Genome-wide Enhancer Maps Differ Significantly in their Genomic Distribution, Evolution, and Function

By

Mary Lauren Benton

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

MASTER OF SCIENCE
in

Biomedical Informatics
August, 31, 2018
Nashville, Tennessee

Approved:

John A. Capra, Ph.D.
Emily Hodges, Ph.D.
Jacob J. Hughey, Ph.D.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction

Motivation

Enhancers are traditionally defined as genomic sequences that regulate the transcription of one or more genes, regardless of orientation or relative distance to the target promoter[1]. These *cis*-regulatory regions can bind specific transcription factors and cofactors to increase transcription, and in current models of enhancer function, they physically interact with their long-range targets via loops in the three-dimensional chromatin structure. Enhancers play a vital role in the regulation of genes during development and cell differentiation[2,3]. Genetic variation in enhancers has been implicated in etiology of complex disease[4,5] and in differences between closely related species[6–8].

Given their significant role in a range of biological processes, enhancers have seen considerable study in recent years. More than 2,300 papers have been published on enhancer biology (MeSH: *Enhancer Elements, Genetic*) since the start of 2015; hundreds of these have focused on the role of enhancers in disease. However, despite the importance of enhancers, they remain difficult to identify[1,9,10]. Experimental assays that directly confirm enhancer activity are time-consuming, expensive, and not always conclusive[1,11]. Although there are recent promising developments in massively parallel reporter assays, current methods are unable to definitively identify and test enhancers on an unbiased genome-wide scale[12]. As a result, many studies use more easily measurable attributes associated with enhancer activity, defining enhancers based on a single set of biochemical and sequence-level proxies for activity.

An evaluation of the robustness of this ‘single definition’ approach through a comprehensive analysis of similarity in the genomic, evolutionary, and functional attributes of enhancers identified by different strategies is essential. Through this comparison, we can assess the stability of conclusions made using only one enhancer identification strategy. While we expect some variation due to differences in the

underlying assays, we found significant differences between enhancer sets identified in the same context which were sufficient to influence downstream conclusions and biological interpretations.

Attributes of Gene Regulatory Enhancers

Enhancer regions are known to be associated with a variety of biochemical, sequence, and functional attributes. Combinations of enhancer-correlated attributes are often used as the enhancer definitions themselves or leveraged as input into more complex computational models.

Biochemical signatures of enhancers

Since they operate through binding to relevant transcription factors, active enhancers localize in regions of open, or accessible chromatin. This is commonly assayed by testing the sensitivity of DNA segments to DNase I nuclease followed by sequencing (DNase-seq)[13,14], identifying nucleosome depleted regions with Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq)[15,16], or more recently with the transposase Tn5 mediated Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq)[17]. Each of these methods leverages the variable susceptibility of nucleosome-occupied versus nucleosome-free DNA to nucleases, formaldehyde cross-linking, or transposase activity, respectively. Despite the differences in DNA-targeting approaches, DNase-seq, FAIRE-seq, and ATAC-seq generate similar output data. The final step in each method is to align the sequencing reads back to a reference genome to generate enrichment 'peaks' thought to be accessible for transcription factor binding[13–17].

Enhancers also often have characteristic sets of histone modifications on surrounding nucleosomes that can be detected using chromatin immunoprecipitation assays followed by next generation sequencing (ChIP-seq)[18]. DNA-bound proteins are cross-linked to the DNA, the genome is fragmented into small pieces, and specific antibodies are used to precipitate out relevant fragments for sequencing. The resulting reads are aligned to the genome, forming peaks where the mark was reliably present. Commonly assessed markers for enhancer identification are monomethylation of lysine 4 on

histone H3 (H3K4me1) and acetylation of lysine 27 on histone H3 (H3K27ac), which often denotes active enhancers[1,19,20]. Combining H3K27ac with the lack of a modification characteristic of promoters, trimethylation of lysine 4 on histone H3 (H3K4me3), is often used to distinguish enhancers from promoters[1,21].

Sequence features of enhancers

Genomic sequence-level features are also often used to locate or characterize potential enhancer regions. Many enhancers, especially those involved in developmental processes, are evolutionarily conserved[22–25]. As a result, evolutionarily conserved sequences are frequently prioritized when identifying putative regulatory elements. Perhaps more directly, the presence of known transcription factor binding motifs[26] or binding of known enhancer associated proteins, such as the histone acetyltransferase p300[19,27,28], have been used to successfully locate enhancer elements[1]. Transcription factor (TF) binding can be elucidated with ChIP-seq assays, resulting in a genome-wide map of locations bound to the factor of interest in the cell at the time the assay was performed[18]. However, if multiple relevant factors exist, this approach can be expensive and time-consuming. Instead, computational methods have been developed for the *de novo* identification of TF binding motifs, or calculating enrichment for computationally derived motifs and known motifs of interest[1,29,30].

Notably, while evolutionary conservation and the presence of relevant TF binding sites are suggestive of enhancer function, they are not a guarantee of activity. Successful binding of TFs is often context-dependent, and the presence of a motif does not require binding[29]. A recent study integrating TF binding motifs with ChIP-seq data in K562 cells reported an average of 430 times the number of unbound motifs as bound motifs, suggesting that actual binding may be rare in comparison with the number of TF motifs[31]. Even positive evidence of binding from a ChIP-seq assay does not confirm enhancer activity. There is a high false-positive rate for enhancer prediction from TF ChIP-seq, possibly because enhancers require specific combinatorial binding patterns[29] or because it is difficult to distinguish transient interactions from critical ones with ChIP-seq data alone[32]. Alternatively, a lack of conserved binding

motifs does not indicate lack of activity; previous studies in mouse embryos have demonstrated that turnover in transcription factor binding sites is quite high, showing that there can be strong conservation of enhancer activity with little to no sequence similarity across evolutionary time[28].

Functional characteristics of enhancers

More recently, other functional attributes related to transcriptional activity and three-dimensional chromatin structure have been used both to identify potential enhancer regions and to predict the activity and target genes of those regions. Some enhancers are transcribed, and it has become possible to map active enhancers by identifying characteristic bi-directionally transcribed enhancer RNAs (eRNAs)[33,34]. The cap analysis of gene expression followed by sequencing (CAGE-seq) has been used to quantify and map eRNAs in a variety of biological contexts, notably by the FANTOM consortium, which generated eRNA measurements across a wide range of tissues and cell lines[33]. These predicted enhancers validate at a relatively high rate (~70%). It is important to note that the function of these noncoding eRNAs is still not well understood. A recent review proposed three potential classes for eRNAs,: (1) eRNAs with no biological function, (2) eRNAs contributing to the activity of the enhancer itself, and (3) eRNAs with independent functionality or effect[34]. However, confirming these classifications will require additional experimental characterization. There is also some question about the specificity of the bi-directional transcription pattern to enhancers[35].

Understanding the three-dimensional structure of chromatin, and its interaction points, is another method used to functionally characterize putative enhancer regions. Chromatin conformation assays are able to generate long-range interaction maps of the genome, suggesting locations that interact with promotors or form three-dimensional compartments that localize regulatory activity[36,37]. Chromatin conformation capture (3C), circular chromatin conformation capture (4C), chromosome conformation capture carbon copy (5C), and Hi-C all generate these maps through sequencing interacting DNA segments after formaldehyde cross-linking[1,36]. A variant of the chromatin conformation capture approaches, chromatin interaction analysis with paired-end tag sequencing (ChIA-PET), uses a similar

methodology to discover interacting regions; however, ChIA-PET has the benefit of allowing the user to pull down interactions involving specific proteins, similar to ChIP-seq[36]. The resulting ChIA-PET maps are used to draw inferences about regions frequently associated with long-range interactions and proteins of interest. Interacting DNA regions can be further combined with alternate biochemical and sequence-level attributes to suggest potential regulatory activity, validate previously predicted enhancer regions, or locate potential gene targets[1,36,38]. Due to the generally low resolution and a decreased ability to detect short (< 10 kb) interactions, these methods require further refinement before they could be applied to reliably identify novel enhancers and their target promoters[1,36].

Limitations of using attributes correlated with enhancer activity to identify enhancers

While informative, none of these attributes are comprehensive, exclusive to enhancers, or completely reliable indicators of enhancer activity. For example, enhancers defined using eRNA based methods are often more restricted sets than those suggested by alternate approaches. They do not capture all sequences that have demonstrated the ability to drive transcription in small-scale transgenic assays, suggesting that they do not provide a complete picture of the regulatory landscape in a given cellular context[33]. Alternately, many of the biochemical and sequence level features account for increased levels of genomic sequence, but these are not exclusive to enhancer regions[39–41]. Previous work describes a potential spectrum of functional genetic elements, including promoters and insulators, that share attributes with enhancers[42,43]. Indeed, a recent study by Dao *et al.* suggests that some mammalian promoters also demonstrate enhancer behavior in certain contexts[44]. Selecting enhancer-associated combinations of histone modifications is another a notable example of this complexity[1,2,39,40]. H3K27ac is known to mark active promoters as well as enhancers, despite being used as a working definition of an enhancer[21,45–49]. The commonly used enhancer-mark, H3K4me1, has been discovered in regions that do not demonstrate enhancer activity[2,39]. The repressive H3K27me3 mark has also been reported to coincide with H3K4me1 marked enhancer regions, referred to as bivalent or poised enhancers, suggesting a mechanism for finer control of enhancer activation[2]. Other recent papers claim additional histone modifications are also

correlated with enhancer states (H3K64ac, H3K122ac, H3K79me3, and H4K16ac), and can mark active enhancer regions lacking the traditional H3K27ac mark, further complicating the idea of a single 'histone code' for enhancer discovery[40,50,51]. Earlier validations of enhancer predictions also suggest that definitions of enhancers based solely on combinations of histone modifications have a low specificity, leading to low validation rates (20-33%)[26,33]. Technical constraints and biases in the experimental assays generating data used to identify enhancers, either alone or in combination with further statistical approaches, are also not fully appreciated[52].

Aside from limitations assaying and using data associated with enhancer activity, biological characteristics of the regulatory architecture add an additional layer of complexity. Enhancer activity is also context- and stimulus-dependent, suggesting that the lack of enhancer activity in a single context or condition may not be sufficient to completely rule out activity[20,53]. There are a number of case studies to explore the enhancer regions with activity only in a specific tissue or developmental time point[20,53]. Other research also suggests that enhancers may occupy multiple states. Sometimes classified as 'poised', 'primed', or 'latent', enhancers may be subject to intermediate states between activity and inactivity, complicating a simpler switch-like model of enhancer activity[1,2,41,54]. Genetic variation between individuals is also known to affect epigenetic modifications and enhancer activity, potentially confounding the generalizability of identification approaches built on epigenetic features. However, recent studies suggest that only a small fraction (1–15%) of epigenetic modifications are influenced by nearby genetic variants[55].

Computational Models for Enhancer Identification

Many complementary computational enhancer identification methods that integrate data correlated with enhancer activity (e.g. histone modification profiles, chromatin accessibility, TF binding, eRNA) in both supervised and unsupervised machine learning approaches have been developed[10,56,57]. These typically

approach the problem by using unsupervised methods to segment the genome into broader functional categories or using supervised learning to classify regions into 'enhancer' and 'non-enhancer' states.

Popular unsupervised chromatin segmentation approaches include ChromHMM, Segway, and the more recent GenoSTAN[58–60]. These methods integrate knowledge of histone modifications ChIP-seq with hidden Markov models or dynamic Bayesian networks, to produce a map of the genome labeled by a user-specified number of states. The levels of enrichment for different combinations of markers can then be assessed by an expert to assign a plausible annotation to each state[56]. Unsupervised approaches are useful since they do not require a positive and negative training set, which, given the small number of experimentally validated enhancers, can be difficult to define and introduce unknown biases into the model[60].

Supervised classification approaches used to distinguish enhancers from other functional regions and genomic background range in complexity. Some of the most widely used approaches involve simple rule-based intersections of combinations histone modifications and other genomic annotations[2,19,66,20,21,39,61–65]. For example, the co-occurrence of H3K4me1 and H3K27ac in the same genomic region, or the presence of H3K27ac without any H3K4me3 signal are often cited as evidence of an enhancer regions[19–21,39,64,65]. Since enhancers are frequently defined as non-protein-coding regions that regulate target genes from a long distance, these simple intersections can also be filtered to exclude coding sequences and regions within a certain distance from the nearest transcription start site (TSS)[61,66–69]. In some cases, however, this filtering may be too stringent since enhancer sequences have been reported in coding sequences and intronic regions close to genes[1,70]. In recent years, the application of supervised machine learning classifiers for enhancer prediction has grown in popularity. These models are trained on similar combinations of input as the unsupervised or simple approaches: histone modifications, regions of open chromatin, transcription factor binding motifs, and other genomic annotations[10,57,67,68]. They use statistical principles to learn higher order patterns and classify the sequence or region by enhancer status.

Limitations of computational enhancer predictions

Despite current and widespread use, an accurate quantification of the model performance is limited by the available validation metrics and the lack of a comprehensive gold standard. Without a gold standard enhancer set, enhancer identification studies and algorithms validate their results via a combination of small-scale transgenic reporter gene assays and enrichment for other functional attributes, such as trait-associated genetic variants, evolutionary conservation, or proximity to relevant genes. This form of validation does provide additional evidence of enhancer function. However, considering that attributes correlated with enhancer activity also have an error rate, the validation can become biased. Furthermore, complex statistical models like many of those employed in an enhancer prediction framework can be difficult to interpret. The 'black box' nature of machine learning makes it difficult to discern generalizable biological and mechanistic insights into the regulatory architecture, even from well-performing models[71].

Experimental Approaches for Enhancer Identification and Validation

Until recently, experimental identification of enhancer sequences was limited to small-scale studies in cell lines or transgenic embryos. In this approach, the sequence of interest is incorporated into a bacterial plasmid upstream of a minimal promoter and reporter gene[1,72]. Sequences with activity sufficient to drive expression of the reporter construct are labeled as enhancers. While informative, transgenic reporter assays are low-throughput and require all sequences to be specified in advance[1]. With the advent of more sophisticated high-throughput experimental approaches, focus turned to the development of genome-scale enhancer identification protocols.

One such technology, massively parallel reporter assays (MPRAs), shows promise for generating large-scale enhancer maps with demonstrated activity[12]. In an MPRA, unique barcodes are incorporated into reporter constructs with the putative enhancer sequences being tested. Libraries with thousands of

barcode labeled plasmids can be tested in a single experiment, allowing for validation of a large number of sequences at one time. MPRAs have been used extensively in recent years, both to test the regulatory potential of a large number of candidate regions and to assess the impact of variation on putative enhancer elements[73–77]. Despite a relatively low validation rate for computationally predicted enhancer sequences (~26%), MPRA results do confirm many long-held beliefs about enhancers[76]. They find enrichment for active regulatory elements in DHSs and evolutionarily conserved regions, and show active enhancers co-occurring with traditional histone modifications and relevant TF binding motifs[12,43,74]. A variation on the original MPRA, self-transcribing active regulatory region sequencing (STARR-seq) has also recently been applied to human enhancer sequences[78,79]. STARR-seq differs from other MPRAs in that the predicted enhancer sequence is placed downstream of the minimal promoter and is thus transcribed if the enhancer is active. This allows for direct quantification of enhancer strength and activity level without secondary barcodes[78].

Limitations of experimental approaches

While significant progress has been made in recent years, experimental enhancer identification and validation retain a number of caveats that prevent the adoption of a single gold-standard enhancer map. The recent focus on MPRAs for identifying enhancers on a genome-wide scale overcomes some of the issues of previous experimental methods, primarily in terms of throughput. However, the sequences derived from this method still represent only a subset of active enhancers in a given cell type[12]. Enhancers are known to be cell-type and stimulus-dependent, which cannot be fully accounted for using transgenic assays or MPRAs as a validation approach. Additionally, MPRAs suffer from several of the same caveats as traditional reporter constructs, including sequence length restrictions and removal from the enhancer's endogenous context[11,12]. The latter has been explored as a source of potential bias by altering an enhancer's ability to drive gene expression[11]. A study by Inoue *et al.* compared the results of a traditional episomal MPRA with that of a novel lentivirus MPRA (lenti-MPRA) that integrates the putative enhancer sequence into the genome[11]. While the two approaches were highly correlated, the authors concluded that

the integrated MPRA resulted in increased reproducibility and positive signals that were more highly correlated with relevant genomic annotations. Poorer performance of the traditional MPRA suggests that noise and a lack of biological context in episomal assays may obscure relevant signals, and that it is important to consider the endogenous context when testing enhancer sequences[11]. Furthermore, a recent study by Muerdter *et al*. claimed that transcription often initiates in the origin-of-replication in plasmid-based reporter constructs, including both small-scale luciferase assays and high-throughput methods like STARR-seq, confounding readouts of enhancer activity. They also describe unintended inflammatory responses induced by plasmid transfection which can also obscure true signals and confound interpretability of such assays[79]. These caveats must be understood and accounted for in order to obtain reliable identification and validation of enhancer sequences. Currently, the use of clustered regularly interspaced short palindromic repeats (CRISPR)-Cas9 is being explored as a way to validate enhancer sequences and assess the impact of genetic variation on regulatory elements in an endogenous context[43].

Chapters

Genome-wide enhancer maps are commonly used in many different applications, including studies of the gene regulatory architecture of different tissues, and the interpretation of variants identified in genome-wide association studies (GWAS)[4,5,45]. However, in these applications, a single assay or computationally predicted enhancer set regularly serves as the working definition of an "enhancer" for all analyses and individuals. We hypothesized that differences between identification strategies are substantial enough to influence biological interpretations and conclusions about enhancer evolution and disease-associated variant function. Chapter II provides a quantification of the genomic differences between nine diverse enhancer identification strategies across four biological contexts. This highlights the level of disagreement between enhancer definitions in common use. Chapter III characterizes the functional implications of the similarities and differences between putative enhancer sets through standardized analyses of each set. Understanding the way identification strategies impact the functional associations of

each enhancer set is vital to both scientific reproducibility and furthering our understanding of the regulatory architecture. Chapter IV contrasts the performance of each identification strategy on experimentally validated enhancer sequences and assesses the extent to which focusing on enhancers supported by multiple identification methods can resolve the disagreement between individual methods. In this work, we highlight a fundamental challenge to studying gene regulatory mechanisms, and to evaluating the functional relevance of thousands of non-coding variants associated with traits, for instance from GWAS. Understanding and incorporating the unique characteristics of different enhancer identification strategies will be essential to ensuring reproducible results, and to furthering our understanding of enhancer biology.

