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

Mary Lauren Benton[1], Sai Charan Talipineni[2], Dennis Kostka[2,3*], and John A. Capra[1,4]*

[1] *Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, 37235, USA*

[2] *Department of Developmental Biology. University of Pittsburgh School of Medicine, Pittsburgh, PA, 15201, USA*

[3] *Department of Computational & Systems Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, 15201, USA*

[4] *Departments of Biological Sciences and Computer Science, Vanderbilt Genetics Institute, Center for Structural Biology, Vanderbilt University, Nashville, TN, 37235, USA*

** Corresponding authors:*

kostka@pitt.edu

tony.capra@vanderbilt.edu

ABSTRACT

Non-coding gene regulatory loci are essential to transcription in mammalian cells. As a result, a large variety of experimental and computational strategies have been developed to identify *cis*-regulatory enhancer sequences. However, in practice, most studies consider enhancer candidates identified by a single method alone. Here we assess the robustness of conclusions based on such a paradigm by comparing enhancer sets identified by different strategies. Because the field currently lacks a comprehensive gold standard, our goal was not to identify the best identification strategy, but rather to quantify the consistency of enhancer sets identified by ten representative identification strategies and to assess the robustness of conclusions based on one approach alone. We found significant dissimilarity between enhancer sets in terms of genomic characteristics, evolutionary conservation, and association with functional loci. This substantial disagreement between enhancer sets within the same biological context is sufficient to influence downstream biological interpretations, and to lead to disparate scientific conclusions about enhancer biology and disease mechanisms. Specifically, we find that different enhancer sets in the same context vary significantly in their overlap with GWAS SNPs and eQTL, and that the majority of GWAS SNPs and eQTL overlap enhancers identified by only a single identification strategy. Furthermore, we find limited evidence that enhancer candidates identified by multiple strategies are more likely to have regulatory function than enhancer candidates identified by a single method. The difficulty of consistently identifying and categorizing enhancers presents a major challenge to mapping the genetic architecture of complex disease, and to interpreting variants found in patient genomes. To facilitate evaluation of the effects of different annotation approaches on studies' conclusions, we developed a database of enhancer annotations in common biological contexts, creDB, which is designed to integrate into bioinformatics workflows. Our results highlight the inherent complexity of enhancer biology and argue that current approaches have yet to adequately account for enhancer diversity.

INTRODUCTION

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[1–3]. Enhancers play a vital role in the regulation of genes during development and cell differentiation[3,4], and genetic variation in enhancers has been implicated in etiology of complex disease[5,6] and in differences between closely related species[7–9].

Given their importance, enhancers have seen considerable study in recent years. More than 1,300 papers have been published on enhancer biology (MeSH: *Enhancer Elements, Genetic*) since the start of 2015, and more than 100 of these have focused on the role of enhancers in disease. Despite the importance of enhancers, they remain difficult to identify[1,10,11]. Experimental assays that directly confirm enhancer activity are time-consuming, expensive, and not always conclusive[1,12]. And, despite recent promising developments in massively parallel reporter assays, current methods are unable to definitively identify and test enhancers *in vivo* on a genome-wide scale[13]. As a result, many studies have identified more easily assayed attributes associated with enhancer activity, including DNA sequence, evolutionary conservation, and biochemical properties[1,12,14–16]. Active enhancers localize in regions of open chromatin, which is commonly assayed by testing the sensitivity of DNA segments to DNase I nuclease followed by sequencing (DNase-seq)[17,18]. Enhancers also often have characteristic sets of histone modifications on surrounding nucleosomes. For example, monomethylation of lysine 4 on histone H3 (H3K4me1) and lack of trimethylation of lysine 4 on histone H3 (H3K4me3) is used to distinguish enhancers from promoters, while acetylation of lysine 27 on histone H3 (H3K27ac) often denotes active enhancers[1,14,19]. Additionally, features such as evolutionary sequence conservation[20–22] and the presence of known transcription factor binding motifs[16] or known enhancer associated proteins, such as the histone acetyltransferase p300[14,23,24], have been used to successfully locate enhancer elements[1]. Furthermore some enhancers are transcribed, and it has become possible to map active enhancers by identifying characteristic bi-directionally transcribed enhancer RNAs (eRNAs)[25,26]. However, while informative, none of these attributes are comprehensive, exclusive to enhancers, or completely reliable indicators of enhancer activity. Furthermore, enhancer activity is context- and stimulus-dependent, creating an additional layer of complexity[19,27].

Many complementary computational enhancer identification methods that integrate these data in both supervised and unsupervised machine learning approaches have been developed[11,15,28]. Since there are no genome-wide gold standard enhancer sets, such 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 SNPs, evolutionary conservation, or proximity to relevant genes. The resulting 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 SNPs identified in genome-wide association studies (GWAS)[5,6,29]. However, in these applications, a single assay or computationally predicted enhancer set regularly serves as the working definition of an "enhancer" for all analyses.

