kb region on human chromosome 1 containing genetic variants associated with LDL cholesterol levels and MI in GWAS and the causal SNP (rs12740374). Here, the region containing the casual SNP is predicted to be an enhancer by four of the seven methods. GWAS tag SNPs are colored in red and LD blocks are shown with a horizontal line. (B) The 60 kb region of human chromosome 9 containing loci associated with coronary artery disease (CAD) in GWAS. Two of the associated variants (rs10811656 and rs4977757) have been shown to contribute to CAD risk. However, the enhancer annotations in this region are generally non-overlapping and do not highlight either functional variant.

As an alternative example, Figure 6B shows a 60 kb region of human chromosome 9 that contains ten loci associated with coronary artery disease (CAD) in GWAS and other variants in high LD with the GWAS tag variants. It also shows the regions identified as enhancers by six representative methods in heart. Two of these variants (rs10811656 and rs4977757) out of 59 evaluated were recently demonstrated to disrupt binding of TEAD transcription factors in vitro and in vivo in primary human aortic smooth muscle cells, which leads to reduced expression of the cell cycle suppressor protein, p16, and contributes to CAD risk[249]. The enhancer annotations in this region are largely non-overlapping and do not highlight these two functional variants. Indeed, neither overlaps any enhancer annotation.

To explore the frequency of these scenarios genome-wide, we evaluated the overlap of GWAS loci with different enhancer identification strategies by intersecting each of the enhancer sets with 20,458 unique loci significantly associated with traits from the NHGRI-EBI GWAS Catalog. Since the GWAS

catalog contains regions associated with diverse traits, we manually curated the set of GWAS SNPs into subsets associated with phenotypes relevant to liver (n = 346) or heart (n = 2,127) (Table S4 in Benton *et al.*[192]). We found 27% (92 / 346) of the liver associated tag SNPs and 24% (503 / 2,127) of the heart associated tag SNPs overlap an enhancer predicted by at least one of strategies we considered in the appropriate context. (We consider variants in high linkage disequilibrium (LD) below.) While the amount of overlap is low, the heart and liver enhancer sets are almost universally more enriched for overlap with GWAS SNPs that influence relevant phenotypes compared to GWAS SNPs overall (Figure 7, Table S5 in Benton *et al.*[192]; 1.74x–2.68x). FANTOM enhancers are the exception to this trend due to the small number of overlapping context-specific SNPs (Table S6 in Benton *et al.*[192]). This suggests that the different methods, in spite of their lack of agreement, all identify regulatory regions relevant to the target context.

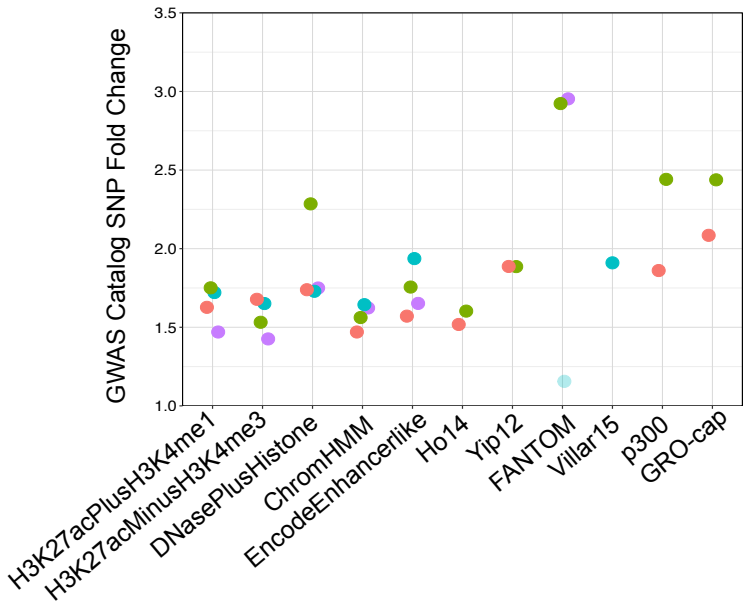

Figure 7. Enhancers have different levels of enrichment with GWAS SNPs.
GWAS SNP enrichment among all enhancer sets for each biological context. All sets are significantly enriched, except FANTOM in K562 and liver contexts due to small sample size.

However, there is variation in the number of overlapping GWAS SNPs between enhancer sets, as is expected given the large variation in the number and genomic distribution of enhancers predicted by different methods (Table S6 in Benton *et al.*[192]). The majority of curated GWAS liver SNPs with any enhancer overlap are overlapped by a single method (53%) and none are shared by all methods (Figure 8A). This trend is also seen in heart, where 58% (293 / 503) of the heart associated SNPs overlapping an enhancer are identified by only a single identification strategy (Figure 8A). This suggests that cases such as the one illustrated in Fig 6B are far more frequent than those like Fig 6A.

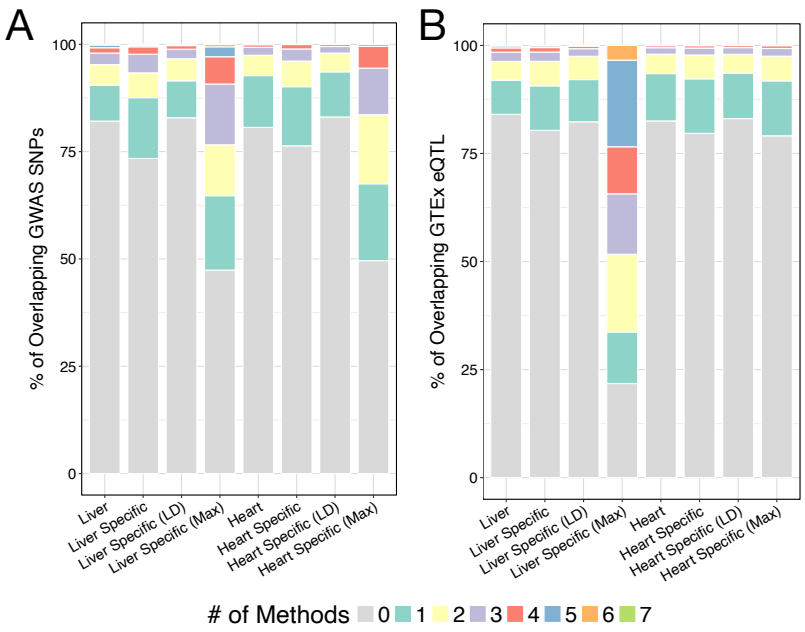

Figure 8. Few genetic variants overlap multiple enhancer sets.
(A) Few GWAS SNPs overlap an enhancer; the colored bars represent the number of methods that identified the region as an enhancer. The majority of these variants are not predicted as enhancers, and very few GWAS variants are overlap enhancers from multiple methods. The conclusions are similar when considering variants in high LD (r^2 > 0.9) with the GWAS tag SNPs in liver (Liver LD; Fig S8 in Benton *et al.*[192]). The pattern is also similar when limiting to SNPs associated with liver or heart related phenotypes (Liver Specific, Heart Specific). When considering the SNP in each LD block with the maximum number of enhancer overlaps there is still a large percentage of SNPs supported by none or only one method (Liver Max). This demonstrates that the situation illustrated in Figure 6B is very common. (B) Among all eQTL that overlap at least one enhancer, the majority is supported by only a single method. This holds for LD- expanded and context-specific sets (Liver LD, Liver Specific, Heart Specific; Fig S8 in Benton *et al.*[192]). Many variants remain unique to a single method, even when limiting to the variant in each LD block overlapping the maximum of enhancer sets (Liver Max).

Since tag SNPs are often not the functional variants, we also considered SNPs in high LD with the GWAS SNPs ($r^2 > 0.9$). The distribution of enhancer overlaps was similar when considering all candidate variants in LD (Figure 8A), although the enrichments were lower (Fig S8 in Benton *et al*.[192]). Even after limiting to GWAS LD blocks with enhancer overlap and selecting the variant with maximum overlap between strategies, 47% (164 / 346) and 50% (1,055 / 2,127) are not predicted as enhancers by any method and 17% (60 / 346) and 18% (381 / 2,127) are only predicted by one enhancer identification method in liver and heart, respectively (Figure 8A). This demonstrates that enhancer maps usually disagree about which variants are likely to be functional, and that the situation illustrated in Figure 6A is rare. Across the entire GWAS Catalog, 33% (6,736 / 20,458) of SNPs overlap an enhancer predicted by at least one of the strategies in one of the contexts we considered. The trends are similar for heart, K562, and Gm12878 (Figure 8A; Fig S8 in Benton *et al*.[192]). This illustrates that the annotation of variants in regions highlighted by GWAS varies greatly depending on the enhancer identification strategies used.

To test if these patterns hold for genetic variants in other functional regions, we analyzed the overlap of enhancer sets with expression quantitative trait loci (eQTL) identified by the GTEx Consortium. Enrichment for overlap with context-specific eQTL in liver or heart is generally higher than enrichment for significant eQTL overall, but the distribution of shared eQTL remains similar (Figure 8B, Figure 9; Table S7 in Benton *et al*.[192]). Within a context, most eQTL do not overlap an enhancer, and there is wide variation in the number of eQTL overlapped by different enhancer sets (Figure 8B; Table S8 in Benton *et al*.[192]). Across liver enhancer sets, 52% (2,925 / 5,585) of all overlapped liver eQTL and 50% (33,941 / 68,563) of general eQTL overlap an enhancer called by only one method (Figure 8B). Considering variants in high LD ($r^2 > 0.9$) does not affect this trend (Figure 8B). After limiting the analysis to the variants with the maximum number of overlaps in each LD block, 15% (3,386 / 22,234) of liver eQTL with enhancer overlap are identified by only one enhancer set (Figure 8B). The lack of overlap is more extreme in heart, where 60% (13,925 / 22,919) heart eQTL overlap a single method, as well as K562 and Gm12878 (Figure 8B; Fig S8 in Benton *et al*.[192]). Thus, in regions known to influence

traits or gene regulation, the interpretation of which variants are causal varies substantially depending on the enhancer identification strategy used.

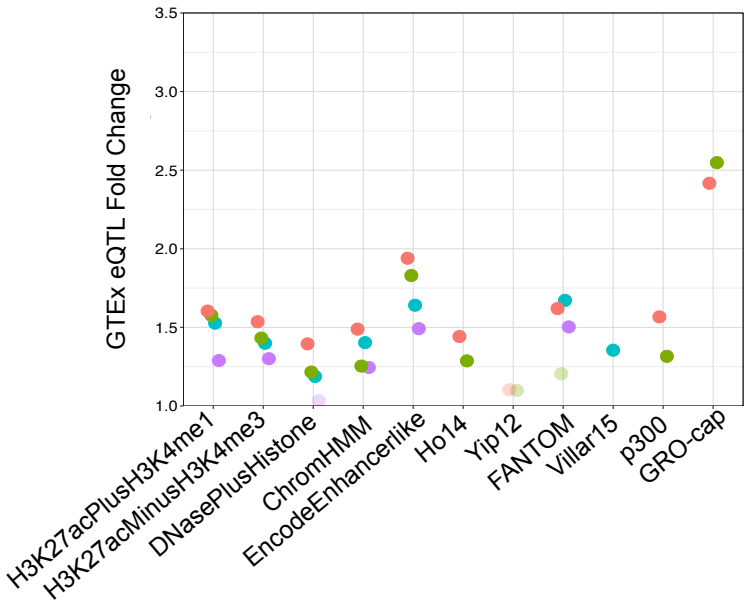

Figure 9. Enhancers have different levels of enrichment with GTEx eQTL.
GTEx eQTL enrichment among all enhancer sets for each biological context. Transparent points indicate nonsignificant enrichment ($p > 0.05$).

Enhancers identified by different strategies have different functional contexts

Given the genomic dissimilarities between enhancer sets, we hypothesized that different enhancer sets from the same context would also vary in the functions of the genes they likely regulate. To test this hypothesis, we identified Gene Ontology (GO) functional annotation terms that are significantly enriched among genes likely targeted by enhancers in each set. We used two different approaches to map to genes and associated GO terms: (*i*) using the joint effect of multiple enhancers (JEME) method for mapping enhancers to putative target genes and then performing gene-based enrichment analyses, and (*ii*) applying the Genomic Regions Enrichment of Annotations Tool (GREAT) (Methods)[155,169]. Many of the GO terms identified by both methods for the enhancer sets are relevant to the associated context (Table 2). However, most of the associated terms for the target-mapping approach were near the root of the

ontologies and thus lacking in functional specificity (Table 2), likely due to the large gene target lists for most enhancer sets (Table S9 in Benton *et al.*[192]). As a result, we focus on the results from GREAT here, and report the results based on JEME target mapping in Figs S9-10 (provided in Benton *et al.*[192]).

Table 2. Top 5 Gene Ontology (Molecular Function) terms for liver enhancer sets.

Enhancer Set	**GO MF Terms (GREAT)**	**GO MF Terms (JEME+WebGestalt)**
H3K27acPlusH3K4me1	cytoskeletal adaptor activity	small molecule binding
	14-3-3 protein binding	anion binding
	leukotriene-C4 synthase activity	nucleoside phosphate binding
	nucleobase-containing compound transmembrane transporter activity	nucleotide binding
	FAD binding	transferase activity
H3K27acMinusH3K4me3	14-3-3 protein binding	oxidoreductase activity
	cytoskeletal adaptor activity	anion binding
	thyroid hormone receptor binding	small molecule binding
	ARF guanyl-nucleotide exchange factor activity	nucleoside phosphate binding
	high-density lipoprotein particle binding	nucleotide binding
DNasePlusHistone	cytoskeletal adaptor activity	small molecule binding
	glucocorticoid receptor binding	anion binding
	nucleobase-containing compound transmembrane transporter activity	transferase activity
	high-density lipoprotein particle binding	nucleotide binding
	14-3-3- protein binding	nucleoside phosphate binding
ChromHMM	high-density lipoprotein particle binding	nucleotide binding
	nucleobase-containing compound transmembrane transporter activity	nucleoside binding
	cytoskeletal adaptor activity	purine nucleoside binding
	14-3-3 protein binding	DNA binding
	retinoid X receptor binding	RNA binding
EncodeEnhancerlike	cytoskeletal adaptor activity	nucleotide binding
	14-3-3 protein binding	transferase activity
	nucleobase-containing compound transmembrane transporter activity	small molecule binding
	apolipoprotein A-I binding	anion binding
	high-density lipoprotein particle binding	carbohydrate derivative binding
FANTOM	glucocorticoid receptor binding	structural constituent of ribosome
	protein kinase binding	receptor binding
	kinase binding	cell adhesion molecule binding
	methylglutaconyl-CoA hydratase activity	molecular function regulator
	vitamin D response element binding	transcription regulatory region DNA binding

Villar15	protease binding	anion binding
	phosphatidylinositol 3-kinase binding	small molecule binding
	14-3-3 protein binding	oxidoreductase activity
	cytoskeletal adaptor activity	cofactor binding
	glucocorticoid receptor binding	oxidoreductase activity, acting on CH-OH group of donors

The majority of the top 30 significant annotations from GREAT for each enhancer set are not enriched in any other set in the same context, and no terms are shared by all of the methods in a given context (Figure 10, lower triangle). In all of these pairwise comparisons, fewer than half of the GO terms are shared between a pair of enhancer sets. Furthermore, many of the terms shared by multiple enhancer sets are near the root of the ontology (e.g., nucleotide binding) and thus are less functionally specific. These results provide evidence that the different enhancer sets influence different functions relevant to the target biological context. These trends hold for both the Biological Process (BP) and Molecular Function (MF) ontologies and considering the top 10 and 50 annotations for each set (Figs S11-13 in Benton *et al.*[192]).

To further compare the enriched GO MF and BP annotations of each enhancer set in a way that accounts for the distance between GO terms in the ontology hierarchy and their specificity, we computed a semantic similarity measure developed for GO annotations[236,237]. The ChromHMM and EncodeEnhancerlike enhancer sets are among the most functionally similar, with similarity scores near 0.80 in most contexts (Figure 10, upper triangle; Fig S12 in Benton *et al.*[192]). This is not surprising given that their underlying assays overlap. The functional similarity scores are lower for comparisons of the other histone modification sets, around 0.50–0.75. In all comparisons, the FANTOM enhancers have the lowest functional similarity with other enhancer sets—below 0.40 in the vast majority of comparisons in K562, liver, and heart (Figure 10; Fig S12 in Benton *et al.*[192]). FANTOM is more similar to other methods in Gm12878, with an average score of 0.59 (Fig S12 in Benton *et al.*[192]). As a benchmark, biological replicates of the Gm12878 H3K27ac ChIP-seq peaks received a similarity of 0.93. This suggests different functional influences for enhancer sets from the same context identified by different methods, with

FANTOM as a particular outlier. We note that enhancer target gene identification remains a challenging problem, and both strategies for mapping enhancers to potential target genes considered here (GREAT and JEME) likely include false positives and negatives. However, insofar as they reflect the genomic context of the different enhancer sets, they reveal significant functional differences between enhancer identification methods.

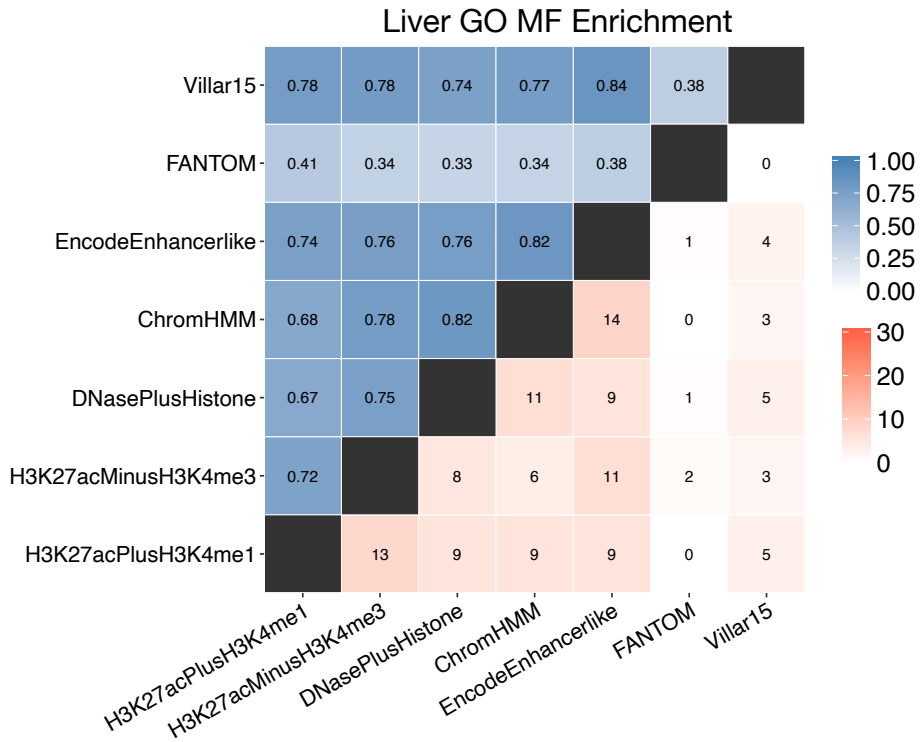

Figure 10. Enhancer sets from the same biological context have different functional associations.
We identified Gene Ontology (GO) functional annotations enriched among genes likely to be regulated by each enhancer set using GREAT. The upper triangle represents the pairwise semantic similarity for significant molecular function (MF) GO terms associated with predicted liver enhancers. The lower triangle shows the number of shared MF GO terms in the top 30 significant hits for liver enhancer sets. Results were similar when using enhancer-gene target predictions from JEME (Figs S9-10 provided in Benton *et al.*[192]).

Genomic and functional clustering of enhancer sets

Our analyses of enhancer sets within the same biological context reveal widespread dissimilarity in both genomic and functional features. To summarize and compare the overall genomic and functional

similarity of the enhancer sets across contexts, we clustered them using hierarchical clustering and MDS based on their Jaccard similarity in genomic space and the GO term functional similarity of predicted target genes.

Several trends emerged from analyzing the genomic and functional distribution within and between biological contexts. First, the FANTOM eRNA and GRO-cap enhancers are consistently distinct from all other enhancer sets in both their genomic distribution and functional associations (Fig 6). Differences between eRNA and non-eRNA enhancer sets appear to dominate any other variation introduced by biological, technical, or methodological differences.

A second trend in these comparisons is that similarity in genomic distribution of enhancer sets does not necessarily translate to similarity in functional space, and vice versa. For example, although EncodeEnhancerlike regions are close to ChromHMM and the histone-derived H3K27acPlusH3K4me1 set and the machine learning models in the genomic-location-based projection (Figure 11A,C), they are located far from those sets in the functional comparisons and hierarchical clustering (Figure 11B,D). Finally, comparing enhancer sets by performing hierarchical clustering within and between biological contexts reveals that genomic distributions are generally more similar within biological contexts, compared to other sets defined by the same method in a different context (Figure 11E). For example, the ChromHMM set from heart is more similar to other heart enhancer sets than to ChromHMM sets from other contexts. In contrast, the enhancer set similarities in functional space are less conserved by biological context (Figure 11F). Here, the heart ChromHMM set is functionally more similar to the H3K27acMinusH3K4me3 set from liver cells than other heart enhancer sets. In general, cell line enhancer sets (red and green) show more functional continuity than heart and liver sets (blue and purple). However, FANTOM enhancers are the exception to these trends; FANTOM enhancers from each context form their own cluster based on their genomic distribution, underscoring their uniqueness. GRO-cap enhancers cluster with FANTOM in the genomic location clustering and with their cellular contexts in the functional clusters.

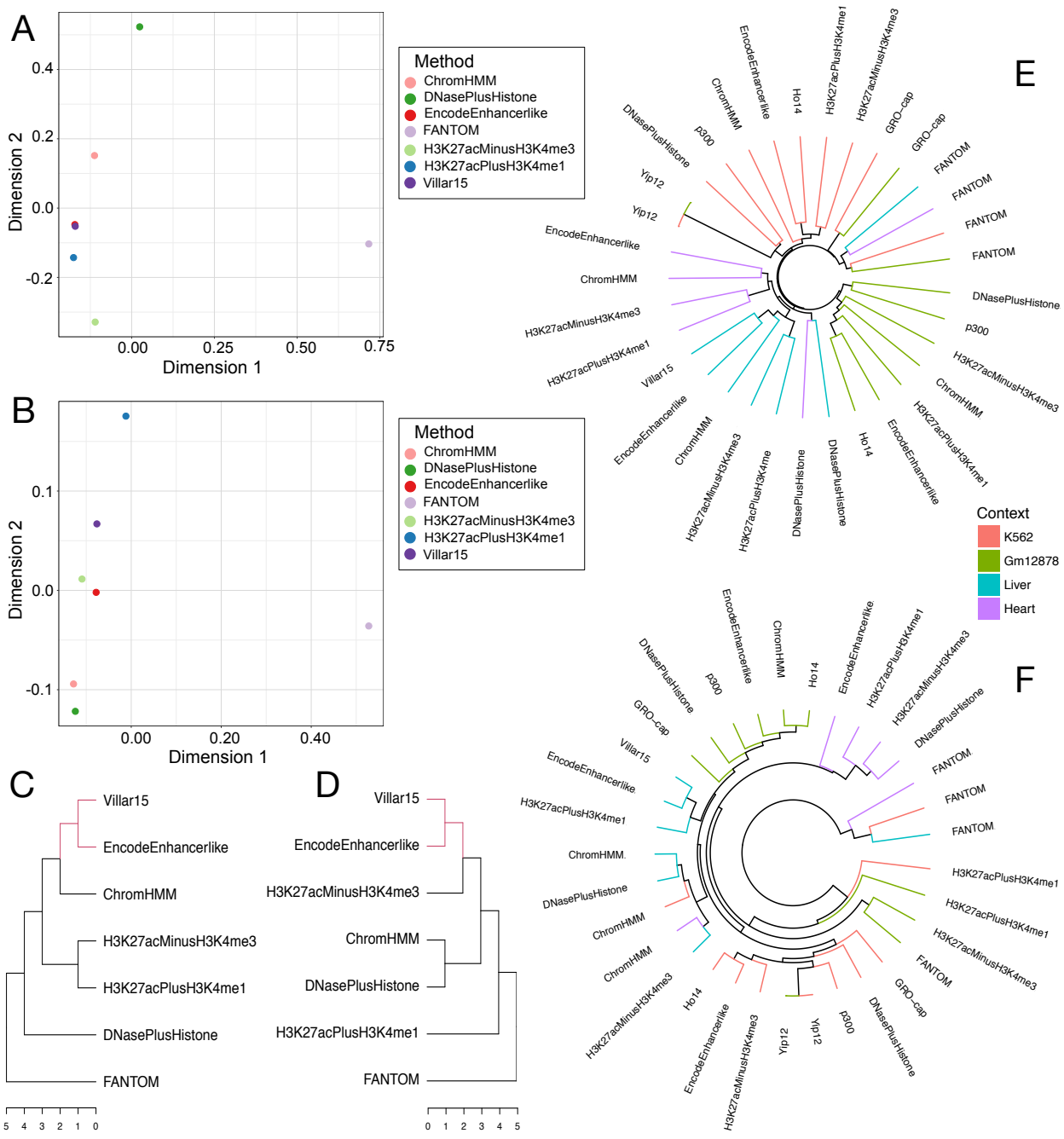

Figure 11. The genomic and functional similarities between enhancer sets are not consistent.
(A) Multidimensional scaling (MDS) plot of liver enhancer sets based on the Jaccard similarity of the genomic distributions (Figure 4). (B) MDS plot for liver enhancers based on distances calculated from molecular function (MF) Gene Ontology (GO) term semantic similarity values with GREAT (Figure 10). (C, D) Ranked hierarchical clustering based on the Jaccard similarities of the genomic distributions (C) of all liver enhancer sets compared to clustering based on GO semantic similarity (D). FANTOM enhancers are the most distant from all other enhancer sets in both genomic and functional similarity, but the relationships between other sets are not conserved. Red branches denote identical subtrees within the hierarchy. (E) Hierarchical clustering based on genomic Jaccard distances for all contexts and methods with annotations in each context. (F) Hierarchical clustering of all available enhancer sets based on GO term distances. Terminal branches are colored by biological context. With the exception of FANTOM enhancers, the enhancer sets' genomic distributions are more similar within than between biological

contexts. Functional similarity does not always correlate with genomic similarity, and the clustering by biological context is weaker in functional space.