CHAPTER II

Quantifying Genomic Differences Between Enhancer Sets

Introduction

The successful identification of gene regulatory enhancers remains a complex problem, due to the large number of activity-correlated attributes and the lack of a gold-standard enhancer map. As a result, many enhancer identification and validation approaches exist. In practice, most studies consider a single identification approach based on a select number of enhancer characteristics and base all downstream analyses on that definition. Given the wide range of identification strategies, we hypothesized that enhancer identification strategies would display significant genome differences. This chapter evaluates the variation in enhancer sets annotated by different strategies using a consistent computational pipeline to compare enhancer sets genome-wide. We consider a representative selection of enhancer sets, contrasting genomic characteristics such as location, length, distance from transcription start sites (TSSs), and level of evolutionary conservation. We then quantify the amount of overlap and genomic similarity between all pairs of enhancer sets.

Methods

Defining a panel of enhancer identification strategies

Our approach is based on publicly available data applied 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[1,10], it is not possible to compare them all; so for each biological context, we consider methods that represent the diversity of experimental and computational approaches in common use. We define "common use" as methods cited as the enhancer identification or definition strategy in at least one high-profile published paper after 2007,

although most strategies have more frequent publications. All enhancer sets were generated and analyses were performed in the context of the GRCh37/hg19 build of the human genome. We used transcription start site (TSS) definitions from Ensembl v75 (GRCh37.p13).

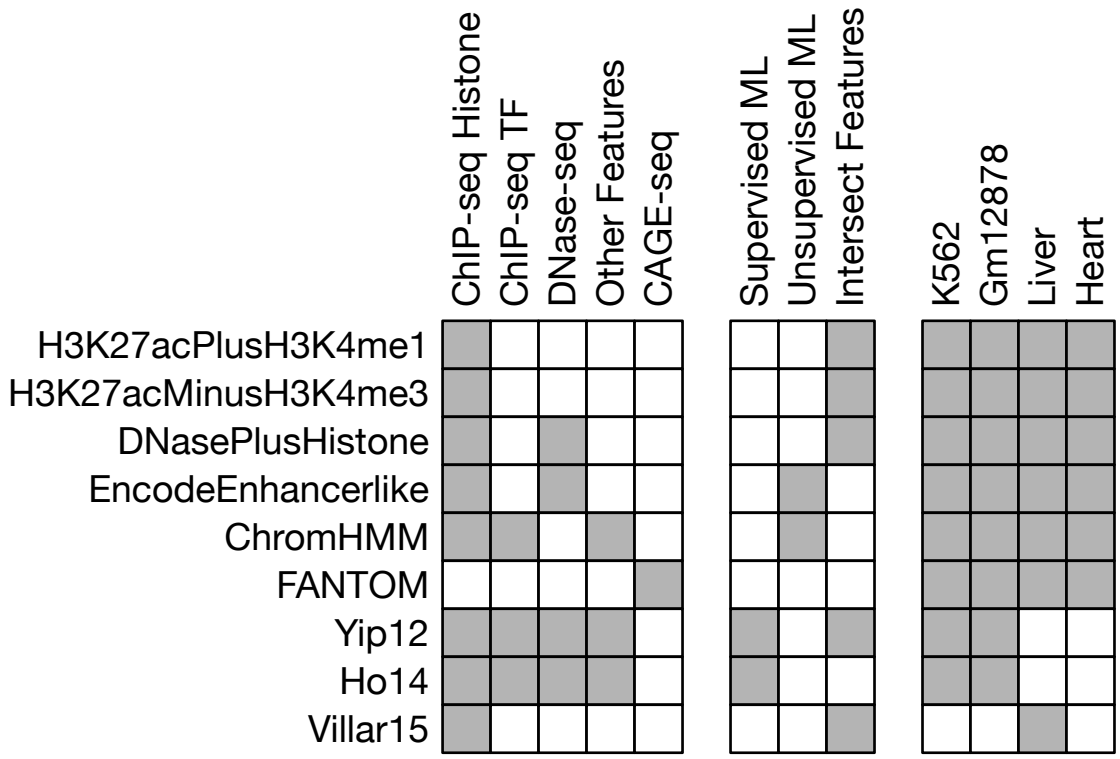

Figure 1: Ten diverse enhancer identification strategies were evaluated across four cellular contexts. Each row summarizes the data sources, analytical approaches, and contexts for the ten enhancer identification strategies we considered. The leftmost columns of the schematic represent the experimental 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. **Error! Reference source not found.** gives the number, length, and genomic coverage of each enhancer set.

For all contexts, we consider two enhancer sets derived solely from ChIP-seq for histone modifications informative about enhancer activity from the Encyclopedia of DNA Elements (ENCODE) Consortium[80]. We downloaded broad peak ChIP-seq data for three histone modifications, H3K27ac, H3K4me1, and H3K4me3 from the ENCODE project[80] for two cell lines, K562 and Gm12878, and from the Roadmap Epigenomics Consortium[81] 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 biological replicates. The "H3K27acPlusH3K4me1" track is a combination of H3K27ac and H3K4me1 ChIP-seq peak files[2,20,39]. 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[19,21,39]. We also downloaded enhancer predictions in liver from Villar et al. 2015[21] which uses an identical histone-modification-derived enhancer definition as the H3K27acMinusH3K4me3 set. In liver, the "Villar15" set provides us with the ability to contrast enhancer sets formed using the same definition and input data type in different laboratories.

To represent an additional enhancer identification strategy in common use, we created another enhancer set for this study using histone modification ChIP-seq peaks and DNase-seq peaks downloaded from ENCODE and Roadmap Epigenomics. The "DNasePlusHistone" track is based on the pipeline described in Hay et al. 2016[61]. 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.

We also curated five representative computationally-defined enhancer sets using more sophisticated machine learning approaches: "EncodeEnhancerlike", "ChromHMM", "Yip12", and "Ho14". We downloaded the "enhancer-like" annotations from ENCODE (version 3.0). These combine DNase-seq and H3K27ac ChIP-seq peaks using an unsupervised ranking model. Each DNase-seq peak is ranked based on the level of combined DNase and H3K27ac signal across a uniform genomic window from the center of the peak (500bp for DNase and 2kb for H3K27ac). The edges of the enhancer-like region are predicted by intersecting the ranked DNase peaks with H3K27ac, and the set is filtered to the top 20,000 regions, excluding those within 2kb of a TSS. However, since there is potential for more

proximal enhancers, any regions within 2kb of a TSS but ranking within the top 20,000 distal enhancer-like elements are included. We retrieved ChromHMM enhancer predictions[58] for the K562 and Gm12878 cell lines from the 25-state segmentation models trained on ENCODE data[56]. This model was trained on ChIP-seq data for eight histone modifications (including H3K4me3, H3K4me1, H3K27me3, and H3K27ac), CTCF, and RNA polymerase II. We downloaded ChromHMM predictions for liver and heart tissues from the additional 15-state segmentation performed by the Roadmap Epigenomics Consortium. This model was trained on H3K9me3 and four of the same histone modifications as the ENCODE model: H3K4me3, H3K4me1, H3K27me3, and H3K36me3. For all ChromHMM sets, we combined the regions labeled as weak and strong enhancer states into a single enhancer set. We considered two enhancer sets for K562 and Gm12878 based on supervised machine learning techniques—one described in Yip et al. 2012[68], and the other in Ho et al. 2014[67]. The Yip12 set predicted 'binding active regions' (BARs) from DNA accessibility data from DNase-seq and FAIRE-seq, and histone modification data on twelve marks using a random forest model. The histone modifications included: H3K27ac, H3K27me3, H3K36me3, H3K4me1, and H3K4me3. To train the random forest, the positive set contained 5,000 randomly sampled BARs overlapping at least one 'transcription-related factor' (TRF), and the negative set contained 5,000 randomly BARs with no TRF peaks. Regions predicted as a BAR with a score above 0.9, a low promoter score, some evidence of evolutionary conservation, and at least 2kb from a TSS became the final BAR set. These positively predicted regions were filtered for relevant TF binding motifs and combined with a second set of putative enhancers. The second set was generated by an additional machine learning model incorporating evolutionary, chromatin, and sequence features to predict TF binding. Proximal regions were excluded and remaining distal regions were required to have used H3K4me1 or H3K4me3 as features during the prediction process. The intersection of the two sets was length restricted (100-700 bp) to create a final set of ~13,000 enhancers[68]. Ho14 was created with a supervised model-based boosting algorithm (mboost). The feature set included fold enrichment of H3K4me1 and H3K4me3 ChIP-seq peaks in conjunction with DHS location and p300 binding sites. The model was trained to predict regions

with regulatory activity both distal (>1 kb) and proximal (< 250 bp) to TSSs. The predicted distal regulatory elements make up the published enhancer set[67].

Finally, 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[33,82,83]. We downloaded enhancer regions predicted by FANTOM for each of the four sample types analyzed[33]. In K562 and Gm12878, CAGE-seq was performed with 3 replicates that were pooled into the final published enhancer sets. In liver and heart, the published predictions were generated from tissue samples from multiple donors[33].

After obtaining or generating each enhancer set, we uniformly processed each one by excluding elements overlapping ENCODE blacklist regions and gaps in the genome assembly[84]. 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. The filtering process removed relatively few annotations (**Error! Reference source not found.**).

Analysis of biological replicates in ChIP-seq

When considering the agreement between biological replicates for K562, Gm12878, and liver H3K27ac ChIP-seq data, we downloaded the FASTQ files from ENCODE[80] and Villar et al. 2015[21], respectively. We aligned the reads from each replicate to GRCh37.p13 using the Burrows-Wheeler Aligner (BWA)[85] (v.0.7.15, default options). We called peaks of broad enrichment using the Model-based Analysis of ChIP-seq (MACS) tool[86] (v.1.4.2, default options). We processed each of the replicate peak files using the same pipeline as the published peak files.

Context	Enhancer Set	Number of Enhancers Removed
K562	H3K27acPlusH3K4me1	200
K562	H3K27acMinusH3K4me3	684
K562	DNasePlusHistone	220
K562	ChromHMM	2117
K562	EncodeEnhancerlike	593
K562	Yip12	109
K562	Ho14	765
K562	FANTOM	21
Gm12878	H3K27acPlusH3K4me1	205
Gm12878	H3K27acMinusH3K4me3	461
Gm12878	DNasePlusHistone	400
Gm12878	ChromHMM	1300
Gm12878	EncodeEnhancerlike	733
Gm12878	Yip12	109
Gm12878	Ho14	766
Gm12878	FANTOM	49
Liver	H3K27acPlusH3K4me1	686
Liver	H3K27acMinusH3K4me3	2342
Liver	DNasePlusHistone	3546
Liver	ChromHMM	1985
Liver	EncodeEnhancerlike	719
Liver	Villar15	225
Liver	FANTOM	14
Heart	H3K27acPlusH3K4me1	590
Heart	H3K27acMinusH3K4me3	3908
Heart	DNasePlusHistone	1693
Heart	ChromHMM	1978
Heart	EncodeEnhancerlike	892
Heart	FANTOM	35
Heart	VISTA	0

Table 1: Number of enhancers removed by length filtering.

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[87].

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[88].

Enrichment for overlap with other genomic regions

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)[89]. We note that this measure of overlap significance is not symmetric, and we confirmed results of our element-wise results for both orderings of the pairs of enhancer sets accordingly.

Enrichment for overlap with evolutionarily conserved elements

In addition to comparing the overlap between pairs of enhancer sets, we also computed enrichment for overlap of evolutionarily conserved regions with each of the enhancer sets. We downloaded evolutionarily conserved regions defined by PhastCons, a two-state hidden Markov model that defines conserved elements from multiple sequence alignments[90]. 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 used the same permutation analysis approach as the genomic comparisons between pairs of enhancer sets, but considered the conserved elements as set *A* and the enhancers as set *B*.

Results

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 (**Error! Reference source not found.**; Figure 2). Different identification methods assay different aspects of enhancer biology (e.g., co-factor binding, histone modification, enhancer RNAs), and therefore we expected to find variation among enhancer sets. Nevertheless, the magnitude of differences we observed is striking. For each attribute we considered, enhancer sets differ by several orders of magnitude (**Error! Reference source not found.**; Figure 2). In liver, FANTOM identifies 326 kilobases (kb) of sequence with enhancer activity, EncodeEnhancerlike identifies 89 megabases (Mb), and H3K27acMinusH3K4me3 identifies almost 138 megabases (Mb). In addition, methods based on similar approaches often differ

substantially; e.g., Villar15, which uses the same enhancer definition as H3K27acMinusH3K4me3, only annotates 86.1 Mb with enhancer function in liver. Overall, methods based on histone modifications tend to identify larger numbers of longer enhancers compared with CAGE data, while machine learning methods are variable. We highlight these trends in liver, but they are similar in other contexts (**Error! Reference source not found.**; Figure 2).

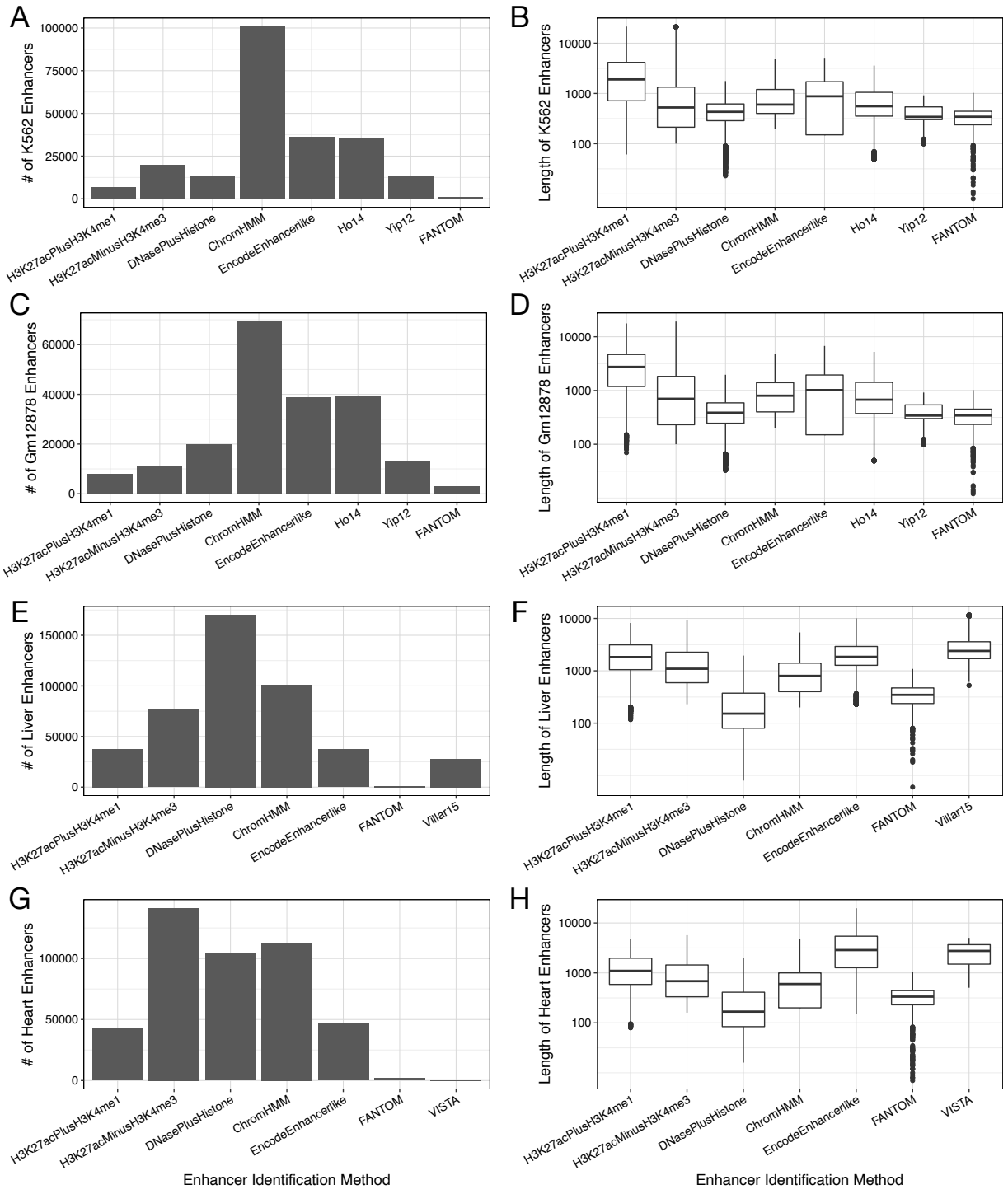

Figure 2: Enhancer identification methods vary in the number and length of predicted enhancers. The number of (A) K562, (C) Gm12878, (E) liver, and (G) heart enhancers identified by each method vary over two orders of magnitude. There is considerable variation even among methods defined based on similar input data, e.g., histone modifications. The length of (B) K562, (D) Gm12878, (F) liver, and (H) heart enhancers identified by different methods shows similar variation. Enhancer lengths are plotted on a $\log_{10}$ scale on the y-axis.

Context	Enhancer Set	Number of Base Pairs (kb)	Number of Enhancers	Median Length	Genomic Fraction
K562	H3K27acPlusH3K4me1	22,113	6,642	1,903	0.0078
K562	H3K27acMinusH3K4me3	34,072	19,698	525	0.0120
K562	DNasePlusHistone	6,620	13,402	431	0.0023
K562	ChromHMM	96,545	100,837	600	0.0339
K562	EncodeEnhancerlike	39,961	36,008	878	0.0140
K562	Ho14	29,027	35,769	556	0.0102
K562	Yip12	5,389	13,303	342	0.0019
K562	FANTOM	390	1,084	344	0.0001
Gm12878	H3K27acPlusH3K4me1	28,355	8,019	2,749	0.0099
Gm12878	H3K27acMinusH3K4me3	20,868	11,238	701	0.0073
Gm12878	DNasePlusHistone	9,286	19,815	386	0.0033
Gm12878	ChromHMM	73,929	69,314	800	0.0259
Gm12878	EncodeEnhancerlike	50,224	38,872	1,018	0.0176
Gm12878	Ho14	41,543	39,550	674	0.0146
Gm12878	Yip12	5,389	13,303	342	0.0019
Gm12878	FANTOM	1,025	2,826	343	0.0004
Liver	H3K27acPlusH3K4me1	87,576	37,644	1,831	0.0307
Liver	H3K27acMinusH3K4me3	137,874	77,014	1,096	0.0484
Liver	DNasePlusHistone	51,292	170,212	152	0.0180
Liver	ChromHMM	108,375	101,260	800	0.0380
Liver	EncodeEnhancerlike	89,129	37,426	1,849	0.0313
Liver	FANTOM	326	869	347	0.0001
Liver	Villar15	86,139	27,725	2,545	0.0302
Heart	H3K27acPlusH3K4me1	59,892	42,910	1,102	0.0210
Heart	H3K27acMinusH3K4me3	157,468	141,162	684	0.0553
Heart	DNasePlusHistone	33,224	103,898	168	0.0117
Heart	ChromHMM	93,067	113,092	600	0.0327
Heart	EncodeEnhancerlike	186,866	47,235	2,872	0.0656
Heart	FANTOM	611	1,720	335	0.0002
Heart	VISTA	261	96	2,772	0.0001

Table 2: Summary of enhancer sets analyzed in this study.