We evaluated the robustness of this "single definition" approach by performing a comprehensive analysis of similarities in genomic, evolutionary, and functional attributes of enhancers identified by different strategies. By comparing the characteristics of different enhancer sets identified in the same biological context, we were able to assess the stability of conclusions made using only one enhancer identification strategy. While we expected some variation due to differences in the underlying assays, we found significant differences between enhancer sets identified in the same context. These differences between identification strategies are substantial enough to influence biological interpretations and conclusions about enhancer evolution and disease-associated variant function. Additionally, enhancers supported by multiple identification methods are not more likely to overlap markers of function, compared with enhancers identified only by a single strategy. These results highlight a fundamental challenge to evaluating the functional relevance of thousands of non-coding variants associated with disease from GWAS, and to studying gene regulatory mechanisms. Incorporating the unique characteristics of different enhancer identification strategies will be essential to ensuring reproducible results, and to furthering our understanding of enhancer biology. As a step toward this goal, we have created creDB, a large, easily queried database of enhancers identified by different methods across common tissues and cellular contexts.

RESULTS

A panel of enhancer identification strategies across four biological contexts

To evaluate the variation in enhancers identified by different genome-wide methods, we developed a consistent computational pipeline to compare enhancer identification strategies. Our approach is based on publically available data, and we applied it to a representative set of methods in four common cell types and tissues (biological contexts): K562,

Gm12878, liver, and heart cells (Figure 1 and Methods). Given the large number of enhancer identification strategies that have been proposed[1,11], it is not possible to compare them all; so for each context, we considered methods that represent the diversity of experimental and computational approaches in common use.

For all contexts, we considered three enhancer sets derived solely from chromatin immunoprecipitation followed by sequencing (ChIP-seq) for histone modifications informative about enhancer activity from the ENCODE Project[30]. The "H3K27ac" set includes all H3K27ac ChIP-seq peaks without additional refinement[31–35]. The "H3K27acPlusH3K4me1" set includes all H3K27ac ChIP-seq peaks that also overlap an H3K4me1 peak, and the "H3K27acMinusH3K4me3" set contains H3K27ac peaks that do not overlap an H3K4me3 peak[36–38]. In liver only, we have an additional set of enhancers identified from H3K27ac ChIP-seq peaks without overlapping H3K4me3 peaks for comparison ("Villar15")[38]. We also considered a method that incorporates DNase I hypersensitive sites (DHSs) with histone modifications to generate the "DNasePlusHistone" enhancer set, which is composed of DHSs where the ratio of H3K4me1 to H3K4me3 is less than one[39]. 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[25,40,41]. We also include several methods that combine machine learning with functional genomics data, such as the ENCODE consortium's "EncodeEnhancerlike" made by combining DHSs and H3K27ac peaks using an unsupervised ranking method and the "ChromHMM" predictions generated by a hidden Markov model trained on ChIP-seq data from eight histone modifications, CTCF, and RNA Pol II[15,42–44]. For the K562 and Gm12878 cell lines we also include enhancer predictions made by two supervised machine learning methods trained to identify enhancers based on ChIP-seq data in conjunction with other functional genomic features. We will refer to these sets as "Yip12" and "Ho14"[45,46]. An overview of the data and computational approaches used by each method is given in Figure 1 and full details are available in the Methods.

Enhancers identified by different strategies in the same context differ substantially

The genomic coverage of different enhancer sets varies by several orders of magnitude: Enhancer regions identified in the same context by different methods differ drastically in the number of enhancers identified, their genomic locations, their lengths, and their coverage of the genome (Table 1; Figure 2A; Supplementary Figure 1). 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, the enhancer sets identified differ by several orders of magnitude (Table 1). For instance, FANTOM identifies 326 kilobases (kb) of sequence with liver enhancer activity, EncodeEnhancerlike identifies 89 megabases (Mb), and H3K27acMinusH3K4me3 identifies almost 138 megabases (Mb). 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. The enhancer sets also vary in their relative distance to other functional genomic features, such as transcription start sites (TSSs). For example, in liver, the average distance to the nearest TSS ranges from 14 kb for EncodeEnhancerlike to 64 kb for DNasePlusHistone (Supplementary Table 1). Overall, methods based on histone modifications tend to identify larger numbers of longer enhancers compared with CAGE data, and machine learning methods are variable. We highlight these trends in liver, but they are similar in other contexts (Table 1).