Combining enhancer sets does not strongly increase evidence for regulatory function

Although there are large discrepancies in genomic and functional attributes between enhancer sets identified by different methods in the same context, we hypothesized that the subset of regions shared by two or more sets would have stronger enrichment for markers of gene regulatory function. To test this, we analyzed whether regions identified by multiple methods have increased "functional support" compared to regions identified by fewer methods. We evaluated three signals of functional importance: *(i)* enrichment for overlap with evolutionarily conserved elements, *(ii)* enrichment for overlap with GWAS SNPs, and *(iii)* enrichment for overlap with GTEx eQTL. For each, there are only small changes as the number of methods identifying a region increases (Figure 12A-C). Regions identified as enhancers by more than one method are slightly more enriched for conserved elements compared to the genomic background, but there is little difference among regions identified by 2–5 methods (Figure 12A). Regions predicted by 6 or more methods are significantly more enriched for conserved elements than those with less support, but effect size is modest (1.36x for 1 vs. 1.62x for 6+). There is a modest increase in the enrichment for overlap with GWAS SNPs among enhancers identified by more identification methods (1.50x for 1 vs. 1.89x for 6+); however, given the relatively small number of GWAS SNP overlaps, none of these differences were statistically significant (Figure 12B). We observed no increase in the enrichment for overlap with eQTL as the support for enhancer activity increased (Figure 12C). Thus, we do not find strong evidence of increased functional importance in enhancers identified by multiple methods compared to enhancers identified by a single method. Importantly, this implies that intersecting enhancer identification strategies will focus on a smaller set of enhancers with only modest evidence for increased functional relevance.

Several enhancer identification methods provide confidence scores that reflect the strength of evidence for each enhancer. We hypothesized that high confidence enhancers from one method would be

more likely to overlap enhancers identified by other methods. To test this, we ranked each enhancer based on its confidence or signal, with a rank of 1 representing the highest confidence in the set. There was no clear trend between the confidence score of an enhancer from one method and the number of methods that identified the region as an enhancer (Figure 12D; Figs S15-18 in Benton *et al*.[192]). Overall, enhancers identified by multiple methods show a similar confidence distribution when compared to regions identified by a single method. Indeed, for some enhancer sets the median score decreases as the regions become more highly shared (Figs S15 and S18 in Benton *et al*.[192]). This provides further evidence that building enhancer sets by simple combinations of existing methods is unlikely to lead to a higher confidence subset, and that filtering based on simple agreement between methods may not substantially improve the specificity of enhancer predictions.

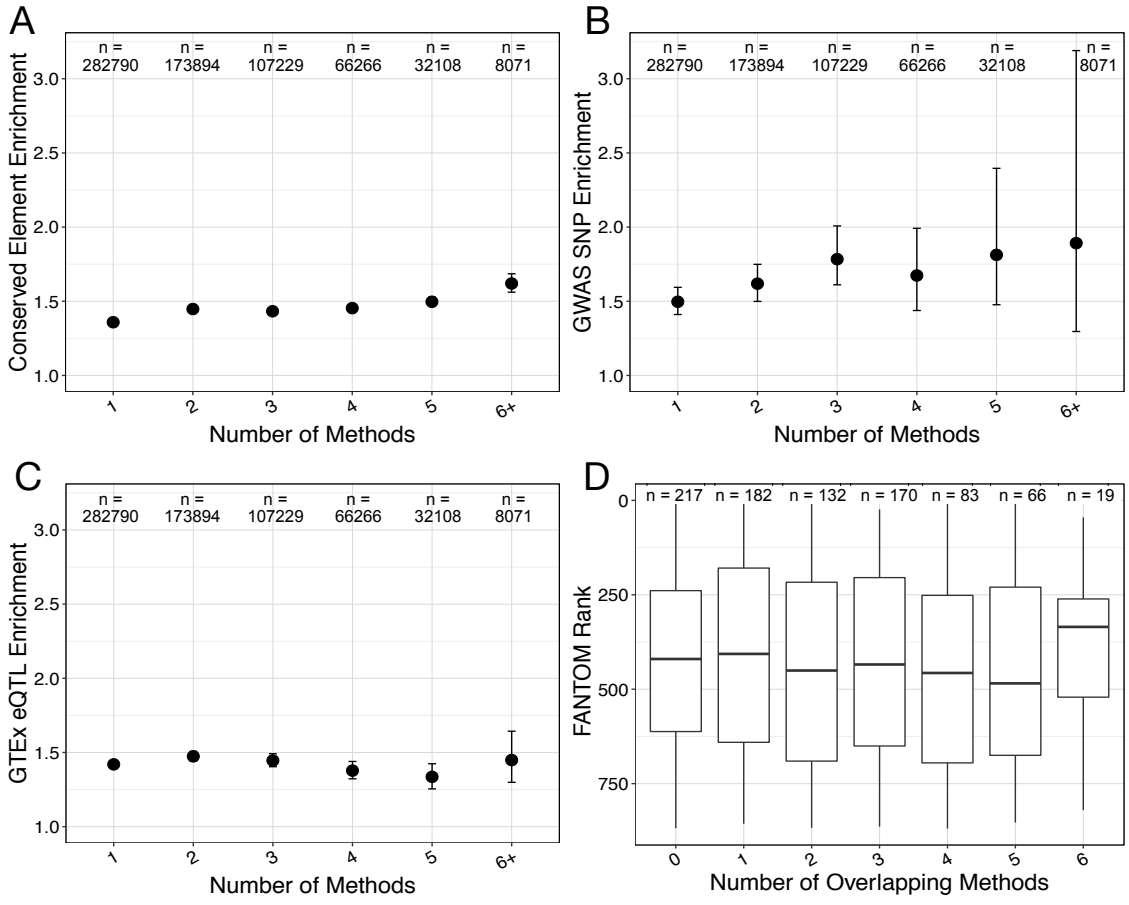

Figure 12. Enhancers identified by multiple methods have little additional evidence of function.

(A) Enrichment for overlap between conserved elements (n = 3,930,677) and liver enhancers stratified by the number of identification methods that predicted each enhancer. (B) Enrichment for overlap between GWAS SNPs (n = 20,458) and liver enhancers stratified by the number of identification methods that predicted each enhancer. (C) Enrichment for overlap between GTEx eQTL (n = 429,964) and liver enhancers stratified by the number of identification methods that predicted each enhancer. In (A-C), the average enrichment compared to 1,000 random sets is plotted as a circle; error bars represent 95% confidence intervals; and n gives the number of enhancers in each bin. The only significant differences are found in the enrichment for evolutionary conservation (A), but the difference is modest in magnitude (1.36x for 1 vs. 1.62x for 6+). (D) Boxplots showing the distribution of confidence score ranks for FANTOM enhancers in liver partitioned into bins based on the number of other methods that also identify the region as an enhancer. Lower rank indicates higher confidence; note that the y-axis is flipped so the high confidence (low rank) regions are at the top. The lack of increase in enhancer score with the number of methods supporting it held across all methods tested (Figs S14-17 provided in Benton *et al*.[192]).

Enhancer sets clustered on similarity of transcription factor binding motifs have different patterns

We calculated the enrichment for TF binding motifs within each enhancer set relative to dinucleotide-frequency matched random sequences. Most enhancer sets were enriched for more than half of the motifs in the database (~300), compared to random (Table S10 in Benton *et al*.[192]). Thus, most pairs have a relatively high Jaccard similarity index (>0.8) for overlap between enriched motifs. For comparison, we calculated the baseline Jaccard similarity for random motif sets of a matched size, which produced average scores between 0.6-0.69. We note that due to the context-dependent nature of TF binding, the presence or even enrichment of a motif does not guarantee function. Alternately, the lack of a significant enrichment for a binding site among a set of sequences, does not necessarily indicated a lack of activity.

We clustered the enhancer sets for each tissue based on the similarity of enriched TF binding site motifs (Figure 13). The FANTOM eRNA and GRO-cap enhancers are consistently distinct from all other enhancer sets in their in predicted TF binding site enrichment profiles. The enhancer sets based on similar combinations of histone modifications (H3K27acMinusH3K4me3, H3K27acPlusH3K4me1, ChromHMM, and Ho14/Villar15 where available) cluster together in most biological contexts. In the two cell lines, the DNasePlusHistone, p300, Yip12, and EncodeEnahncerlike sets are more clustered. These results suggest that while there are similarities in the sequence-level characteristics, this may not translate to similar TF binding profiles for enhancers defined by different approaches in the same cellular context.

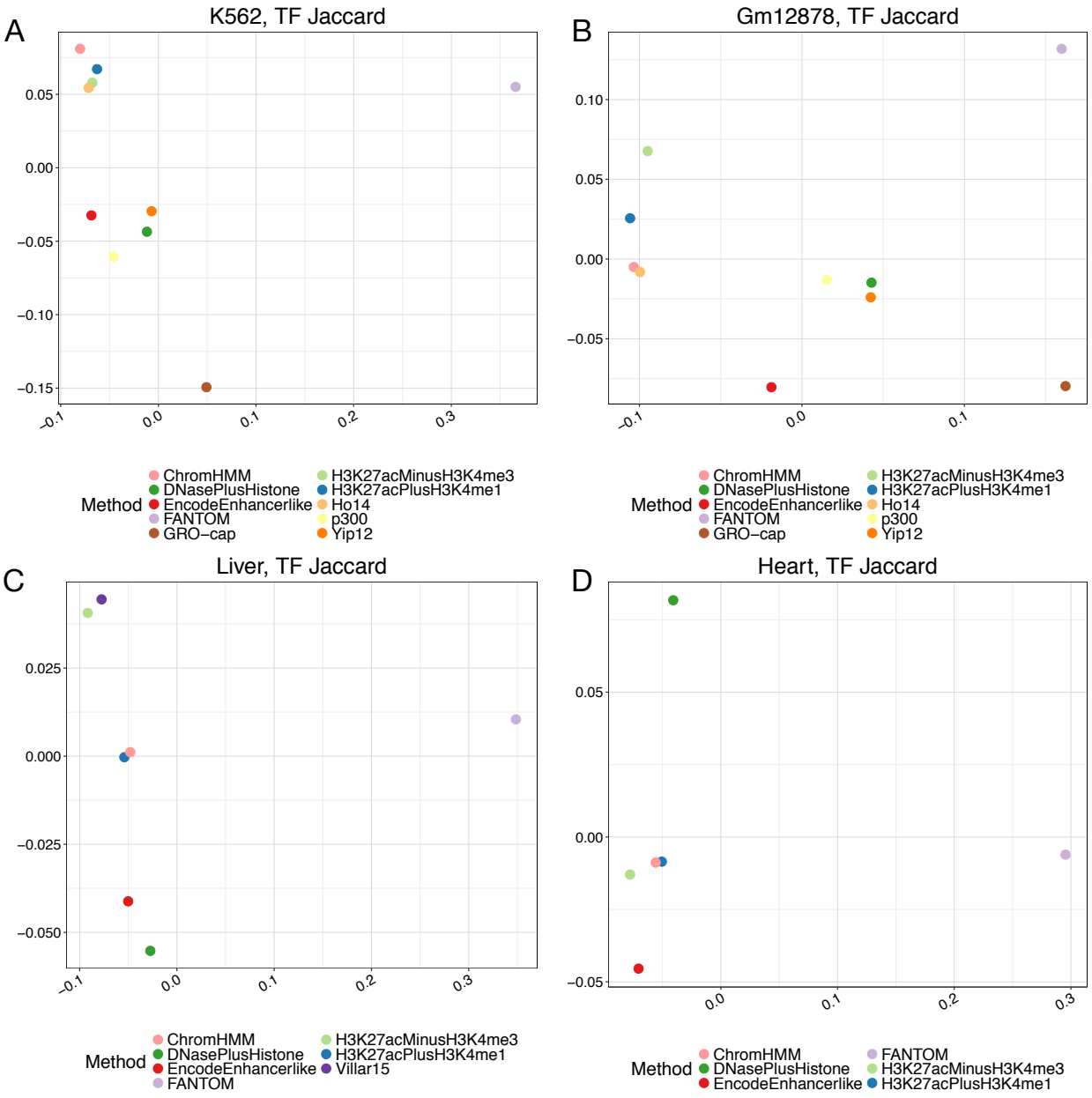

Figure 13. Enhancer sets have different enriched TF binding motifs.
We computed the enrichment of 402 TF binding site motifs in different enhancer sets and clustered the enrichment profiles for each enhancer set based on Jaccard similarities using multidimensional scaling (MDS) (Methods). FANTOM and GRO-cap are consistent outliers with the largest differences in predicted TF binding site enrichment, as in the genomic and GO analyses reported in Figure 11A-B; however, the clustering of other methods varies. Most enhancer sets are enriched for more than half of the motifs in the database (~300) compared to dinucleotide frequency matched random sequences, and thus most pairs of sets have Jaccard similarity >0.8. Random motif sets of a matched size produce average Jaccard similarities between 0.6–0.69. When combined with the observed dissimilarity of GO terms, these results suggest that similarities in sequence-level characteristics may not translate into similar regulatory targets. (A) K562, (B) Gm12878, (C) liver, and (D) heart.

Using the most confident predictions does not significantly improve results

We do not find evidence that analyzing only the top enhancer predictions from each method changes the results reported here. Using the five enhancer identification strategies with confidence or signal rankings in each context we compared the top 100, 500, and 800 enhancer predictions. There remains little sharing between these subsets, with most enhancer regions remaining unique to a single method. Furthermore, the level of enrichment for overlap with functional attributes remains largely similar, but less significant in many cases (Figure 14; Figure S20-S21 in Benton *et al*.[192]).

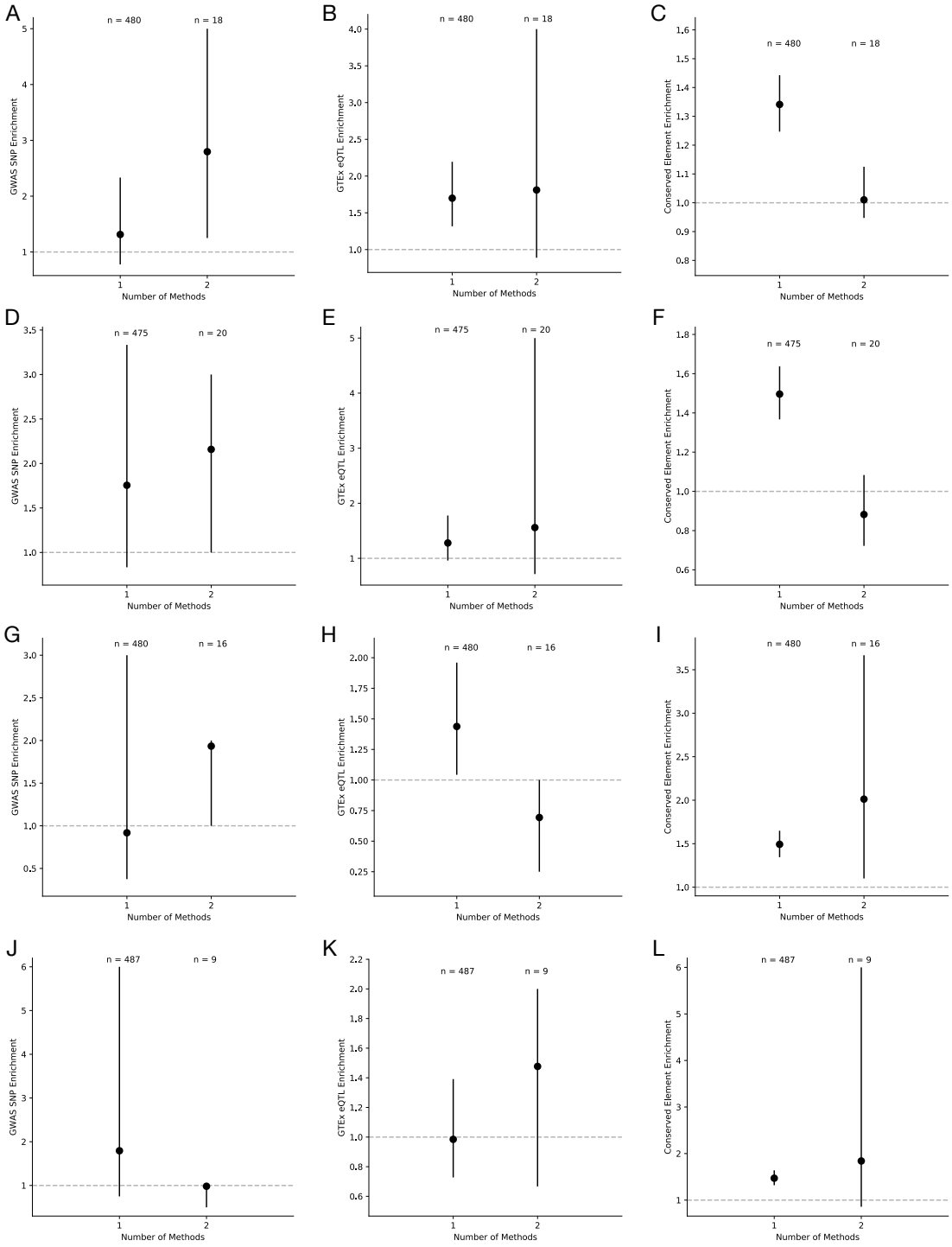

Figure 14. Enrichment for functional attributes is similar for top 100 predictions.
For enhancer sets in (A–C) K562, (D–F) Gm12878, (G–I) liver, and (J–L) heart, we consider the top 100 regions per method ranked by confidence or signal scores. This analysis was limited to enhancer sets that could be ranked (DNasePlusHistone, EncodeEnhancerlike, H3K27acPlusH3K4me1, H3K27acMinusH3K4me3, FANTOM). The dotted line represents the level of enrichment expected under a random null distribution; error bars show empirical 95% confidence intervals. This analysis does not include p300 or GRO-cap data for K562 or Gm12878.

Short sequence motifs are not predictive of reproducibility across identification approaches

We used a machine learning approach to test the hypothesis that enhancers that are predicted by multiple methods have consistent sequence differences from those that do not replicate. We trained a support vector machine (SVM) classifier to distinguish between enhancers that are reproducible across methods versus those that are unique to a single method. The feature set for the SVM classifier was the k-mer spectrum for enhancer sequences in each category. We chose a k-value of 6 to capture potential TF binding sites within the enhancers, where similar k-mer spectra may indicate similar TF binding profiles.

We evaluated the approach on our enhancer datasets using ten-fold cross validation. The SVM model performs poorly at distinguishing reproducible enhancers (ROC AUC = 0.6, Figure 15). This suggests that short sequence patterns within shared enhancers are not substantially different from enhancers that are uniquely identified by one method. Our previous TF binding motif analyses indicate that there is some sharing of enriched motifs between enhancer sets, but this does not distinguish between shared and unique enhancer sequences.

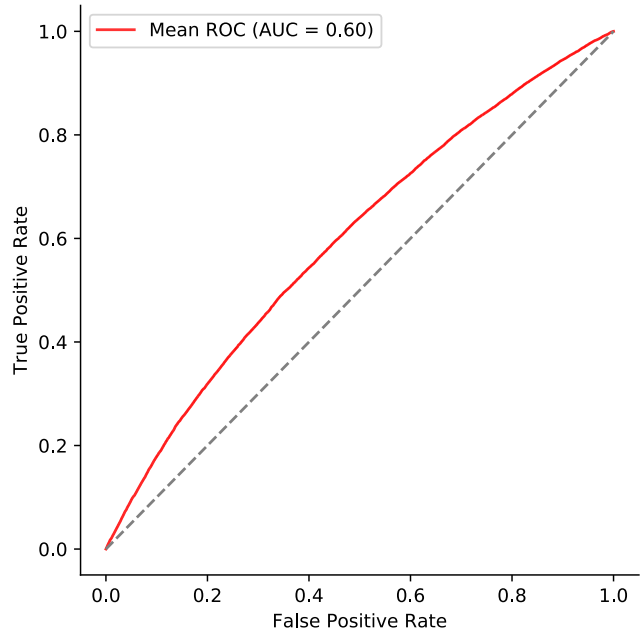

Figure 15. SVMs do not distinguish reproducible enhancers from unique enhancers.

We trained an SVM classifier to predict enhancers that are reproducible across methods versus those that are unique using short sequence motifs (k = 6). The model performs poorly, suggesting that the sequence patterns within shared enhancers are not different from those within enhancers defined by only one method.

Conclusion

Accurate enhancer identification is a challenging problem, and recent efforts have produced a variety of experimental and computational approaches. This chapter provides a quantification of the genomic and functional differences between enhancer sets identified by eleven of these approaches. We find that the enhancer sets analyzed here differ significantly in their genomic, evolutionary, and functional attributes. Although all enhancer sets analyzed here agree more than would be expected by chance, and are enriched for markers of functional relevance, the differences we observe are striking. Our results show that the choice of enhancer identification strategy has the ability to influence downstream biological conclusions about transcription factor binding potential, evolutionary history of enhancer elements, and regulatory mechanisms of complex disease. Simple approaches to generate more confident sets of enhancers by combining methods or using subsets of more well-supported elements does not improve performance or increase evidence of function. Furthermore, machine learning fails to distinguish short sequence motifs that are predictive of reproducibility, complicating attempts to explain overlap between existing enhancer sets.

Ultimately, this work suggests that different strategies contribute unique information towards the identification of functionally important enhancers. Using different strategies can yield substantially different biological interpretations and conclusions, e.g., about the gene regulatory potential of a genetic variant or the degree of evolutionary constraint on enhancers. Thus, our findings in this chapter complicate the use of annotated enhancers to study the mechanisms of gene regulation and to elucidate the molecular underpinnings of disease, most notably in non-coding variant prioritization.

CHAPTER III

Characterizing Gene Expression Consequences of Structural Variants Disrupting Gene Regulation[2]

Introduction

Structural variants (SVs) can have a profound impact on gene expression and have been implicated in phenotypes ranging from developmental abnormalities to cancer to neuropsychiatric disorders[124,200,206,250]. While SVs may affect protein sequence, their effects are often mediated through changes to the regulatory architecture[198]. In this context, regulatory architecture refers to both the individual regulatory elements involved in the maintaining the expression of genes and the three-dimensional chromatin structures that facilitate regulatory interactions. Effect of SVs have been demonstrated in many disease contexts, where duplications, deletions, translocations, inversions, and topologically associated domain (TAD) boundary disruptions have all been shown to cause enhancer dysfunction associated with disease risk[124].

This chapter includes a novel cohort of genome- and RNA-sequencing from the CommonMind Consortium—a large effort that brings together more than 1,000 brain samples. These samples include many individuals with schizophrenia or bipolar disorder, making this a valuable resource for examining the contribution of genetic variation to neuropsychiatric disease. We integrate genome- and RNA-sequencing from 629 samples with publicly available functional genomics data to identify SVs that putatively alter gene regulatory architecture. We then use regulation-associated data to functionally annotate SVs and quantify the effects of these SVs on gene expression. We find that SVs that affect regulatory elements and features of the chromatin architecture are at significantly lower frequencies than expected, consistent with the hypothesis that SVs impacting regulatory architecture are deleterious. We also find that SVs altering regulatory elements have a clear effect on gene expression. Finally, we use

[2] Adapted from Han, L., Zhao, X., Benton, M.L. *et al.* Functional annotation of rare structural variation in the human brain. *Nat Commun* 11, 2990 (2020). https://doi.org/10.1038/s41467-020-16736-1

these results to develop a model of regulatory disruption for SVs, allowing us to annotate and prioritize pathogenic variants based on the inferred changes to gene regulation. This model improves the current approach for SV prioritization, highlighting the importance of regulatory SVs in disease.

Methods

Cohort description

Samples were included from two different cohorts. The CommonMind Consortium (CMC) study is a combined collection of brain tissues from the Mount Sinai NIH Brain Bank and Tissue Repository (n = 127), The University of Pennsylvania Brain Bank of Psychiatric Illnesses and Alzheimer's Disease Core Center (n = 62) and The University of Pittsburgh NIH NeuroBioBank Brain and Tissue Repository (n = 139). Tissue for the collection was dissected at each brain bank and shipped to the Icahn School of Medicine at Mount Sinai (ISMMS) for nucleotide isolation and data generation in one facility in order to reduce site-specific sources of technical variation. Postmortem tissue from schizophrenia and bipolar disorder cases were included if they met the diagnostic criteria in DSM-IV for schizophrenia or schizoaffective disorder, or for bipolar disorder, as determined in consensus conferences after review of medical records, direct clinical assessments, and interviews of care providers. Cases that had a history Alzheimer's disease, and/or Parkinson's disease, or acute neurological insults (anoxia, strokes, and/or traumatic brain injury) immediately prior to death, or were on ventilators near the time of death, were excluded. The CMC_HBCC study includes brain samples from the NIMH Human Brain Collection Core (n = 445). All specimens were characterized neuropathologically, clinically and toxicologically. A clinical diagnosis was obtained through family interviews and review of medical records by two psychiatrists based on DSM-IV criteria. Nonpsychiatric controls were defined as having no history of a psychiatric condition or substance use disorder. Among the 773 samples used here, there are 505 males and 268 females. Self-reported ancestries consisted of 484 European, 264 African, 15 Hispanic, 9 Asian and 1 other. Forty-eight percent of the samples had a psychiatric diagnosis (287 schizophrenia, 83 bipolar

disorder) and the remaining 403 were considered controls[251]. All research complied with ethical regulations and was approved by the Vanderbilt University Medical Center Institutional Review Board (IRB#161488).

Genome sequencing pipeline

Tissue was sampled from the dorsal lateral prefrontal cortex (DLPFC) and DNA isolated for all 773 individuals in the cohort. Genome sequencing was performed by the New York Genome Center and reads were aligned to the GRCh37/hg19 human reference sequence.

Structural variant discovery and description

SVs were called using a previously described ensemble approach to maximize sensitivity followed by a refinement step to reduce the number of false positive calls[252]. The approach was applied to all 773 of the individuals from the CMC cohort, with a 99.9% (n = 772) success rate. Outliers related to technical aspects of the pipeline were excluded, leaving 755 (97.8%) individuals with SV data. In total, the pipeline yielded 125,260 SVs, including 62,948 deletions, 30,547 duplications, 31,155 insertions, 268 simple inversions, 341 complex SVs, and 1 reciprocal translocation. On average, 6220 SVs were identified per sample, consisting of 3579 deletions, 755 duplications, 1839 insertions, 15 inversions, and 14 complex SVs. In this chapter we focus primarily on deletions and duplications, although other SV types are discussed elsewhere[253].

RNA sequencing pipeline

For all samples, we used RNA-sequencing (RNA-seq) data processed by the pipeline described previously[253]. RNA-sequencing reads were aligned to GRCh37 with STAR (v2.4.0g1)[254] from the original FASTQ files. Uniquely mapping reads overlapping genes were counted with featureCounts (v1.5.2)[255] using annotations from Ensembl v75. RNA-seq samples were processed separately for each cohort and

normalized to adjust for batch effects and technical variation. To evaluate the impact of SVs on gene expression, the RNA-seq and genome sequencing were matched by individual.

Measures of gene expression consequences of SVs

To quantify expression changes associated with SVs, we considered two values: the relative expression and an expression *z*-score. Relative expression was defined as the average expression of SV carriers divided by noncarriers. The expression *z*-scores were calculated using only data from noncarriers for the mean and standard deviation.