Enhancer sets also vary in their relative distance to other 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 (**Error! Reference source not found.**). In general, the EncodeEnhancerlike regions are the closest to a TSS, possibly due to inclusion of promoter proximal regions with strong H3K27ac and DNase signal.

Context	**Enhancer Set**	**Average Distance to TSS (kb)**
Gm12878	EncodeEnhancerlike	12.71
Gm12878	H3K27acPlusH3K4me1	25.99
Gm12878	Yip12	34.58
Gm12878	Ho14	37.67
Gm12878	FANTOM	39.84
Gm12878	H3K27acMinusH3K4me3	40.29
Gm12878	DNasePlusHistone	42.54
Gm12878	ChromHMM	46.75
Heart	EncodeEnhancerlike	24.73
Heart	FANTOM	34.04
Heart	ChromHMM	40.69
Heart	H3K27acPlusH3K4me1	54.93
Heart	H3K27acMinusH3K4me3	63.35
Heart	VISTA	67.44
Heart	DNasePlusHistone	76.04
K562	EncodeEnhancerlike	13.85
K562	H3K27acPlusH3K4me1	21.38
K562	ChromHMM	35.63
K562	DNasePlusHistone	36.21
K562	Ho14	37.67
K562	Yip12	38.04
K562	H3K27acMinusH3K4me3	38.04
K562	FANTOM	45.26
Liver	EncodeEnhancerlike	14.18
Liver	H3K27acPlusH3K4me1	32.93
Liver	FANTOM	35.21
Liver	H3K27acMinusH3K4me3	43.80
Liver	ChromHMM	46.31
Liver	Villar15	35.36
Liver	DNasePlusHistone	64.41

Table 3: Average distance (kb) to the closest TSS for all enhancer sets.

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 bp overlap between them. All pairs of enhancer sets overlap more than one could expect if they were randomly distributed across the genome (Figure 3A; $p < 0.001$ for all pairs). 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).

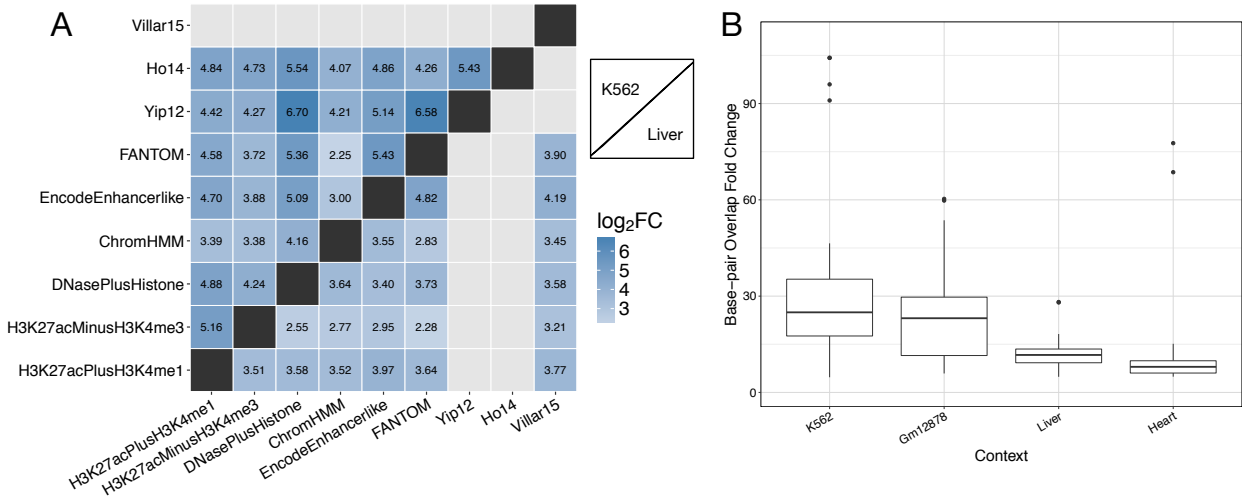

Figure 3: Enhancer sets have more overlap than expected by chance. (A) Pairwise base pair enrichment values (log_2 fold change) for overlap between each K562 (upper triangle) or liver (lower triangle) enhancer set, compared to the expected overlap between randomly distributed, length-matched regions. (B) The enrichment for base pair overlap compared to a random genomic distribution for each pair of enhancer sets within each context. The fold changes for the primary tissues—liver and heart—are significantly lower than the cell lines—K562 and Gm12878 ($p = 4.11E-21$ Kruskal-Wallis test, followed by Dunn's test with Bonferroni correction for pairwise comparisons).

However, the magnitude of overlap between enhancer sets is low: less than 50% for nearly all pairs of methods across contexts (Figure 4). Indeed, 54% of all annotated regions are "singletons" that are annotated by only a single enhancer identification strategy. 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 nearly the same when evaluating overlap on an

element-wise basis, although the percentage of overlap was higher (Table 4). This is not surprising given that only 1 bp of overlap is required for a shared element. Nonetheless, even with more lenient criterion, we still see low percentages overall.

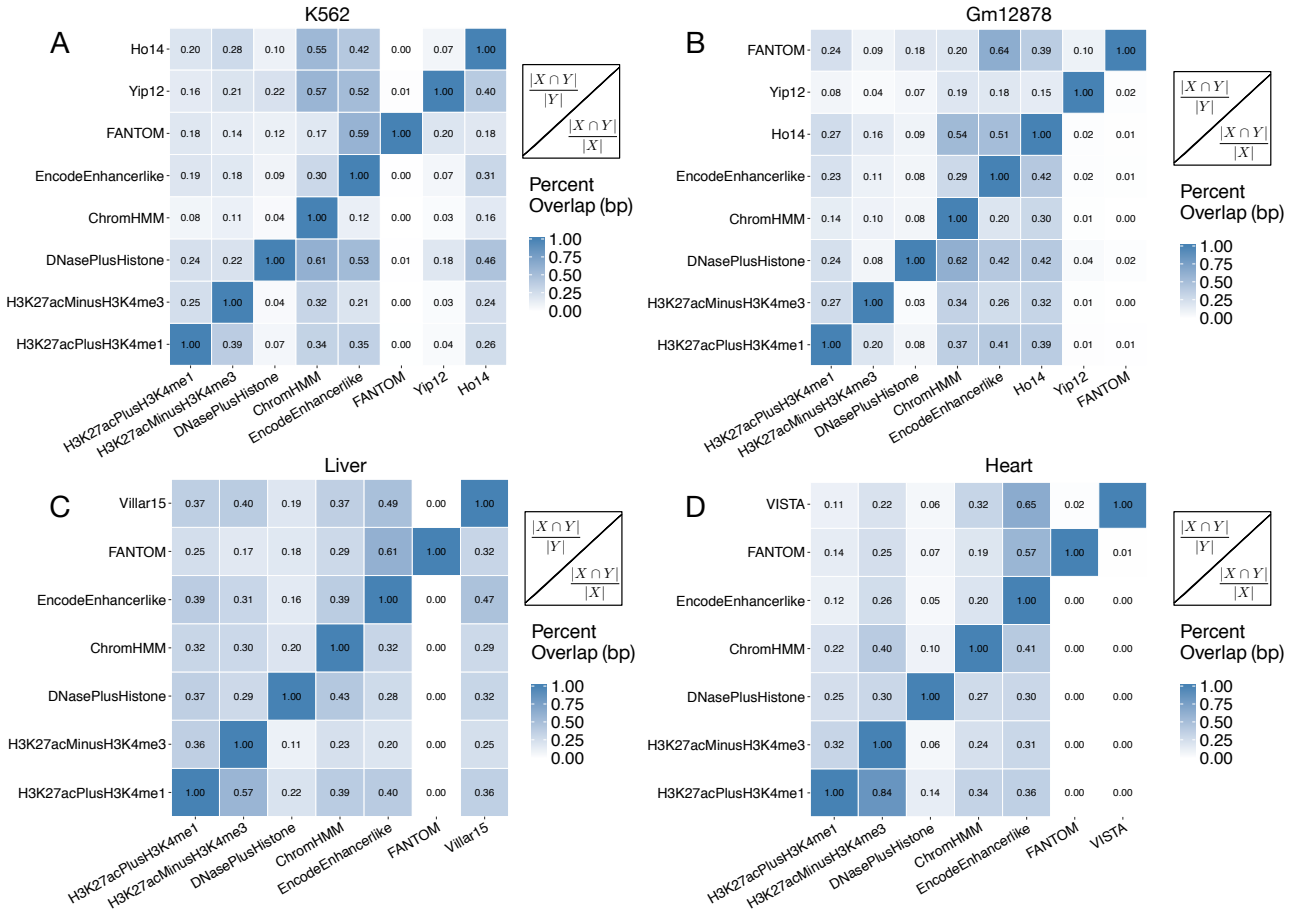

Figure 4: Percent overlap (bp) between enhancer sets. The percent base pair (bp) overlap between all pairs of (A) K562, (B) Gm12878, (C) liver, and (D) heart enhancer sets. Percent overlap for each pair was calculated by dividing the number of shared bp between the two sets by the total number of base 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.

Context	Comparison Type	Minimum	Maximum	Median	Mean
K562	Base-pair	0.00	0.61	0.18	0.21
	Element	0.00	0.72	0.23	0.29
Gm12878	Base-pair	0.00	0.64	0.16	0.19
	Element	0.01	0.79	0.19	0.25
Liver	Base-pair	0.00	0.61	0.30	0.28
	Element	0.00	0.71	0.34	0.34
Heart	Base-pair	0.00	0.84	0.17	0.19
	Element	0.00	0.83	0.22	0.24

Table 4: Summary statistics for pairwise percent overlap.

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 extremely 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 5, upper triangles). 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 5, lower triangles). In these comparisons, FANTOM and EncodeEnhancerlike had among the highest relative similarity scores, suggesting they localize in similar genomic regions.

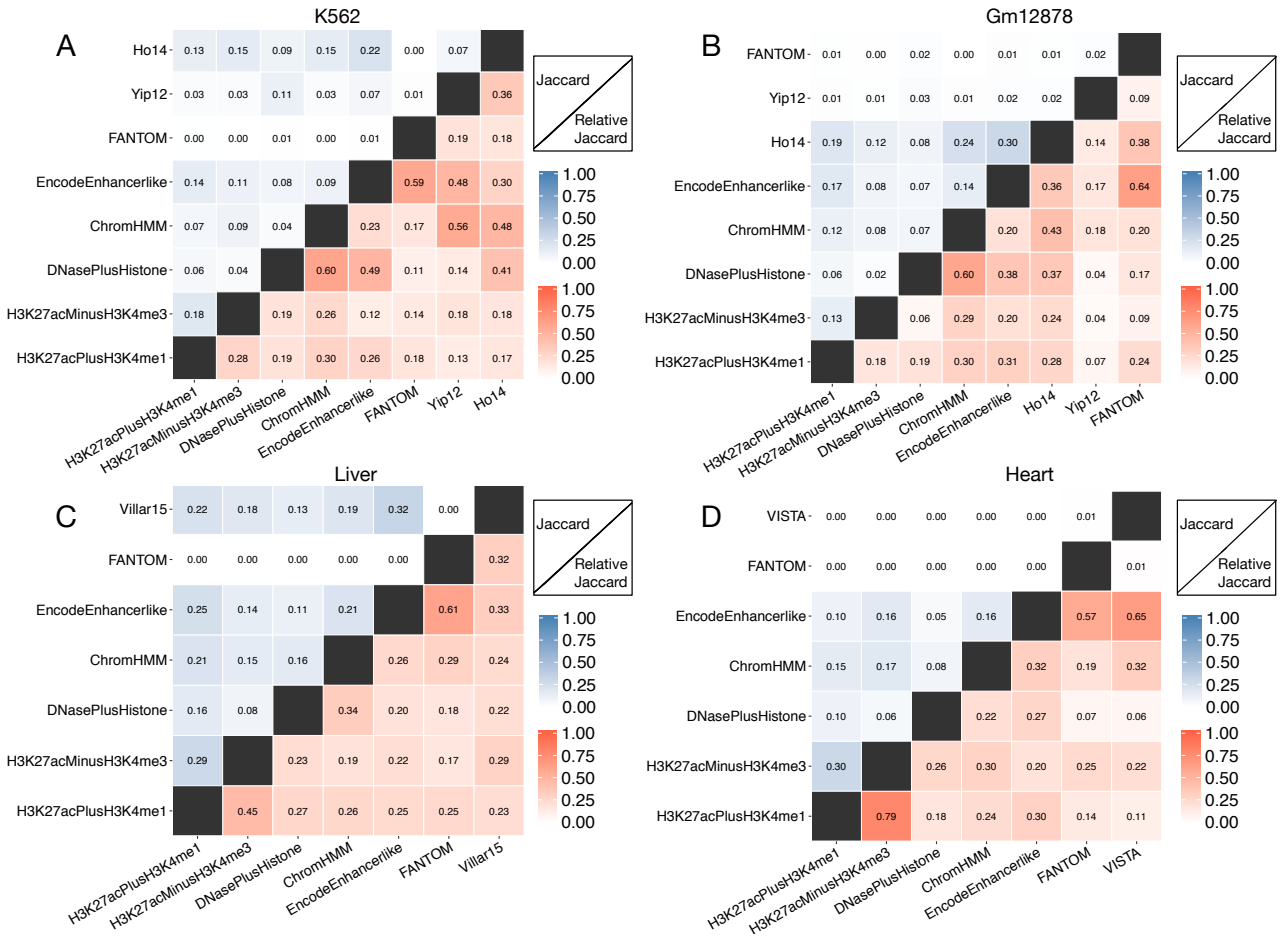

Figure 5: Jaccard similarity (bp) between enhancer sets. The Jaccard similarity between all pairs of (A) K562, (B) Gm12878, (C) liver, and (D) heart enhancer sets. The upper triangle gives the Jaccard similarity, and the lower triangle gives the relative Jaccard similarity in which the observed similarity is divided by the maximum possible similarity for the pair of sets.

To assess the influence of technical 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. H3K27ac ChIP-seq data is used in the formation of the majority of the enhancer sets considered here, so high technical variability could impact many of the predictions. In practical applications, we expect the replicates to have high overlap and serve as an "upper bound" of similarity. On average, the replicates overlap 76% at the bp level. 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[90], 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 6). Across each context, the histone-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 and Yip12 enhancers are nearly twice as enriched for conserved regions as the histone-based sets (2.7x and 3.3x, respectively). Evolutionary conservation was directly considered in the definition of the Yip12 set, but not in FANTOM. Here we considered enhancer elements overlapped by conserved elements; enrichment trends are similar when we consider the number of conserved base pairs overlapped by each enhancer set (Figure 6).

Figure 6: Enhancer sets vary in their degree of evolutionary conservation. (A) Base pair enrichment for conserved elements. (B) Element-wise enrichment for conserved elements. 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.

Conclusion

Despite attempting to annotate the same regulatory element, different identification strategies produce enhancer sets that have low levels of genomic similarity. This chapter provides a formal quantification of similarity in genomic locations, demonstrating low overlap between enhancer sets. Furthermore, the enhancer sets differ in other genomic attributes such as distance to a TSS and level of evolutionary conservation.

CHAPTER III

Characterizing Functional Similarity Between Enhancer Sets

Introduction

While some variation due to differences in the underlying assays or computational processing is expected, Chapter II identifies significant differences between enhancer sets defined in the same context. Since it is well established that the dysregulation of genes due to perturbations of enhancer sequences lead to disease phenotypes, a common goal for genome-wide enhancer maps is to inform our functional and mechanistic understanding of gene regulation. The ability to accurately map and interpret enhancer annotations is critical to achieving this goal. Chapter III explores the differences in biological interpretations caused by differences in enhancer identification strategies. We begin by quantifying the stability of conclusions about disease and expression associated variation between enhancer sets. We then use several enhancer-target mapping approaches coupled with Gene Ontology (GO) enrichment analyses to contrast the mechanistic and functional attributes associated with each enhancer set. Finally, we cluster enhancer sets based on their dissimilarity in genomic or functional space to understand the relationships between enhancer sets from multiple perspectives.

Methods

GWAS Catalog SNPs and GTEx eQTL

We downloaded the full list of 20,458 unique GWAS SNPs from the NHGRI-EBI GWAS Catalog (v1.0, downloaded 08-10-2016)[91]. From this set we manually curated the GWAS SNPs into two subsets associated with phenotypes relevant to liver (n = 50) or heart (n = 169), respectively, for context-specific analyses (Appendix). We also downloaded all GTEx eQTL from the GTEx Portal (v6p, downloaded 09-07-2016)[92]. 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 using the same significance threshold ($p < 10^{-10}$). 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)[93].

Enrichment for overlap with GWAS catalog SNPs and GTEx eQTL

We computed enrichment for overlap with GWAS SNPs and GTEx eQTL with each of the enhancer sets described in Chapter II using the same permutation framework. For GWAS tag SNPs, we considered each variant as a region in set *A* and the enhancer regions as set *B*. We used an identical 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. We computed enrichments for the eQTL SNP sets using the same strategy as described above for GWAS SNPs.

Enhancer set Gene Ontology annotation and similarity

We used GREAT to find Gene Ontology (GO) annotations enriched among genes nearby each enhancer set. 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. The regulatory domain was defined using the 'basal plus extension' rule, which includes base pairs 5 kb upstream and 1 kb downstream of a gene (e.g. the basal domain) plus up to 1000 kb on either side. The domain ends after extending 1000 kb or at the TSS of the next gene[94]. 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 ($p < 0.05$). We downloaded up to 1,000 significant terms for each enhancer set from both the Molecular Function (MF) and Biological Process (BP) GO ontologies.

To calculate the similarity between lists of GO terms we used the GOSemSim package in R[95]. GOSemSim uses a sematic similarity metric that accounts for the hierarchical organization of the ontology and relatedness between terms when calculating the similarity score[96]. For each pair of enhancer sets, we calculated the similarity between their associated GO terms. We converted the resulting similarity matrix into a dissimilarity matrix by subtracting each score from 1. 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 across long distances, we also considered target predictions generated by the JEME algorithm to assign enhancers to potential target genes in each context. 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 using a regression model. The first step creates regression models for all enhancers within 1Mb of a potential target gene, using error terms from those models to inform the second step. The second step trains random forest models to predict the final enhancer targets[97]. By intersecting each enhancer set with 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[98]. 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. We calculated the pairwise similarity between lists of GO terms using the same semantic similarity metric as above.

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 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. We

calculated similarity using the semantic similarity metric described above and converted it to dissimilarity by subtracting the score from 1. For both, we used agglomerative hierarchical clustering in R with the default complete linkage method to iteratively combine clusters[99]. We visualized the cluster results as dendrograms using ggplot2 and dendextend[88,100]. We also performed multidimensional scaling (MDS) on the Jaccard and GO term dissimilarity matrices using default options in R[99].