Enhancer sets have low similarity: Given the diversity of the enhancer sets identified by different methods, we evaluated the extent of overlap between different sets. The magnitude of overlap between enhancer sets is typically low: less than 50% for nearly all pairs of methods (median 31% bp overlap for liver; Figure 3A). Furthermore, the largest overlaps are in comparisons including one set with high genome coverage, or in comparisons of sets that were identified based on similar data. To further quantify overlaps, we calculated the Jaccard similarity index— the number of shared bp between two enhancer sets divided by the size of their union—for each pair of methods. Overall, the Jaccard similarities are also extremely low, with an average of 0.17 and the majority of pairwise comparisons below 0.30 (Figure 3B, upper triangle). Since the Jaccard similarity is sensitive to differences in set size, we also computed a "relative" Jaccard similarity by dividing the observed value by the maximum value possible given the set sizes. The relative similarities were also consistently low (Figure 3B, lower triangle). In spite of the large differences between enhancer sets, all pairs of enhancer sets overlap more than expected if they were randomly distributed across the genome (Figure 3C,D, $p < 0.001$ for all paris). However, this result should not detract from the fact that most annotated enhancers are "singletons", annotated by only a single identification strategy. The comparisons described here are based on bp overlap, but trends are similar when evaluating the overlap on an element-wise basis (Supplementary Figure 2-3, Supplementary Table 2).

To assess the influence of technical variation on the observed overlaps, we compared the overlap of replicates from H3K27ac in the K562 and Gm12878 cell lines ChIP-seq data generated by in same laboratory. We expected the

replicates to have high of overlap and serve as an "upper bound" of similarity in practical applications. The replicates overlap 65–80% at the bp level. Thus, while there is some variation, the amount of overlap observed between enhancers identified by different methods almost always falls far below the variation between 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[47], we calculated the percentage of conserved base pairs present in each enhancer set. All enhancer sets have more conserved base pairs than expected from length-matched regions drawn at random from the genome. However, enhancers identified by some methods are more conserved than others (Figure 4A). Across each context, approximately 10% of the base pairs in histone-based, ChromHMM, and Ho14 enhancer sets are conserved. Adding DNaseI hypersensitivity data, as in the DNasePlusHistone and EncodeEnhancerlike sets, increases the level of conservation a few percent compared to solely histone-derived enhancers to ~14%. In contrast, the FANTOM and Yip12 enhancers have nearly twice as many conserved base pairs as the histone-based sets (17.8–21.6%). This is similar to levels of conservation among experimentally validated enhancers from the VISTA database (20.5%)[48]. Evolutionary conservation was not directly considered in the definition of the FANTOM and Yip12 sets; however, VISTA explicitly considered conservation in the selection of some regions to test.

Enhancer sets have different overlaps with validated enhancers: Nearly two thousand human sequences have been tested for enhancer activity *in vivo* in transgenic mice at E11.5 by VISTA[49]. Of 1757 tested human regions, 925 were validated as enhancers (VISTA positives), and 832 did not exhibit enhancer activity (VISTA negatives). All genome-wide liver enhancer sets are significantly enriched for overlap with the VISTA positives (Figure 4B; $p < 0.001$ for all), and with the exception of FANTOM, each set is ~2x more likely to overlap validated enhancers than expected if they were randomly distributed across the genome. The FANTOM set is 7x enriched; however, given its smaller size, this was based on 6 overlaps compared to an expected ~0. We also tested for depletion of VISTA negatives among the enhancer sets; we found that all liver sets, except FANTOM and DNasePlusHistone, are also enriched for overlap with VISTA negatives ($p < 0.001$). This is not surprising as the regions tested by VISTA were largely selected based on some evidence of enhancer activity, and they may have enhancer activity in other contexts not tested by VISTA. Nonetheless, to evaluate the ability of different methods to distinguish VISTA positives from negatives, we computed the relative enrichment for VISTA positives vs. negatives. Again, with the exception of FANTOM, the enhancer sets from all contexts do not demonstrate clear ability to distinguish the positives and negatives (Supplementary Figure 4). For all contexts, except GM12878, the FANTOM enhancers are more than 2x more enriched for VISTA positives than negatives, but again we note that these FANTOM/VISTA results are based on relatively small numbers of enhancers (Supplementary Table 3). Furthermore, among VISTA positives overlapped by one or more enhancers in liver, 37.9% are unique to a single method and 28.8% are overlapped by two methods, suggesting that different strategies identify different subsets of validated enhancers.

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

Genome-wide enhancer sets are commonly used to interpret the potential function of genetic variants observed in GWAS and sequencing studies[35,36,41,44,50–55]. Functional genetic variants—in particular mutations associated with complex disease—are enriched in gene regulatory regions[5,6]. We evaluated the sensitivity of this pattern to the enhancer identification strategy used by intersecting each of the enhancer sets with 20,458 unique tag SNPs significantly associated with traits from the NHGRI-EBI GWAS Catalog. Overall, 34.9% (7,133 of 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 (Supplementary Table 4). 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 (Figure 4C). Since the tag SNPs may not be 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 (Supplementary Figure 5).