Genomic and cis-regulatory annotation sources

All data were downloaded in the GRCh37/hg19 build of the human genome. We used TSS definitions from Ensembl v75. To map regions of open chromatin, we used a set of DNase hypersensitive sites (DHSs) downloaded from Roadmap Epigenomics[191]. We mapped the three-dimensional chromatin architecture using TAD domains identified by PsychENCODE from Hi-C contact matrices with 40 kb resolution in the prefrontal cortex (PFC)[256]. As a proxy for TAD boundaries or other insulated regions, we used a set of CTCF binding sites from ChIP-seq data downloaded from ENCODE in brain-relevant cell types[9]. We merged overlapping CTCF peaks from each tissue into a single consensus region.

We downloaded PFC enhancer annotations from the PsychENCODE project[256]. These were generated by overlapping cross-tissue DNase-seq and ATAC-seq assay information with H3K27ac ChIP-seq peaks. Regions overlapping H3K4me3 peaks and within 2 kb of a TSS were excluded from the set of putative enhancers. All ChIP-seq, ATAC-seq, and DNase-seq data were filtered to include only high-signal peaks with a z-score greater than 1.64. We also downloaded the high confidence set of enhancer annotations which, in addition to the criteria above, require high PFC H3K27ac ChIP-seq signal (z-score > 1.64) in both the PsychENCODE and Roadmap Epigenomics experiments. We generated a set of promoter annotations by using 2 kb windows upstream from each TSS. We intersected these 2 kb windows with PFC H3K27ac from PsychENCODE and PFC H3K4me3 from Roadmap Epigenomics to

create a set of high confidence promoters[191,256]. As in the enhancer definition, the H3K27ac and H3K4me3 ChIP-seq data included only high signal peaks with a z-score > 1.64. For comparison with other sites with potential regulatory activity, we used the set of DNase hypersensitivity sites downloaded from Roadmap Epigenomics, and TAD boundary elements defined from Hi-C data[136,191]. For comparison across tissues, we downloaded H3K27ac and H3K4me3 ChIP-seq peaks in liver and Gm12878 from Roadmap Epigenomics and defined enhancers elements using the procedure outlined above[191].

Definition of gene targets for regulatory SVs

We tested multiple approaches to define putative gene targets for SVs that impact enhancer sequences. We considered the nearest gene, all genes within the same TAD, all genes within 1 Mb, and genes linked to enhancers through Hi-C[256]. We chose the Hi-C linked target genes for published analyses, although we include some results for genes within 1 Mb here as well.

Comparison of allele frequencies across SV annotation classes

In order to assess whether SVs affecting particular functional elements (genes, enhancers, etc.) were present at lower frequencies than nonfunctional SVs we needed to account for differences in lengths. Because longer SVs are more likely to affect functional elements and, thus, are rarer (Supplementary Figure 1, provided in Han *et al.*[253]) we wanted to break the dependence on length to assess the role of the functional elements on AF. For each class of SVs affecting a particular annotation we required a matching intergenic SV of the same class that affected none of the annotations with a length within 500 base pairs or 10% of the length of the functional SV, whichever was shorter. We then tested for differences in AF distributions between the two equal numbers of functional and intergenic SVs using the non-parametric Wilcoxon test.

Regression model to predict gene expression changes

We applied a generalized linear model to test the relationship between SVs overlapping annotated regulatory annotations and gene expression *z*-scores. For models using the 1 Mb window strategy to link enhancers to genes, we included the proportion of exonic sequence, the number of affected regulatory annotations (enhancers, promoters, CTCF binding sites, etc.), and the SV length. For all other models we used a joint linear model including the proportion of exonic sequence, SV length, and whether SV and gene were within the same TAD; we replaced the number of affected regulatory annotations with sum proportion of affected enhancers and promoters.

Results

Evidence for selection against SVs affecting gene regulatory elements

We characterized how frequently SVs putatively alter gene dosage— the number of gene copies— based on overlap with genes or regulatory elements. We defined a set of regulatory elements that included CTCF sites (n = 100,894), enhancers (n = 79,056) derived from brain tissue and promoters (2 kb upstream of the TSS). Genes were defined as those in Ensembl v75 (n = 57,773); where noted we split protein-coding genes (coding) from others which we label as other transcribed products. For comparison, we defined two nonfunctional categories of SVs that did not overlap any annotation including those falling within introns (intronic) or those falling outside of any gene (intergenic). We note that these nonfunctional SV categories will include some proportion of SVs altering functional elements that were either not included, or that have not yet been identified, which should make our comparisons conservative.

The allele frequency (AF) of SVs affecting protein-coding genes (AF = 0.00168, p = 7.42E−15), enhancers (AF = 0.00123, p = 6.42E−30), and CTCF sites (AF = 0.00161, p = 1E−16) were significantly lower and singleton proportions were significantly higher than intergenic SVs (mean AF = 0.00193, Figure 16) after matching on SV length to account for the known relationship between frequency and SV

length (Supplementary Fig. 1 and Supplementary Table 1, Wilcoxon test of AF distributions, provided in Han *et al.*[253]). These results were consistent across both deletions and duplications (Supplementary Table 1, provided in Han *et al.*[253]).

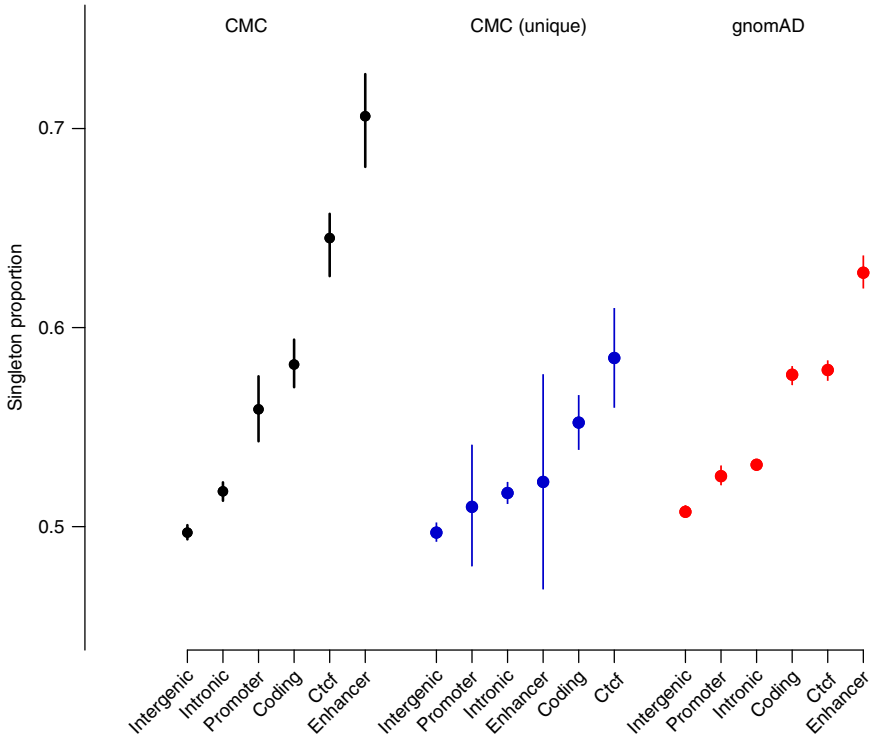

Figure 16. Genic and regulatory SVs occur at significantly lower frequencies.
Proportion of variants that are seen only a single time with bootstrapped 95% confidence interval in the sample stratified by overlap with any annotation, allowing for multiple (CMC), only a single annotation (CMC unique) and any annotation in gnomAD SV.

To explore the contributions of different functional elements to this result, we stratified SVs based on the specific annotations (e.g., coding and enhancer, Figure 17) to isolate those that alter combinations of annotations classes and those that uniquely alter a single annotation class (Supplementary Table 2, provided in Han *et al.*[253]). We identified a significant negative correlation between the total number of annotation classes affected and AF indicating that SVs with more potential to alter dosage are less likely to be tolerated (Figure 18). Further, we show that SVs exclusively affecting

CTCF sites (AF = 0.00175, p = 1.48E-4) when compared to intergenic variation showed comparable frequencies and significance to SVs that only affected protein-coding genes (AF = 0.00179, p = 1.48E-5). These results are consistent across SV type and this difference in AF is seen when performing the same annotation of the gnomAD SV dataset of ~15 k samples called from genome-sequencing using the same pipeline (Figure 16)[257]. These results suggest a strong selection against SVs that alter CTCF sites, consistent with previous work[204].

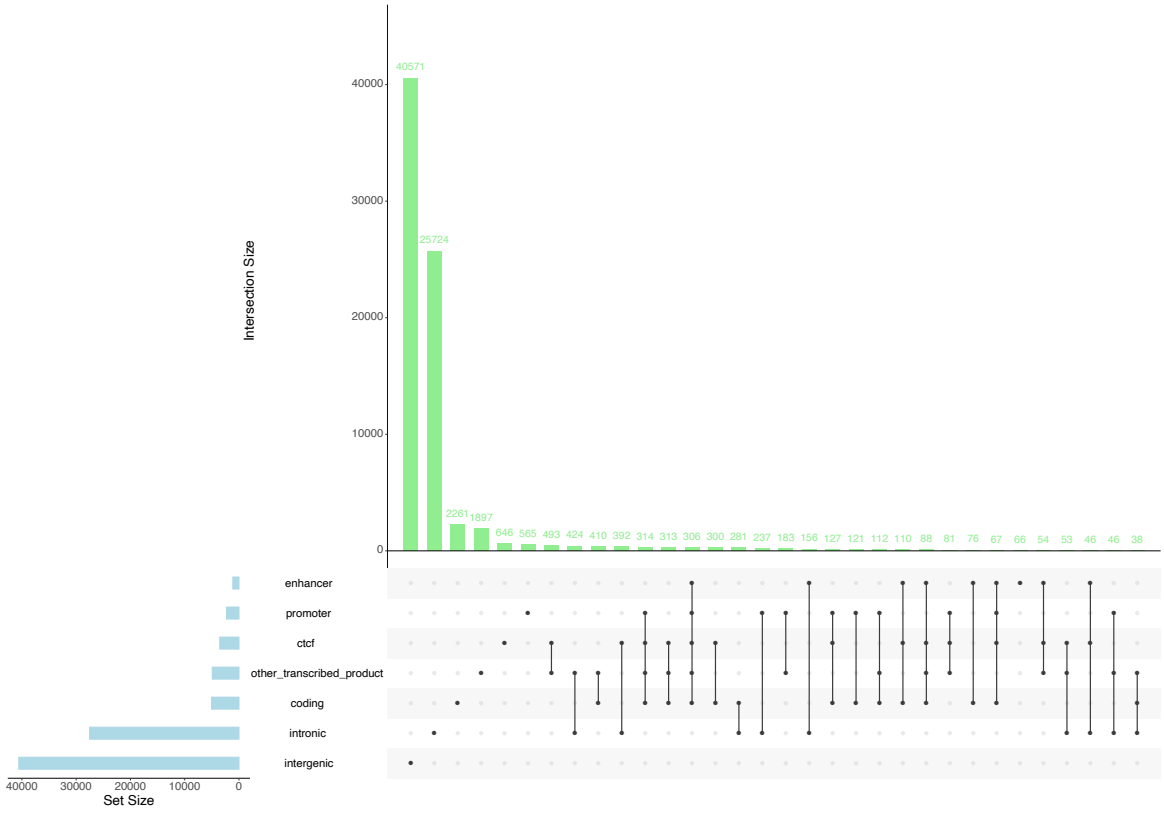

Figure 17. Counts of SVs by overlapping genomic annotations.
Blue histogram (left) shows the counts of annotations for each genomic annotation considered here. The green histogram (top) shows the counts of SVs for each combination of genomic annotation. Annotations included in the combination are designated with filled points.

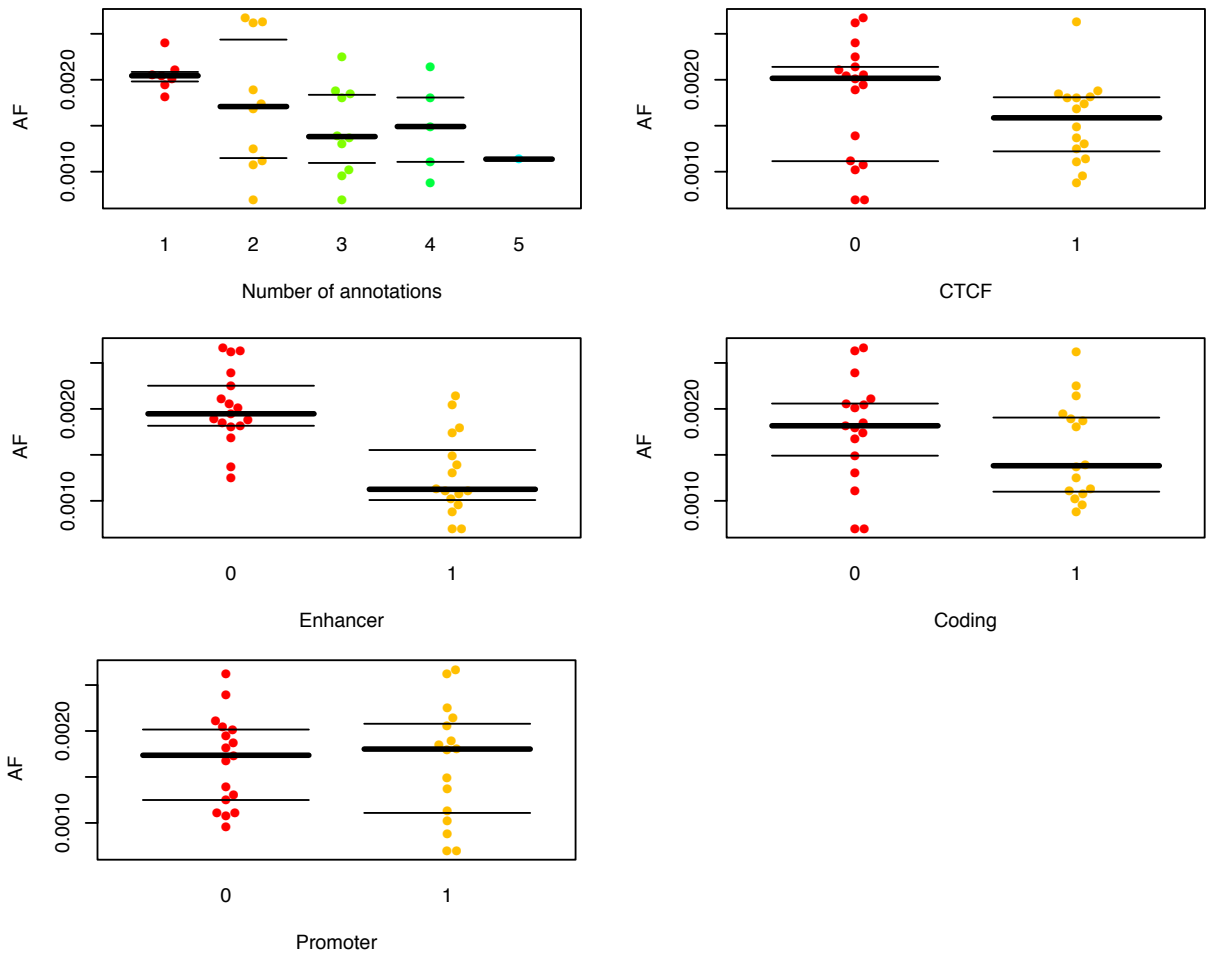

Figure 18. SVs altering regulatory annotations are observed at lower frequencies.
Mean allele frequency for each of 33 combinations of annotations split by the number of unique annotations and then by SVs that affect all combinations that include CTCF sites, enhancers, coding genes and promoters. Each point represents an annotation combination (or unique annotation). The thick bar is the median and the thin bars are 95% confidence interval.

These results are also consistent across regulatory element annotations defined using brain samples from other publicly available datasets. We downloaded enhancers and DNase I hypersensitivity sites (DHSs) from Roadmap Epigenomics and the ENCODE Consortium in brain tissues. SVs deleting these regulatory elements are observed at significantly lower frequencies than SVs that do not, providing additional evidence that regulatory SVs are deleterious (MWU, enhancer p = 1.19E-16, DHS p = 3.61E-90; Figure 19A-B). The three-dimensional chromatin architecture plays an important role in gene regulation, and previous results suggest that variants affecting both CTCF and TAD boundaries are under

selection. We downloaded a set of TAD boundaries defined using Hi-C in brain to test whether these algorithmically defined elements show similar evidence of selection[136]. However, we do not observe the same difference in SV frequency (MWU, p = 0.73), suggesting that these boundary elements do not contain additional functionally relevant information.

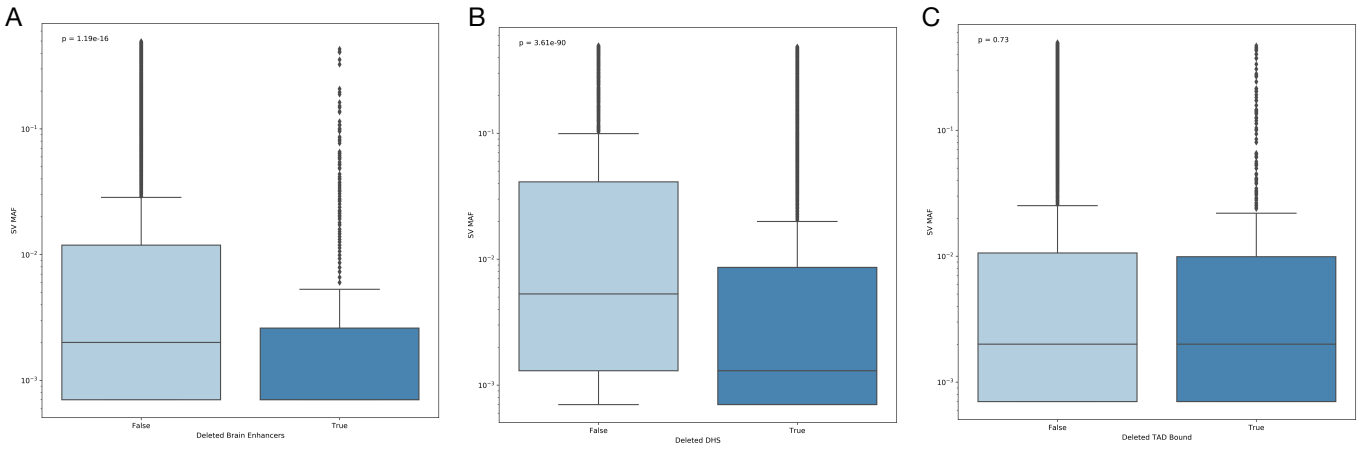

Figure 19. SVs deleting regulatory elements in independent samples are observed at low frequencies. Boxplots show the allele frequency of deletions affecting (A) brain enhancers defined using histone modifications from Roadmap Epigenomics, (B) DHSs in the brain, and (C) algorithmically defined TAD boundary elements from Schmitt *et al.*[136] Deletions are significantly less frequent for brain enhancers (MWU, p = 1.19E-16) and DHSs (MWU, p = 3.61E-90), but not for the TAD boundary elements (MWU, p = 0.73).

SVs altering gene regulatory elements predict changes in the expression of nearby genes

Annotating the influence of SVs on genomic sequence remains a challenging task. Although there is some variability in the effects of coding SVs, we expect deletions and duplications of exonic sequence to change the gene dosage and be reflected in the level of expression. For these SVs, we also expect changes gene expression to be related to the proportion of coding sequence affected. Using RNA-sequencing from DLPFC of individuals with SV information, we demonstrated that the SVs altering coding sequence had clear, directional effects on that gene's expression[253]. However, this relationship is more complicated when considering SVs that alter non-coding or regulatory, regions. We hypothesized that many of these SVs would have an effect on gene expression by perturbing the regulatory architecture, either by disrupting regulatory elements or altering chromatin structure and enhancer-gene interactions.

Quantifying the transcriptional consequences of SVs that affect regulatory elements required a set of genes to analyze for each element. We used gene proximity to define promoters, making them gene-specific by default. However, assigning enhancers to genes is a more complex task. We tested multiple approaches including the nearest gene, all genes within the same TAD, all genes within a 1 Mb window, and gene-targets predicted from Hi-C interactions. We determined that approach leveraging Hi-C interactions to link enhancers with genes was more accurate and interpretable[256]. The other approaches we considered showed similar results. Therefore, for downstream analyses we included 90,015 enhancer-gene pairs covering 6,535 genes and 32,803 enhancers predicted from PsychENCODE Hi-C data[256].

To capture the relative contributions of all annotations, we tested the relationship between gene expression z-scores and SV annotations with a joint linear model that included proportion of exonic sequence, promoter proportion, sum proportion of all affected enhancers, whether SV and gene were within the same TAD and SV length. The most significant contributor to expression was the proportion of the exonic sequence affected (deletions: beta = −1.78, p = 9.9E-158; duplications: beta = 0.78, p = 3E-109). Expression was significantly and positively correlated with the proportion of a promoter that was affected by CNVs with deletions leading to lower expression (beta = −0.17, p = 3.4E-3) and duplications leading to higher expression (beta = 0.37, p = 2.5E-30). Further, expression was significantly correlated with the cumulative sum of enhancer sequence that was affected by an SV only in duplications, but both deletions and duplications led to decreased expression (deletions: beta = −0.02, p = 0.067; duplications: beta = −0.02, p = 8.1E-9). The presence of the SV and the gene within the same TAD contributed significantly and directionally to expression in deletions (beta = −0.009, p = 5.7E-5) but not duplications (beta = 0.005, p = 0.21). The effects of these variables on expression were consistent across cohort (Table 3; coefficients stratified by cohort available in Table 2 of Han *et al.*[253]) and while proportion of exonic sequence provided the strongest contributor, the effects of cis-regulatory elements remained significant in duplications and to a lesser extent in deletions after excluding all genic SVs (Supplementary Table 3, provided in Han *et al.*[253]).

Table 3. Genic and regulatory features significantly contribute to predicting transcriptional consequences of CNVs. Coefficients of linear regression model to predict expression *z*-scores in deletions and duplications across samples.

CNV class	Variable	Beta	SE	T	P
Deletions	Exonic Proportion	−1.7762	0.0664	−26.77	9.9E−158
	Enhancer sum	−0.0152	0.0083	−1.83	6.7E−02
	Promoter proportion	−0.1726	0.0589	−2.93	3.4E−03
	SV Length	−2.14E−07	1.46E-08	−14.70	6.9E−49
	Within TAD	−0.0090	0.0022	−4.03	5.7E−05
Duplications	Exonic Proportion	0.7825	0.0352	22.22	3.0E−109
	Enhancer sum	−0.0157	0.0027	−5.77	8.1E−09
	Promoter proportion	0.3735	0.0326	11.45	2.5E−30
	SV Length	3.99E-07	2.60E-08	15.34	4.6E−53
	Within TAD	0.0046	0.0036	1.26	2.1E−01

Effects of regulatory SVs are weaker for enhancers active in other tissues

Due to the cell-type specific nature of enhancers, we expected that these results were dependent on the appropriate matching of regulatory annotations to the context where gene expression was measured. Using earlier models focused on the deletion of regulatory annotations, we considered the robustness of our results to the cell type of the enhancer annotation. We downloaded enhancers identified using the same procedure in two other human cell or tissue types, Gm12878 (n = 96,941) and liver (n = 97,588). Although SVs that impact Gm12878 or liver enhancer annotations are less frequent (MWU, p = 2.23E-11, p = 6.89E-17; Figure 20) than those that do not, the amount of deleted enhancer sequence is not a significant predictor of expression *z*-score (Gm12878 p = 0.328, liver p = 0.554; Table 4). Instead, the lower frequency for these SVs is likely related to other cell-type-relevant annotations and the length of the SVs. These results further support the need to consider cell-type specific regulatory architecture when interpreting non-coding SVs in order to define meaningful trends.

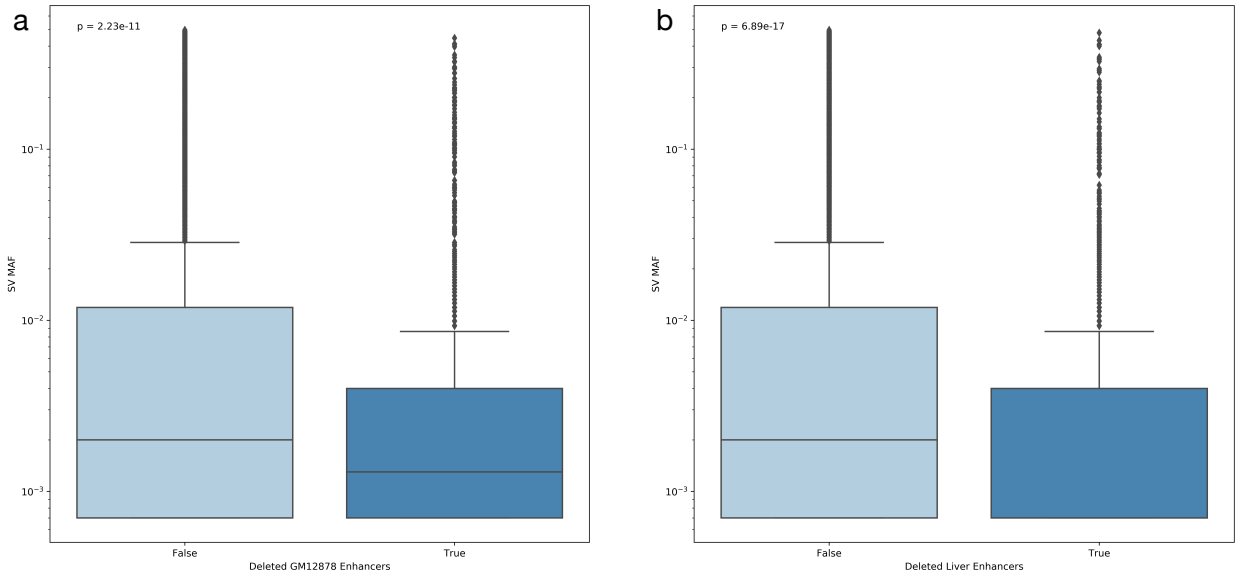

Figure 20. SVs disrupting enhancer regions from other biological contexts are rare.
Boxplots show the allele frequency of deletions affecting (a) Gm12878 and liver (b) enhancers defined using histone modifications from Roadmap Epigenomics. Deletions are significantly less frequent for Gm12878 enhancers (MWU, p = 2.23E-11) and liver enhancers (MWU, p = 6.89E-17), although the amount of deleted sequence does not significantly predict changes in expression of nearby genes.

Table 4. Coefficients of Gaussian GLM model to predict z-score of genes within 1 Mb of deleted enhancers from Gm12878 and liver.