Results

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

Genome-wide enhancer sets are commonly used to interpret the potential function of genetic variants observed in GWAS and sequencing studies[49,62,64,65,83,101–105]. Functional genetic variants—in particular mutations associated with complex disease—are enriched in gene regulatory regions[4,5]. We evaluated the sensitivity of this pattern to enhancer identification strategy by intersecting each of the enhancer sets with 20,458 unique tag SNPs significantly associated with traits from the GWAS Catalog. Overall, 32.9% (6,736 / 20,458) of GWAS SNPs overlap an enhancer identified by at least one of the strategies in one of the contexts we considered. However, there is wide variation in the number of overlapping GWAS Catalog SNPs between enhancer sets, as is expected given the large variation in the number and genomic distribution of enhancers predicted by different methods (Table 5).

Context	Enhancer Set	Number of GWAS SNPs	Number of Context-Specific GWAS SNPs
K562	H3K27acPlusH3K4me1	269	
	H3K27acMinusH3K4me3	420	
	DNasePlusHistone	88	
	ChromHMM	1081	
	EncodeEnhancerlike	476	
	Ho14	332	
	Yip12	79	
	FANTOM	2	
Gm12878	H3K27acPlusH3K4me1	371	
	H3K27acMinusH3K4me3	235	
	DNasePlusHistone	161	
	ChromHMM	865	
	EncodeEnhancerlike	666	
	Ho14	499	
	Yip12	79	
	FANTOM	24	
Liver	H3K27acPlusH3K4me1	1102	25
	H3K27acMinusH3K4me3	1658	36
	DNasePlusHistone	654	12
	ChromHMM	1303	34
	EncodeEnhancerlike	1268	36
	FANTOM	3	0
	Villar15	1203	26
Heart	H3K27acPlusH3K4me1	644	91
	H3K27acMinusH3K4me3	1633	222
	DNasePlusHistone	428	47
	ChromHMM	1126	155
	EncodeEnhancerlike	2284	302
	FANTOM	16	3
	VISTA	3	0

Table 5: Number of overlapping GWAS SNPs per enhancer identification method.

Nonetheless, GWAS tag SNPs are significantly enriched at similar levels in most enhancer sets and contexts, with the exception of FANTOM, which has roughly twice the enrichment of other methods

in Gm12878 and heart (Figure 7A). Since the tag SNPs are often not the functional variants, we also considered SNPs in high linkage disequilibrium (LD) with the GWAS SNPs ($r^2 > 0.9$). While the overall enrichments were lower, the variability between enhancer sets remained small (Figure 7B). We also identified the variant in each LD block with the maximum number of overlaps with distinct enhancer sets. Even when limiting our analysis to this biased set, 58% of GWAS SNPs (11,854 / 20,458) have no enhancer overlap (Figure 8A).

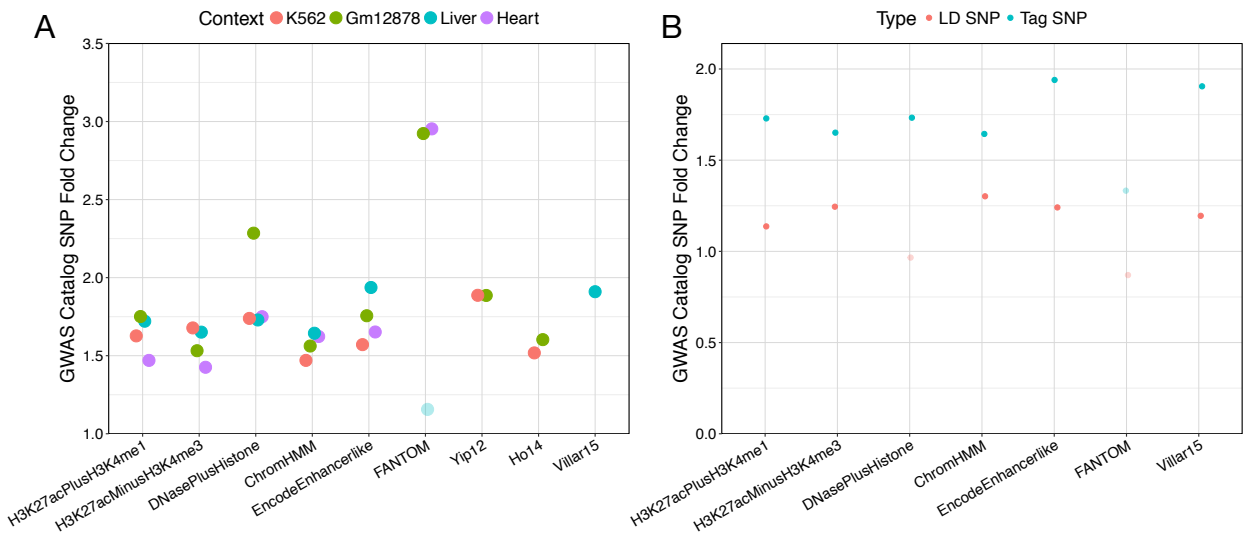

Figure 7: Different GWAS SNP enrichment between enhancer sets. (A) GWAS SNP enrichment among all enhancer sets for each biological context. All sets are significantly enriched, except FANTOM in K562 and liver contexts due to small sample size. (B) Enhancer sets are less enriched for variants in LD with GWAS tag SNPs than with the tag SNPs, but, more importantly, a similar magnitude of enrichment is observed across enhancer identification methods. Transparent points indicate non-significant enrichment values.

Furthermore, GWAS SNPs with enhancer overlap are commonly predicted to overlap an enhancer by only a single identification strategy (Figure 8B). For example, in liver, 47% (1710 / 3660) of the GWAS SNPs that overlapped an enhancer are unique to a single set, and only 27% (982 / 3660) overlap enhancers from more than two sets. The distribution of enhancer overlaps was similar when considering all candidate variants in LD (Figure 8B). Even after limiting to GWAS LD blocks with enhancer overlap and selecting the variant with maximum overlap, 30% (2620 / 8604) are still only

predicted by one enhancer identification method (Figure 8B). This demonstrates that the annotation of variants in regions highlighted by GWAS varies greatly depending on the enhancer identification strategy used. 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 or heart (Appendix). As in the full GWAS set, 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 8B). 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 (Table 6; 1.74x–2.68x). FANTOM enhancers are the exception to this trend due to the small number of overlapping context-specific SNPs (Table 5). This suggests that the different methods, in spite of their lack of agreement, all identify regulatory regions relevant to the target context.

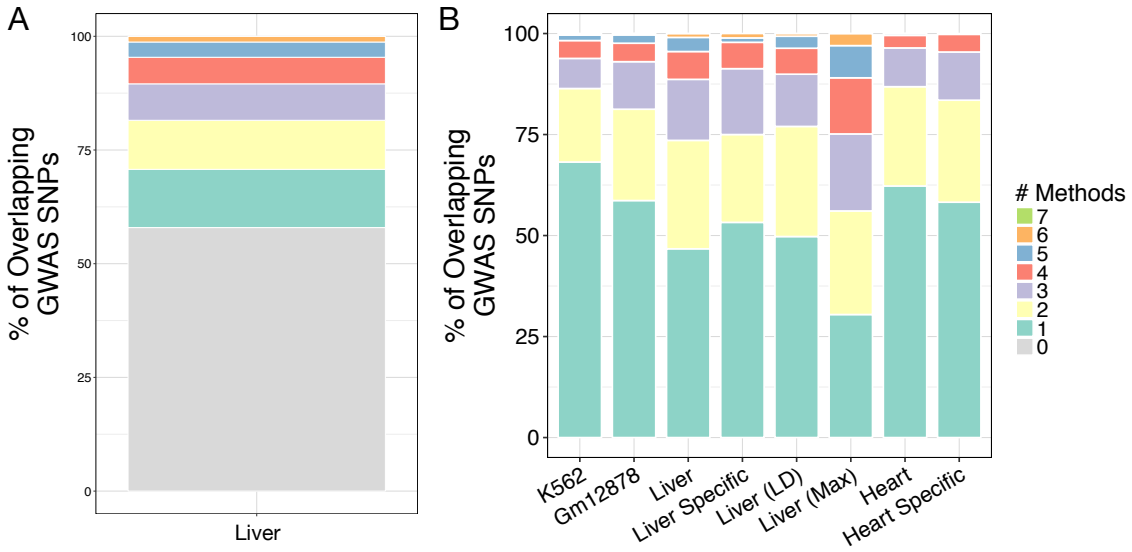

Figure 8: Most GWAS SNPs overlapped by a single enhancer set. (A) For each LD block, the majority of variants with the maximum amount of overlap do not overlap enhancers identified by the studied methods. (B) Among all GWAS SNPs that overlap at least one enhancer in a context, the colored bars represent the number of methods that identified the region as an enhancer. The majority of these variants are supported by a single method; very few GWAS variants are shared among all methods. The conclusions are similar when considering variants in high LD (r^2 > 0.9) with the GWAS tag SNPs in liver (Liver LD), when limiting to SNPs associated with liver or heart related phenotypes (Liver Specific, Heart Specific), and when considering the SNP in each LD block with the maximum number of enhancer overlaps (Liver Max).

Context	Method	Fold Change	P Value
Liver	H3K27acPlusH3K4me1	2.12	0.002
	H3K27acMinusH3K4me3	2.00	0.001
	DNasePlusHistone	1.74	0.035
	ChromHMM	2.38	0.001
	EncodeEnhancerlike	2.95	0.001
	FANTOM	0.945	1.000
	Villar15	2.24	0.001
Heart	H3K27acPlusH3K4me1	2.00	0.001
	H3K27acMinusH3K4me3	1.87	0.001
	DNasePlusHistone	1.83	0.001
	ChromHMM	2.16	0.001
	EncodeEnhancerlike	2.10	0.001
	FANTOM	2.68	0.023

Table 6: Enrichment for overlap with context-specific SNPs in liver and heart.

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. These analyses yielded similar results as for the GWAS Catalog variants (Figure 9). Within a context, most eQTL are identified as enhancers by a single enhancer prediction method only, and there is wide variation in the number and enrichment of eQTL overlapped by different enhancer sets (Figure 9; Table 8). Across liver enhancer sets, 50% (33,941 / 68,563) of all overlapped eQTL are called an enhancer by only a single method (Figure 9). Considering variants in high LD ($r^2 > 0.9$) does not affect this trend (Figure 9B). Similarly, after limiting the analysis to the variants with the maximum number of overlaps in each LD block, 24% (64871 / 271732) of the eQTL with enhancer overlap are identified by only one enhancer set (Figure 9B). Furthermore, restriction to context-specific eQTL in liver or heart does increase the enrichment for eQTL across most methods, but the distribution of shared eQTL remains similar (Figure 9B; Table 8). Thus, the interpretation of variants in regions known to influence gene regulation varies substantially depending on the enhancer identification strategy used.

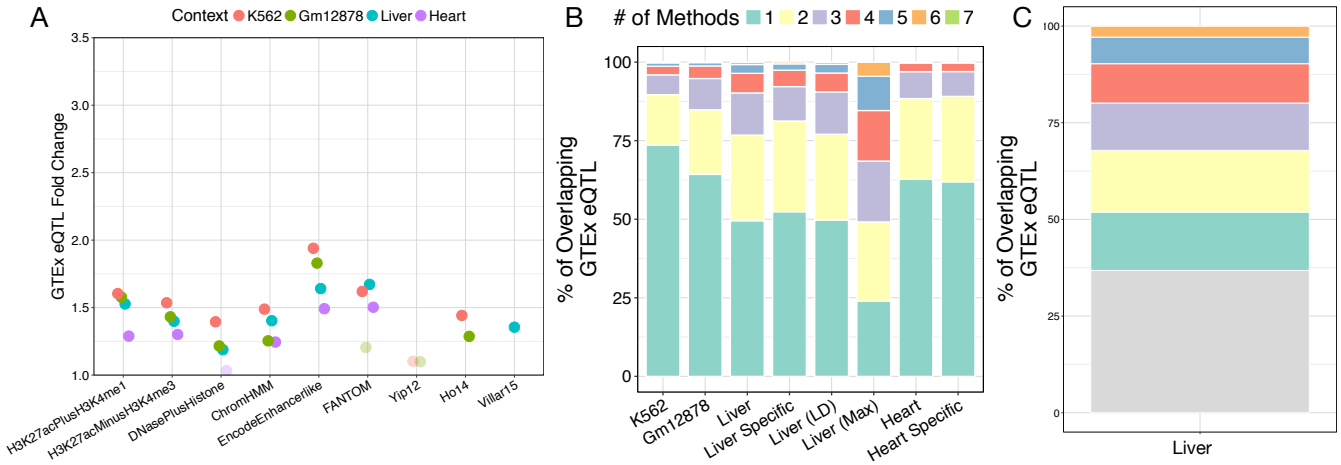

Figure 9: Enhancer set differ in amount of GTEx eQTL overlap. (A) GTEx eQTL enrichment among all enhancer sets for each biological context. Transparent points indicate nonsignificant enrichment ($p > 0.05$). (B) Among eQTL that overlap at least one enhancer, the majority is supported by only a single method. This holds for LD-expanded and context-specific sets (Liver LD, Liver Specific, Heart Specific). 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). These trends are similar to what is seen for GWAS SNPs in Figure 8. (C) Number of enhancer sets intersected by the variant with the maximum number of overlap for each eQTL LD block; 37% do not overlap an enhancer identified by the studied methods.

Context	Enhancer Set	Number of GTEx eQTL	Number of Context-specific GTEx eQTL
K562	H3K27acPlusH3K4me1	5768	
	H3K27acMinusH3K4me3	8076	
	DNasePlusHistone	1561	
	ChromHMM	24072	
	EncodeEnhancerlike	13142	
	Ho14	6910	
	Yip12	988	
	FANTOM	112	
Gm12878	H3K27acPlusH3K4me1	7254	
	H3K27acMinusH3K4me3	4653	
	DNasePlusHistone	1881	
	ChromHMM	14794	
	EncodeEnhancerlike	15423	
	Ho14	8672	
	Yip12	988	
	FANTOM	207	
Liver	H3K27acPlusH3K4me1	21426	1603
	H3K27acMinusH3K4me3	30704	2477
	DNasePlusHistone	9698	596
	ChromHMM	24279	2037
	EncodeEnhancerlike	23962	1904
	FANTOM	95	2
	Villar15	18648	1281
Heart	H3K27acPlusH3K4me1	12193	3466
	H3K27acMinusH3K4me3	32068	9165
	DNasePlusHistone	5380	1412
	ChromHMM	18665	5269
	EncodeEnhancerlike	45916	14061
	FANTOM	162	72
	VISTA	14	23

Table 7: Number of GTEx eQTL overlap per enhancer set.

Context	Method	Fold Change	P Value
Liver	H3K27acPlusH3K4me1	1.63	0.002
	H3K27acMinusH3K4me3	1.62	0.001
	DNasePlusHistone	1.05	0.534
	ChromHMM	1.68	0.001
	EncodeEnhancerlike	1.84	0.001
	FANTOM	0.58	0.575
	Villar15	1.34	0.001
Heart	H3K27acPlusH3K4me1	1.42	0.001
	H3K27acMinusH3K4me3	1.45	0.001
	DNasePlusHistone	1.07	0.234
	ChromHMM	1.35	0.001
	EncodeEnhancerlike	1.76	0.001
	FANTOM	2.49	0.006

Table 8: Number of overlapping GTEx eQTL per enhancer set.

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 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 discover 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 enrichment analyses, and (*ii*) applying the Genomic Regions Enrichment of Annotations Tool (GREAT)[94,97]. Many of the GO terms identified by both methods for the enhancer sets are relevant to the associated context (Table 9). 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 9), likely due to the large gene target lists for most enhancer sets (Table 10).

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

Table 9: Top five GO terms for liver enhancer sets from GREAT and JEME target-mapped WebGestalt enrichments.

Context	Enhancer Set	Number of Genes
K562	H3K27acPlusH3K4me1	3444
	H3K27acMinusH3K4me3	4837
	DNasePlusHistone	3001
	ChromHMM	10676
	EncodeEnhancerlike	10004
	Yip12	2754
	Ho14	7064
	FANTOM	3152
Gm12878	H3K27acPlusH3K4me1	4626
	H3K27acMinusH3K4me3	3407
	DNasePlusHistone	5014
	ChromHMM	10710
	EncodeEnhancerlike	11303
	Yip12	1941
	Ho14	8947
	FANTOM	5352
Liver	H3K27acPlusH3K4me1	6871
	H3K27acMinusH3K4me3	5964
	DNasePlusHistone	7176
	ChromHMM	11788
	EncodeEnhancerlike	8066
	Villar15	3626
	FANTOM	1796
Heart	H3K27acPlusH3K4me1	2121
	H3K27acMinusH3K4me3	3940
	DNasePlusHistone	1771
	ChromHMM	7144
	EncodeEnhancerlike	5124
	FANTOM	1779
	VISTA	32

Table 10: Number of target genes mapped to each enhancer set by JEME.

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 triangles). 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 and thus are less functionally specific. These results provide evidence that the different enhancer sets influence different functions relevant to the target biological context. Using the MF GO enrichment analyses calculated with the JEME target-mapped genes, we see similar trends, although the numbers of overlapping terms are higher, especially for enhancer sets relying on histone modification data from ENCODE (Figure 11). We note that since the JEME target maps were built using ENCODE ChromHMM enhancer tracks, it is likely that some of this additional similarity is due to bias in the predictions.