Most GWAS tag SNPs are predicted to overlap an enhancer by only a single method (Figure 4D). For example, across four liver enhancer sets (H3K27acPlusH3K4me1, H3K27acMinusH3K4me3, EncodeEnhancerlike, and ChromHMM), 55.9% (1833/3279) of the GWAS SNPs that overlapped an enhancer are unique to a single set, and only 3.7% (120/3279) are shared by all four sets. The distribution of enhancer overlaps was similar when considering all candidate SNPs, with or without LD expansion (Figure 4D). We also identified the SNP in each LD block with the maximum number of overlaps with distinct enhancer sets. Even when limiting our analysis to this set, only 42% of the GWAS SNP LD blocks with any enhancer overlap are identified by more than two enhancer sets considered

(Supplementary Figure 5). This illustrates that the annotation of sequences harboring GWAS SNPs varies greatly depending on the enhancer identification strategies 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 (Supplementary Table 5). As in the full GWAS set, the majority of curated liver SNPs are overlapped by a single method (58.5%) and only 3.7% are shared by all four methods. 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 ($p = 0.002$ for liver, $p = 0.026$ for heart, Mann-Whitney *U* test). FANTOM enhancers are the exception to this trend due to the small number of overlapping context-specific SNPs (Supplementary Table 4). This suggests that the different methods, in spite of their lack of agreement, all identify regulatory regions relevant to the target context.

To test if these patterns hold for functional genetic variants that are not tied to complex disease, 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 SNPs. Within a context, most eQTL are identified as enhancers by a single enhancer prediction method only, and there is wide variation in the number of eQTL overlapped by different enhancer sets (Figure 4E; Supplementary Table 6). Across the same four representative liver enhancer sets, 3.0% (1,864/62,058) of eQTL that overlapped at least one enhancer are shared by all sets, and 49.5% (33941/68563) of all overlapped eQTL are called an enhancer by only a single set (Figure 4D). Thus, the interpretation of these known functional SNPs varies substantially depending on the enhancer identification strategy used.

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 near enhancers in each set using the Genomic Regions Enrichment of Annotations Tool (GREAT) (Figure 4F, and Methods). Many of the GO terms identified for the enhancer sets are relevant to the associated context (Table 2). However, the majority of the top 50 significant annotations 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 4F, lower triangle). In most of the pair-wise comparisons, fewer than half of the GO terms are shared between a pair of enhancer sets. Furthermore, many of the terms, especially those shared by multiple enhancer sets, are near the root of the ontology and thus are less functionally specific and more broadly applicable. These results provide evidence that the enhancer sets are influencing different functions relevant to the target biological context. These trends hold for both the Biological Process (BP) and Molecular Function (MF) ontologies and considering the top 10 and 50 annotations for each set (Supplementary Figure 6).

To 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[56,57]. The histone-derived H3K27ac and H3K27acMinusH3K4me3 enhancer sets are among the most functionally similar, with similarity scores of at least 0.80 in most contexts (Figures 4F, upper triangle; Supplementary Figure 7). This is not surprising given their underlying assays overlap. The functional similarity scores are lower for comparisons of the other histone modification sets, around 0.50–0.80. 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 each context (Figure 4F). As a benchmark, biological replicates for the Gm12878 H3K27ac enhancer set received a similarity of 0.93. This suggests different functional influences for enhancer sets from the same context identified by different methods, with FANTOM as a particular outlier.

Genomic and functional comparison 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 their GO term functional similarity (Methods).

Several trends emerged from analyzing the genomic and functional distribution within and between biological contexts. First, the FANTOM enhancers are consistently distinct from all other enhancer sets in both their genomic distribution and functional associations (Figure 5). 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 DNasePlusHistone enhancers are close to the other histone-derived enhancer sets and the machine learning models in the functional comparisons (Figure 5B, D), they are located far from those sets in the genomic-location-based projection and hierarchical clustering (Figure 5A, C). Histone-derived enhancer sets are also farther from

the machine-learning-based sets in genomic comparisons than in functional analyses. 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 5E). 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 5F). 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, the FANTOM enhancers are exceptions to these trends; FANTOM enhancers from each context form their own cluster based on their genomic distribution, underscoring their uniqueness.

Combining enhancer sets does not 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 enhancers to determine whether those identified by multiple methods have increased "functional support" compared to regions identified by only a single method. We evaluated three signals of functional importance for each enhancer: *(i)* the percentage of conserved base pairs, *(ii)* the number of GWAS SNPs per kilobase (kb), and *(iii)* the number of GTEx eQTL per kb. For each signal, there are only marginal changes as the number of methods identifying a region increases, and the evidence for function is lower among enhancers predicted by multiple methods in some settings (Figure 6A-C). Thus, we do not find increased evidence of functional importance in enhancers identified by multiple methods, compared to enhancers identified by a single method.