Tissue	Predictor	Beta	SE	Z	P
Gm12878	Number of deleted enhancers	-0.0018	0.002	-0.979	0.328
	Number of deleted genes	-0.0010	9.2E-05	-10.580	<0.001
	SV length	-2.0277	0.038	-52.676	<0.001
	SV MAF	5.004E-07	4.08E-08	12.269	<0.001
Liver	Number of deleted enhancers	-0.0011	0.002	-0.592	0.554
	Number of deleted genes	-0.0010	9.19E-05	-10.611	<0.001
	SV length	4.943E-07	5.01E-08	9.859	<0.001
	SV MAF	-2.0272	0.038	-52.668	<0.001

Integrating transcriptional consequence and gene intolerance[3]

To better understand the relationship between our variant annotations in the context of the genes affected, we incorporated two distinct measures of genic intolerance to variation: (1) gene intolerance to CNVs

[3] Analysis developed and conducted by other authors, leveraging my contribution of gene regulatory element annotations in the brain.

defined empirically from exome-sequencing in nearly 60,000 individuals[258], and (2) a measure of gene intolerance to LoF variation generated from a sample of ~141,000 individuals[259]. Several significant relationships between the functional effects of SVs and the intolerance of the genes affected existed. SVs that disrupted intolerant genes were significantly more likely to alter a smaller proportion of the exonic sequence (pLoF = 2.42E-38, pCNV = 1.31E-33, Spearman correlation test of intolerance and proportion of exonic sequence affected, Figure 21a). Intolerant genes were also significantly less likely to have a genic SV (pLoF = 2.4E-38, pCNV = 9.36E-34, Wilcoxon test of gene metric by whether SV affects exonic sequence or not, Figure 21b). Consistent with previous literature showing intolerance to dosage changes in either direction25 we saw intolerant genes less likely to be affected by both deletions (pLoF = 7.41E-29, pCNV = 3.02E-26) and duplications (pLoF = 3.81E-22, pCNV = 1.57E-46). Further, when restricting to SVs that only alter regulatory elements and not exonic sequence, we identified a significant decrease in the number of enhancers affected by SVs in genes with higher intolerance, although this was only observed for the CNV intolerance metric (pLoF = 0.62, pCNV = 7.23E-22, Figure 21d). We did not find any effects from promoter SVs in either metric (pLoF = 0.35, pCNV = 0.37, Figure 21c). Combined with the differences seen by CNV type these results may indicate unique properties of these metrics and what they reflect (e.g., haploinsufficiency vs. dosage sensitivity). In general, as with single nucleotide variation, genic measures of intolerance should help functionally annotate SVs.

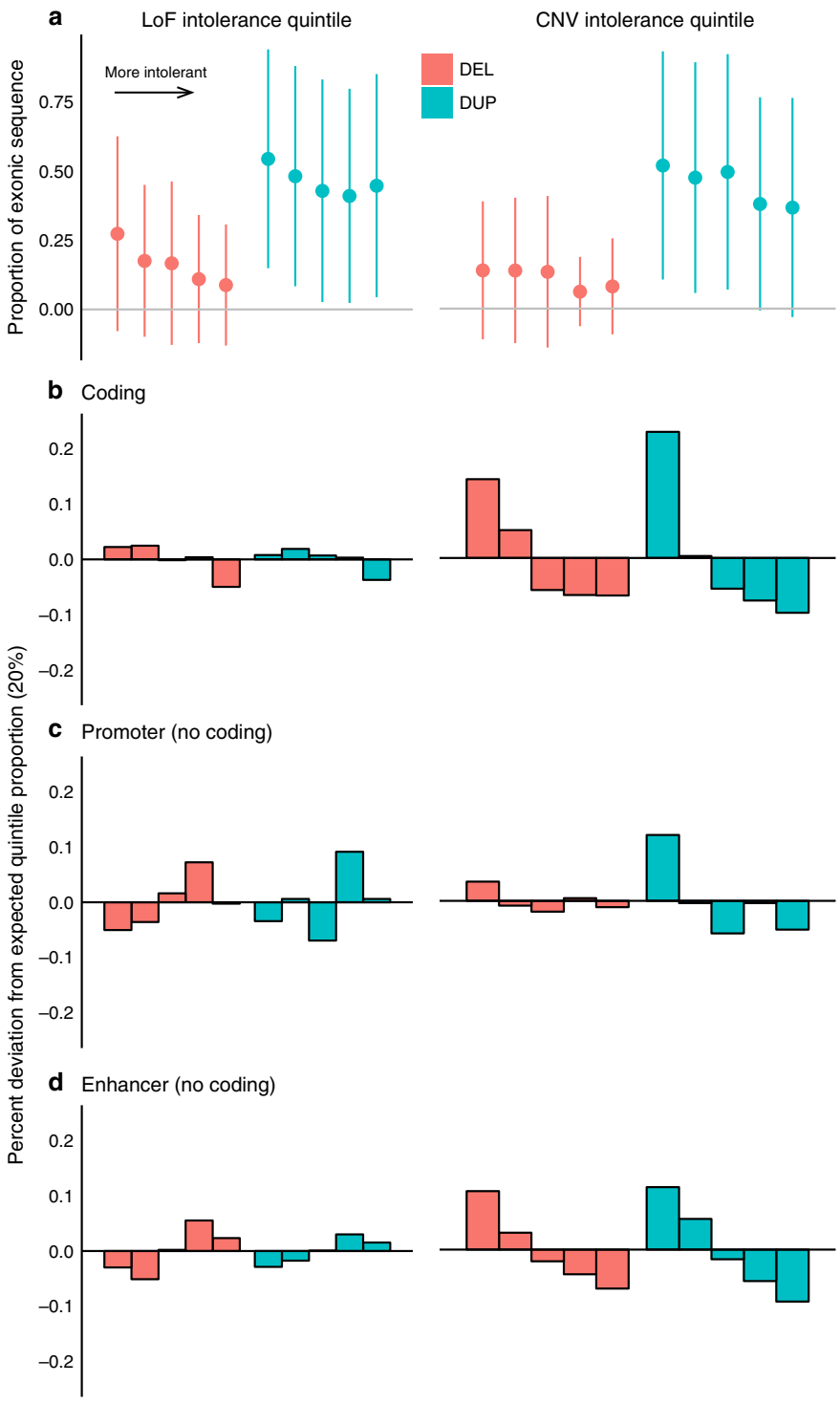

Figure 21. Genes intolerant to variation are less likely to be affected by genic or regulatory SVs.
Each plot stratifies genes using either the LoF intolerance metric or the CNV intolerance metric that have been split into quintiles (20% bins) ordered left to right from least to most intolerant genes and by deletion (red) and duplication (blue). The plots show the effect of this stratification on (a) the proportion of the exonic sequence that is affected showing mean and standard deviation, (b) the deviation from the expected 20% of CNV that alter exonic sequence, (c) the deviation from expected for noncoding CNV that alter promoters, and (d) the deviation from expected for noncoding CNV that alter enhancers.

A model to annotate SVs form predicted dosage and gene intolerance[4]

Having demonstrated a significant role for SVs in altering expression, we sought to test whether this model could be used to predict expression effects of SVs in independent samples. We split our DLPFC sample by cohort (CMC and CMC_HBCC, Methods, provided in Han *et al.*[253]) and constructed the linear model described previously in each subset and then applied that model to SVs in the other set to infer expression effects. We identified significant correlation between the true expression value and the predicted value across all four pairwise comparisons (R^2 CMC_HBCC into CMC = 0.35, R^2 CMC into CMC_HBCC = 0.17, R^2 CMC into CMC = 0.36, R^2 CMC_HBCC into CMC_HBCC = 0.17, Figure 22) with deletions (particularly when tested in CMC) consistently performing better.

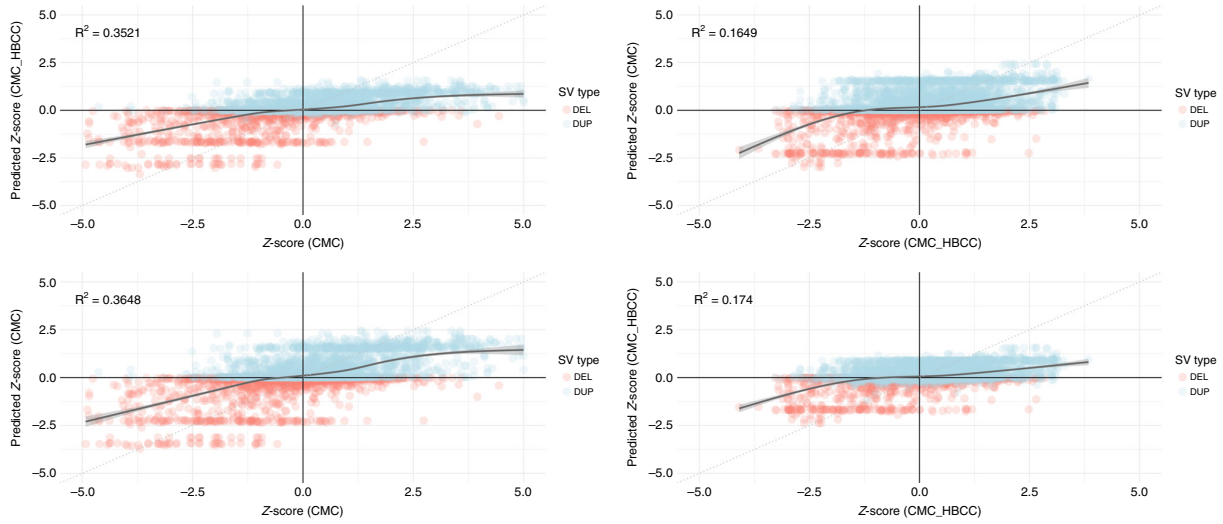

Figure 22. Transcriptional consequences of rare CNVs can be significantly predicted.
SV expression prediction performance and associated R2 from building the same linear model using different training and test datasets. a CMC into CMC_HBCC, b CMC_HBCC into CMC, c CMC into CMC, and d CMC_HBCC into CMC_HBCC. The best fit line with confidence interval was produced using generalized additive model smoothing.

[4] Analysis developed and conducted by other authors, leveraging my contribution of gene regulatory elements in the brain.

Leveraging this model and the previously used measure of genic intolerance to LoF variation, we built an aggregate regulatory disruption score that was the sum of the predicted expression *z*-scores for each gene weighted by the gene's intolerance metric (normalized between 0 and 1 with 1 being most intolerant) to annotate SVs. We then applied our model to annotate 210,244 variants in the gnomAD SV dataset[260] after restricting to CNVs that were below 1% frequency. Of those, 31,492 (15%) were predicted to alter the expression of at least one protein-coding gene where we had an intolerance metric, 20,236 of these variants were deletions and 11,256 were duplications. We considered a deletion or duplication in gnomAD as pathogenic if it overlapped at least 50% of a CNV of the same type (3454 deletions and 1894 duplications) labeled pathogenic for neurodevelopmental disorders (developmental delay, intellectual disability, or autism) in ClinGen (downloaded from UCSC Genome Browser June 2019). There were 84 deletions and 84 duplications that met this criterion (39 deletions and 33 duplications overlapped 100% of the pathogenic ClinGen variant, as gnomAD includes some individuals with neuropsychiatric disorders). This set of pathogenic CNVs had significantly larger regulatory disruption scores in the direction of the dosage change with deletions having a more severe reduction in expression among intolerant genes due to these deletions (p = 1.78E-26, mean score in pathogenic deletions = −5.21, mean score in nonpathogenic deletions = −0.18, Wilcoxon test) and duplications having a dramatic increase ($p = 9.76 \times 10^{-39}$, mean score in pathogenic duplications = 12.67, mean score in nonpathogenic duplications = 0.52). Despite ascertainment bias leading to longer CNVs being more likely to overlap pathogenic variants, prioritizing variants by regulatory disruption would identify more pathogenic deletions than prioritizing by length, with four of the top ten variants being pathogenic if ranked by length (two complete overlaps) and seven by regulatory disruption (five complete overlaps). The regulatory disruption score also better prioritized pathogenic deletions than number of all genes affected, number of intolerant genes (top 10%) affected and AF, which has limited utility since most deletions (53% or 10,776) have the same frequency, as singletons. (Figure 23a, Supplementary Data 1 provided in Han *et al.*[253]). For duplications, the regulatory disruption score performs similarly to length but still outperforms other measures (Figure 23b,

Supplementary Data 2 provided in Han *et al.*[253]). These results indicate the potential of this metric to contribute to improved prioritization of disease causing CNVs, particularly deletions.

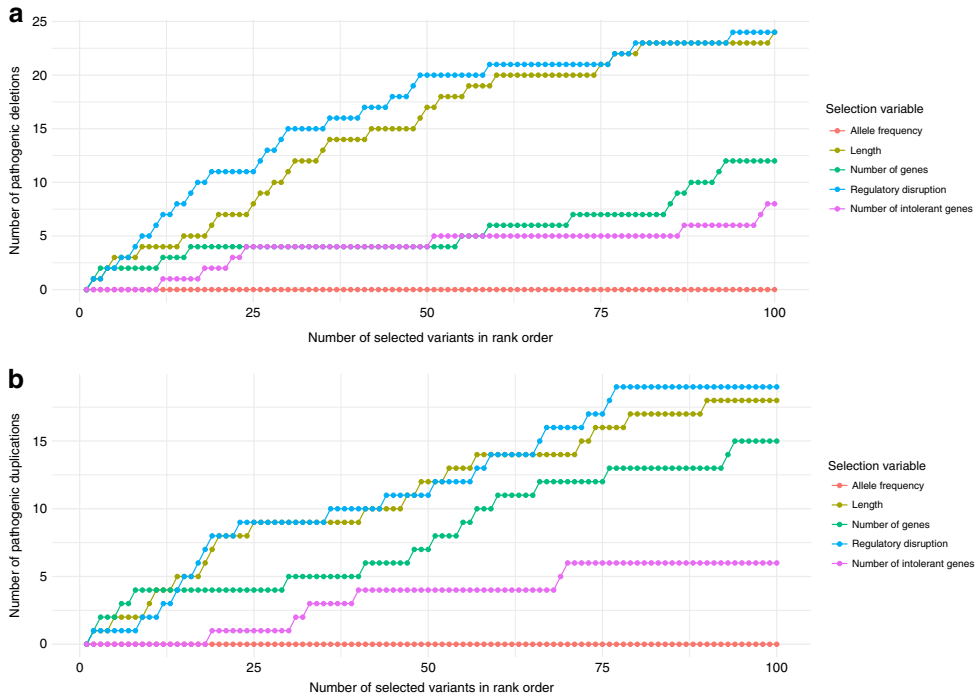

Figure 23. Regulatory disruption scores prioritize pathogenic CNVs better than standard annotations. Number of pathogenic variants defined as 50% overlap with known pathogenic variant in ClinGen (84 deletions and 84 duplications) identified based on rank ordering deletions a and duplications b by length (yellow), number of genes deleted (green), number of intolerant genes deleted (purple) allele frequency (red), and regulatory disruption (blue). Where multiple variants had the same value, the order was random.

Conclusion

Although previous work has highlighted the role of SVs in a wide range of diseases[124,200,206,250], predicting the influence of an SV on expression or disease risk is a challenge. SVs can disrupt gene regulatory mechanisms and cause substantial changes to the regulatory architecture of the genome. Interpreting the effects of regulatory changes is difficult, as is disentangling the relative contributions of protein-coding and non-protein-coding sequence disruptions.

This chapter leverages a novel dataset of genome and RNA sequencing across hundreds of individuals with publicly available functional genomics annotations to characterize the features of

regulatory SVs and quantify their effects on expression. We conclude that both genic and regulatory SVs can substantially alter gene expression. In our models, the level of disruption is proportional to the amount of altered functional sequence. We also show that SVs impacting regulatory elements, such as enhancers, promoters, and CTCF binding sites are significantly lower in frequency than those that impact unannotated sequence. This suggests that perturbation of the gene regulatory architecture is deleterious and provides evidence that regulatory SVs have been selected against because of their potential to negatively influence fitness. This trend is especially striking for CTCF binding sites. CTCF has been implicated in the proper maintenance of 3D genomic architecture and enhancer-promoter contacts; our results suggest a crucial role for this factor that should be explored in future variant effect prediction efforts.

CHAPTER IV

Quantifying the Role of Enhancer Landscapes in Maintaining Gene Expression Patterns

Introduction

Non-protein-coding distal regulatory elements, often referred to as enhancers, control the regulation of gene expression by binding to transcription factors and modulating transcription. These elements play crucial roles in the development and maintenance of transcription across cell types. Although many definitions of enhancers exist in the literature, most are associated with regions of open chromatin and key histone modifications and are thought to regulate genes through three-dimensional chromatin looping. It has been widely shown that multiple enhancers can act additively, synergistically, or redundantly to mediate gene expression[12,147–149,153,181,182]. In this work, we define "enhancer landscapes" as the set of enhancer elements involved in regulating the expression of a gene in a specific biological context. We consider many attributes of these enhancer landscapes, including the number of active enhancers and the physical interactions between enhancers and a given target gene.

Some aspects of enhancer landscapes are well established in *Drosophila* where studies have demonstrated that the presence of multiple enhancers for a gene often provides robustness to genetic variation. These redundant "shadow enhancers" maintain appropriate gene expression when other enhancers regulating the same target gene are inactivated[152]. Other early investigations of enhancer landscapes in humans suggest that the number and DNA sequence conservation of enhancers are associated with differences in target gene function and partially explain the highly correlated expression patterns of orthologous genes across mammalian species[160,164]. Thus, consideration of the landscape of enhancers is crucial to understanding a gene's expression level, dynamics, and the effects of variation across individuals.

Despite growing evidence that enhancer landscape will be important when interpreting the effect genetic variation will have on gene expression, it is not frequently considered. Furthermore, much of the previous work on enhancer landscapes is confined to studies on model organisms and does not leverage

the growing amount of three-dimensional (3D) chromatin interaction data now available for a range of cellular contexts. We do not yet understand how features of the enhancer landscape relate to constraints on gene expression, alter the impact of genetic variation, or the dynamics of enhancer turnover across species. In this chapter, we develop a framework using 3D chromatin interaction data to define enhancer landscapes in multiple human tissues. We then assess how different enhancer landscapes influence gene regulation. We observe that differences in gene function and constraint on gene expression are reflected in features of enhancer landscapes, including the number and tissue-specificity of associated enhancers. We also find that enhancer landscape attributes are associated with differences in enrichment for non-coding genetic variants. Our results demonstrate that the enhancer landscape of a gene should be considered when studying gene expression dynamics and interpreting regulatory genetic variation.

Methods

Genomic annotations and chromatin interaction data

All analyses were conducted using the GRCh37/hg19 build of the human genome. We used gene and transcription start site (TSS) definitions from Ensembl v75 (GRCh37.p13).

We obtained liver enhancer annotations for six mammalian species (human, macaque, marmoset, mouse, rat, rabbit) defined using histone modification peaks from ChIP-seq[29]. H3K27ac peaks were called as enhancers if the peak was present in that species and it overlapped by less than 50% of its length with an H3K4me3 peak, as defined previously[29]. We exclude H3K4me3 peaks since these are considered markers of promoters. We filtered these putative enhancer regions to remove any that overlap an ENCODE blacklist or UCSC gap region[226]. For cross-tissue analyses, we downloaded H3K27ac and H3K4me3 ChIP-seq peaks from the Roadmap Epigenomics Consortium for 10 tissues: brain (prefrontal cortex, hippocampus), heart (left ventricle), liver, lung, ovary, pancreas, psoas muscle, spleen, and small intestine[191]. We used the same protocol to identify putative enhancer regions (H3K27ac peaks without an overlapping H3K4me3 peak) across tissues.

For each of the 10 tissues with enhancer data, we downloaded normalized, 40 kb resolution Hi-C interaction frequency matrices from human samples[136]. The matrices were normalized using FitHiC[261]. The locations of topologically associating domain (TAD) regions were derived from the same Hi-C interaction data by the 3D Genome Browser using the approach described in Dixon *et al.*[110,262].

Definition of genome-wide enhancer landscapes

We defined a gene-level enhancer landscape using chromatin conformation data to associated putative enhancer elements and the transcription start sites of genes. For each gene, the landscape definition is based on the combination of enhancer, gene, and Hi-C annotations. The number of enhancer elements with evidence of a significant interaction from the Hi-C data ($Q < 0.05$) is one main attribute of the landscape. The significance of a Hi-C interaction is determined by comparing the frequency of the observed interaction to an empirical null model adjusted for known technical biases. The Q-value is the p-value of the Hi-C interaction adjusted for a false discovery rate of 5%. Enhancers elements that overlap the anchor of a significant Hi-C interaction ($Q < 0.05$) are assigned to all genes with a TSS inside the other anchor and considered part of the landscape. Where there are multiple enhancers or TSSs within a single anchor, all enhancers are linked to all potential gene targets. To account for the known role of TADs in constraining regulatory interactions, we limit the enhancer-gene assignment to intra-TAD interactions.

We also define a region-level enhancer landscape that is not focused on a single gene. This includes the number of enhancers, number of genes, and the number of interactions in a region. To create genomic regions, we tiled the genome into nonoverlapping 1 Mb windows. These were filtered to exclude any windows that are comprised of more than 5% ENCODE blacklist or UCSC gap regions. Within the chosen windows, we consider the number of genes and enhancer annotations that overlap that window by at least 1 bp. Based on the evolutionary history of each enhancer element, they are divided into ‘ancestral’ and ‘gained’ enhancers. This classification is described in more detail below. We limited the number of genes in the window to the number of protein-coding genes. We also counted the number of significant

Hi-C interactions (Q < 0.05) with at least one anchor point within the region. If both anchor points are within the region, the interaction was counted twice. We also considered other methods for defining genomic regions of interest, including TADs and the gene-regulatory domain strategy proposed by GREAT[155].

Evolutionary model to infer ancestral enhancer state

We classify enhancers into two categories: ancestral and gained. We leverage cross-species ChIP-seq data to define these categories using the Wagner evolutionary parsimony model implemented by Count[263]. We set the penalty (g) for a gain of enhancer function to 2, where $g = 1$ represents an equal likelihood of independent gain and loss. Under this model, gained enhancers are those that are inferred to have gained activity in the human lineage based on the observed patterns of activity at terminal branches of the species tree. This model was also used to derive the ancestral enhancer state at each branchpoint. We consider the ancestral enhancer state for humans as the inferred activity of the enhancer at the most recent common ancestor of humans and macaque. Ancestral enhancers have activity at that branchpoint.

Calculating tissue-specificity of genes

We downloaded RNA-seq gene expression data from GTEx (v7) in transcripts per million (TPM) for ten tissues with matching Hi-C data: prefrontal cortex, hippocampus, heart, liver, lung, ovary, pancreas, skeletal muscle, spleen, and small intestine[136,198]. We calculated the tissue-specificity using the relative entropy of each gene's expression profile across tissues compared to the median gene expression distribution across tissues. We then scaled the resulting value to between 0 and 1, where genes closer to 0 are broadly expressed and genes closer to 1 are tissue specific. Another tissue-specificity metric, τ, and tissue-specific genes classified using the tissue-specificity score (TSPS) from Ravasi *et al.*, produced similar results[264,265].

Although most genes are more broadly expressed under this metric, we defined a set of "tissue-specific" genes using a threshold on the relative entropy score. We tested multiple cutoffs at varying

levels of stringency (Table 5). For our final cutoff we selected the most conservative score (> 0.8) which classifies 483 genes (2.8%) as tissue-specific and has been used previously[266].

Table 5. Entropy thresholds for tissue-specific genes.

Threshold	# Genes	% Genes
>= 0.8	483	2.8
>= 0.6	1,172	6.8
>= 0.4	2,752	15.9
>= 0.2	5,277	30.5

Calculating tissue-specificity of enhancers

We also calculated the tissue-specificity of enhancers using the scaled entropy score. However, because enhancers do not have consistent lengths or locations across tissues, we standardized the enhancer lengths before calculating the number of tissues where each enhancer was active. We tested three possible standard lengths: 1) the median enhancer length across tissues with lower quality ChIP-seq data (220 bp), 2) the median enhancer length in the liver (460 bp; no quality flags), and 3) the median enhancer length for the histone-modification-defined liver enhancers from Villar *et al.* (2500 bp)[27]. We centered the standardized enhancer on the midpoint of the existing annotation and either expanded or truncated each region to the desired length. We selected the most conservative standard length of 220 bp for our final entropy calculations; the scores for the other two were strongly to moderately correlated with our chosen conservative threshold (ρ = 0.82 for 460 bp, ρ = 0.53 for 2500 bp). We then intersected the standardized enhancers from all tissues and, for each enhancer, counted the number of tissues where it had activity. We assigned this number of active tissues back to the original enhancer element and used these to calculate the entropy. This value was scaled to create a score between 0 and 1, with low scores corresponding to broad activity and high scores corresponding with tissue-specific activity.

Defining relevant gene sets with constraint on expression

We created three gene sets that we expected to experience higher levels of constraint on their expression than genes overall: housekeeping genes, essential genes, and loss-of-function intolerant genes. The housekeeping genes were downloaded from earlier study which defined them based on expression at the similar levels across sixteen tissues in RNA-seq (n = 3804)[267]. We downloaded a set of essential genes from a study that used CRISPR screens in five human cell lines to define genes required for growth and proliferation across cell types (n = 1580)[268]. Finally, we downloaded a set of likely loss-of-function (LoF) intolerant genes from gnomAD (v2)[259]. Following the gnomAD threshold, we defined LoF intolerant genes as those with a 90% confidence interval upper bound of the observed/expected (o/e) metric less than 0.35. Lower o/e scores indicate greater intolerance to protein variants.

Gene Ontology enrichment stratified by enhancer landscape attributes

We calculated enrichment for Gene Ontology (GO) annotations enrichments for gene sets of interest using WebGestalt[238] with default options. We quantified enrichment for Biological Process (BP) terms against a background of protein-coding genes. We considered significant terms with FDR <= 0.05.

Identifying transcription factor binding motifs in enhancer sequences

We used the FIMO[269] tool from the MEME suite to scan for the presence of known TF motifs from the HOCOMOCO (v11)[240] core database in all human liver enhancers defined by Villar *et al.*[27] We ran FIMO with default options. We calculated the number of motifs, unique motifs, and the density of motifs for each scanned enhancer element.

Enrichment for GWAS variants and GTEx eQTL in different enhancer landscapes

We used permutation testing to determine whether the enhancers in different enhancer landscapes are enriched for overlap with genetic variation compared to genomic background. We calculated an empirical p-value for the observed number of overlapping variants compared to a null distribution of length-

matched random genomic regions. For a comparison of variant set *A* with enhancer set *B*, we calculate the number of observed overlaps between A and B. To generate the null distribution, we then randomly shuffle the locations of enhancer set B, maintaining the length distribution of B, and repeat the overlap process 1000 times. We use this empirical null distribution to compute a two-sided empirical p-value.