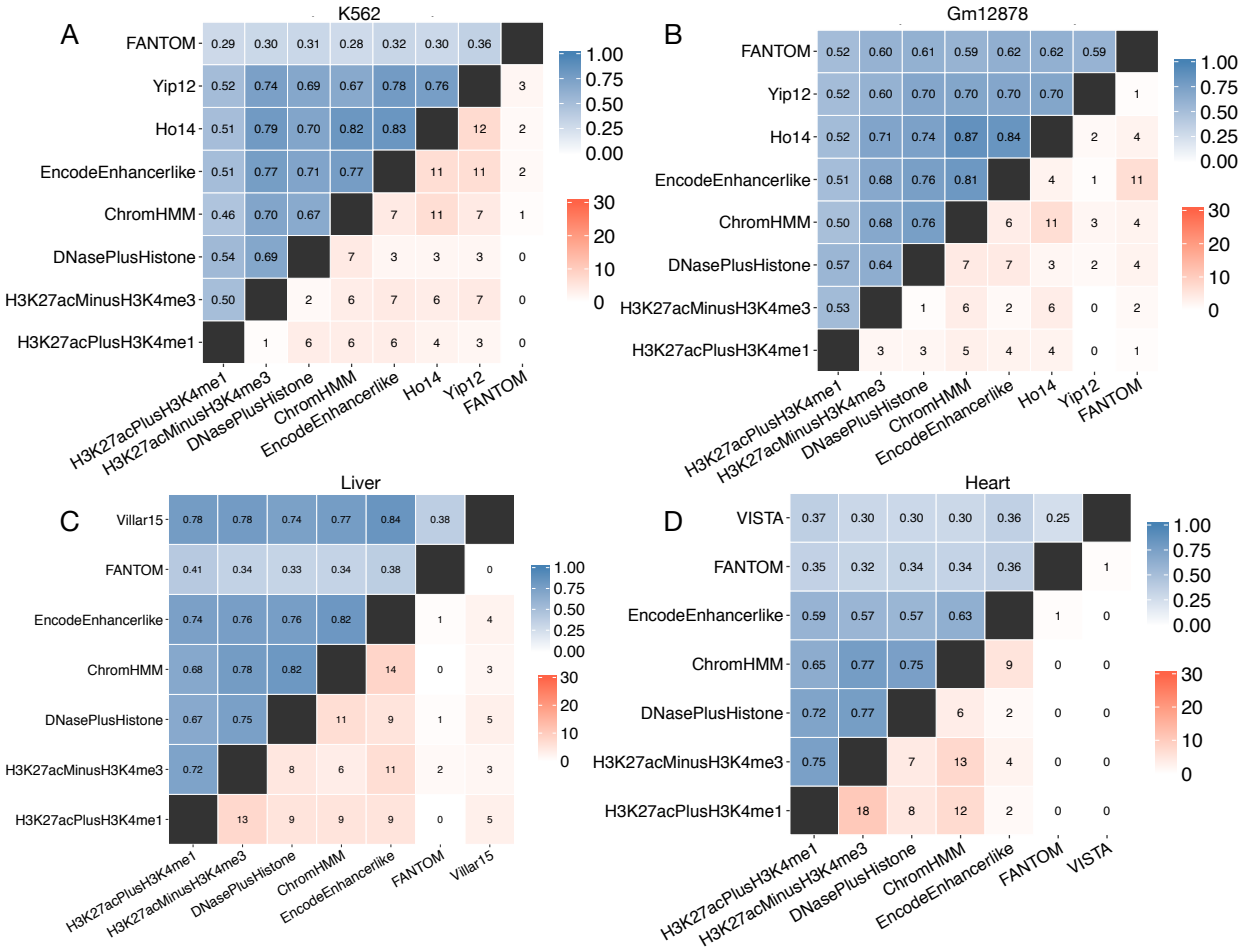

Figure 10: GO term similarity for GREAT (MF). Enhancer sets from the same biological context have different functional associations. The upper triangle shows the semantic similarity from GoSemSim; the lower triangle shows the number of top 30 most significant GO MF terms shared by each pair of enhancer sets in K562 (A), Gm12878 (B), liver (C), and heart (D).

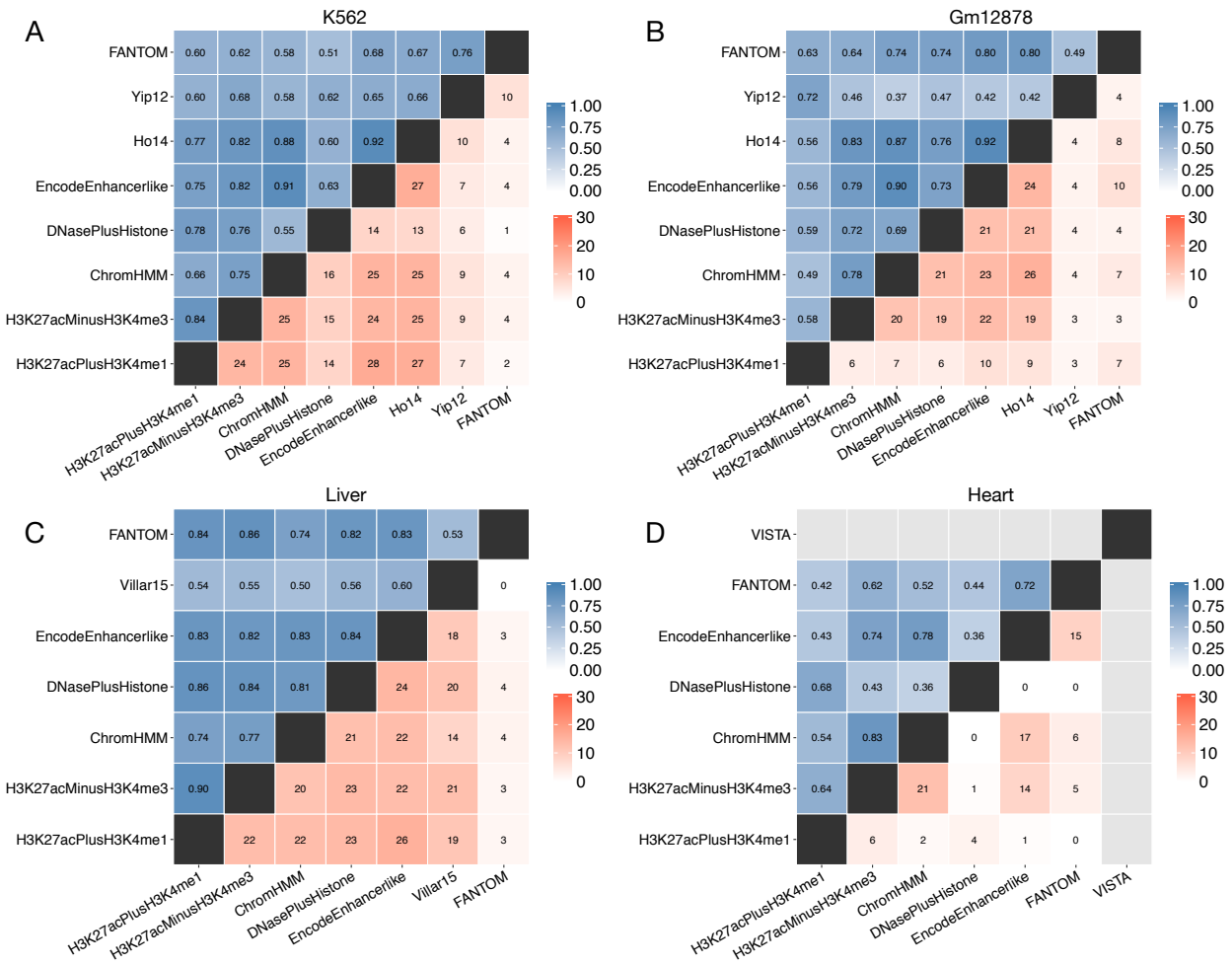

Figure 11: GO term similarity for JEME-mapped genes (MF). Pairwise similarity for GO Molecular Function (MF) enrichments for enhancer sets based on JEME's putative mappings to target genes in K562 (A), Gm12878 (B), liver (C), and heart (D). The upper triangle shows the semantic similarity calculated using GoSemSim, and the lower triangle shows the number of shared terms of the top 30 most significantly enriched. Gray squares indicate that the analysis found no significantly enriched terms. There is greater similarity between these associations compared to GREAT (Figure 10) although the similarity remains relatively low and many of the matched terms are high in the hierarchy.

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[95,96]. The EncodeEnhancerlike and ChromHMM enhancer sets are among the most functionally similar, with similarity scores near 0.80 in most contexts (Figure 10, Figure 11, upper triangle; Figure 12). 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, Figure 11). FANTOM is more similar to other methods in Gm12878, with an average score of 0.59, and in the JEME mappings (Figure 10, Figure 11). Since the JEME target predictions used FANTOM enhancers as input, it is difficult to know if the increased similarity represents shared function or a technical artifact. In general, these trends hold for both the Biological Process (BP) ontologies (Figure 12). However, the JEME mappings for heart do not result in many significantly enriched terms for the BP ontology which makes it difficult to make detailed comparisons.

As a benchmark, biological replicates of the Gm12878 H3K27ac ChIP-seq peaks received a similarity of 0.93. This is much lower than most of the functional similarities across target mapping approached and ontologies. It suggests there are 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. However, insofar as they reflect the regulatory context of the different enhancer sets, they reveal significant functional differences between enhancer identification methods.

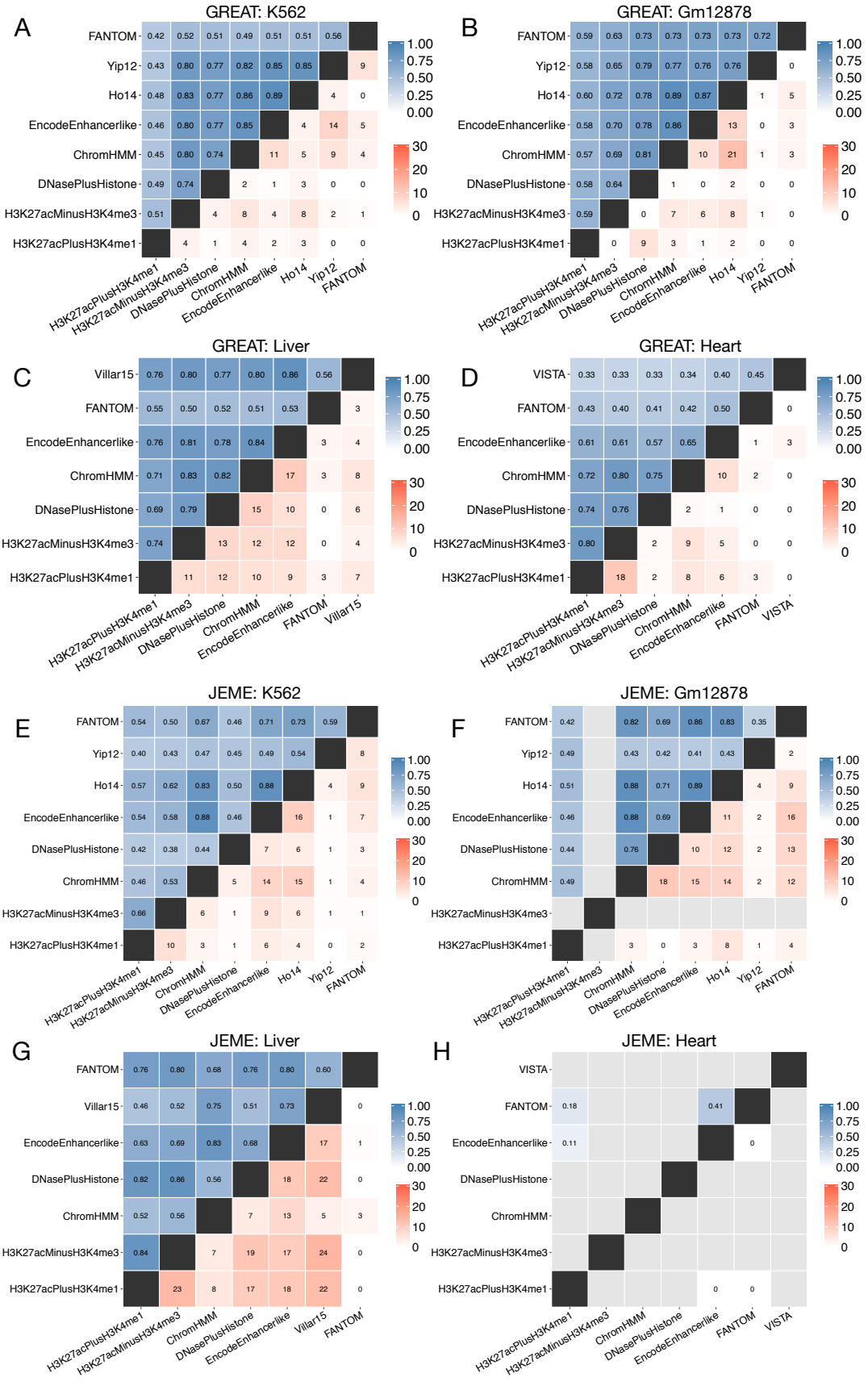
A
GREAT: K562
FANTOM 0.42 0.52 0.51 0.49 0.51 0.51 0.56
Yip12 0.43 0.80 0.77 0.82 0.85 0.85 9
Ho14 0.48 0.83 0.77 0.86 0.89 4 0
EncodeEnhancerlike 0.46 0.80 0.77 0.85 4 14 5
ChromHMM 0.45 0.80 0.74 11 5 9 4
DNasePlusHistone 0.49 0.74 2 1 3 0 0
H3K27acMinusH3K4me3 0.51 4 8 4 8 2 1
H3K27acPlusH3K4me1 4 1 4 2 3 0 0
H3K27acPlusH3K4me1 H3K27acMinusH3K4me3 DNasePlusHistone ChromHMM EncodeEnhancerlike Ho14 Yip12 FANTOM
1.00 0.75 0.50 0.25 0.00
30 20 10 0
B
GREAT: Gm12878
FANTOM 0.59 0.63 0.73 0.73 0.73 0.73 0.72
Yip12 0.58 0.65 0.79 0.77 0.76 0.76 0
Ho14 0.60 0.72 0.78 0.89 0.87 1 5
EncodeEnhancerlike 0.58 0.70 0.78 0.86 13 0 3
ChromHMM 0.57 0.69 0.81 10 21 1 3
DNasePlusHistone 0.58 0.64 1 0 2 0 0
H3K27acMinusH3K4me3 0.59 0 7 6 8 1 0
H3K27acPlusH3K4me1 0 9 3 1 2 0 0
H3K27acPlusH3K4me1 H3K27acMinusH3K4me3 DNasePlusHistone ChromHMM EncodeEnhancerlike Ho14 Yip12 FANTOM
1.00 0.75 0.50 0.25 0.00
30 20 10 0
C
GREAT: Liver
Villar15 0.76 0.80 0.77 0.80 0.86 0.56
FANTOM 0.55 0.50 0.52 0.51 0.53 3
EncodeEnhancerlike 0.76 0.81 0.78 0.84 3 4
ChromHMM 0.71 0.83 0.82 17 3 8
DNasePlusHistone 0.69 0.79 15 10 0 6
H3K27acMinusH3K4me3 0.74 13 12 12 0 4
H3K27acPlusH3K4me1 11 12 10 9 3 7
H3K27acPlusH3K4me1 H3K27acMinusH3K4me3 DNasePlusHistone ChromHMM EncodeEnhancerlike FANTOM Villar15
1.00 0.75 0.50 0.25 0.00
30 20 10 0
D
GREAT: Heart
VISTA 0.33 0.33 0.33 0.34 0.40 0.45
FANTOM 0.43 0.40 0.41 0.42 0.50 0
EncodeEnhancerlike 0.61 0.61 0.57 0.65 1 3
ChromHMM 0.72 0.80 0.75 10 2 0
DNasePlusHistone 0.74 0.76 2 1 0 0
H3K27acMinusH3K4me3 0.80 2 9 5 0 0
H3K27acPlusH3K4me1 18 2 8 6 3 0
H3K27acPlusH3K4me1 H3K27acMinusH3K4me3 DNasePlusHistone ChromHMM EncodeEnhancerlike FANTOM VISTA
1.00 0.75 0.50 0.25 0.00
30 20 10 0
E
JEME: K562
FANTOM 0.54 0.50 0.67 0.46 0.71 0.73 0.59
Yip12 0.40 0.43 0.47 0.45 0.49 0.54 8
Ho14 0.57 0.62 0.83 0.50 0.88 4 9
EncodeEnhancerlike 0.54 0.58 0.88 0.46 16 1 7
DNasePlusHistone 0.42 0.38 0.44 7 6 1 3
ChromHMM 0.46 0.53 5 14 15 1 4
H3K27acMinusH3K4me3 0.66 6 1 9 6 1 1
H3K27acPlusH3K4me1 10 3 1 6 4 0 2
H3K27acPlusH3K4me1 H3K27acMinusH3K4me3 ChromHMM DNasePlusHistone EncodeEnhancerlike Ho14 Yip12 FANTOM
1.00 0.75 0.50 0.25 0.00
30 20 10 0
F
JEME: Gm12878
FANTOM 0.42 0.82 0.69 0.86 0.83 0.35
Yip12 0.49 0.43 0.42 0.41 0.43 2
Ho14 0.51 0.88 0.71 0.89 4 9
EncodeEnhancerlike 0.46 0.88 0.69 11 2 16
DNasePlusHistone 0.44 0.76 10 12 2 13
ChromHMM 0.49 18 15 14 2 12
H3K27acMinusH3K4me3
H3K27acPlusH3K4me1 3 0 3 8 1 4
H3K27acPlusH3K4me1 H3K27acMinusH3K4me3 ChromHMM DNasePlusHistone EncodeEnhancerlike Ho14 Yip12 FANTOM
1.00 0.75 0.50 0.25 0.00
30 20 10 0
G
JEME: Liver
FANTOM 0.76 0.80 0.68 0.76 0.80 0.60
Villar15 0.46 0.52 0.75 0.51 0.73 0
EncodeEnhancerlike 0.63 0.69 0.83 0.68 17 1
DNasePlusHistone 0.82 0.86 0.56 18 22 0
ChromHMM 0.52 0.56 7 13 5 3
H3K27acMinusH3K4me3 0.84 7 19 17 24 0
H3K27acPlusH3K4me1 23 8 17 18 22 0
H3K27acPlusH3K4me1 H3K27acMinusH3K4me3 ChromHMM DNasePlusHistone EncodeEnhancerlike Villar15 FANTOM
1.00 0.75 0.50 0.25 0.00
30 20 10 0
H
JEME: Heart
VISTA
FANTOM 0.18 0.41
EncodeEnhancerlike 0.11 0
DNasePlusHistone
ChromHMM
H3K27acMinusH3K4me3
H3K27acPlusH3K4me1 0 0
H3K27acPlusH3K4me1 H3K27acMinusH3K4me3 ChromHMM DNasePlusHistone EncodeEnhancerlike FANTOM VISTA
1.00 0.75 0.50 0.25 0.00
30 20 10 0

Figure 12: GO Enrichment for BP ontology using GREAt and JEME target-mapping. There is low pairwise similarity between GO Biological Process (BP) enrichments calculated with GREAT for enhancer sets in the same context. Pairwise similarity for GO MF terms for enhancer sets in K562 (A), Gm12878 (B), liver (C), and heart (D). Pairwise similarity for GO Biological Process (BP) for enhancer sets based on JEME's putative mappings to target genes in K562 (E), Gm12878 (F), liver (G), and heart (H). The upper triangle shows the semantic similarity calculated using GoSemSim, and the lower triangle shows the number of shared terms of the top 30 most significantly enriched. Scores for the BP ontology are noticeably lower than those for the MF ontology. There are few significantly enriched terms for genes mapped from heart enhancers.

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 multidimensional scaling (MDS) based on their Jaccard similarity in genomic space and the GO term functional similarity of predicted target genes.