Several of the 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 6D, Supplementary Figure 8). 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 confidence score decreases as the regions become more highly shared (Supplementary Figure 8A-C). This provides further evidence that enhancer sets formed by a combination of methods do not necessarily represent a high confidence subset, and that filtering based on simple agreement between methods may not improve the specificity of enhancer predictions.

A tool for evaluating of the robustness of conclusions across enhancer identification strategies

Our findings imply that results of studies of enhancers are sensitive to the enhancer identification strategy used. To facilitate evaluation of the sensitivity and robustness of results with respect to this choice, we provide an integrated database (creDB) of annotated enhancer sets. It contains all enhancer sets used in this study, and enhancer sets from multiple primary and meta-analyses annotating enhancers across cellular contexts. In contrast to other efforts providing large compilations of enhancer sets[52,53], creDB is not a web-server or "enhancer browser". Rather, it is an SQL database with a convenient interface to the R programming language and distributed as an R package. creDB is designed to enable users to easily consider many different genome-wide sets of enhancers in their bioinformatics analyses and to evaluate the robustness of their findings. The current version of creDB is available at http://www.kostkalab.net/software.html.

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. Further on, 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. We found substantial differences in the genomic, evolutionary, and functional characteristics of identified enhancers *within* similar tissues and cell types. Furthermore, annotated enhancer sets vary significantly in their overlap with GWAS SNPs and eQTL, and the vast majority of overlapped GWAS SNPs and eQTL are *unique to a single strategy*. In addition, regions identified as enhancers by multiple methods *do not* have increased 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. Importantly, 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[34–36,41,44,54,55,58]. 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[29,59,60].

Our goal was not to identify the best method for identifying enhancers, but rather 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; however, the degree of difference we observe is striking. The consistent lack of agreement between methods demonstrates that the respective working definitions of "enhancer" have little overlap. Focusing on functional annotations, we find agreement between methods about basic functions, but substantial differences in more specific annotations. This suggests that different strategies contribute unique information towards the identification of functionally important enhancers.

In 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 robust view of the regulatory landscape. To facilitate this, we have developed creDB, a comprehensive, easily queried database of diverse enhancer annotations.

Second, our study highlights the need for more sophisticated models of the architecture and dynamics of *cis*-regulatory regions. Many different classes of regions with enhancer-like regulatory activities have been discovered[4,14,19,29,37,61,62]. We argue that collapsing the diversity of vertebrate distal gene regulatory regions into a single category is overly restrictive. Simply calling all of the regions identified by these diverse approaches "enhancers" obscures functionally relevant complexity and creates false dichotomies. While there is appreciation of this subtlety, the field requires more precise terminology and improved statistical and functional models of the diversity of *cis*-regulatory "enhancer-like" sequences.

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 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 properties in multiple, potentially uncharacterized, dimensions [2,63,64]. 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.

METHODS

Enhancer set definition

In this section, we summarize how we defined 34 human enhancer sets across four biological contexts. All analyses were performed in the context of the GRCh37/hg19 build of the human genome. We used TSS definitions from Ensembl v75 (GRCh37.p13).

We downloaded broad peak ChIP-seq data for three histone modifications, H3K27ac, H3K4me1, and H3K4me3 from the ENCODE project[30] for two cell lines, K562 and Gm12878, and from the Roadmap Epigenomics Consortium[65] for two primary tissues, liver and heart. Also from the ENCODE Encyclopedia (version 3.0), we downloaded the "enhancer-like" annotations; these combine DHSs and H3K27ac ChIP-seq peaks using an unsupervised machine learning model. We retrieved ChromHMM enhancer predictions[66] for the K562 and Gm12878 cell lines from the models trained on ENCODE data[15]. We downloaded ChromHMM predictions for liver and heart tissues from the 15-state segmentation performed by the Roadmap Epigenomics Consortium. For all ChromHMM sets, we combined the weak and strong enhancer states. We considered two enhancer sets for K562 and Gm12878 based on supervised machine learning techniques—one described in Yip et al. 2012[45], and the other in Ho et al. 2014[46]. Yip12 predicted 'binding active regions' (BARs) using DNA accessibility and histone modification data; the positive set contained BARs overlapping a 'transcription-related factor' (TRF), and the negative set contained BARs with no TRF peaks. The predicted regions were

filtered using other genomic characteristics to determine the final set of enhancers[45]. Ho14 used H3K4me1 and H3K4me3 ChIP-seq peaks in conjunction with DHSs and p300 binding sites to predict regions with regulatory activity both distal and proximal to TSSs. The distal regulatory elements make up their published enhancer set[46]. We downloaded enhancer regions predicted by the FANTOM consortium for the four sample types analyzed[25]. For the primary tissues, we also downloaded enhancer predictions in liver from Villar et al. 2015[38] and experimentally verified heart enhancers from VISTA[48].