We performed this analysis for overlap between trait-associated variants from the GWAS Catalog and GTEx eQTL for enhancers in regions with a large number of enhancer elements (n > 5), and enhancers in regions with few other enhancer elements (n <= 5). We chose this cutoff based on the median number of ancestral enhancers in a 1 Mb window. We also stratified the enhancers based on the level of enhancer activity conservation across species. We considered an enhancer conserved if it had evidence of activity in at least 3 of the 6 species.

We downloaded expression quantitative trait loci (eQTL) from GTEx (v6p) using the liver eQTL (p-value < 10E-10)[231]. We also downloaded variants associated with phenotypes in the GWAS Catalog (n = 20,458; v1.0, downloaded 08-10-2016)[230]. We manually curated the subset of GWAS Catalog variants into those associated with phenotypes relevant to the liver (n = 346). To capture variants tagged by the downloaded GWAS SNPs and eQTL, we included SNPs in high LD ($r^2 > 0.9$) in individuals of European ancestry from the 1000 Genomes Project phase 3 and calculate overlap with haplotype blocks[232].

Regression model to predict enhancer gain in a regional enhancer landscape

We regressed the number of gained enhancers in a region (1 Mb window, TAD, or GREAT domain) on the other regional enhancer landscape features using a negative binomial model. The features considered are the number of genes in the region, the number of ancestral enhancers in the region, and the number of significant Hi-C interactions with at least one anchor point in the region and within the same TAD. We used a negative binomial model rather than a Poisson model to account for overdispersion in the data.

Quantifying enhancer conservation in regional enhancer landscapes

We quantified enhancer conservation in two ways: DNA sequence conservation and enhancer activity conservation across species. We used conserved elements from vertebrates and primates defined by the two-state hidden Markov model, PhastCons[229]. We merged the two sets of PhastCons conserved elements using Bedtools mergeBed[233], and then calculated the proportion of each enhancer that overlaps one of these elements. To measure enhancer activity conservation, we aligned enhancer sequences across species using cross-species histone modification data from six mammalian species. Enhancers were scored from 1 to 6 based on the number of species where that enhancer showed evidence of activity.

Results

Definitions of genome-wide enhancer landscapes

We define two approaches to characterize enhancer landscapes genome-wide. The first is a gene-level definition based on using chromatin conformation data to define interactions between putative enhancer elements and the transcription start sites (TSSs) of their target genes (Figure 24). This approach links enhancer elements with putative target genes by overlapping the anchor points of significant Hi-C interactions with enhancer and TSS annotations. Interactions that overlap an enhancer with one anchor and a TSS with the other anchor are paired. If more than one enhancer or TSS exists in the anchor, we consider all pairs of annotations.

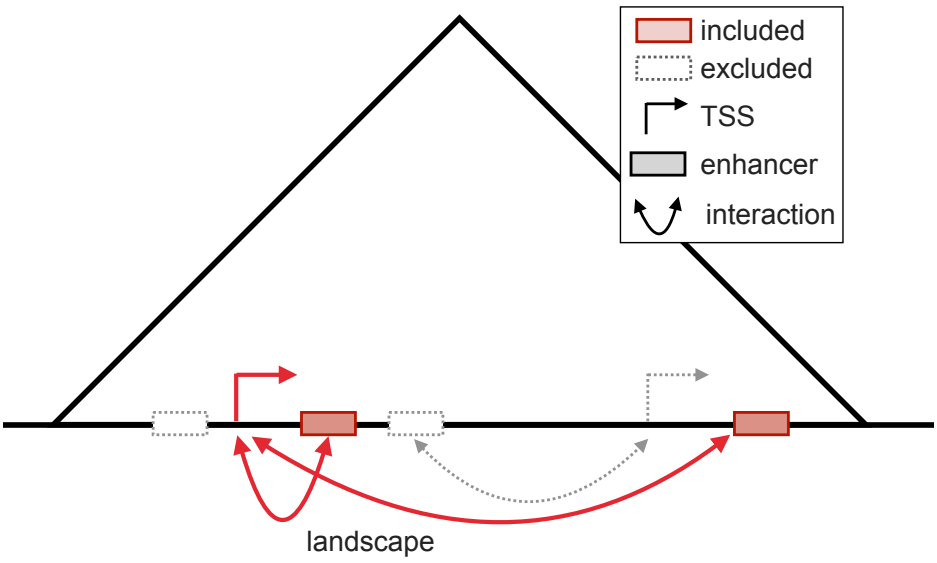

Figure 24. Schematic illustrating gene-level enhancer landscape definition.
Gene-level enhancer landscapes are defined as all enhancers with a Hi-C interaction with the TSS of a gene. The black horizontal line represents the genome, and triangle represents a TAD. Enhancers are shown using filled (included) and dotted (excluded) rectangles, and the significant Hi-C interactions within the TAD are shown as arcs underneath the genome line. The red arcs highlight within TAD interactions with the gene of interest.

Genes do not exist in isolation. The density of genes and regulatory elements varies across the genome. To account for this, we also define region-level enhancer landscapes that account for the number of enhancers, number of genes, and the number of interactions in a given genomic region (Figure 25). More specifically, for each region we count the number of ancestral enhancers, the number of expressed genes, and the number of significant Hi-C interactions with at least one interaction anchor overlapping the region. We considered multiple ways of defining the genomic regions of interest, including 1 Mb non-overlapping windows tiled across the genome, dynamically-size windows centered on TSSs[155], and TADs defined by the 3D Genome Browser[262]. We focus on 1 Mb windows, although other approaches provide similar results.

We infer ancestral enhancer state using an evolutionary parsimony model applied to cross-species histone modification profiles (Figure 25, inset; Methods). This allows us to explore both the relationship between the ancestral state of the human enhancer landscape and current attributes and the enhancers that

gained activity along the human lineage. We incorporate the ancestral or gained enhancer annotations into the region-level liver enhancer landscapes.

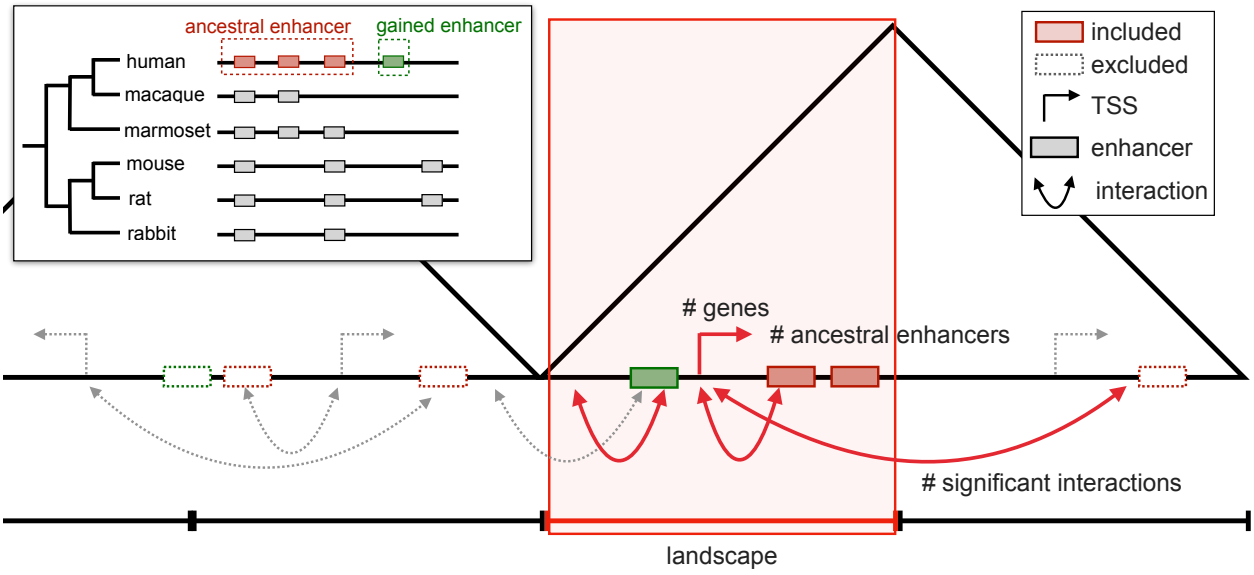

Figure 25. Schematic illustrating region-level enhancer landscape definition.
Gene-level enhancer landscapes are defined as the number of ancestral enhancers, the number of genes, and the number of significant Hi-C interactions in a 1 Mb genomic window. The black horizontal line represents the genome, and triangles represent TADs. Enhancers are shown using filled (included) and dotted (excluded) rectangles, and the significant Hi-C interactions are shown underneath the genome line. The ancestral (red) and recently gained (green) enhancers are defined using an evolutionary parsimony model (inset, Methods).

Region and gene-level enhancer landscapes vary across the genome

Using the 1 Mb region definition of an enhancer landscape in the liver, 64% of the windows contain both multiple liver enhancers and multiple genes, and 98% contain at least one significant Hi-C interaction (Figure 26A-C). Furthermore, most of the significant Hi-C interactions occur between functional regions (e.g. genes, enhancers; Figure 26D).This supports the use of Hi-C interactions to link enhancers to genes and provides evidence that the significant interactions identify biologically meaningful functional connections in a region.

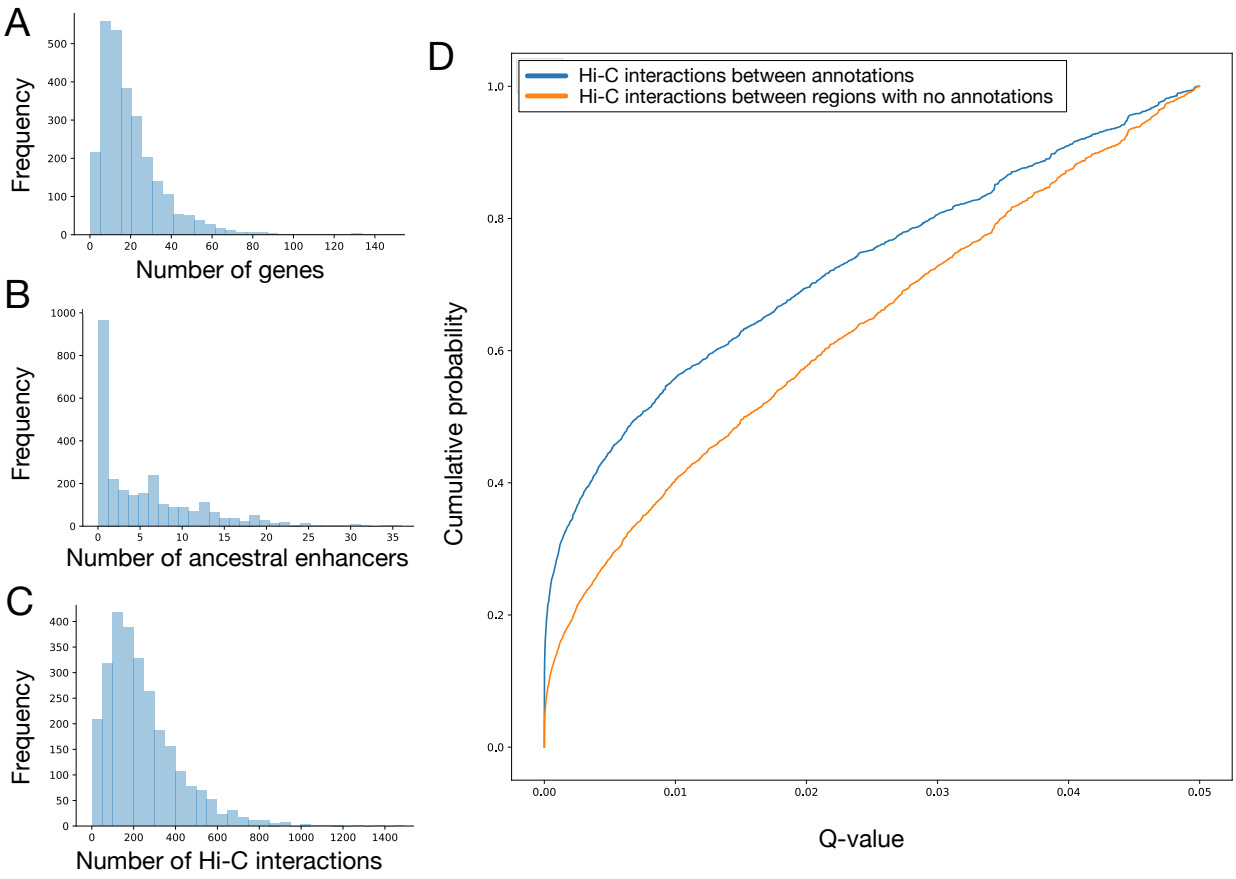

Figure 26. Distribution of regulatory attributes in region-level enhancer landscapes.
Distribution of the (A) number of genes, (B) number of enhancers, (C) number of Hi-C interactions within non-overlapping 1 Mb bins tiled across the genome. (D) Plot of the cumulative density function of significant Hi-C interactions in the liver with anchors that link functional elements (blue; e.g. genes, enhancers) versus those that do not (orange). Most of the interactions linking enhancers and genes are highly significant.

Using the gene-level definition of an enhancer landscape in the liver, 71% of genes are linked to enhancers and 68% are associated with multiple enhancers. Across tissues, most genes that are able to be linked with enhancers are linked with multiple, supporting previous claims that many genes are regulated by groups of enhancer elements (Figure 27; Table 6).

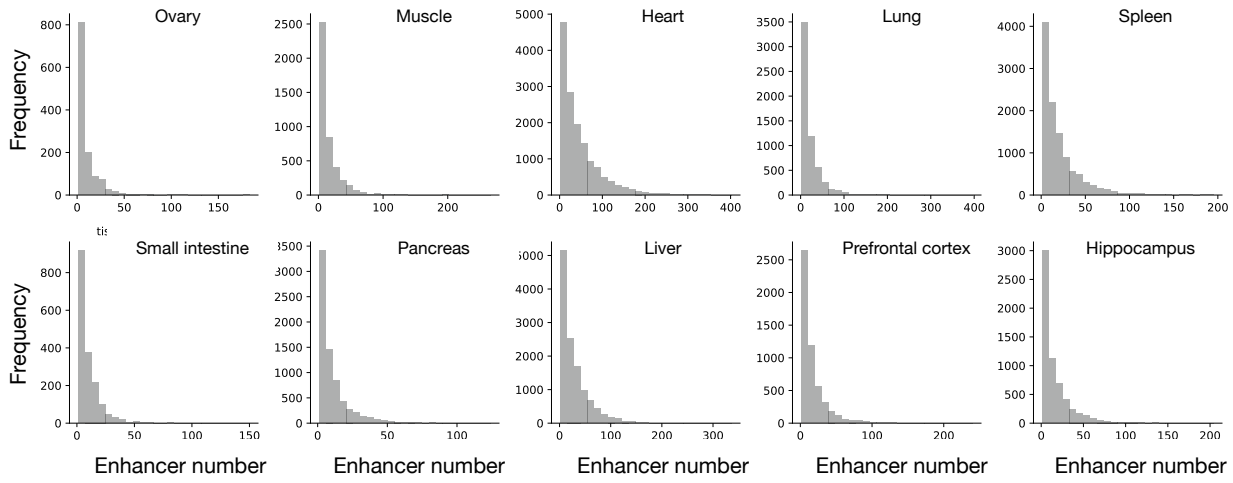

Figure 27. Genes have variable numbers of linked enhancers across tissues.
Histograms of the number of enhancers per gene in gene-level enhancer landscape by tissue. Across tissues, most genes that can be linked to enhancers have multiple enhancers. However, the proportion of genes linked varies by tissue due to differences in the number of Hi-C interactions.

Table 6. Proportion of genes linked to multiple enhancers in the gene-level enhancer landscape.

Tissue	Proportion of genes linked to 1+ enhancers	Proportion of genes linked to 2+ enhancers	Percent linked to multiple enhancers
Ovary	0.07	0.06	80%
Muscle	0.25	0.22	89%
Heart	0.84	0.82	97%
Lung	0.34	0.31	92%
Spleen	0.62	0.58	94%
Small intestine	0.1	0.09	88%
Pancreas	0.41	0.35	86%
Liver	0.71	0.68	95%
Prefrontal cortex	0.3	0.28	91%
Hippocampus	0.35	0.32	91%

It is important to note, however, that the number of enhancers and genes able to be linked varies widely across cell types. This is likely due to differences in the number of significant Hi-C contacts ascertained in each cell type (Figure 28). For example, in the liver there are 492,187 significant contacts that link 72% of the enhancers and 73% of the expressed genes in that context. By contrast, in the prefrontal cortex there are only 37,889 significant contacts. These link only 25% of enhancers and 32% of

genes expressed in the prefrontal cortex. The large differences in the number of Hi-C contacts and potential enhancer-gene links are consistent across multiple significance thresholds, suggesting that the threshold choice of $Q < 0.05$ for identifying significant connections is not responsible for differences in the distribution of connections across tissues (Figure 28). This may be due to both biological differences in the number of chromatin interactions between tissues and the known technical limitations of chromatin interaction mapping across Hi-C experiments with varying read depth. This result complicates direct comparisons across tissues, so we will instead focus on trends within tissues and the consistency of these across tissues.

Table 7. Number of linked enhancers, genes, and Hi-C contacts per tissue.

Tissue	Number of enhancers		% enh linked	Number of genes		% genes linked		Number of contacts
	total	linked		linked (tot)	linked (exp)	total	exp	
Heart	136,844	116,396	85.1	14,575	10,003	84.2	85.1	868,904
Liver	100,060	71,794	71.8	12,247	7,818	70.7	72.6	492,187
Spleen	119,958	68,299	56.9	10,777	8,182	62.2	65.6	87,828
Muscle	91,424	33,248	36.4	4,276	2,900	24.7	27.4	32,962
Hippocampus	124,590	43,928	35.3	6,065	4,757	35.0	38.6	49,052
Lung	199,237	57,575	28.9	5,799	4,684	33.5	35.0	30,600
Prefrontal cortex	184,426	45,502	24.7	5,237	4,147	30.2	32.3	37,889
Pancreas	153,214	30,106	19.6	7,139	4,560	41.2	41.8	79,924
Small intestine	181,197	11,808	6.5	1,738	1,400	10.0	10.6	14,483
Ovary	188,908	9,237	4.9	1,251	953	7.2	7.7	12,762

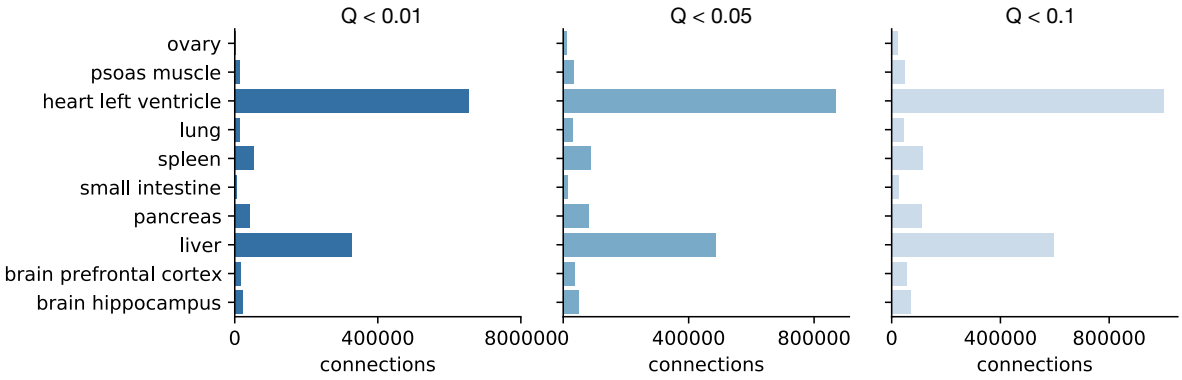

Figure 28. Variability in the number of Hi-C contacts by tissue is not sensitive to the significance cutoff. Histograms showing variability in the number of Hi-C contacts (connections) across tissues at three significance cutoffs. We used a cutoff of Q < 0.05; the tissue variability exists at both more lenient and more conservative cutoffs.

Expressed genes are associated with the number of enhancers in the gene-level enhancer landscape

We evaluated whether properties of genes' enhancer landscapes associated with gene-level attributes. We found that genes expressed in a tissue have a larger number of active enhancers in that tissue than genes that are not expressed. For example, in the liver, the median number of enhancers for an expressed gene is 12, compared to 6 for gene not expressed in the liver (Figure 29; MWU p = 3.01e-50). Due to the resolution of the Hi-C data, some enhancers are linked to multiple TSSs in the corresponding anchor point. Because we are unable to disentangle the precise number of enhancers associated with each of these TSSs in this case, we also stratified this analysis by TSS number. Although the trend is weaker as the number of potential gene targets increases, we still observe that genes expressed in the liver are associated with more liver enhancers than those that are not expressed (Figure 29B). This trend is consistent across the ten tissues we considered. We observed a higher number of enhancers associated with expressed genes compared to those that are not expressed in each tissue (Figure 29C).

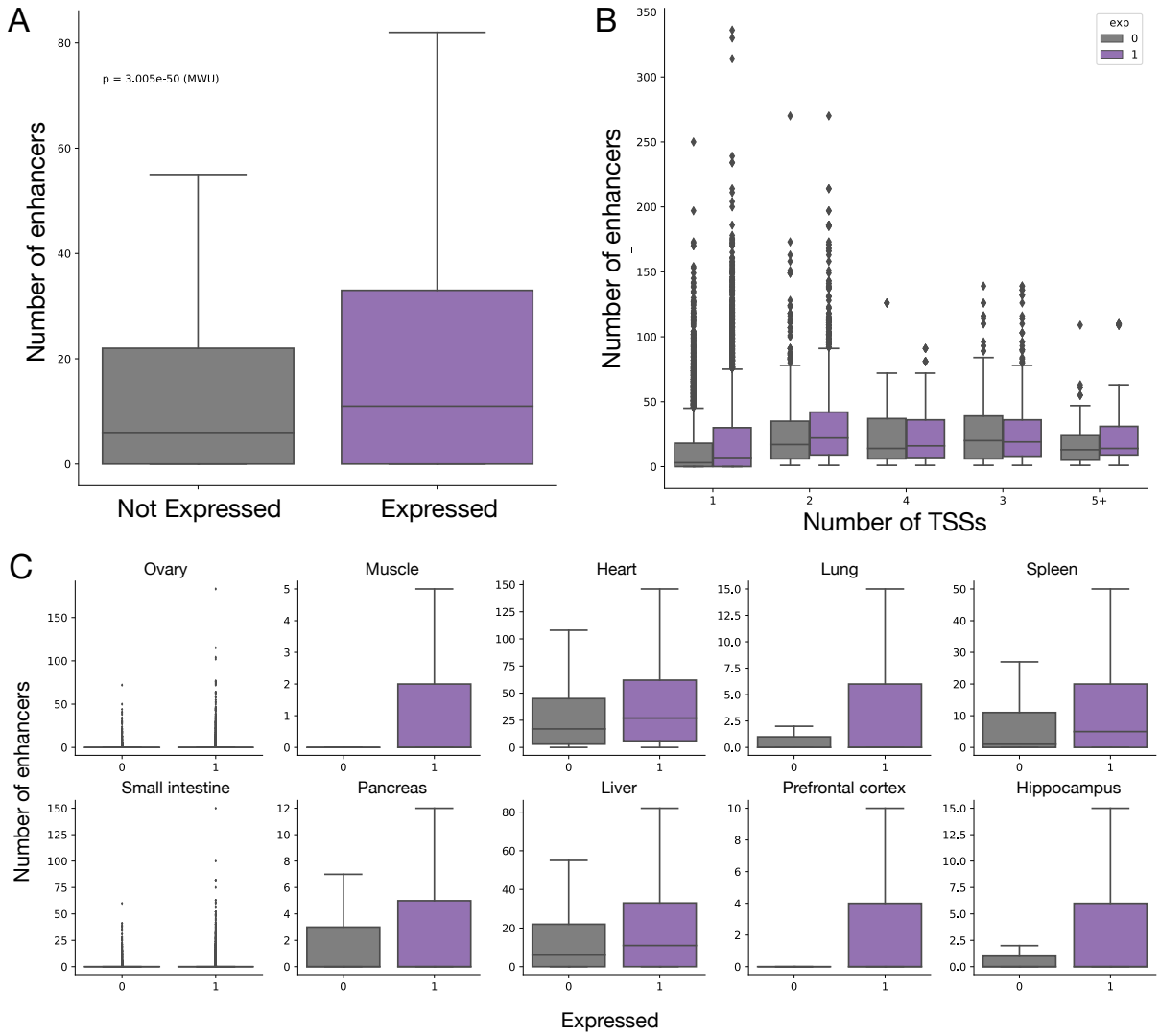

Figure 29. Expressed genes are linked to more enhancers than non-expressed genes.
(A) Genes expressed in the liver have more linked enhancers than non-expressed genes. Liver genes with expression in the liver have a median of 12 liver enhancers, while genes without expression in the liver have a median of 6 liver enhancers (MWU, p = 3.01e-50). Expressed genes are shown in purple and non-expressed genes are shown in gray. (B) Liver genes with expression in the liver have more linked enhancers than those that do not, even when controlling for the number of TSSs in the 40 kb Hi-C anchor points. This suggests that the difference between expressed and non-expressed genes is robust, despite our inability to separate interactions that occur within 40 kb. (C) Across tissues, genes with expression (1, purple) in a tissue have a larger number of enhancers active in the same tissue than non-expressed genes (0, gray). Outliers only shown for ovary and small intestine.

Features of the gene-level enhancer landscape are associated with tissue-specific genes

Genes vary in their expression across tissues. When considering expressed genes, we observe wide variability in the number of associated enhancers. The tissue-specificity of the gene and the activity

patterns of the enhancer elements should contribute to this variability. For example, we hypothesized that genes with tissue-specific expression patterns would have a higher proportion of tissue-specific enhancers. We find that the proportion of tissue-specific enhancers in a landscape is higher for genes with tissue-specific expression than for expressed genes overall (Figure 30). In some contexts, such as the liver, tissue-specific genes also have a greater number of associated enhancers (Figure 30). However, this is not true across all tissues we considered, suggesting that the proportion of the tissue-specific elements may be more important than the number of elements overall. We also observe that some tissues have more genes associated with completely tissue-specific landscapes, although this may be partially due to the differences in the number of enhancers linked to genes across tissues (Figure 31).