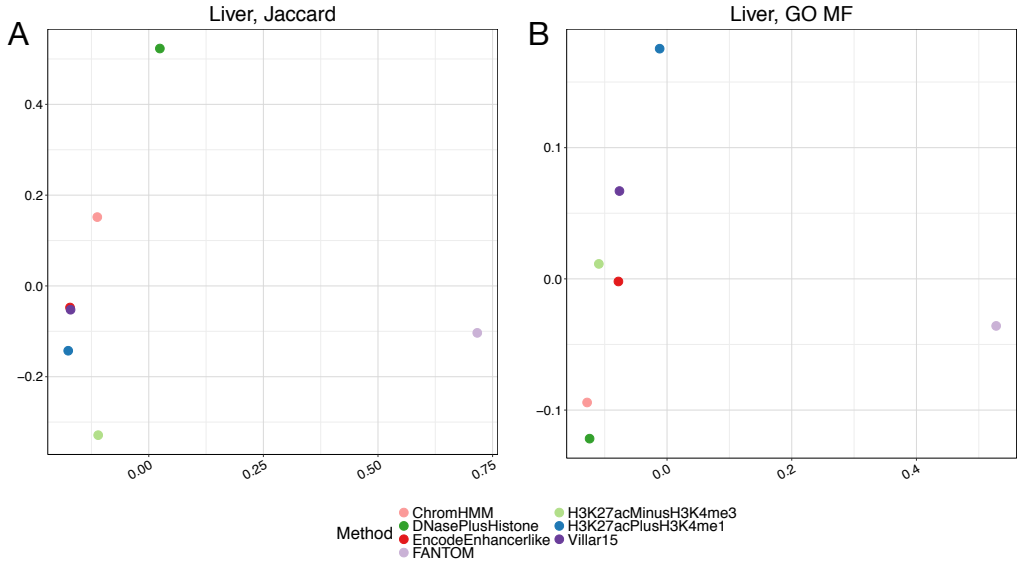

Figure 13: Multidimensional scaling (MDS) projections of enhancer sets. (A) MDS plot of liver enhancer sets based on the Jaccard similarity of the genomic distributions. (B) MDS plot for liver enhancers based on distances calculated from molecular function (MF) Gene Ontology (GO) term semantic similarity values with GREAT.

Several trends emerged from analyzing the genomic and functional distributions within and between biological contexts. First, the FANTOM eRNA enhancers are consistently distinct from all other

enhancer sets in both their genomic distribution and functional associations (Figure 13, Figure 14, Figure 15). Differences between eRNA and non-eRNA enhancer sets appear to dominate any other variation introduced by biological, technical, or methodological differences. Second, 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 13, Figure 14), they are located far from those sets in the functional comparisons and hierarchical clustering (Figure 13, Figure 14). 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 15). 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 15). 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.

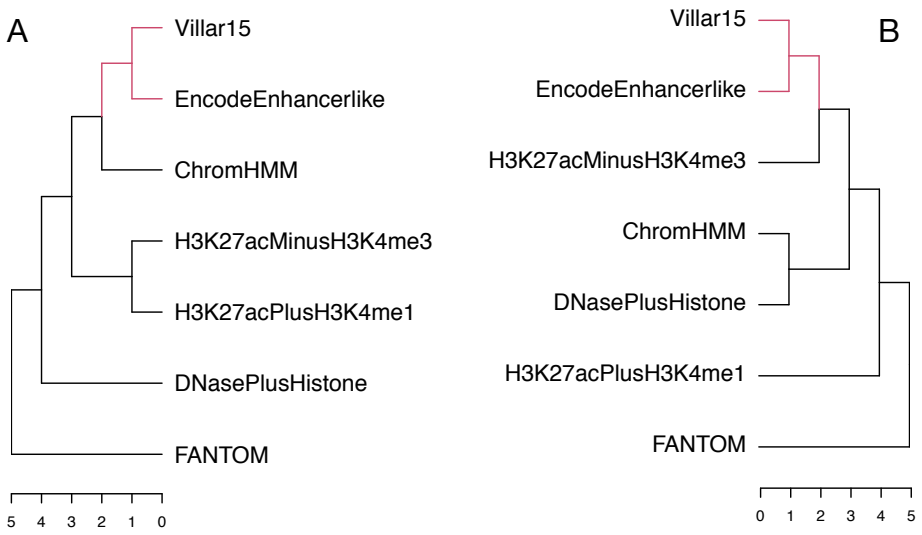

Figure 14: Ranked hierarchical clustering of enhancer sets. Clustering based on the Jaccard similarities of the genomic distributions (A) of all liver enhancer sets compared to clustering based on GO semantic similarity (B). FANTOM enhancers are the most distant from all other enhancer sets in both genomic and functional similarity, but the relationships between other sets are not conserved. Red branches denote identical subtrees within the hierarchy.

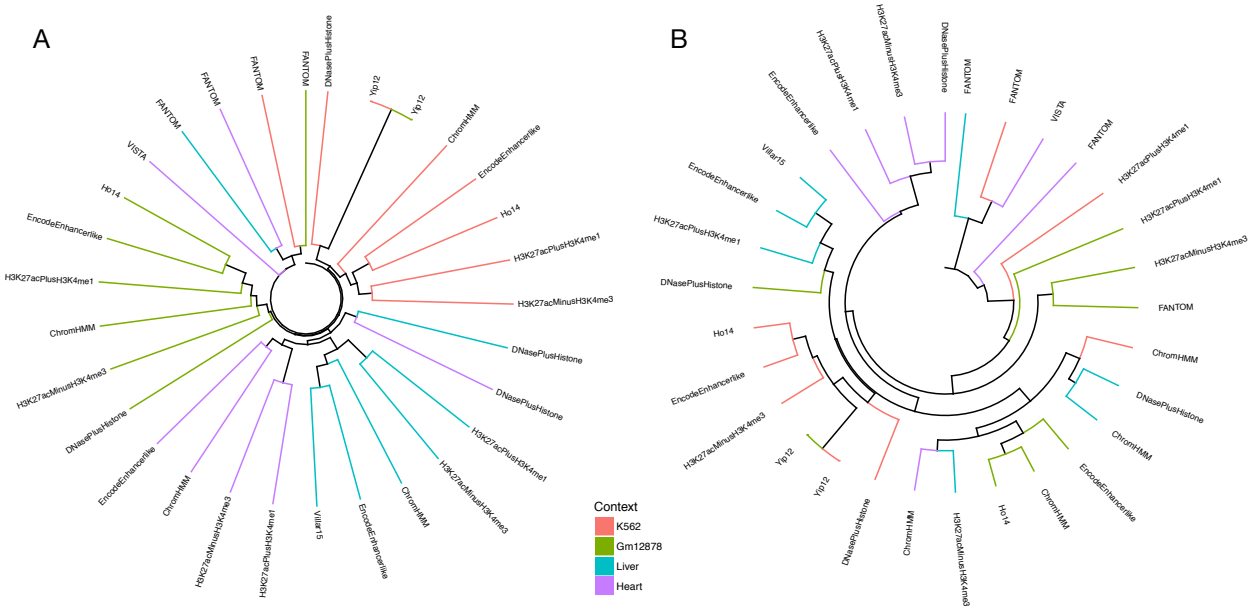

Figure 15: Hierarchical clustering of enhancer sets across biological contexts. (A) Hierarchical clustering based on genomic Jaccard distances for all methods and all contexts. (B) Hierarchical clustering of all available enhancer sets based on GO term distances. Terminal branches are colored by biological 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.

Conclusion

Although all enhancer sets considered here display enrichment for relevant functional attributes, the level of enrichment varies between sets. Additionally, when multiple gene-target mapping strategies are employed, enhancer sets identified by different methods are enriched for functional terms with little overlap and low similarity. This chapter demonstrates significant functional differences between enhancer sets that are substantial enough to influence biological interpretations and conclusions about disease-associated variant in enhancer regions. These results have implications for the interpretation of enhancer annotations and the use of enhancer sets for the interpretation of disease or expression associated genetic variants.

CHAPTER IV

Assessing Performance of Integrated Enhancer Identification Methods

Introduction

The work in Chapters II and III characterizes the differences between enhancer sets in both genomic and functional space These disagreements show that enhancer identification strategies result in putative enhancer annotations in disparate locations throughout the genome associated with different biological functions. Chapter IV builds on these observations to explore potential solutions. We begin by looking at enrichment of the nine representative enhancer sets with experimentally validated sequences from small-scale transgenic reporter assays and a high-throughput MPRA. The chapter concludes with an analysis of the enrichment of regions predicted by multiple enhancer identification strategies for functional proxies and higher confidence scores.

Methods

Experimentally validated enhancer sets: VISTA and Sharpr-MPRA

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)[72]. 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)[77]. 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[77]. Regions were selected for the tiling Sharpr-MPRA based on previous evidence of regulatory function in one of four cell lines, including

K562. Evidence included DNase-seq peaks and enhancer states from a 25-state ChromHMM model trained on histone modifications, DHSs, CTCF and RNA polymerase II.

To evaluate the ability of different methods to distinguish VISTA or MPRA positives from negatives, we computed the relative enrichment for positive/activating regions vs. negative/repressive regions. A positive value indicates higher enrichment for enhancers with demonstrated activity in the relevant context, and a negative value indicates more enrichment for non-enhancers or repressive base pairs. Equal enrichment in both sets yields a score of 0.

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 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.

Results

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[72] and thousands more have been tested in cell lines via massively parallel reporter assays (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 (Table 11), and for each annotated K562 enhancer with 15,720 regions tested in K562 cells by Sharpr-MPRA[77].

Context	Enhancer Set	Observed VISTA Positive Overlaps	Observed VISTA Negative Overlaps
Heart	H3K27acPlusH3K4me1	19	79
Heart	H3K27acMinusH3K4me3	36	152
Heart	DNasePlusHistone	25	106
Heart	ChromHMM	71	168
Heart	EncodeEnhancerlike	87	179
Heart	FANTOM	17	5

Table 11: Number of VISTA enhancer overlaps.

All heart enhancer sets are significantly enriched for overlap with the 126 VISTA heart positives (Figure 16; $p < 0.001$ 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. The FANTOM set is 16x enriched; however,

given its smaller size, this was based on 17 overlaps compared to an expected ~0. 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. The methods in heart demonstrate some ability to distinguish between the positives and negatives; however, there are only small differences in relative enrichment between the histone derived enhancer sets. Overall, FANTOM heart enhancers have the highest enrichment for experimentally validated enhancers relative to the negative set, but again we note that the FANTOM results are based on relatively small numbers of enhancers (Supplementary Table 3)

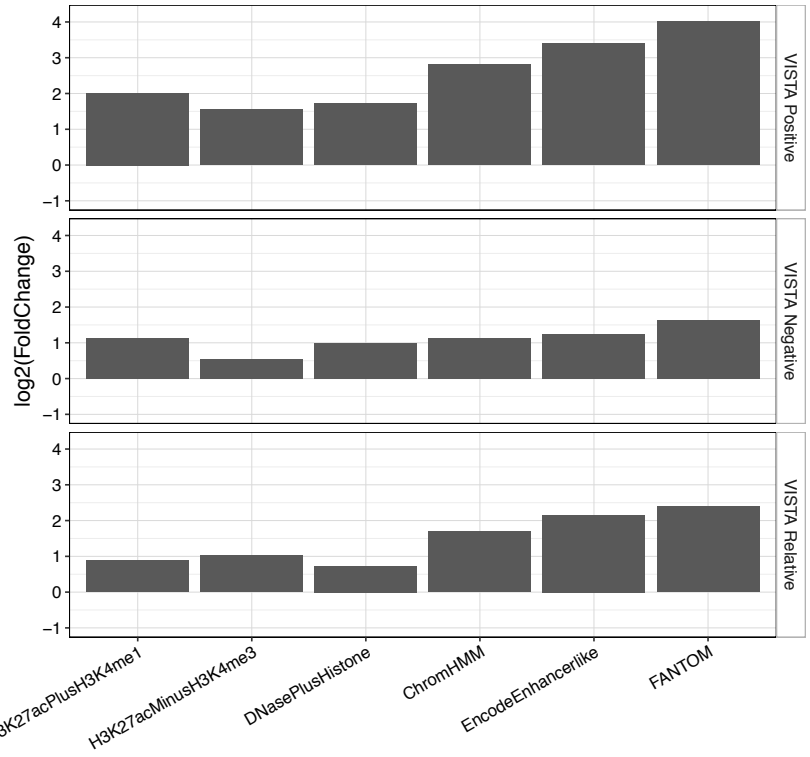

Figure 16: Enrichment for VISTA enhancers in heart. (A) Plot of the element-wise enrichment for 126 positive VISTA heart enhancers (upper panel) and 882 negative VISTA regions (middle panel). All heart enhancer sets are significantly enriched for overlap with the VISTA positives ($p < 0.001$ for all), and each set is at least ~3x more likely to overlap validated enhancers than expected if it was randomly distributed across the genome. The bottom panel is the $\log_2$ of the relative enrichment ratio for heart enhancer sets with VISTA heart positives compared to VISTA negatives.

Furthermore, 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 (Figure 17). Of the top three relatively enriched methods, nearly 40% (41/104) of the VISTA heart positives identified by the top methods are unique to one method (Figure 17A). Perhaps more striking, out of the validated VISTA heart enhancers 17% are identified by a single method. Less than 5% of the positives are identified by all methods, suggesting that different methods identify different subsets of validated enhancers (Figure 17B).

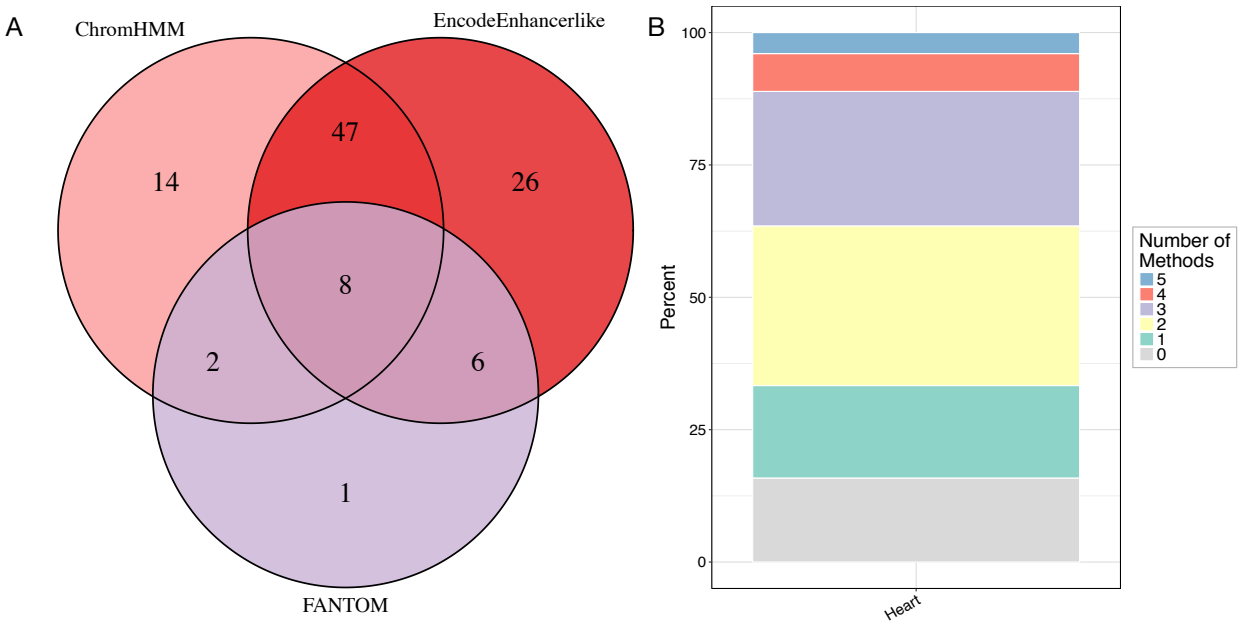

Figure 17: Enhancer sets share relatively few overlaps with the same VISTA enhancers. (A) Venn diagram of shared VISTA positives between enhancer sets with the top three relative enrichment methods. (B) Stacked bar chart showing the proportion of VISTA positives overlapped by zero or more enhancer sets.

Similarly, all of the enhancer sets in K562 are significantly enriched for overlap with both activating and repressive regions characterized by Sharpr-MPRA (Figure 18; $p < 0.001$). There is little variation between the methods in terms of overall enrichment, with most being ~8x enriched for activating regions. FANTOM has the highest relative enrichment (4.2x; $p < 0.001$). Overall, the enrichment values for the Sharpr-MPRA activating regions are greater than those in the VISTA analysis.

This could be due to the heterogenous nature of primary tissue samples, like heart, compared to cell lines, like K562. It is also likely related to the way regions were selected for testing in the Sharpr-MPRA: DHS sites and ChromHMM enhancer states from a 25-state model trained on ENCODE data[77]. We also note that the enrichment for repressive regions is also higher, leading to relative enrichment values that are similar in magnitude to the VISTA enhancers.

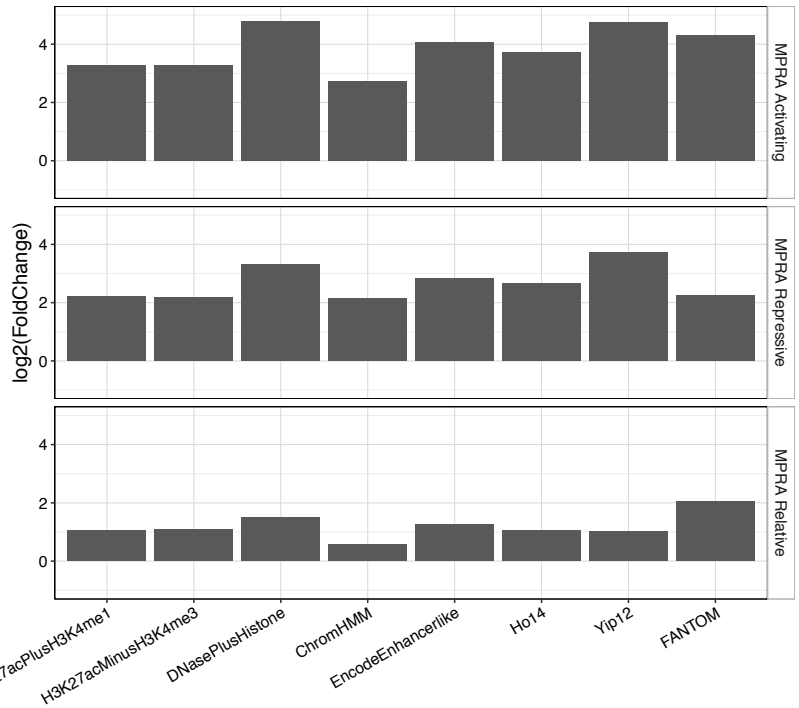

Figure 18: Enrichment for Sharpr-MPRA activating and repressive regions. (A) Enrichment (log_2 fold change) for activating (top) and repressive (middle) regions as defined by the Sharpr-MPRA assay in K562. The bottom panel shows the relative enrichment for activating regions of each enhancer set.