To represent enhancer identification strategies in common use, we created three additional enhancer sets for this study using histone modification ChIP-seq peaks and DNase-seq peaks downloaded from ENCODE and Roadmap Epigenomics. The H3K27acPlusH3K4me1 track is a combination of H3K27ac and H3K4me1 ChIP-seq peak files[4,19,37]. If both types of peaks were present (i.e., the regions overlap by at least 50% of the length of one of the regions) the intersection was classified as an enhancer. Similarly, to create the H3K27acMinusH3K4me3 set for each context, we intersected H3K27ac and H3K4me3 ChIP-seq peak files and kept regions where H3K27ac regions did not overlap a H3K4me3 peak by at least 50% of their length. We derived the combination of H3K27ac and H3K4me3 and the 50% overlap criterion from previous studies[14,37,38]. The DNasePlusHistone track is based on the pipeline described in Hay et al. 2016[39]. It combines H3K4me1, H3K4me3, DNaseI hypersensitive sites (DHSs), and transcription start site (TSS) locations. We filtered a set of DHSs, as defined by DNase-seq, for regions with an H3K4me3 / H3K4me1 ratio less than 1, removed regions within 250 bp of a TSS, and called the remaining regions enhancers.

For all enhancer sets, we excluded elements overlapping ENCODE blacklist regions and gaps in the genome assembly[67]. Additionally, due to the presence of extremely long regions in some enhancer sets, likely caused by technical artifacts, we removed any regions more than three standard deviations above or below the mean length of the dataset. This filtering process removed relatively few annotations (Supplementary Table 7).

Enhancer set overlap and similarity

We calculated the base pair overlap between enhancer set *A* and enhancer set *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 overlapping by at least 1 bp between two sets, *A* and *B*, 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 tools from the BEDtools v2.23.0 suite[68].

We also evaluated the similarity of each pair of enhancer sets using the Jaccard similarity index. The Jaccard index is equal to 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 enhancer sets, 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[69].

Enhancer set overlap enrichment analysis

We used the Genomic Association Tester (GAT) to evaluate whether the observed base pair overlap between pairs of enhancer sets was significantly greater than would be expected by chance[70]. GAT uses simulations to calculate an empirical *p*-value for the observed overlap from the distribution of overlaps expected under a null model of random placement of length-matched regions from the second enhancer set throughout the genome, excluding ENCODE blacklist regions. We ran GAT on all pairs of enhancer sets using the default options with 1,000 iterations. We used the same framework to evaluate element-wise comparisons.

Enhancer conservation analysis

We intersected each enhancer set with evolutionarily conserved regions defined by PhastCons, a two-state hidden Markov model that defines conserved elements from multiple sequence alignments[47]. We concatenated primate and vertebrate PhastCons elements defined over the UCSC alignment of 45 vertebrates with human into a single set of conserved genomic regions. We then calculated the fraction of all base pairs in each enhancer set overlapping the merged conserved elements.

Enhancer GWAS catalog SNP and GTEx eQTL enrichment

We downloaded the full list of 20,458 unique GWAS SNPs from the GWAS Catalog (v1.0, downloaded 08-10-2016)[71]. We intersected the GWAS Catalog SNPs with each of the enhancer sets and calculated the total number of SNPs overlapped as well as the total number of regions overlapped by at least one SNP. We compared those values between enhancer sets and contexts. To determine whether there is enrichment of GWAS SNPs in a given enhancer set, we used

the following protocol: count the number of SNPs that overlap enhancers; generate 1,000 random sets of regions that maintain the length distribution of the enhancer set and avoid ENCODE blacklist regions; count the number of SNPs that overlap each of the random regions; compare the observed SNP count with the SNP counts from the simulations and calculate an empirical p-value. We expanded the set of GWAS Catalog SNPs to include SNPs in high LD ($r^2 > 0.9$) using individuals of European ancestry from the 1000 Genomes Project (phase 3)[72]. We intersected this larger set of SNPs with each of the enhancer sets and calculated the amount of overlap. We also identified the SNP with the maximum number of enhancer set overlaps for each GWAS SNP's LD block.

We downloaded all GTEx eQTL from the GTEx Portal (v6p, downloaded 09-07-2016)[73]. 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 interpreted the results from the p-value threshold of 10^{-10}. To determine enrichment of eQTL in a given enhancer annotation set, we used the same protocol as for the GWAS SNP enrichment above. We also performed a tissue-specific enrichment analysis with only eQTL from liver and heart at p-value thresholds of 10^{-6} and 10^{-10}, presenting the results from $p < 10^{-10}$, although our main results are consistent across thresholds.