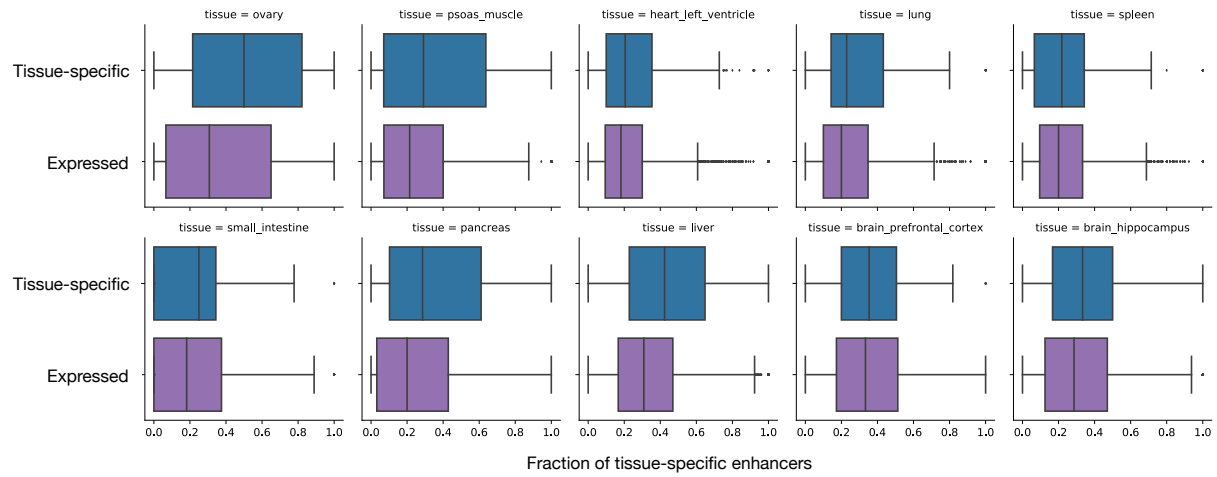

Figure 30. Tissue-specific genes have a higher proportion of tissue-specific enhancers.
Boxplots of the fraction of tissue-specific enhancers in gene-level enhancer landscapes for all ten tissues considered. Plots are stratified by gene type: tissue-specific genes and all expressed genes. Outliers are shown.

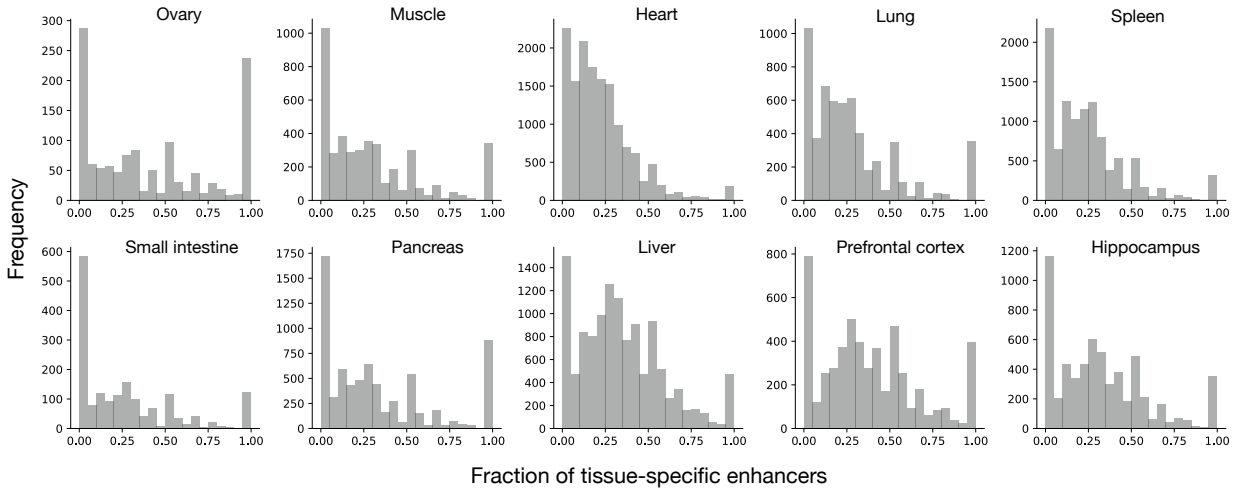

Figure 31. Some tissues have a higher proportion of genes associated with tissue-specific enhancer landscapes. Some tissues (e.g. ovary, psoas muscle, pancreas) have more genes associated with completely tissue-specific landscapes than others (e.g. heart, liver, spleen), although this may be due to the differences in the number of enhancers linked to genes across tissues.

Although tissue-specific genes generally have more tissue-specific enhancer landscapes, some genes do not follow this pattern (n = 1391). At the extremes, some highly tissue-specific genes are linked to only broadly active enhancers, or broadly active genes have only tissue-specific enhancers. The first case occurs only once in our dataset with *SLC22A25*, a transmembrane transport gene. This gene is expressed specifically in the liver, but all of the linked enhancers show evidence of activity across multiple tissues. It is difficult to draw any conclusions from a single example, but this will be an interesting category to revisit as enhancer-gene assignments improve. The second case (broadly active genes, tissue-specific enhancers) is more common, occurring for 1390 unique genes. These cases in particular are interesting because they represent instances where different enhancer landscapes can lead to similar gene expression patterns across tissues. These genes are enriched for GO annotations related to membrane fission, endomembrane system organization, regulation of GTPase activity, and central nervous system development. This may suggest a role for tissue-specific enhancers in regulating processes that are not limited to a single cell type.

Features of the enhancer landscape are not strongly associated with gene-level constraint

Previous work in mouse suggested that the number of enhancers regulating a gene was associated with gene function. Housekeeping genes were shown to have few enhancers, while genes coding for important developmental transcription factors had a median of three enhancers[147]. We sought to evaluate this trend in human tissues and hypothesized that other gene attributes would be also be associated with gene-level enhancer landscape features because of constraint on their expression. We expected that genes under stronger constraint would have more associated enhancers because this provides the potential for regulatory buffering or finer control of expression. We curated three gene sets likely to be under higher levels of constraint than expressed genes overall: housekeeping genes, essential genes, and loss-of-function intolerant genes. We then asked whether these subsets of genes were associated with different numbers of enhancers. We found that housekeeping, loss-of-function intolerant, and essential genes are not associated with a higher number of enhancers than expressed genes outside of these categories (Figure 32). This trend was similar across tissues, although the median number of enhancers associated with the gene subsets varied. This is likely due to differences in Hi-C resolution across tissues.

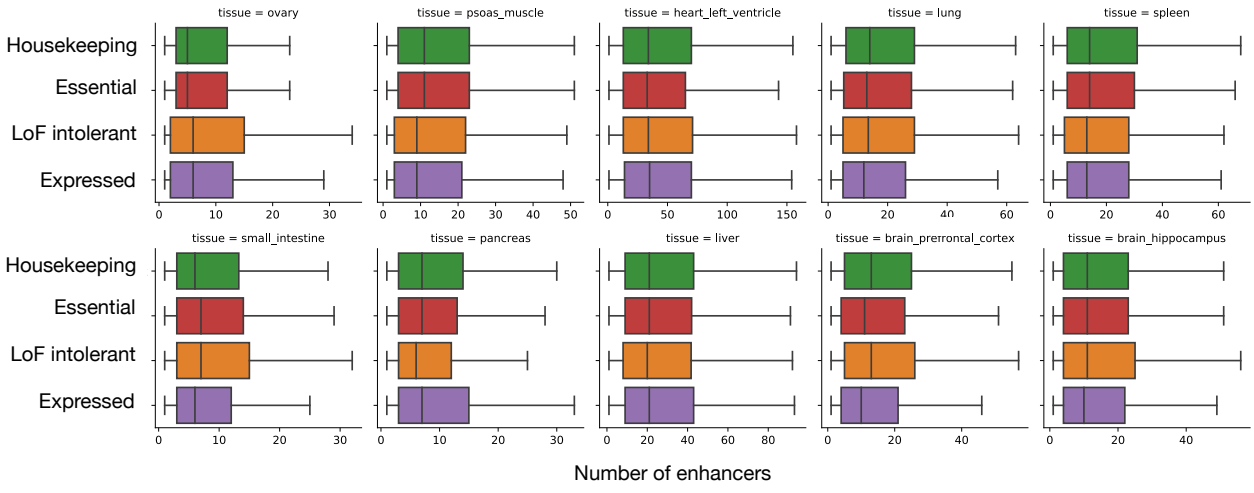

Figure 32. Number of enhancers in a gene-level landscape is not strongly associated with gene constraint. We curated three gene sets we expected to be under stronger constraint than genes overall: housekeeping genes, essential genes, and loss-of-function intolerant genes. Boxplots for each tissue show only modest differences in the median number of enhancers associated with genes in each of these categories. We conclude that housekeeping,

loss-of-function intolerant, and essential genes are not strongly associated with a higher number of enhancers than expressed genes outside of these categories.

The result that housekeeping genes do not have fewer associated enhancers is contradictory to the previously published results discussed earlier[147]. However, that result was described using a different enhancer-gene linkage approach that is based on patterns of co-activity across tissues. In contrast, our approach uses physical evidence of proximity in the tissue of interest from Hi-C to link enhancers and genes. Since there is no gold standard set of enhancer-gene links, it is difficult to evaluate individual methods. To reconcile these contradictory results, we downloaded an additional set of published enhancer-gene links that uses third strategy to make predictions[169]. This approach, called JEME, predicts enhancer-gene links using a two-step process. First, JEME uses lasso regression to filter the set of all enhancers within 1 Mb of a TSS to those that best predict TSS activity. Second, it uses the regression output in combination with histone modification ChIP-seq to train a random forest model to predict enhancer-gene links derived from ChIA-PET, Hi-C, and eQTL studies. Our previous result holds when using this approach to define enhancer-gene links (Figure 33), suggesting that our conclusions are robust and reflect of real attributes of the enhancer landscapes of housekeeping genes.

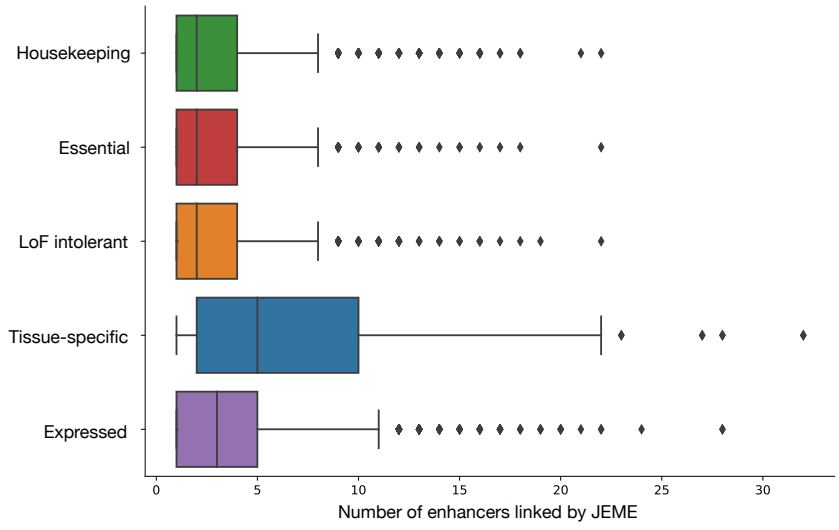

Figure 33. The distribution of the number of enhancers stratified by gene-level attributes is reproducible across enhancer-gene linking strategies.

We applied a second approach to link liver enhancers with their target genes that uses an activity-linking approach described previously[169]. This figure shows boxplots of the number of enhancers in each gene-level enhancer landscape, stratified by gene type. We still observe no difference between the number of enhancers in the enhancer landscapes associated with housekeeping, essential, and LoF intolerant genes. We also find that tissue-specific genes are associated with more enhancers using this approach, which is also consistent with our previous results (Figure 32).

Gain of enhancer activity is correlated with features of region-level enhancer landscapes

Although gene expression is highly correlated across species, the individual enhancer elements regulating gene expression are known to turnover rapidly[27,160]. This may be due to robustness provided by redundancy in the region-level enhancer landscape. We hypothesized that a large number of active enhancers decrease the chance that an individual enhancer gain or loss would disrupt gene expression levels. Therefore, the gain of new enhancer activity would be positively correlated with the number of enhancers in a region. To test this hypothesis, we leveraged a dataset of genome-wide profiling of enhancer activity and gene expression in liver across mammalian species to evaluate the stability of enhancer activity in different landscape contexts over time. In these analyses, we quantify enhancer landscapes using 1 Mb regions tiled across the genome; this enables us to account for the varying gene density and varying potential for a gained enhancer to influence the regulation of multiple genes across the genome.

The gain of human enhancer activity in liver is positively correlated with both the number of enhancers and the number of genes in a landscape (Figure 34; $\rho = 0.72$ for enhancers, $\rho = 0.61$ for genes). This suggests that enhancers are more likely to be gained in regions with a higher level of existing regulatory activity and with more potential gene targets. We also observed a negative correlation between the number of Hi-C interactions in the landscape and the number of gained enhancers ($\rho = -0.26$). This may be related to physical constraints on the number of chromatin interactions possible in a 1 Mb window at one time, where new enhancers are more likely to be gained in regions that have space to acquire new functional interactions. All of these features of the region significantly predict the number of gained

enhancers in a negative binomial model (Table 8), providing evidence that greater numbers of enhancers in a region is associated with increased turnover of enhancer activity.

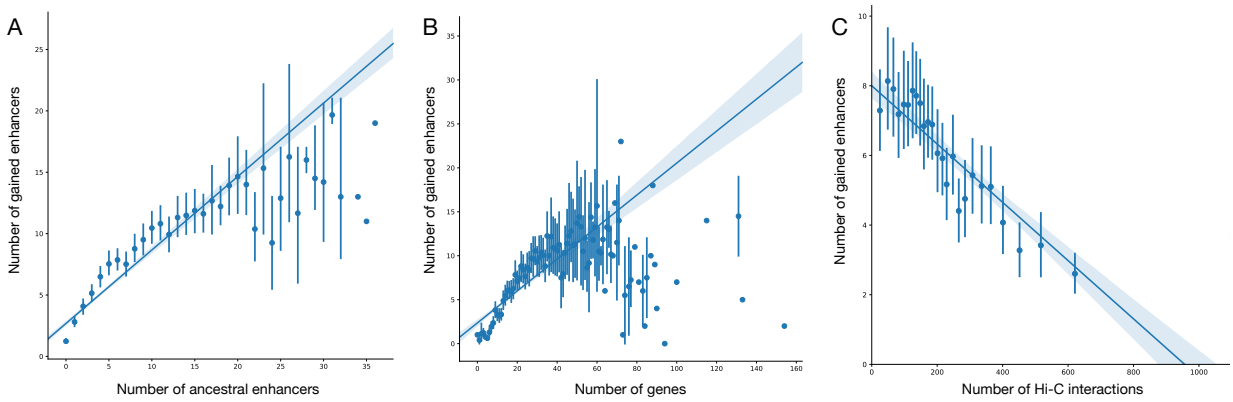

Figure 34. Gain of enhancer activity is associated with a larger number of ancestral enhancers and genes in the region-level enhancer landscape.
Plots show the association between the number of ancestral enhancers (A), number of genes (B), and number of Hi-C interactions (C) versus the number of gained enhancers in region-level enhancer landscapes. Data are binned into evenly sized bins along the x-axis with points representing the median and vertical lines displaying bootstrapped 95% confidence intervals. The plotted linear regression is fit to the original data.

Table 8. Region-level enhancer landscape features predict the number of gained enhancers.

Predictor	Beta	SE	Z	P
Number of ancestral enhancers	0.0867	0.0037	23.517	2.7E-122
Number of genes	0.0198	0.0016	12.630	1.4E-36
Number of Hi-C interactions	-0.0006	0.0001	-4.164	3.1E-05

The number of enhancers in a landscape is correlated with the level of evolutionary conservation

In the previous section, we showed that enhancers are more likely to be gained in regions with a larger number of ancestral enhancers. Although we can profile the dynamics of enhancer activity across species using ChIP-seq, conserved activity is not necessarily correlated with the underlying DNA sequence conservation in these enhancer elements. DNA sequence conservation of enhancer regions in different enhancer landscapes has yet to be explored. We hypothesized that ancestral enhancers would have a higher proportion of evolutionarily conserved bp than enhancers that are gained in the human lineage. In

the liver, ancestral enhancers have a higher proportion of overlap with conserved PhastCons elements than gained enhancers (Figure 35A; MWU, p = 2.3e-168). Across 1 Mb regions, this translates into a slight positive relationship between the total proportion of conserved enhancer bp and the number of ancestral enhancers in the window (Figure 35B). This suggests more enhancer-dense regions also have a greater proportion of conserved sequence, which supports previous work showing that evolutionary conservation is associated with enhancer function. The gained enhancers may be evidence of exaptation of less conserved or younger sequences to create novel regulatory pathways.

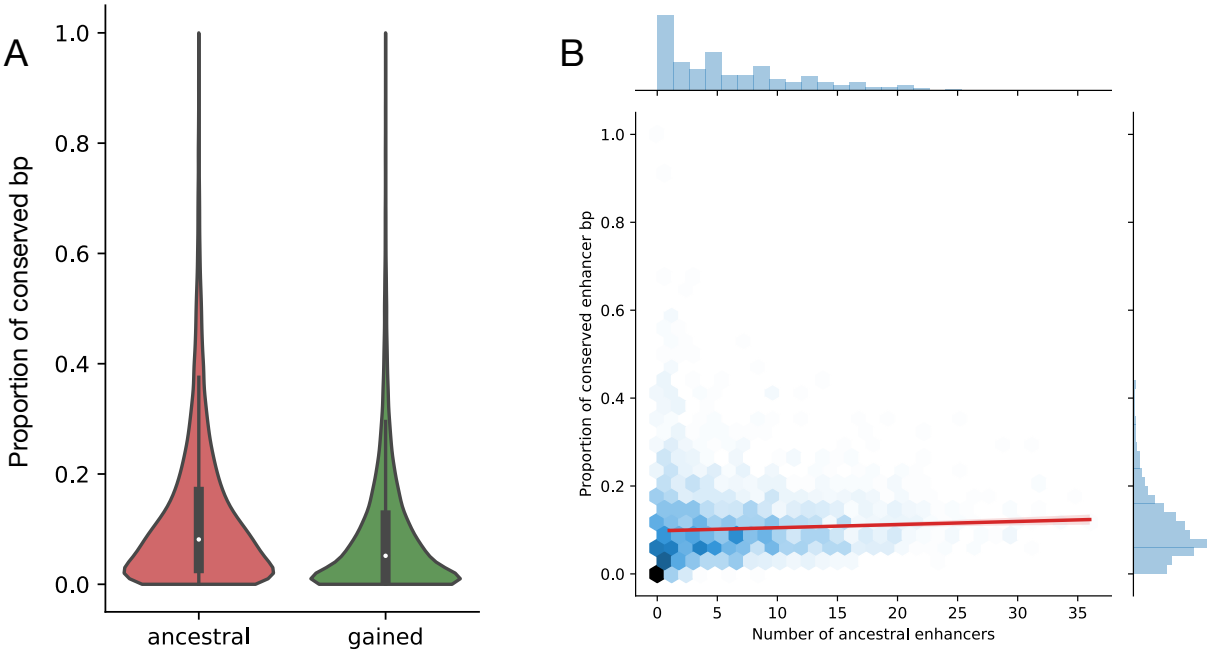

Figure 35. Ancestral enhancers are correlated with conserved sequences in region-level enhancer landscapes. (A) Ancestral enhancers have a higher proportion of overlap with evolutionarily conserved sequences than enhancers that have gained activity after divergence from macaque. (B) The number of ancestral enhancers in a region-level enhancer landscape is modestly correlated with the total proportion of conserved enhancer bp in the region. The proportion of conserved enhancer sequence in the region-level landscape is quantified as the number of enhancer bp in the region overlapping a PhastCons element divided by the total number of enhancer bp in the region. The color of the hexbins represent increasing density of observations, scaled from white to dark blue. On the outside of the plot are histograms of the x-axis (top) and y-axis (right) values. The red line is a linear regression line fit to the original data with bootstrapped 95% confidence intervals (number of bootstraps = 500).

We next evaluated whether the relationship between the number of enhancers and the proportion of evolutionarily conserved sequences holds for a the Hi-C-based enhancer landscape definition. We regressed the proportion of conserved bp on the number of enhancers associated with each gene using our

Hi-C based landscape definition. Across the ten tissues we considered, we observed a positive relationship between the number of enhancers linked to a gene and the proportion of conserved sequence (Figure 36). This is consistent with the previous finding that region-level enhancer landscapes with more elements are also more conserved and with the known relationship between evolutionary conservation and enhancer function.

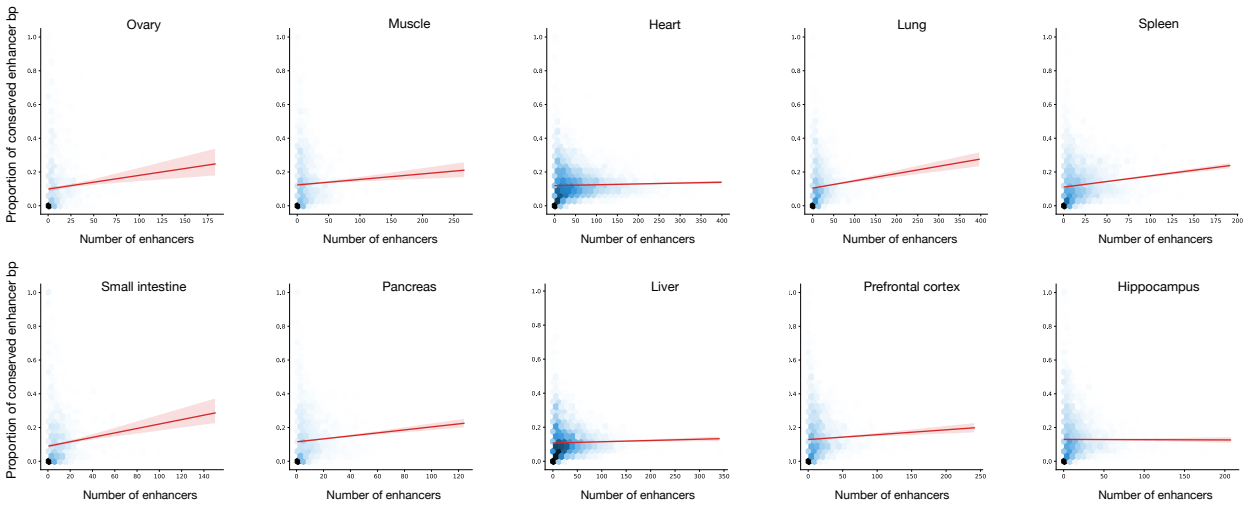

Figure 36. The number of enhancers in the gene-level enhancer landscape is correlated with the proportion of evolutionarily conserved enhancer sequence.
Across the ten tissues considered in this study, the number of enhancers in a gene-level enhancer landscape is positively correlated with the proportion of evolutionarily conserved enhancer sequence in the enhancer landscape. The proportion of conserved enhancer sequence in the gene-level landscape is quantified as the number of enhancer bp in the landscape overlapping a PhastCons element divided by the total number of enhancer bp in the landscape. The color of the hexbins represent increasing density of observations, scaled from white to dark blue. The red line is a linear regression line fit to the original data with bootstrapped 95% confidence intervals (number of bootstraps = 500).

Region-level enhancer landscapes with more enhancers have higher TFBS density

Enhancers bind to specific transcription factors in order to regulate the expression of their target genes. Recent work using synthetic enhancer sequences in mouse suggests that the density of transcription factor binding sites (TFBSs) is the best indicator of the strength of enhancer activity[270]. In these analyses we use the region-level definition of enhancer landscapes to test whether the number of enhancers in a region is associated with the TFBS density of those enhancers. We identified putative TFBSs using a motif

scanning program, FIMO, that looks for matches with canonical TF motifs from HOCOMOCO core database (v11). We hypothesized that regions with a greater number of active enhancers would also have a higher TFBS density. We observe a positive relationship between the number of enhancers in a region and the TFBS density of the enhancers in that region (Figure 37A). Limiting to the density of unique TFBSs diminished the strength of the trend. This is consistent with the hypothesis that regulatory activity is related to the amount of potential TF binding in a region and not necessarily the diversity of binding sites. Stratifying enhancers by evolutionary history, we observe that, on average, ancestral enhancers have higher TFBS densities than recently gained enhancers (Figure 37B). This may contribute to the maintenance of activity in these sequences over evolutionary time.

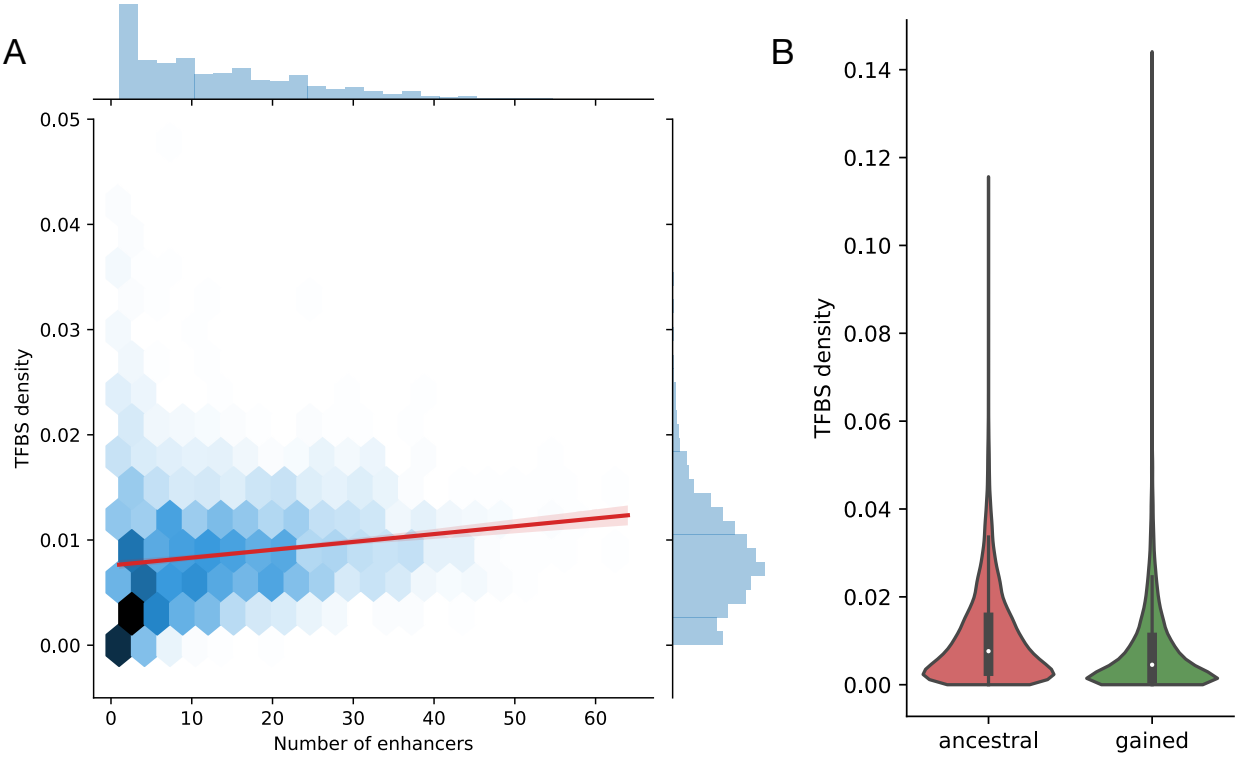

Figure 37. The number of enhancers in a region-level enhancer landscape is associated with TFBS density.
(A) In the liver, the number of enhancers in a region-level enhancer landscape is positively correlated with the TFBS density of the enhancers in that region. The color of the hexbins represent increasing density of observations, scaled from white to dark blue. On the outside of the plot are histograms of the x-axis (top) and y-axis (right) values. The red line is a linear regression line fit to the original data with bootstrapped 95% confidence intervals (number of bootstraps = 500). (B) Ancestral enhancers have higher TFBS densities than enhancer that have gained activity in the human lineage.