Many of the MPRA validated regions are not shared between the top three enriched sets (FANTOM, DNasePlusHistone, EncodeEnhancerlike). Among the activating regions identified by the top three methods, 74% are unique to a single set and only 9 are shared by all three (Figure 19A). Nearly half of the activating regions in the MPRA (49%; 2,606 / 5,373) were not identified by any of the enhancer sets, and 40% of activating regions overlapping a predicted enhancer are unique to a single set (Figure 19B; 1,098 / 2,747). Thus, comparison 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. Additionally, like lower-throughput reporter assays, MPRA approaches also suffer from inaccuracies induced by experimental variation, length restrictions on the tested sequences, and removal of the tested element from its endogenous context[11].

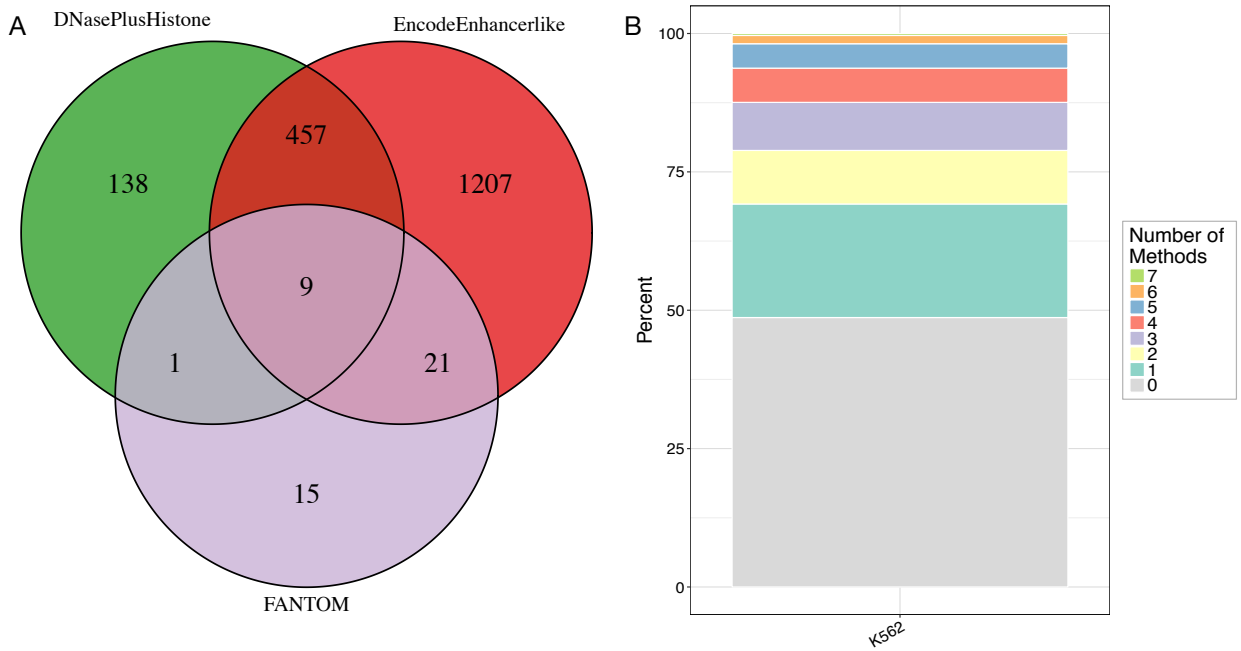

Figure 19: Number of activating Sharpr-MPRA regions overlapped by enhancer sets. (A) Venn diagram of shared Sharpr-MPRA activating regions from the top 3 relatively enriched sets. (B) Stacked bar chart showing the proportion of Sharpr-MPRA activating regions overlapped by zero or more enhancer sets.

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 20). 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 20A). 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; however, given the relatively small number of GWAS SNP overlaps, none of these differences were statistically significant (Figure 20B). We observed no increase in the enrichment for overlap with eQTL as the support for enhancer activity increased (Figure 20C). 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 little evidence for increased functional relevance.

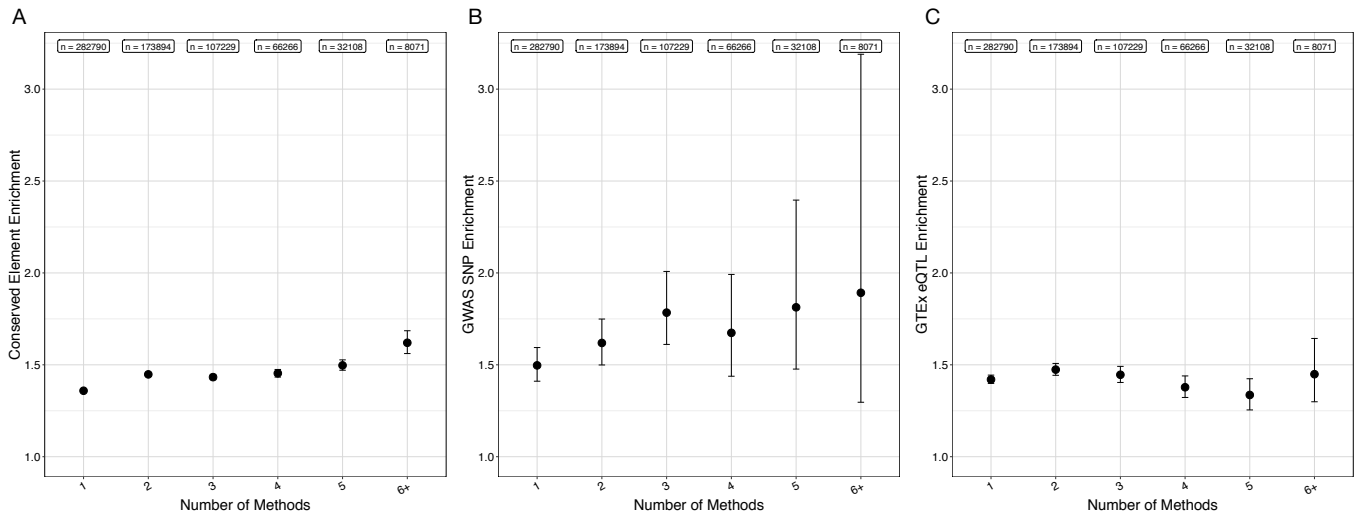

Figure 20: Enrichment for signals of functional importance in shared enhancer regions. (A) Enrichment for overlap between conserved elements (n = 3,930,677) and liver enhancers stratified by the number of identification methods that predicted each enhancer. (B) Enrichment for overlap between GWAS SNPs (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.

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 21-24). 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 (Figure23A-C, Figure 24A-C). While possibly a sign of poor specificity in shared enhancer regions, this trend may also be explained by transcription factor binding within or near histone acetylation sites. Transcription factor binding has been previously associated with the local minima of acetylation ChIP-seq binding profiles, so the decrease in peak signals may indicate that some of these shared regions are correlated with TF binding activity[106]. The lack of increase in enhancer score with the number of methods

supporting it held across all methods tested, providing further evidence that building enhancer sets by simple combinations of existing methods is unlikely to lead to a higher confidence subset (Figure 21-24). Similarly, filtering based on simple agreement between methods may not improve the specificity of enhancer predictions.

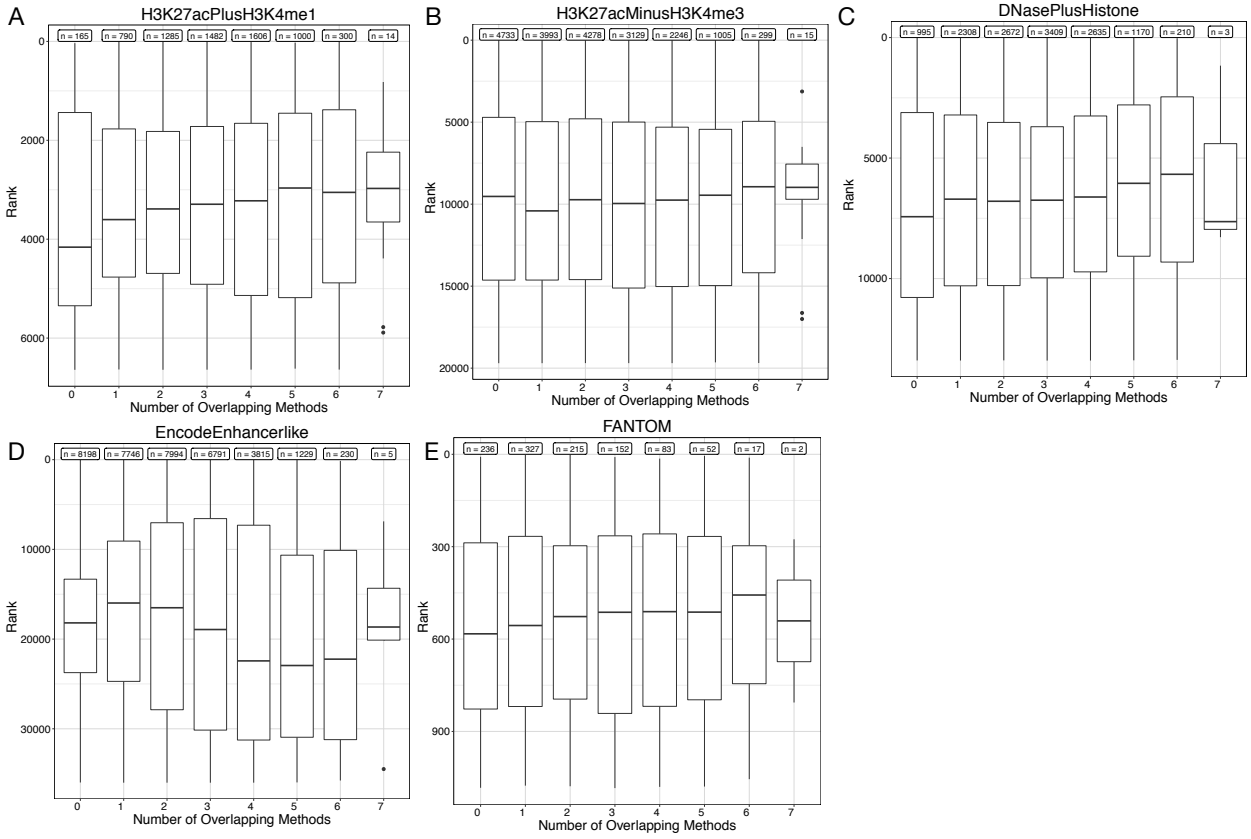

Figure 21: Confidence distributions for K562 enhancer sets. Score distributions for K562 enhancer sets are similar between regions identified as enhancers by a single method and those identified by multiple methods: (A) H3K27acPlusH3K4me1, (B) H3K27acMinusH3K4me3, (C) DNasePlusHistone, (D) EncodeEnhancerlike, and (E) FANTOM.

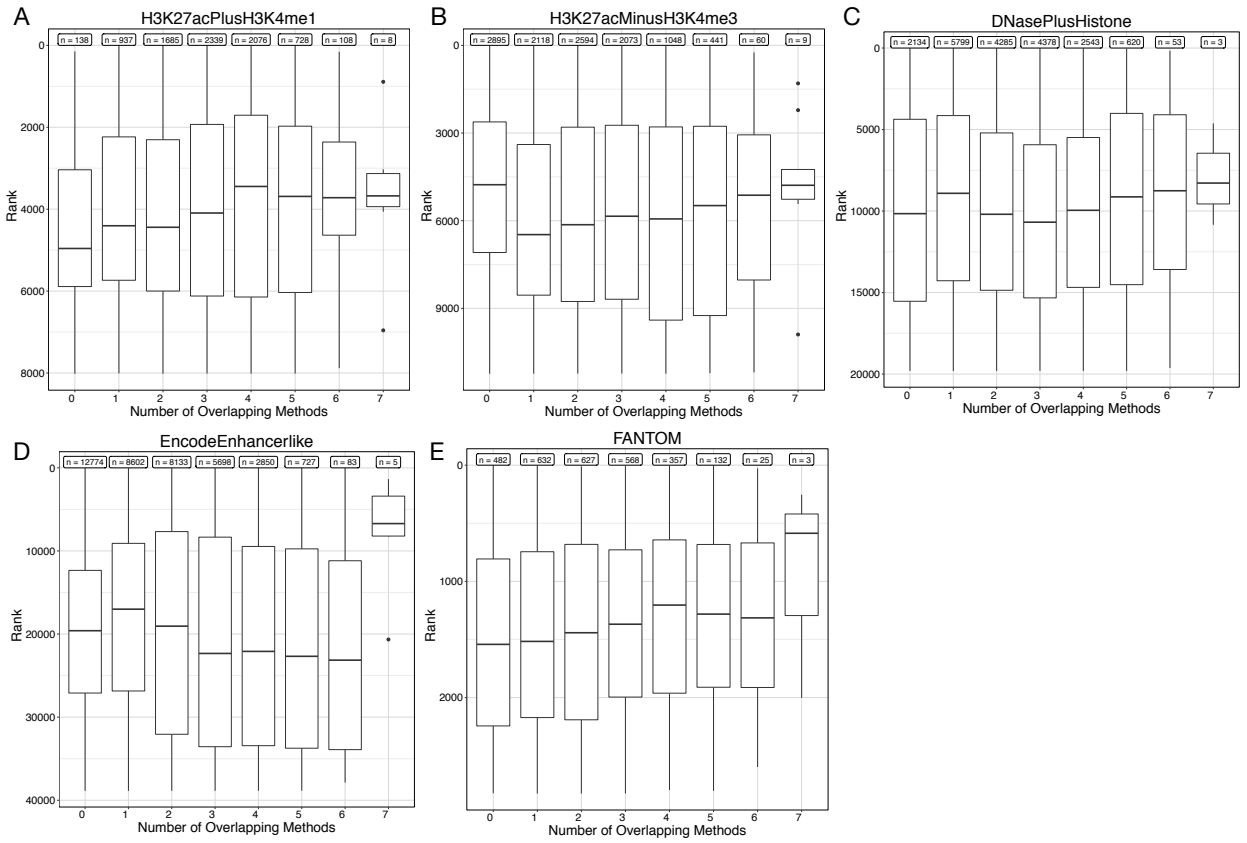

Figure 22: Confidence distributions for Gm12878 enhancer sets. Score distributions for Gm12878 enhancer sets are similar between regions identified as enhancers by a single method and those identified by multiple methods: (A) H3K27acPlusH3K4me1, (B) H3K27acMinusH3K4me3, (C) DNasePlusHistone, (D) EncodeEnhancerlike, and (E) FANTOM.

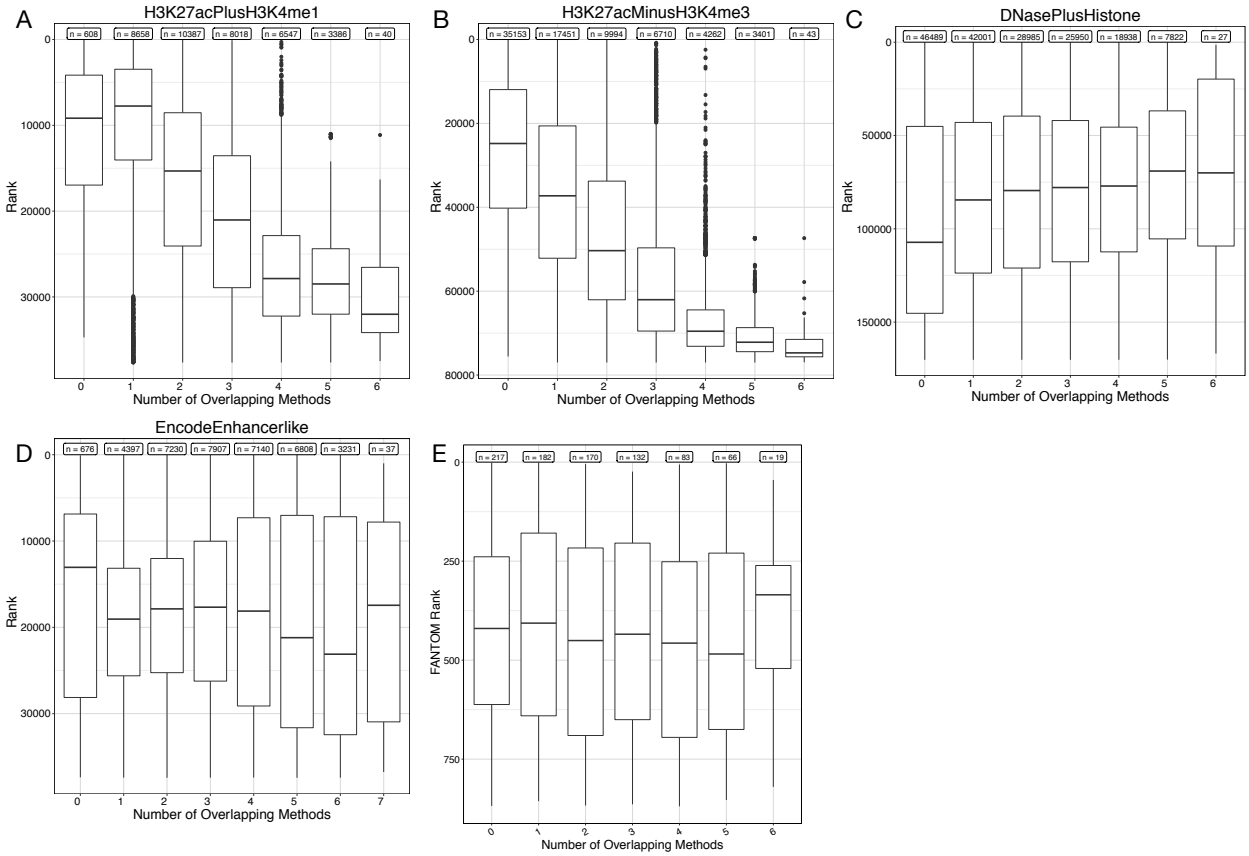

Figure 23: Confidence distributions for liver enhancer sets. The confidence distributions for regions shared between multiple enhancer sets are similar to the confidence distributions of regions unique to a single set: (A) H3K27acPlusH3K4me1, (B) H3K27acMinusH3K4me3, (C) DNasePlusHistone, (D) EncodeEnhancerlike, and (E) FANTOM.