Relative enrichment for functionally validated enhancers

We downloaded all positive and negative enhancer sequences from the VISTA enhancer browser. For each method, we separately determined enrichment for the positive and negative enhancer sets using the random permutation approach described in the GWAS SNP and GTEx eQTL analyses. The relative enrichment value is defined as the ratio of enrichment in the set of positive enhancers versus the enrichment in the set of negative enhancers. Thus, values greater than 1 indicate increased enrichment for positive enhancers and values below 1 indicate increased enrichment for enhancer sequences from the negative set.

Enhancer set GO term annotation and similarity

We used GREAT to find Gene Ontology annotations enriched among genes nearby the enhancer sets. GREAT assigns each input region to regulatory domains of genes and uses both a binomial and a hypergeometric test to discover significant associations between regions and associated genes' ontology terms[74]. Due to the large number of reported regions in each enhancer set, we consider significance based only on the binomial test with the Bonferroni multiple testing correction (<0.05). We downloaded up to 1,000 significant terms for each enhancer set from two of the GO ontologies: Molecular Function (MF) and Biological Process (BP). We calculated the similarity between lists of GO terms using the GOSemSim package in R[57]. GOSemSim uses Wang's sematic similarity metric which accounts for the hierarchical organization and relatedness of ontology terms when calculating the similarity score[56]. For each pair of enhancer sets, we calculated the similarity between GO terms, and converted the resulting similarity matrix into a distance matrix for hierarchical clustering. 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.

Genomic and functional clustering of enhancers sets

To identify groups of similar enhancers in genomic and functional space, we performed hierarchical clustering on the enhancer sets. For functional similarity, we clustered the lists of GO terms returned for each enhancer set. We calculated Wang's similarity metric using the GoSemSim package in R and converted it to distance by subtracting the similarity score from 1. For genomic similarity, we converted the pairwise Jaccard similarity to distance in the same way and clustered the enhancer methods based on this measure. For both distance matrices we used agglomerative hierarchical clustering in R function with the default complete linkage method to iteratively combine clusters[75]. We visualized the cluster results as dendrograms using ggplot2 and dendextend[69,76]. We performed multidimensional scaling (MDS) on the Jaccard and GO term distance matrices using default options in R[75].

Combinatorial analysis of enhancer sets

We stratified enhancers by the number of identification strategies that annotate them in order to determine whether enhancers predicted by increasing numbers of methods show a higher percentage of evolutionarily conserved base pairs or a higher fraction of SNPs/eQTL per bp. For each predicted enhancer, we calculated the number of base pairs that overlap a given reference set, and used that to compute the relevant statistics. We considered three different reference sets: evolutionarily conserved base pairs as defined by PhastCons elements, GWAS SNPs, and GTEx eQTL to calculate the percent bases conserved and GWAS SNPs and GTEx eQTL per base pair for predicted enhancers.

Confidence distribution of enhancer sets

For enhancer sets with region-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 (H3K27ac, 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.

Enhancer database

We created creDB, a database of *cis*-regulatory elements. It currently contains all 1,371,867 enhancers analyzed in this study and representative sets from other common contexts. It is implemented as an SQLite database with an R interface and distributed as an R package. Our design facilitates the integration of creDB into a wide range of genome-wide bioinformatics workflows, alleviating the vetting and processing that is necessary with flat file downloads. This sets it apart from other efforts that collect enhancers but focus on providing interfaces for searching a small number of candidates or for "browsing" enhancers on a case-by-case basis. For all enhancers creDB contains information about genomic location, identification strategy, and tissue or cell type of activity. See Supplementary Text S1 for details about database content, design, and use. We envision creDB as a growing community resource and encourage other researchers to contribute.

Data Availability

All data analyzed in this manuscript are available in the creDB R package and database: http://www.kostkalab.net/software.html.

ACKNOWLEDGEMENTS

M.L.B. was supported by the National Institutes of Health [5T32LM012412]. D.K. was supported by the National Institutes of Health [1R01GM115836] and a Basil O'Connor Starter Scholarship from the March of Dimes. J.A.C. was supported by the National Institutes of Health [1R01GM115836], a March of Dimes Innovation Catalyst award, and the Burroughs Wellcome Fund.