Enhancer landscapes with more enhancers are depleted for eQTL, but enriched for GWAS variants

After establishing that enhancer landscapes vary across genes and are associated with attributes of gene-level constraint, we asked whether the enhancer landscape influences a gene's robustness to genetic variation. Focusing on the liver, enhancer sequences in regions with a large number of enhancers ($n > 5$; Methods) are slightly depleted for overlap with genetic variation associated with the expression of genes (eQTL; Figure 38A). This is possibly because redundancy in the enhancer landscape buffers against changes in gene expression, although further study is required to confirm this hypothesis. However, enhancers with conserved activity across species in regions with a large number of enhancers are enriched for variants influencing liver traits identified by genome-wide association studies (Figure 38B). It is possible that, rather than providing redundancy, some enhancer landscapes have multiple enhancer elements that require cooperation. We hypothesize that disruption of these landscapes would be more likely to result in disease phenotypes. Enhancers in landscapes with few enhancers ($n <= 5$) have even less overlap with eQTL, although the trend is not statistically significant. Similarly, these enhancers do not show significant evidence of depletion for variants associated with relevant GWAS traits (Figure 38). The non-significant GWAS results may be due to a lack of power because of the relatively small number of liver-relevant GWAS variants.

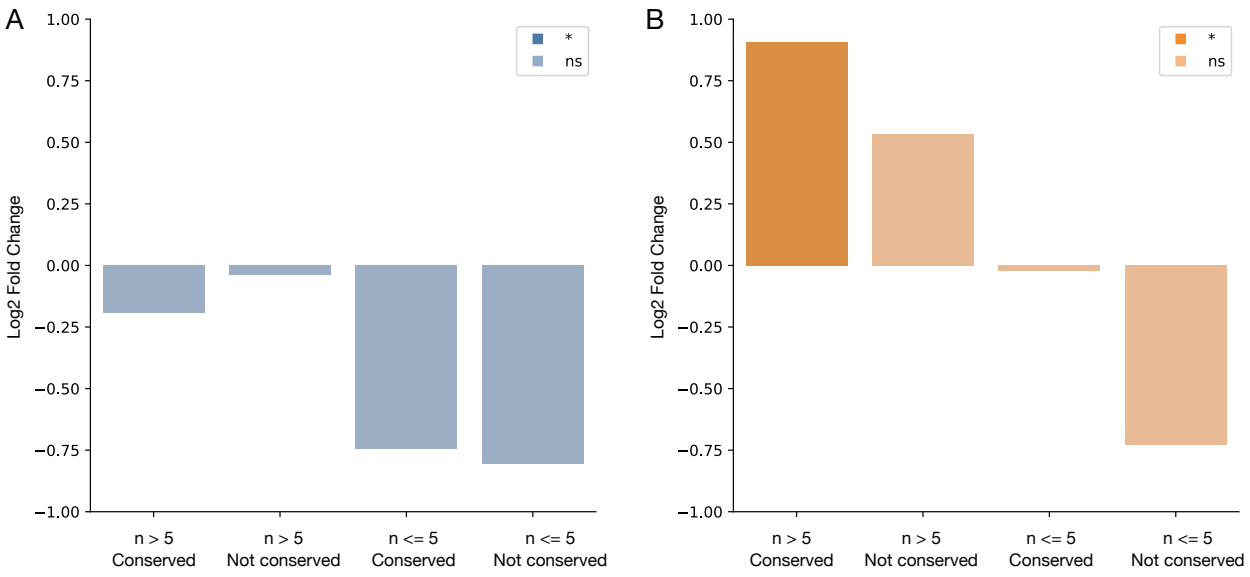

Figure 38. Region-level liver enhancer landscapes have variable enrichment for non-coding variants.
We tested liver enhancers for enrichment with genetic variation stratified along two axes: (1) the number of enhancers in region-level enhancer landscapes; (2) whether the activity of the enhancer was conserved across at least two mammalian species. We expected that the number of enhancers in a region could buffer the effects of genetic variation, and that the enhancers with conserved activity would be more likely to be enriched for trait-associated variation. (A) For all groups of enhancers considered, there is not significant depletion for overlap with GTEx eQTL, although enhancer landscapes with fewer enhancers have less overlap. (B) Enhancers with conserved activity in landscapes with more regulatory elements are significantly enriched for overlap with liver-relevant GWAS variants.

Conclusion

Enhancers are crucial for the proper regulation of gene expression; the disruption of these elements has been linked to a range of disease phenotypes. Despite new work showing that multiple enhancers can work in combination to regulate the expression of their target genes, enhancer elements are often studied in isolation. Considering the enhancer landscape of a gene is required to fully understand the dynamics of gene expression and the influence of genetic variation on regulatory elements.

In this chapter we leverage experimental data across six mammalian species and ten different human tissues to define enhancer landscapes on a genome-wide scale. Using both a region- and gene-level definition of an enhancer landscape, we show that most genes are associated with multiple regulatory elements. We quantify the number of regulatory elements in these landscapes and show this is

correlated with other measures of functional activity, such as sequence conservation and transcription factor binding site density. We also find that the size and tissue-specificity of the enhancer landscape is associated with the tissue-specificity of the gene target. Using the enhancers profiled across mammalian species, we explore the relationship between enhancer landscapes attributes and enhancer turnover; we find that the ancestral enhancer landscape is predictive of the amount of enhancer turnover in that region. Finally, we show that features of the enhancer landscape are associated with differences in enrichment for expression- and trait-associated genetic variation. This chapter highlights the features of enhancer landscapes associated with differences in gene expression across tissues and enhancer turnover across species. In order to more comprehensively capture the relationship between enhancers and genes, future work must consider enhancer landscapes, especially when interpreting regulatory variation.

CHAPTER V

Discussion

In this dissertation, we explore strategies for identifying enhancers and incorporating them into models of gene regulatory architecture. We begin with a comprehensive comparison of existing enhancer identification strategies in order to understand the features and limitations of these elements. We then build on our current understanding of individual enhancers by integrating experimentally derived genomic interaction data to define regulatory units, which we refer to as enhancer landscapes. The enhancer landscapes provide a more complex view of enhancer function than simply considering individual elements. This broader context enables us to interpret gene expression dynamics and the impact of genetic variation on these elements. Throughout the dissertation, we quantify the enrichment and gene expression consequences of single nucleotide and structural variation on regulatory elements. The advances made here improve our understanding of gene regulation and our ability to interpret non-coding genetic variation that impacts enhancer elements.

Enhancer identification remains a challenging and unsolved problem, despite the wide range of existing experimental and computational approaches. Each method, either explicitly or implicitly, represents a different perspective on what constitutes an enhancer and which identifiable signatures are most informative about enhancer activity. For example, histone modifications characteristic of enhancers are found on histones that flank active enhancers, while eRNA is thought to be bidirectionally transcribed from the active sequence itself. As a result, despite the use of the term “enhancer” to describe all these regions in the literature, we expected different assays and algorithms to identify somewhat different sets of regions. However, given the lack of comprehensive genome-wide gold standard enhancer sets, evaluation of the accuracy of these approaches is challenging. Thus, in Chapter II, we compared existing strategies with respect to one another and to proxies for regulatory function. All pairs of enhancer sets overlap more than expected by chance, but we found substantial differences in the genomic, evolutionary, and functional characteristics of identified enhancers *within* similar tissues and cell types. Enhancer sets

vary significantly in their overlap with conserved genomic elements, GWAS loci, and eQTL. Furthermore, the majority of GWAS loci and eQTL have inconsistent evidence of enhancer function across enhancer sets. In addition, regions identified as enhancers by multiple methods *do not* have significantly stronger evidence of regulatory function.

The consistent lack of agreement between methods demonstrates that many working definitions of "enhancer" have very low overlap. Focusing on functional annotations, we find agreement between methods about basic functions, but substantial differences in more specific annotations. This suggests that different strategies contribute unique information towards the identification of functionally important enhancers. In general, enhancers defined by eRNA (FANTOM, GRO-cap) have modestly more enrichment for proxies of functional activity than other methods, but this comes at the expense of low sensitivity. Our results argue that, given the lack of a clear gold standard and the substantial disagreement between strategies, it does not make sense to identify a single "best" method given current knowledge. Furthermore, because enhancer identification strategies have such substantial differences, one strategy cannot and should not be used as a proxy for another. Different strategies may produce substantially different conclusions, especially when predicting whether a genetic variant will alter regulatory function or quantifying the level of evolutionary constraint on enhancer regions. Understanding this is particularly important given that studies of gene regulation commonly use only a single approach to define enhancers. GWAS have identified thousands of non-coding loci associated with risk for complex disease, and a common first step in the interpretation of a trait-associated locus is to view it in the context of genome-wide maps of regulatory enhancer function[54,55,78,80,81,244,247,271]. Our work complicates the standard application of genome-wide enhancer predictions to understand the molecular mechanisms underlying disease and highlights the need for more precise terminology[51,272,273].

We must acknowledge both the biological and technical differences between enhancer sets, especially when applying them to study non-coding genetic variants. Enhancer sets identified by different approaches rely on different underlying assumptions about what constitutes an active enhancer element in a given context. When choosing which identification strategy to employ, we must weigh the tradeoffs

between sensitivity and specificity. For example, enhancers defined using DNase hypersensitivity or histone modification ChIP-seq generate predictions with higher genome coverage than other methods, but at the expense of more false positive predictions. These methods may be more appropriate for hypothesis generation where a more inclusive definition is beneficial. Alternatively, enhancers defined using eRNA or specific TF binding profiles generate enhancer sets with fewer putative regions and stronger evidence of active transcription or relevant protein binding in a given cell type. These methods may be more useful when studying molecular mechanisms in biological contexts that are well-defined. Finally, it is useful to examine the robustness of all downstream conclusions to the enhancer definition used. Our results suggest that many conclusions will differ based on the chosen enhancer identification strategy; it is important to understand how specific conclusions change based on the enhancer set and interpret downstream biological conclusions with this additional context.

Technical and biological variation in the underlying experimental assays and data processing pipelines contribute to the variation between putative enhancer sets. However, we minimized technical variation by calling and comparing enhancers using consistent computational pipelines. Furthermore, comparisons of biological replicates of histone modification ChIP-seq data suggest that the level of difference we observe between enhancer sets is larger than among biological replicates. Genetic variation between individuals could also explain some of the discordance. Previous work shows that chromatin states associated with weak enhancer activity exhibit some variation between individuals, and QTL associated with changes in epigenetic modifications and enhancer activity between individuals have been identified[59,274]. However, the proportion of epigenetic modifications that are variable across individuals is estimated to be small (1–15%)[61]. Thus, variation between individuals is unlikely to be the main cause of the lack of agreement we observe between methods, in particular for enhancer sets defined from cell lines. Furthermore, there are strong similarities between enhancers and other regulatory elements, like promoters, and some promoters even have enhancer activity[46,47,49]. We focus on methods designed to distinguish enhancers to reduce the impact of disagreement due to the comparison of different elements from the broader regulatory spectrum; nevertheless, some identification strategies may include or exclude

functional elements with variable activities. However, while the proxies we use for regulatory function (evolutionary conservation, GWAS loci, and eQTL) each have weaknesses, we observe similar disagreement across each proxy. This supports the functional relevance of the differences we demonstrate between enhancer sets. Taking each of these limitations into account, the disagreements we observe remain striking.

Despite substantial differences in the enhancers identified by different strategies, disruption of predicted enhancers regions is known to contribute to a range of disease phenotypes[189,193–196]. SVs in particular have been previously shown to alter individual enhancer elements as well as the larger three-dimensional chromatin architecture[121,124,198]. These alterations can lead to regulatory disruption by removing important enhancer elements or changing the interactions between enhancers and their target genes. The latter often refers to enhancer hijacking, where inappropriate enhancer-gene contacts can lead to ectopic gene expression and disease[121,199–20]. In Chapter III, we used a novel cohort with both genome and transcriptome data to quantify the gene expression effects of SVs on gene regulatory elements and architecture. We demonstrate that SVs altering regulatory elements, including enhancers and promoters, have a clear impact on the expression of associated genes. These expression effects are present even when coding sequence is unaffected and are proportional to the amount of regulatory sequence disrupted, underscoring the importance of proper gene regulatory function. Our results provide an additional layer of information that is crucial for more accurate functional annotation of SVs. Indeed, the regulatory disruption scores developed using these data successfully distinguished known pathogenic SVs. Deletions and duplications with the most extreme regulatory disruption scores were enriched for overlap with pathogenic variants identified by gnomAD. Many of these were uniquely identified using our approach, suggesting that integrating regulatory annotations provided useful orthogonal information for variant prioritization.

In addition to gene expression consequences, we also provide evidence of selection on SVs that influence regulatory elements and binding sites of an architectural protein, CTCF. Since variants affecting fitness are subject to selection, we expect that SVs with negative fitness effects will be observed at lower

frequencies. We find that SVs altering enhancer and CTCF binding sites are observed at significantly lower frequencies than SVs in other non-coding (intronic and intergenic) contexts. The low frequency of SVs affecting CTCF sites is particularly notable since it is similar to that of SVs affecting coding sequence. We hypothesize that the disruption of CTCF has significant potential to be deleterious because CTCF plays a large role in the establishment and maintenance of three-dimensional chromatin loops. These loops are important for creating regulatory domains, such as TADs, bringing regulatory elements into close proximity to genes, and insulating certain enhancer-promoter interactions[117,129]. Disruption of these functions may be more likely to have an effect on multiple enhancer-promoter interactions and gene expression than the disruption of individual regulatory elements. Broadly, our results suggest that the disruption of regulatory elements and CTCF sites is often deleterious and SVs altering these elements are under negative selection.

Quantifying the gene expression consequences of SVs requires several assumptions that may impact the results presented here. As discussed in Chapter II, the accurate identification of enhancer elements is non-trivial. Differences in enhancer identification strategy have the ability to influence downstream conclusions. We chose enhancers defined using histone modification ChIP-seq data in a relevant brain cell-type in order to provide a broad set of putative elements. However, there are likely to be both false-positive and false-negative predictions in this set; additional work is required to determine whether our results are robust to the enhancer identification strategy used. Similarly, regulatory elements must be linked to their putative target genes in order to test for gene expression changes from SVs. Enhancer-gene mapping remains open problem in the regulatory genomics, and CTCF binding sites are not easily linked with the genes. For enhancers, we chose a commonly adopted approach based on Hi-C interactions in the same biological context. Using this method, we were able to link enhancers with approximately 30% of genes. As experimental and computational approaches for enhancer-gene mapping improve, we can expand our original analyses to include a greater proportion of genes. Additionally, better models of the regulatory architecture will allow us to make predictions about the potential gene-specific effects of CTCF disruption.

While individual enhancer elements make important contributions to the proper regulation of gene expression, there is a growing body of evidence to support the more complex view of enhancers as part of a broader enhancer landscape[12,147–149,184]. In Drosophila, enhancer landscapes have been shown to provide robustness by buffering the effects of genetic variation in enhancer elements[150–152]. Literature describing super-enhancers, enhancer domains, and early studies of enhancer landscapes in mice and humans suggest that similar mechanisms exist in mammalian species[12,147,148,160,164,183]. Chapter IV leverages Hi-C interaction data and ChIP-seq profiling across six mammalian species and ten human tissues to develop a framework to define human enhancer landscapes. The features of enhancer landscapes vary in multiple dimensions, including the number of enhancers, level of sequence or activity conservation, TFBS density, and enrichment for genetic variants. Across tissues, we find that these enhancer landscape features are reflective of the gene expression dynamics of associated genes. For example, we observe that tissue-specific genes have a larger proportion of tissue-specific enhancer elements. Aside from providing a framework to define gene regulatory landscapes in human tissues, these results inform future work in genetic variant interpretation by highlighting differences in enrichment for variants with different effects between enhancer landscapes.

Previous work has demonstrated that, although gene expression is highly conserved across species, regulatory elements turnover quickly between species[27,160]. We hypothesized that this phenomenon is partially due to robustness provided by redundancy in enhancer landscapes that allows for the emergence of novel regulatory elements while maintaining overall gene expression levels. By integrating enhancer profiling across six mammalian species with our definition of an enhancer landscape, we find that enhancers are more likely to be gained in regions that have a greater number of existing enhancers and gene targets. This is consistent with our hypothesis that multiple regulatory elements can provide stability of expression as novel elements emerge. These ancestral enhancers overlap a higher proportion of evolutionarily conserved elements and have a greater TFBS density, further suggesting that they serve an important regulatory role.

As in Chapter III, the results in Chapter IV are limited by our ability to accurately identify active enhancer regions in a given biological context. We again use a histone-modification-derived approach, which likely provides sensitivity at the expense of specificity. Furthermore, we rely on interaction matrices from Hi-C to assign putative enhancers to their target genes. The 40 kb resolution of the Hi-C data is a limiting factor. It prevents us from distinguishing interactions within a 40 kb window, which precludes us from recognizing proximal regulatory interactions or disentangling the precise gene target when multiple TSSs fall into the same window. Because enhancers are thought to act as distal regulatory elements, we expect to capture the majority of the functional interactions. As Hi-C and other technologies continue to advance, higher resolution datasets will become available to refine our current results.

This dissertation highlights several key limitations in the field of regulatory genomics and areas of future work. First, we must resist the convenience of ignoring the lingering complexity of enhancer identification. When interpreting non-coding variants of interest or characterizing the enhancer landscape in a new biological context, we must be mindful that using a single identification strategy is insufficient to comprehensively catalog enhancers. Different assays and algorithms have different attributes, and we suggest employing a range of approaches to obtain a more robust view of the regulatory landscape. However, simply focusing on variants with multiple lines of evidence of enhancer activity will not solve the problem, especially when our ability to quantify the false positive rate in a genome-wide enhancer map is limited. More sophisticated statistical models of enhancers and their properties are needed in order to interpret non-coding variants of interest. Previous work has shown that integrating diverse genomic, evolutionary, and functional data can improve the ability to distinguish validated enhancers from the genomic background[70], but obtaining a concordant and functionally relevant set of enhancers remains challenging. We are hopeful that new experimental techniques, like MPRAs, and biologically motivated machine learning methods for integrating different definitions of enhancers will yield more consistent and specific annotations of regions with regulatory functions. Furthermore, functional genomics datasets and three-dimensional chromatin interaction assays performed in the same samples are required to improve our ability to determine accurate links between regulatory elements and genes. Because gene regulation is

a dynamic and context-dependent process, experimental assays performed in the same samples and across multiple time points will provide more robust information about the underlying molecular mechanisms and improve computational models trained on these data.

Second, this work highlights the need for more refined models of the architecture and dynamics of *cis*-regulatory regions. Many different classes of regions with enhancer-like regulatory activities have been discovered[14,28,30,33,44,51,275]. We argue that collapsing the diversity of vertebrate distal gene regulatory regions into a single category is overly restrictive. Simply calling all of the regions identified by these diverse approaches "enhancers" obscures functionally relevant complexity and creates false dichotomies. While there is appreciation of this subtlety within the functional genomics community, there is still a need for more precise terminology and improved statistical and functional models of the diversity of *cis*-regulatory "enhancer-like" sequences and their architectures. Given this diversity, we should not expect all results to be robust to the enhancer identification strategy used. Furthermore, gene regulatory regions do not act in isolation. While this has been explored extensively in model organisms and through case studies of specific enhancer clusters, we still require more comprehensive modeling of human regulatory landscapes. Our work begins this process using computational approaches and publicly available datasets, but more precise data on chromatin interactions across tissues will refine and validate our framework.

Finally, we believe that ignoring enhancer diversity impedes research progress and replication, since "what we talk about when we talk about enhancers" includes diverse sequence elements across an incompletely understood spectrum, all of which are likely important for proper gene expression. Efforts to stratify enhancers into different classes, such as poised and latent, are steps in the right direction, but are too coarse given our incomplete current knowledge. We suspect that a more flexible model of distal regulatory regions is appropriate, with some displaying promoter-like sequence architectures and modifications and others with distinct regulatory properties in multiple, potentially uncharacterized, dimensions[46,276,277]. Consistent and specific definitions of the spectrum of regulatory activity and landscapes are necessary for further progress in enhancer identification, successful replication, and accurate genetic variant interpretation.

REFERENCES

1. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* (1981). doi:10.1016/0092-8674(81)90413-X
2. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–86 (2014).
3. Gasperini, M., Tome, J. M. & Shendure, J. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nat. Rev. Genet.* **21**, 292–310 (2020).
4. Crawford, G. E. *et al.* Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16**, 123–131 (2006).
5. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
6. Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17**, 877–885 (2007).
7. Giresi, P. G. & Lieb, J. D. Isolation of Active Regulatory Elements from Eukaryotic Chromatin Using FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements). in *Tag-Based Next Generation Sequencing* 243–255 (2012). doi:10.1002/9783527644582.ch14
8. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **2015**, 21.29.1-21.29.9 (2015).
9. The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **489**, 57–74 (2012).
10. Yip, K. Y. K. K. Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13**, R48 (2012).
11. Ho, J. W. K. *et al.* Comparative analysis of metazoan chromatin organization. *Nature* **512**, 449–52 (2014).

12. Hay, D. *et al.* Genetic dissection of the α-globin super-enhancer in vivo. *Nat. Genet.* 1–12 (2016).

13. Dogan, N. *et al.* Occupancy by key transcription factors is a more accurate predictor of enhancer activity than histone modifications or chromatin accessibility. *Epigenetics Chromatin* **8**, 1–21 (2015).

14. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–8 (2007).

15. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–8 (2009).

16. Blow, M. J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* **42**, 806–810 (2010).

17. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science (80-.).* **316**, 1497–1502 (2007).

18. Slattery, M. *et al.* Absence of a simple code: How transcription factors read the genome. *Trends in Biochemical Sciences* (2014). doi:10.1016/j.tibs.2014.07.002

19. Grossman, S. R. *et al.* Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc. Natl. Acad. Sci.* **114**, E1291–E1300 (2017).

20. Teytelman, L., Thurtle, D. M., Rine, J. & van Oudenaarden, A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl. Acad. Sci.* (2013). doi:10.1073/pnas.1316064110

21. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res.* (2014). doi:10.1093/nar/gkt1249

22. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* (2012). doi:10.1101/gr.139105.112

23. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, (2005).

24. Pennacchio, L. a *et al.* In vivo enhancer analysis of human conserved non-coding sequences.

Nature **444**, 499–502 (2006).

25. Bejerano, G. *et al.* Ultraconserved Elements in the Human Genome. *Science (80-.).* **1321**, 1321–1326 (2007).
26. Visel, A. *et al.* Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40**, 158–160 (2008).
27. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
28. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
29. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
30. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–21936 (2010).
31. Pekowska, A. *et al.* H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J.* **30**, 4198–4210 (2011).
32. Bonn, S. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat. Genet.* **44**, 148–156 (2012).
33. Pradeepa, M. M. *et al.* Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat. Genet.* **48**, 681–686 (2016).
34. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–61 (2014).
35. Li, W., Notani, D. & Rosenfeld, M. G. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nat. Rev. Genet.* **17**, 207–223 (2016).
36. Core, L. J. *et al.* Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.* (2014). doi:10.1038/ng.3142
37. Hirabayashi, S. *et al.* NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. *Nat. Genet.* **51**, 1369–1379 (2019).
38. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how

promoters direct initiation and pausing. *Science (80-.).* (2013). doi:10.1126/science.1229386

39. Mahat, D. B. *et al.* Base-pair-resolution genome-wide mapping of active RNA polymerases using precision nuclear run-on (PRO-seq). *Nat. Protoc.* (2016). doi:10.1038/nprot.2016.086

40. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science (80-.).* (2008). doi:10.1126/science.1162228

41. Young, R. S., Kumar, Y., Bickmore, W. A. & Taylor, M. S. Bidirectional transcription marks accessible chromatin and is not specific to enhancers. *Genome Biol.* (2017). doi:10.1186/s13059-017-1379-8

42. Gil, N. & Ulitsky, I. Production of Spliced Long Noncoding RNAs Specifies Regions with Increased Enhancer Activity. *Cell Syst.* (2018). doi:10.1016/j.cels.2018.10.009

43. Sartorelli, V. & Lauberth, S. M. Enhancer RNAs are an important regulatory layer of the epigenome. *Nat. Struct. Mol. Biol.* **27**, 0–7 (2020).

44. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–83 (2011).

45. Sakabe, N., Savic, D. & Nobrega, M. a. Transcriptional enhancers in development and disease. *Genome Biol.* **13**, 238 (2012).

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

47. Catarino, R. R. & Stark, A. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. *Genes Dev.* **32**, 202–223 (2018).

48. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* **21**, 71–87 (2020).

49. Dao, L. T. M. *et al.* Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nat. Genet.* **49**, 1073–1081 (2017).

50. Diao, Y. *et al.* A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods* 1–11 (2017). doi:10.1038/nmeth.4264

51. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–9 (2011).

52. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–47 (2013).

53. Dickel, D. E. *et al.* Genome-wide compendium and functional assessment of in vivo heart enhancers. *Nat. Commun.* **7**, 12923 (2016).

54. Yao, L., Tak, Y. G., Berman, B. P. & Farnham, P. J. Functional annotation of colon cancer risk SNPs. *Nat. Commun.* **5**, 5114 (2014).

55. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–43 (2015).

56. Meyer, C. A. & Liu, X. S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics* **15**, 709–721 (2014).

57. Arner, E. *et al.* Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science (80-.).* **347**, 1010 LP – 1014 (2015).

58. Ostuni, R. *et al.* Latent enhancers activated by stimulation in differentiated cells. *Cell* (2013). doi:10.1016/j.cell.2012.12.018

59. McVicker, G. *et al.* Identification of genetic variants that affect histone modifications in human cells. *Science (80-.).* **342**, 747–749 (2013).

60. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science (80-.).* **342**, 750–752 (2013).

61. Taudt, A., Colomé-Tatché, M. & Johannes, F. Genetic sources of population epigenomic variation. *Nat. Rev. Genet.* **17**, 319–332 (2016).

62. Ernst, J. *et al.* Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. *Nat. Biotechnol.* **34**, 1180–1190 (2016).