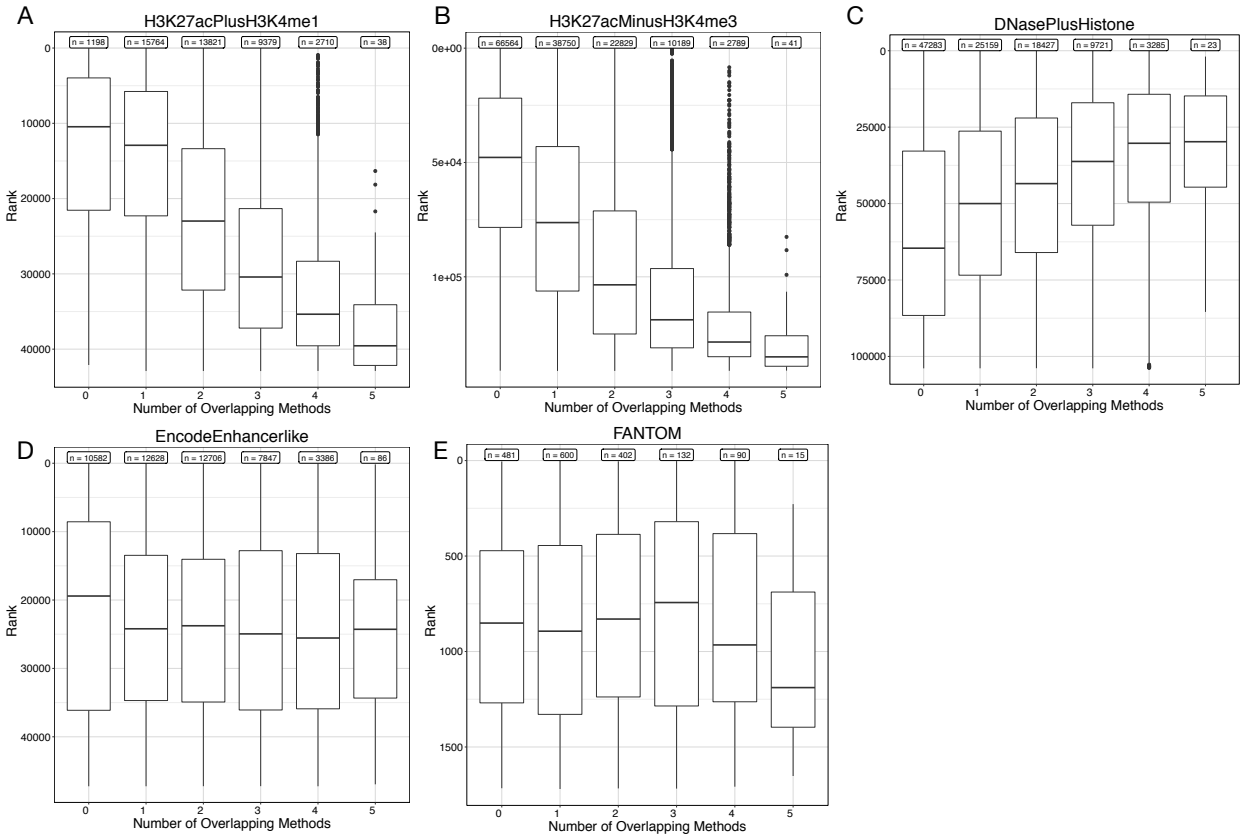

Figure 24: Confidence distributions for heart enhancer sets. The confidence distributions for regions shared between multiple enhancer sets are similar to the confidence distributions of regions unique to a single set: (A) H3K27acPlusH3K4me1, (B) H3K27acMinusH3K4me3, (C) DNasePlusHistone, (D) EncodeEnhancerlike, and (E) FANTOM. As in liver enhancer sets (Figure 23), in some cases (A-B) the median score decreases as the regions are more highly shared. This trend may be a result of poor specificity, or is potentially a sign of transcription factor binding activity in the region

Conclusion

Chapter IV shows that, while each enhancer strategy is enriched for experimentally validated enhancers, each is also enriched for overlap with confirmed negatives. This analysis begins to quantify the false positive and negative rates for different enhancer sets; however, without a gold standard we are unable to definitively quantify those values. Surprisingly, filtering enhancer sets to only include high confidence or highly shared predictions does not increase evidence of function. This suggests that simple combinations of existing enhancer sets will not sufficiently improve predictions for many practical applications.

CHAPTER V

Discussion

Accurate enhancer identification is a challenging problem, and recent efforts have produced a variety of experimental and computational approaches. Each method, either explicitly or implicitly, represents a different perspective on what constitutes an enhancer and which genomic features are most informative about enhancer activity. The lack of comprehensive genome-wide "gold standard" enhancer sets makes comparisons and evaluation challenging. Thus, 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.

Because enhancer identification strategies have such substantial differences, one strategy cannot and should not be used as a proxy for another. Using different strategies can yield substantially different biological interpretations and conclusions, e.g., about the gene regulatory potential of a SNP or the degree of evolutionary constraint on enhancers. This is particularly important, given that studies of gene regulation commonly use only a single approach to identify enhancers. For example, 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[48,49,62,64,65,83,105,107]. Thus, our findings 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[45,108,109].

Our main goal was to evaluate the congruence of the diverse strategies in use today. Given their differences in assumptions, motivations, and protocols, it is not surprising that different assays and algorithms identify somewhat different sets of enhancers. We would expect such diversity in enhancer definitions to produce different sets of annotations, despite the use of identical terminology to describe these regions in the literature. Technical biases in the underlying experimental assays or data processing pipelines may lead to variation between putative enhancer sets preventing high levels of agreement. However, comparisons between biological replicates of histone modification ChIP-seq data suggests that the level of difference we observe between enhancer sets is greater than this potential technical bias. Individual genetic variation may 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 leading to variation in enhancer activity between individuals have been identified[110,111]. However, the proportion of epigenetic modifications that are variable across individuals is estimated to be small (1–15%)[55], and thus 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. Taking these limitations into account, the differences we observe remain striking.

The consistent lack of agreement between methods demonstrates that many working definitions of "enhancer" have 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. 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. In general, enhancers defined by FANTOM have modestly more enrichment for proxies of functional activity than other methods, but this comes at the expense of low sensitivity. Methods derived from combinations of histone modifications and chromatin accessibility profiles generate large and likely inclusive sets of candidate regions. However, since the input attributes are also correlated with genomic elements other than enhancers, the specificity of such methods may suffer. Integrative machine learning models could

reduce the noise in a prediction set, although it is difficult to quantify the true prediction accuracy with current validation approaches.

In light of this complexity, what should we do? First, we must resist the convenience of ignoring it. 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. The most appropriate identification strategy is likely dependent on a number of application-specific factors, preventing simple 'one-size-fits-all' recommendations. To facilitate exploration of different strategies, we developed creDB, a comprehensive, easily queried database of over 3.5 million putative enhancer annotations. 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. Indeed, we find little evidence that regions with higher levels of agreement between identification strategies are more enriched for functional signals. Ultimately, we need more sophisticated statistical models of enhancers and their properties 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[57], but obtaining a concordant and functionally relevant set of enhancers remains challenging. We are hopeful that new experimental techniques and biologically motivated machine learning methods for integrating different definitions of enhancers will yield more consistent and specific annotations of regions with regulatory functions.

Second, our study 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[2,19,20,39,40,45,112]. 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 some appreciation of this subtlety, there is still a need for more precise terminology and improved statistical and functional models of the diversity of *cis*-regulatory "enhancer-like" sequences. Given this diversity, we should not expect all results to be robust to the enhancer identification strategy used.

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 likely 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[42,113,114]. Consistent and specific definitions of the spectrum of regulatory activity and architecture are necessary for further progress in enhancer identification, successful replication, and accurate genome annotation. In the interim, we must remember that genome-wide enhancer sets generated by current approaches should be treated as what they are—incomplete snapshots of a dynamic process.

APPENDIX

List of Relevant Phenotypes in Liver

Aspartate aminotransferase
Autoimmune hepatitis type-1
Biliary atresia
Bilirubin levels
Bilirubin levels in extreme obesity
Butyrylcholinesterase levels
CYP3A4 enzyme activity
Drug-induced liver injury
Drug-induced liver injury (amoxicillin-clavulanate)
Drug-induced liver injury (flucloxacillin)
Gamma gluatamyl transferase levels
Gamma gluatamyl transferase levels (interaction with age)
Gamma glutamyl transpeptidase
Gaucher disease severity
Hematological and biochemical traits
Hematology traits
Hepatitis
Hepatitis B
Hepatitis B (viral clearance)
Hepatitis B vaccine response
Hepatitis C induced liver cirrhosis
Hepatitis C induced liver fibrosis
Hepatocellular carcinoma
Hepatocellular carcinoma (hepatitis B virus related)
Hepatocellular carcinoma
Hepatocellular carcinoma (hepatitis B virus related)
IFN-related cytopenia
Lapatinib-induced hepatotoxicity
Lipid levels in hepatitis C treatment
Liver disease in chronic hepatitis B virus infection
Liver enzyme levels
Liver enzyme levels (alanine transaminase)
Liver enzyme levels (alkaline phosphatase)
Liver enzyme levels (aspartate transaminase)
Liver enzyme levels (gamma-glutamyl transferase)
Lumiracoxib-related liver injury
Non-albumin protein levels
Non-alcoholic fatty liver disease
Non-alcoholic fatty liver disease histology (AST)
Non-alcoholic fatty liver disease histology (lobular)
Non-alcoholic fatty liver disease histology (other)
Nonalcoholic fatty liver disease
Primary biliary cirrhosis
Primary sclerosing cholangitis
Response to hepatitis C treatment
Response to protease inhibitor treatment in hepatitis c (bilirubin toxicity)

Response to protease inhibitor treatment in hepatitis c (peak serum total bilirubin levels)
Serum albumin level
Serum alkaline phosphatase levels
Total bilirubin levels in HIV-1 infection

List of Relevant Phenotypes in Heart

AR-C124910XX levels in individuals with acute coronary syndromes treated with ticagrelor
Abdominal aortic aneurysm
Aortic root size
Aortic stiffness
Aortic-valve calcification
Arterial stiffness
Arterial stiffness (pulse-wave velocity)
Atrial Septal Defect
Atrial fibrillation
Atrial fibrillation/atrial flutter
Atrioventricular conduction
Atrioventricular septal defects in Down syndrome
B-type natriuretic peptide
Blood pressure
Blood pressure (age interaction)
Blood pressure (anthropometric measures interaction)
Blood pressure (response to antihypertensive medication)
Blood pressure (smoking interaction)
Blood pressure measurement (cold pressor test)
Blood pressure measurement (high sodium and potassium intervention)
Blood pressure measurement (high sodium intervention)
Blood pressure measurement (low sodium intervention)
Blood pressure response to hydrochlorothiazide in hypertension
Blood pressure variability
Brugada syndrome
Cardiac Troponin-T levels
Cardiac hypertrophy
Cardiac muscle measurement
Cardiac repolarization
Cardiac structure and function
Cardio vascular disease (drug interaction; BB)
Cardioembolic ischaemic stroke
Cardiovascular disease (drug interaction, BB)
Cardiovascular disease (drug interaction, CCB)
Cardiovascular disease (drug interaction, diuretics)
Cardiovascular disease (drug interaction; ACE)
Cardiovascular disease risk factors
Cardiovascular heart disease in diabetics
Carotid artery intima media thickness (sex interaction)
Carotid atherosclerosis (smoking interaction)
Carotid atherosclerosis in HIV infection
Carotid intima media thickness
Carotid plaque burden (smoking interaction)
Cervical artery dissection
Chagas cardiomyopathy in Tripanosoma cruzi seropositivity
Cholesterol
Cholesterol and Triglycerides
Cholesterol, total

Circulating vasoactive peptide levels
Clozapine-induced agranulocytosis
Clozapine-induced cytotoxicity
Congenital heart disease
Congenital heart malformation
Congenital left-sided heart lesions
Congenital left-sided heart lesions (maternal effect)
Conotruncal heart defects
Coronary arterial lesions in patients with Kawasaki disease
Coronary artery calcification
Coronary artery calcification (smoking interaction)
Coronary artery disease
Coronary artery disease or ischemic stroke
Coronary artery disease or large artery stroke
Coronary artery disease-related phenotypes
Coronary heart disease
Coronary heart disease event reduction in response to statin therapy (interaction)
Coronary heart disease in familial hypercholesterolemia
Coronary restenosis
Coronary restenosis
Coronary spasm
Cystatin C
Dilated cardiomyopathy
Drug-induced torsades de pointes
Echocardiographic traits
Electrocardiographic conduction measures
Electrocardiographic traits
Factor VII
Factor VII levels
Factor VIII levels
Factor XI
HDL cholesterol
Heart failure
Heart rate
Heart rate variability traits
Hemostatic factors and hematological phenotypes
Hypertension
Hypertension (pulmonary)
Hypertension (young onset)
Hypertension risk in short sleep duration
Hypertrophic cardiomyopathy
IgE levels
Ischemic stroke
LDL (oxidized)
LDL cholesterol
LDL cholesterol subfractions
LDL peak particle diameter (total fat intake interaction)
Large artery atherosclerosis ischaemic stroke
Large artery stroke
Left ventricular mass
Life threatening arrhythmia

Lipoprotein-associated phospholipase A2 activity and mass
Lipoprotein (a) - cholesterol levels
Lp (a) levels
Major CVD
Mitral annular calcification
Mitral valve prolapse
Mortality among heart failure patients
Mortality in heart failure
Myocardial infarction
Myocardial infarction (drug interaction; ACE)
Myocardial infarction (drug interaction; BB)
Myocardial infarction (drug interaction; CCB)
Myocardial infarction (drug interaction; diuretics)
Myocardial infarction (early onset)
Myocardial infarction in coronary artery disease
Nonobstructive coronary artery disease
Oleic acid (18:1n-9) plasma levels
P wave duration
PR interval
PR interval in Tripanosoma cruzi seropositivity
PR segment
Palmitic acid (16:0) plasma levels
Palmitoleic acid (16:1n-7) plasma levels
Pericardial fat
Perioperative myocardial infarction in coronary artery bypass surgery
Peripartum cardiomyopathy
Plasma cystastin c levels in acute coronary syndrome
Plasma omega-6 polyunsaturated fatty acid levels (adrenic acid)
Plasma omega-6 polyunsaturated fatty acid levels (arachidonic acid)
Plasma omega-6 polyunsaturated fatty acid levels (dihomo-gamma-linolenic acid)
Plasma omega-6 polyunsaturated fatty acid levels (gamma-linolenic acid)
Plasma omega-6 polyunsaturated fatty acid levels (linoleic acid)
Postoperative atrial fibrillation in coronary artery bypass grafting surgery
Postoperative ventricular dysfunction
Pulse pressure in young-onset hypertension
QRS duration
QRS duration in Tripanosoma cruzi seropositivity
QT interval
QT interval (interaction)
QT interval in Tripanosoma cruzi seropositivity
RR interval (heart rate)
Red blood cell count
Red blood cell fatty acid levels
Red blood cell traits
Renal sinus fat
Response to Dalcetrapib treatment in acute coronary syndrome
Response to rate control therapy in atrial fibrillation
Response to statin therapy
Response to statin therapy (LDL cholesterol subfractions)
Response to statin therapy (LDL-C)
Response to statins (LDL cholesterol change)

Resting heart rate
Serum dimethylarginine levels (asymmetric/symmetric ratio)
Sick sinus syndrome
Stearic acid (18:0) plasma levels
Subclinical atherosclerosis traits (other)
Sudden cardiac arrest
Symmetrical dimethylarginine levels
Tetralogy of Fallot
Thoracic aortic aneurysms and dissections
Ticagrelor levels in individuals with acute coronary syndromes treated with ticagrelor
Triglycerides
Vascular constriction
Vein graft stenosis in coronary artery bypass grafting
Venous thromboembolism
Venous thromboembolism (SNP x SNP interaction)
Ventricular conduction
Ventricular fibrillation
vWF and FVIII levels
vWF levels

REFERENCES

1. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15,** 272–86 (2014).
2. 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).
3. Ong, C. & Corces, V. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nat. Rev. Genet.* **12,** 283–93 (2011).
4. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science (80-.).* **337,** 1190–1195 (2012).
5. Corradin, O. & Scacheri, P. C. Enhancer variants: evaluating functions in common disease. *Genome Med.* **6,** 85 (2014).
6. Sholtis, S. J. & Noonan, J. P. Gene regulation and the origins of human biological uniqueness. *Trends Genet.* **26,** 110–118 (2010).
7. 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
8. Mack, K. L. & Nachman, M. W. Gene Regulation and Speciation. *Trends Genet.* **33,** 68–80 (2016).
9. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14,** 288–95 (2013).
10. Kleftogiannis, D., Kalnis, P. & Bajic, V. B. Progress and challenges in bioinformatics approaches for enhancer identification. *Brief. Bioinform.* **17,** 967–979 (2016).
11. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* **27,** 38–52 (2017).
12. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106,** 159–164 (2015).

13. Crawford, G. E. *et al.* Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16,** 123–131 (2006).

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

15. 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).

16. Giresi, P. G. & Lieb, J. D. in *Tag-Based Next Generation Sequencing* 243–255 (2012). doi:10.1002/9783527644582.ch14

17. 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).

18. 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).

19. 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).

20. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).

21. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160,** 554–566 (2015).

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

23. Pennacchio, L. a *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444,** 499–502 (2006).

24. Visel, A. *et al.* Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40,** 158–160 (2008).

25. Bejerano, G. *et al.* Ultraconserved Elements in the Human Genome. *Science (80-.).* **1321,** 1321–1326 (2007).

26. 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).

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

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

29. 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

30. 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

31. 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

32. 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

33. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–61 (2014).

34. 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).

35. 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

36. 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

37. Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of

chromatin interactions. *Nature* (2012). doi:10.1038/nature11082

38. 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).

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

40. Pradeepa, M. M. *et al.* Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat. Genet.* **48,** 681–686 (2016).

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

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

43. 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).

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

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

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

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

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

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

50. 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).

51. Pekowska, A. *et al.* H3K4 tri-methylation provides an epigenetic signature of active enhancers. *EMBO J.* **30,** 4198–4210 (2011).

52. 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).

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

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

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

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

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

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

59. Hoffman, M. M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9,** 473–6 (2012).

60. 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).

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

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

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

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

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

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

67. Ho, J. W. K. *et al.* Comparative analysis of metazoan chromatin organization. *Nature* **512,** 449–52 (2014).

68. 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).

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

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

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

72. 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).

73. 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).

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

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

76. 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).

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

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

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

80. The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **489,** 57–74 (2012).

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

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

83. 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).

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

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

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

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

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

89. 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).

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

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

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

93. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

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

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

96. 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).

97. 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).

98. 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).

99. 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).

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

101. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30,** 1095–106 (2012).

102. Edwards, S. L., Beesley, J., French, J. D. & Dunning, M. Beyond GWASs: Illuminating the dark road from association to function. *American Journal of Human Genetics* **93,** 779–797 (2013).

103. Ashoor, H., Kleftogiannis, D., Radovanovic, A. & Bajic, V. B. DENdb: Database of integrated human enhancers. *Database* **2015,** (2015).

104. Gao, T. *et al.* EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics* btw495 (2016). doi:10.1093/bioinformatics/btw495

105. 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).

106. Ramsey, S. A. *et al.* Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics* **26,** 2071–2075 (2010).

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

108. 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).

109. 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

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

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

112. 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).

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

114. Andersson, R. Promoter or enhancer, what's the difference? Deconstruction of established

distinctions and presentation of a unifying model. *BioEssays* **37,** 314–323 (2015).