63. Perreault, A. A., Benton, M. L., Koury, M. J., Brandt, S. J. & Venters, B. J. Epo reprograms the epigenome of erythroid cells. *Exp. Hematol.* (2017). doi:10.1016/j.exphem.2017.03.004

64. Osmala, M. & Lähdesmäki, H. Enhancer prediction in the human genome by probabilistic modeling of the chromatin feature patterns. *bioRxiv* 804625 (2019). doi:10.1101/804625

65. Jhanwar, S., Ossowski, S. & Davila-Velderrain, J. Genome-wide active enhancer identification using cell type-specific signatures of epigenomic activity. *bioRxiv* 421230 (2018). doi:10.1101/421230

66. Zehnder, T., Benner, P. & Vingron, M. Predicting enhancers in mammalian genomes using supervised hidden Markov models. *BMC Bioinformatics* **20**, 1–12 (2019).

67. Rajagopal, N. *et al.* RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State. *PLoS Comput. Biol.* (2013). doi:10.1371/journal.pcbi.1002968

68. Le, N. Q. K. *et al.* iEnhancer-5Step: Identifying enhancers using hidden information of DNA sequences via Chou's 5-step rule and word embedding. *Anal. Biochem.* (2019). doi:10.1016/j.ab.2019.02.017

69. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41**, 827–841 (2013).

70. Erwin, G. D. *et al.* Integrating Diverse Datasets Improves Developmental Enhancer Prediction. *PLoS Comput. Biol.* **10**, (2014).

71. Kleftogiannis, D., Kalnis, P. & Bajic, V. B. Progress and challenges in bioinformatics approaches for enhancer identification. *Brief. Bioinform.* **17**, 967–979 (2016).

72. Zacher, B. *et al.* Accurate promoter and enhancer identification in 127 ENCODE and roadmap epigenomics cell types and tissues by GenoSTAN. *PLoS One* **12**, 1–25 (2017).

73. Mammana, A. & Chung, H. R. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. *Genome Biol.* (2015). doi:10.1186/s13059-015-0708-z

74. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9**, 215–6 (2012).

75. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through

genomic segmentation. *Nat. Methods* **9**, 473–6 (2012).

76. Libbrecht, M. W. *et al.* A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types. *Genome Biol.* **20**, 1–14 (2019).

77. Reilly, S. K. *et al.* Evolutionary changes in promoter and enhancer activity during human coticogenesis. *Science (80-.).* **347**, 1155–1159 (2015).

78. Hazelett, D. J. *et al.* Comprehensive Functional Annotation of 77 Prostate Cancer Risk Loci. *PLoS Genet.* **10**, e1004102 (2014).

79. Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat. Genet.* **46**, 61–4 (2014).

80. Harismendy, O. *et al.* 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature* **470**, 264–268 (2011).

81. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* **46**, 136–143 (2014).

82. Rhie, S. K. *et al.* Comprehensive Functional Annotation of Seventy-One Breast Cancer Risk Loci. *PLoS One* **8**, (2013).

83. Birnbaum, R. Y. *et al.* Coding exons function as tissue-specific enhancers of nearby genes. *Genome Res.* (2012). doi:10.1101/gr.133546.111

84. Kleftogiannis, D., Kalnis, P. & Bajic, V. B. DEEP: A general computational framework for predicting enhancers. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gku1058

85. Liu, F., Li, H., Ren, C., Bo, X. & Shu, W. PEDLA: Predicting enhancers with a deep learning-based algorithmic framework. *Sci. Rep.* (2016). doi:10.1038/srep28517

86. Min, X. *et al.* Predicting enhancers with deep convolutional neural networks. *BMC Bioinformatics* (2017). doi:10.1186/s12859-017-1878-3

87. Li, Y., Shi, W. & Wasserman, W. W. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. *BMC Bioinformatics* (2018). doi:10.1186/s12859-018-2187-1

88. Nguyen, Q. H. *et al.* IEnhancer-ECNN: Identifying enhancers and their strength using ensembles

of convolutional neural networks. *BMC Genomics* **20**, 1–10 (2019).

89. Bu, H., Hao, J., Gan, Y., Zhou, S. & Guan, J. DEEPSEN: A convolutional neural network based method for super-enhancer prediction. *BMC Bioinformatics* (2019). doi:10.1186/s12859-019-3180-z

90. Chen, L. & Capra, J. A. Learning and interpreting the gene regulatory grammar in a deep learning framework. *bioRxiv Prepr.* (2019). doi:https://doi.org/10.1101/864058

91. Yip, K. Y., Cheng, C. & Gerstein, M. Machine learning and genome annotation: a match meant to be? *Genome Biol.* **14**, 205 (2013).

92. 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).

93. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).

94. Klein, J. C., Keith, A., Agarwal, V., Durham, T. & Shendure, J. Functional characterization of enhancer evolution in the primate lineage. *Genome Biol.* **19**, 1–13 (2018).

95. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–7 (2012).

96. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265–270 (2012).

97. Kheradpour, P. *et al.* Systematic dissection of motif instances using a massively parallel reporter assay. *Genome Res.* **23**, 800–811 (2013).

98. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **24**, gr.173518.114- (2014).

99. Tewhey, R. *et al.* Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell* **165**, 1519–1529 (2016).

100. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27**, 38–52 (2017).

101. Li, Y. *et al.* Genome-wide regulatory model from MPRA data predicts functional regions, eQTLs, and GWAS hits. *bioRxiv* (2017). doi:10.1101/110171

102. van Arensbergen, J. *et al.* High-throughput identification of human SNPs affecting regulatory element activity. *Nat. Genet.* (2019). doi:10.1038/s41588-019-0455-2

103. Wang, X. *et al.* High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun.* (2018). doi:10.1038/s41467-018-07746-1

104. Arnold, C. D. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science (80-.).* **339**, 1074–1077 (2013).

105. Muerdter, F. *et al.* Resolving systematic errors in widely used enhancer activity assays in human cells. *Nat. Methods* (2017). doi:10.1038/nmeth.4534

106. Liu, Y. *et al.* Functional assessment of human enhancer activities using whole-genome STARR-sequencing. *Genome Biol.* (2017). doi:10.1186/s13059-017-1345-5

107. Maricque, B. B., Chaudhari, H. G. & Cohen, B. A. A massively parallel reporter assay dissects the influence of chromatin structure on cis-regulatory activity. *Nat. Biotechnol.* (2019). doi:10.1038/nbt.4285

108. Zabidi, M. A. *et al.* Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature* **518**, 556–559 (2014).

109. Wit, E. De & Laat, W. De. A decade of 3C technologies-insights into nuclear organization. *Genes Dev.* (2012). doi:10.1101/gad.179804.111.GENES

110. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* (2012). doi:10.1038/nature11082

111. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-.).* (2009). doi:10.1126/science.1181369

112. Zhang, J. *et al.* ChIA-PET analysis of transcriptional chromatin interactions. *Methods* (2012). doi:10.1016/j.ymeth.2012.08.009

113. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution

capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).

114. Mumbach, M. R. *et al.* HiChIP: Efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).

115. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* (2015). doi:10.1101/gr.185272.114

116. Dekker, J. & Mirny, L. The 3D Genome as Moderator of Chromosomal Communication. *Cell* (2016). doi:10.1016/j.cell.2016.02.007

117. Dixon, J. R., Gorkin, D. U. & Ren, B. Chromatin Domains: The Unit of Chromosome Organization. *Molecular Cell* (2016). doi:10.1016/j.molcel.2016.05.018

118. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

119. Vietri Rudan, M. *et al.* Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. *Cell Rep.* **10**, 1297–1309 (2015).

120. McArthur, E. & Capra, J. A. Topologically associating domain (TAD) boundaries stable across diverse cell types are evolutionarily constrained and enriched for heritability. *bioRxiv Prepr.* (2020). doi:https://doi.org/10.1101/2020.01.10.901967

121. Lupiáñez, D. G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012–1025 (2015).

122. Won, H. *et al.* Chromosome conformation elucidates regulatory relationships in developing human brain. *Nature* **538**, 523–527 (2016).

123. Norton, H. K. & Phillips-Cremins, J. E. Crossed wires: 3D genome misfolding in human disease. *Journal of Cell Biology* (2017). doi:10.1083/jcb.201611001

124. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the human genome. *Nat.Rev.Genet.* **7**, 85–97 (2018).

125. Kentepozidou, E. *et al.* Clustered CTCF binding is an evolutionary mechanism to maintain topologically associating domains. *Genome Biol.* **21**, 1–38 (2020).

126. Dowen, J. M. *et al.* Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes. *Cell* (2014). doi:10.1016/j.cell.2014.09.030

127. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171**, 305-320.e24 (2017).

128. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* (2016). doi:10.1016/j.celrep.2016.04.085

129. Sun, F. *et al.* Promoter-Enhancer Communication Occurs Primarily within Insulated Neighborhoods. *Mol. Cell* **73**, 250-263.e5 (2019).

130. Jia, Z. *et al.* Tandem CTCF sites function as insulators to balance spatial chromatin contacts and topological enhancer-promoter selection. *Genome Biol.* **21**, 1–24 (2020).

131. Gong, Y. *et al.* Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. *Nat. Commun.* **9**, (2018).

132. Khoury, A. *et al.* Constitutively bound CTCF sites maintain 3D chromatin architecture and long-range epigenetically regulated domains. *Nat. Commun.* **11**, (2020).

133. Norton, H. K. *et al.* Detecting hierarchical genome folding with network modularity. *Nat. Methods* (2018). doi:10.1038/nmeth.4560

134. Matthews, B. J. & Waxman, D. J. Computational prediction of CTCF/ cohesin-based intra-TAD loops that insulate chromatin contacts and gene expression in mouse liver. *Elife* **7**, 1–40 (2018).

135. Nora, E. P. *et al.* Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* (2017). doi:10.1016/j.cell.2017.05.004

136. Schmitt, A. D. *et al.* A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome. *Cell Rep.* **17**, 2042–2059 (2016).

137. Yardımcı, G. G. *et al.* Measuring the reproducibility and quality of Hi-C data. *Genome Biol.* **20**, 1–19 (2019).

138. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods* **72**, 65–75 (2015).

139. Finn, E. H. *et al.* Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization.

Cell **176**, 1502-1515.e10 (2019).

140. Greenwald, W. W. *et al.* Subtle changes in chromatin loop contact propensity are associated with differential gene regulation and expression. *Nat. Commun.* **10**, 1–17 (2019).

141. Schwessinger, R. *et al.* DeepC: Predicting chromatin interactions using megabase scaled deep neural networks and transfer learning. *bioRxiv* 724005 (2019). doi:10.1101/724005

142. Fudenberg, G., Kelley, D. R. & Pollard, K. S. Predicting 3D genome folding from DNA sequence. *bioRxiv Prepr.* (2019). doi:https://doi.org/10.1101/800060

143. Beagan, J. A. & Phillips-Cremins, J. E. On the existence and functionality of topologically associating domains. *Nat. Genet.* **52**, 8–16 (2020).

144. Cubenas-Potts, C. & Corces, V. G. Topologically Associating Domains : An invariant framework or a dynamic scaffold? 430–434 (2015).

145. Hansen, A. S., Pustova, I., Cattoglio, C., Tjian, R. & Darzacq, X. CTCF and cohesin regulate chromatin loop stability with distinct dynamics. *Elife* (2017). doi:10.7554/eLife.25776

146. Despang, A. *et al.* Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nat. Genet.* (2019). doi:10.1038/s41588-019-0466-z

147. Osterwalder, M. *et al.* Enhancer redundancy provides phenotypic robustness in mammalian development. *Nature* (2018). doi:10.1038/nature25461

148. Shin, H. Y. *et al.* Hierarchy within the mammary STAT5-driven Wap super-enhancer. *Nat. Genet.* **48**, 904–911 (2016).

149. Dukler, N., Gulko, B., Huang, Y.-F. & Siepel, A. Is a super-enhancer greater than the sum of its parts? *Nat. Genet.* **49**, 2–3 (2016).

150. Cannavò, E. *et al.* Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Curr. Biol.* **26**, 38–51 (2016).

151. Hobert, O. Gene regulation: Enhancers stepping out of the shadow. *Curr. Biol.* (2010). doi:10.1016/j.cub.2010.07.035

152. Hong, J. W., Hendrix, D. A. & Levine, M. S. Shadow enhancers as a source of evolutionary

novelty. *Science* (2008). doi:10.1126/science.1160631

153. Will, A. J. *et al.* Composition and dosage of a multipartite enhancer cluster control developmental expression of Ihh (Indian hedgehog). *Nat. Genet.* **49**, 1539–1545 (2017).

154. Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T. & Flicek, P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat. Ecol. Evol.* **2**, 152–163 (2018).

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

156. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* (2012). doi:10.1038/nature11279

157. Li, G. *et al.* Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* (2012). doi:10.1016/j.cell.2011.12.014

158. Jin, F. *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* (2013). doi:10.1038/nature12644

159. Lettice, L. A. *et al.* Development of five digits is controlled by a bipartite long-range cis-regulator. *Dev.* (2014). doi:10.1242/dev.095430

160. Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T. & Flicek, P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat. Ecol. Evol.* **2**, 152–163 (2018).

161. Yao, L., Berman, B. P. & Farnham, P. J. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. *Cox Crit Rev Biochem Mol Biol* **50**, 1549–7798 (2015).

162. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

163. Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* (2013). doi:10.1016/j.cell.2012.12.034

164. Wang, X. & Goldstein, D. B. Enhancer Domains Predict Gene Pathogenicity and Inform Gene

Discovery in Complex Disease. *Am. J. Hum. Genet.* **106**, 215–233 (2020).

165. Wu, Z., Ioannidis, N. M. & Zou, J. Predicting target genes of noncoding regulatory variants with ICE. *Bioinformatics* (2020). doi:10.1093/bioinformatics/btaa254

166. Schoenfelder, S. & Fraser, P. Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.* (2019). doi:10.1038/s41576-019-0128-0

167. Mumbach, M. R. *et al.* Enhancer connectome in primary human cells identifies target genes of disease-associated DNA elements. *Nat. Genet.* **49**, 1602–1612 (2017).

168. Fulco, C. P. *et al.* Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. *Nature Genetics* (2019). doi:10.1038/s41588-019-0538-0

169. Cao, Q. *et al.* Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines. *Nat. Genet.* **49**, 1428–1436 (2017).

170. Okonechnikov, K., Erkek, S., Korbel, J. O., Pfister, S. M. & Chavez, L. InTAD: Chromosome conformation guided analysis of enhancer target genes. *BMC Bioinformatics* **20**, 1–7 (2019).

171. Whalen, S., Truty, R. M. & Pollard, K. S. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* **48**, 488–96 (2016).

172. Gao, T. & Qian, J. Eagle: An algorithm that utilizes a small number of genomic features to predict tissue/ cell type-specific enhancer-gene interactions. *PLoS Comput. Biol.* **15**, (2019).

173. Belokopytova, P. S., Nuriddinov, M. A., Mozheiko, E. A., Fishman, D. & Fishman, V. Quantitative prediction of enhancer–promoter interactions. *Genome Res.* (2020). doi:10.1101/gr.249367.119

174. Yang, Y., Zhang, R., Singh, S. & Ma, J. Exploiting sequence-based features for predicting enhancer-promoter interactions. in *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx257

175. Zeng, W., Wu, M. & Jiang, R. Prediction of enhancer-promoter interactions via natural language processing. *BMC Genomics* (2018). doi:10.1186/s12864-018-4459-6

176. Hait, T. A., Elkon, R. & Shamir, R. CT-FOCS: a novel method for inferring cell type-specific enhancer-promoter maps. *bioRxiv* 707158 (2019). doi:10.1101/707158

177. Cao, F. & Fullwood, M. J. Inflated performance measures in enhancer–promoter interaction-prediction methods. *Nature Genetics* (2019). doi:10.1038/s41588-019-0434-7

178. Xi, W. & Beer, M. A. Local epigenomic state cannot discriminate interacting and non-interacting enhancer–promoter pairs with high accuracy. *PLoS Comput. Biol.* (2018). doi:10.1371/journal.pcbi.1006625

179. Moore, J. E., Pratt, H. E., Purcaro, M. J. & Weng, Z. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biol.* **21**, 1–16 (2020).

180. Scholes, C., Biette, K. M., Harden, T. T. & DePace, A. H. Signal Integration by Shadow Enhancers and Enhancer Duplications Varies across the Drosophila Embryo. *Cell Rep.* **26**, 2407-2418.e5 (2019).

181. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Mol. Cell* **66**, 285-299.e5 (2017).

182. Moorthy, S. D. *et al.* Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome Res* **27**, 246–258 (2017).

183. Huang, J. *et al.* Dissecting super-enhancer hierarchy based on chromatin interactions. *Nat. Commun.* **9**, (2018).

184. Delaneau, O. *et al.* Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science (80-.).* **364**, (2019).

185. Xu, D., Gokcumen, O. & Khurana, E. Loss-of-function tolerance of enhancers in the human genome. *PLoS Genet.* **16**, e1008663 (2020).

186. Sholtis, S. J. & Noonan, J. P. Gene regulation and the origins of human biological uniqueness. *Trends Genet.* **26**, 110–118 (2010).

187. Reilly, S. K. & Noonan, J. P. Evolution of Gene Regulation in Humans. *Annu. Rev. Genom. Hum. Genet* 1–23 (2016). doi:10.1146/annurev-genom-090314-045935

188. Mack, K. L. & Nachman, M. W. Gene Regulation and Speciation. *Trends Genet.* **33**, 68–80 (2016).

189. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science (80-.).* **337**, 1190–1195 (2012).

190. Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.* **94**, 559–573 (2014).

191. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

192. Benton, M. L., Talipineni, S. C., Kostka, D. & Capra, J. A. Genome-wide enhancer annotations differ significantly in genomic distribution, evolution, and function. *BMC Genomics* **20**, 511 (2019).

193. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).

194. Guturu, H., Chinchali, S., Clarke, S. L. & Bejerano, G. Erosion of Conserved Binding Sites in Personal Genomes Points to Medical Histories. *PLoS Comput. Biol.* **12**, 1–19 (2016).

195. Corradin, O. & Scacheri, P. C. Enhancer variants: evaluating functions in common disease. *Genome Med.* **6**, 85 (2014).

196. Mansour, M. R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science (80-.).* **346**, 1373–1377 (2014).

197. Corradin, O. *et al.* Modeling disease risk through analysis of physical interactions between genetic variants within chromatin regulatory circuitry. *Nat. Genet.* **48**, 1313–1320 (2016).

198. Chiang, C. *et al.* The impact of structural variation on human gene expression. *Nat. Genet.* **49**, 692–699 (2017).

199. Short, P. J. *et al.* De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018).

200. Bhagwat, A. S., Lu, B. & Vakoc, C. R. Enhancer dysfunction in leukemia. *Blood* **131**, 1795–1804

(2018).

201. Krijger, P. H. L. & de Laat, W. Regulation of disease-associated gene expression in the 3D genome. *Nat. Rev. Mol. Cell Biol.* **17**, 771–782 (2016).

202. Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).

203. Lettice, L. A. *et al.* Enhancer-adoption as a mechanism of human developmental disease. *Hum. Mutat.* **32**, 1492–1499 (2011).

204. Fudenberg, G. & Pollard, K. S. Chromatin features constrain structural variation across evolutionary timescales. *Proc. Natl. Acad. Sci. U. S. A.* (2019). doi:10.1073/pnas.1808631116

205. Fiddes, I. T. *et al.* Human-Specific NOTCH2NL Genes Affect Notch Signaling and Cortical Neurogenesis. *Cell* **173**, 1356-1369.e22 (2018).

206. Coe, B. P. *et al.* Refining analyses of copy number variation identifies specific genes associated with developmental delay. *Nat. Genet.* **46**, 1063–1071 (2014).

207. Turner, T. N. *et al.* Genome Sequencing of Autism-Affected Families Reveals Disruption of Putative Noncoding Regulatory DNA. *Am. J. Hum. Genet.* **98**, 58–74 (2016).

208. Antonacci, F. *et al.* A large and complex structural polymorphism at 16p12.1 underlies microdeletion disease risk. *Nat. Genet.* **42**, 745–751 (2010).

209. Karnuta, J. M. & Scacheri, P. C. Enhancers: Bridging the gap between gene control and human disease. *Hum. Mol. Genet.* **0**, 1–9 (2018).

210. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

211. Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15**, 480 (2014).

212. Wells, A. *et al.* Ranking of non-coding pathogenic variants and putative essential regions of the human genome. *Nat. Commun.* **10**, (2019).

213. Zhang, S. *et al.* regBase: whole genome base-wise aggregation and functional prediction for

human non-coding regulatory variants. *Nucleic Acids Res.* **47**, e134 (2019).

214. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat. Genet.* **49**, 618–624 (2017).

215. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1016

216. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* (2015). doi:10.1038/nmeth.3547

217. Li, M. J. *et al.* Cepip: Context-Dependent Epigenomic Weighting for Prioritization of Regulatory Variants and Disease-Associated Genes. *Genome Biol.* **18**, 52 (2017).

218. Backenroth, D. *et al.* FUN-LDA: A Latent Dirichlet Allocation Model for Predicting Tissue-Specific Functional Effects of Noncoding Variation: Methods and Applications. *Am. J. Hum. Genet.* **102**, 920–942 (2018).

219. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).

220. Dong, S. & Boyle, A. P. Predicting functional variants in enhancer and promoter elements using RegulomeDB. *Hum. Mutat.* **40**, 1292–1298 (2019).

221. Nott, A. *et al.* Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science (80-.).* **366**, 1134–1139 (2019).

222. Zhou, J. *et al.* Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk. *Nat. Genet.* **51**, 973–980 (2019).

223. Kircher, M. *et al.* Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.* **10**, 1–15 (2019).

224. Kvon, E. Z. *et al.* Comprehensive In Vivo Interrogation Reveals Phenotypic Impact of Human Enhancer Variants. *Cell* **180**, 1262-1271.e15 (2020).

225. Zhang, G. *et al.* DiseaseEnhancer: A resource of human disease-associated enhancer catalog.

Nucleic Acids Res. **46**, D78–D84 (2018).

226. Kundaje, A. A comprehensive collection of signal artifact blacklist regions in the human genome. *… Site/Anshulkundaje/Projects/Blacklists (Last Accessed 30 …* (2013).

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

228. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, (2008).

229. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).

230. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, (2014).

231. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–5 (2013).

232. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

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

234. Wickham, H. *ggplot2*. *Elegant Graphics for Data Analysis* (2009). doi:10.1007/978-0-387-98141-3

235. Heger, A., Webber, C., Goodson, M., Ponting, C. P. & Lunter, G. GAT: A simulation framework for testing the association of genomic intervals. *Bioinformatics* **29**, 2046–2048 (2013).

236. Yu, G. *et al.* GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010).

237. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C.-F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**, 1274–81 (2007).

238. Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: A more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res.* **45**,

W130–W137 (2017).

239. McLeay, R. C. & Bailey, T. L. Motif Enrichment Analysis: A unified framework and an evaluation on ChIP data. *BMC Bioinformatics* (2010). doi:10.1186/1471-2105-11-165

240. Kulakovskiy, I. V. *et al.* HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx1106

241. R Core Team. R Core Team (2015). R: A language and environment for statistical computing. *R Found. Stat. Comput. Vienna, Austria. URL http//www.R-project.org/.* R Foundation for Statistical Computing (2015).

242. Galili, T. dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31**, 3718–3720 (2015).

243. Zhernakova, D. V *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).

244. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46**, 1160–5 (2014).

245. Parker, S. C. J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 17921–17926 (2013).

246. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).

247. Onengut-Gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–6 (2015).

248. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* (2010). doi:10.1038/nature09266

249. Almontashiri, N. A. M. *et al.* 9p21.3 coronary artery disease risk variants disrupt TEAD transcription factor-dependent transforming growth factor β regulation of p16 expression in human aortic smooth muscle cells. *Circulation* (2015). doi:10.1161/CIRCULATIONAHA.114.015023

250. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science (80-.).* (2008). doi:10.1126/science.1155174

251. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).

252. Werling, D. M. *et al.* An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat. Genet.* (2018). doi:10.1038/s41588-018-0107-y

253. Han, L. *et al.* Functional annotation of rare structural variation in the human brain. *Nat. Commun.* **11**, 2990 (2020).

254. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* (2013). doi:10.1093/bioinformatics/bts635

255. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btt656

256. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science (80-.).* **362**, (2018).

257. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* (2020). doi:10.1038/s41586-020-2287-8

258. Ruderfer, D. M. *et al.* Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat. Genet.* (2016). doi:10.1038/ng.3638

259. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* (2020). doi:10.1038/s41586-020-2308-7

260. Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* (2014). doi:10.1038/nature12818

261. Ay, F., Bailey, T. L. & Noble, W. S. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Res.* (2014). doi:10.1101/gr.160374.113

262. Wang, Y. *et al.* The 3D Genome Browser: A web-based browser for visualizing 3D genome

organization and long-range chromatin interactions. *Genome Biol.* (2018). doi:10.1186/s13059-018-1519-9

263. Csurös, M. Count: Evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics* (2010). doi:10.1093/bioinformatics/btq315

264. Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* (2005). doi:10.1093/bioinformatics/bti042

265. Ravasi, T. *et al.* An Atlas of Combinatorial Transcriptional Regulation in Mouse and Man. *Cell* (2010). doi:10.1016/j.cell.2010.01.044

266. Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinform.* (2017). doi:10.1093/bib/bbw008

267. Eisenberg, E. & Levanon, E. Y. Human housekeeping genes, revisited. *Trends in Genetics* (2013). doi:10.1016/j.tig.2013.05.010

268. Wang, T. *et al.* Identification and characterization of essential genes in the human genome. *Science (80-.).* (2015). doi:10.1126/science.aac7041

269. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* (2011). doi:10.1093/bioinformatics/btr064

270. King, D. M. *et al.* Synthetic and genomic regulatory elements reveal aspects of Cis-regulatory grammar in mouse embryonic stem cells. *Elife* (2020). doi:10.7554/eLife.41279

271. Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet* **46**, 61–64 (2014).

272. Tak, Y. G. & Farnham, P. J. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics Chromatin* **8**, 57 (2015).

273. Chatterjee, S. & Ahituv, N. Gene Regulatory Elements , Major Drivers of Human Disease. *Annu. Rev. Genom. Hum. Genet* 1–19 (2017). doi:10.1146/annurev-genom-091416-035537

274. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science (80-.).* **342**,

750–752 (2013).

275. Zhou, J. & Troyanskaya, O. G. Probabilistic modelling of chromatin code landscape reveals functional diversity of enhancer-like chromatin states. *Nat Commun* **7**, 1–9 (2016).

276. Kim, T. K. & Shiekhattar, R. Architectural and Functional Commonalities between Enhancers and Promoters. *Cell* **162**, 948–959 (2015).

277. Andersson, R. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays* **37**, 314–323 (2015).