AUTHOR CONTRIBUTIONS

D.K. and J.A.C. designed and supervised the study. M.L.B. carried out the analyses. M.L.B, D.K., and J.A.C. wrote the manuscript. S.C.T. and D.K. designed and implemented the creDB database and software.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

FIGURES

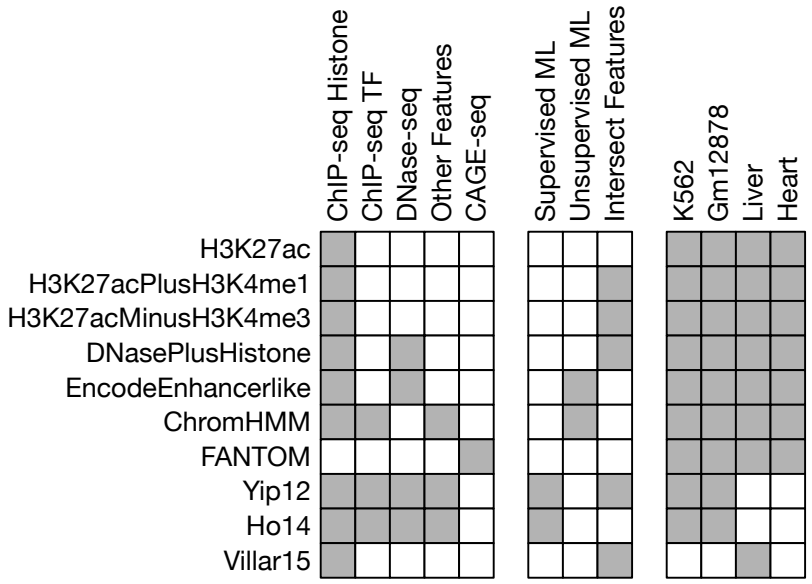


Figure 1. Ten diverse enhancer identification strategies were evaluated across four cellular contexts. The leftmost columns of the schematic represent the experimental assays and sources of the data used by each identification strategy. The middle columns describe the computational processing (if any) performed on the raw data (ML: machine learning). The rightmost columns give the contexts in which the sets were available. Table 1 gives the number, length, and genomic coverage of each enhancer set.

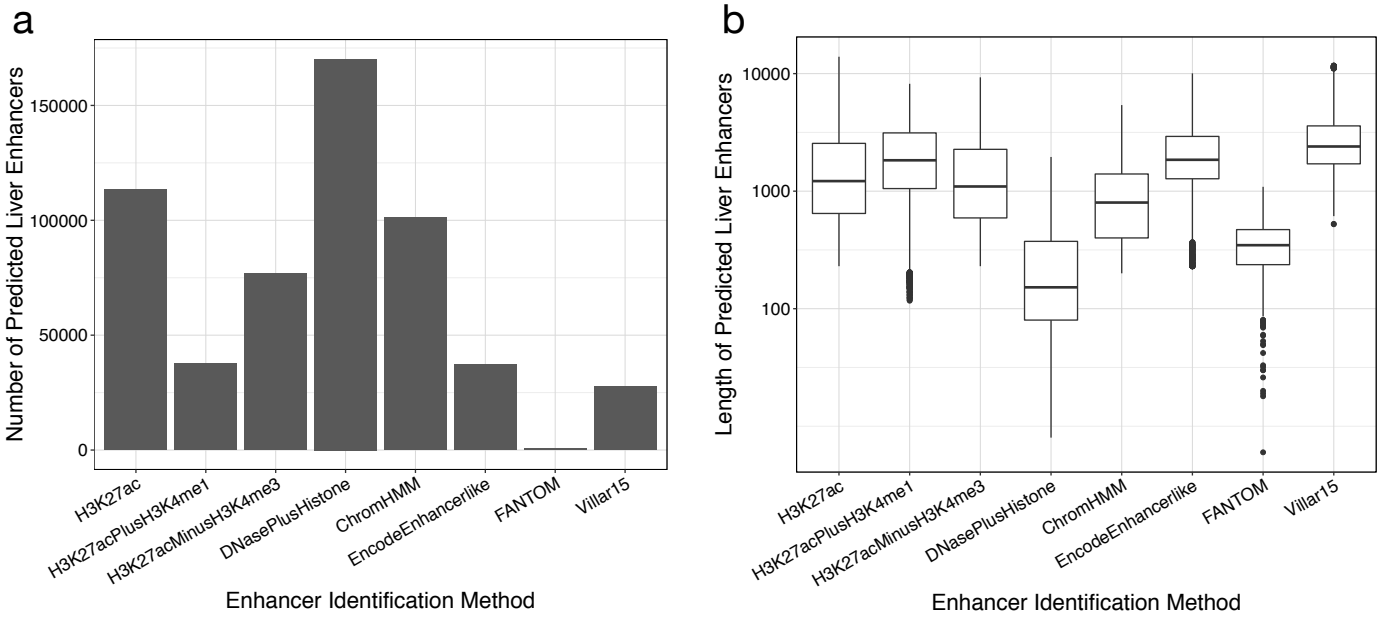


Figure 2. Enhancer identification methods vary in number and length of predicted enhancers. (a) Number of liver enhancers identified by each method varies over two orders of magnitude. There is considerable variation even among methods defined based on similar input data, e.g., histone modifications. (b) Length of liver enhancers identified by different methods shows similar variation. Enhancer lengths are plotted on a log10 scale on the y-axis. Data for other contexts are available in Table 1 and Supplementary Figure 1.

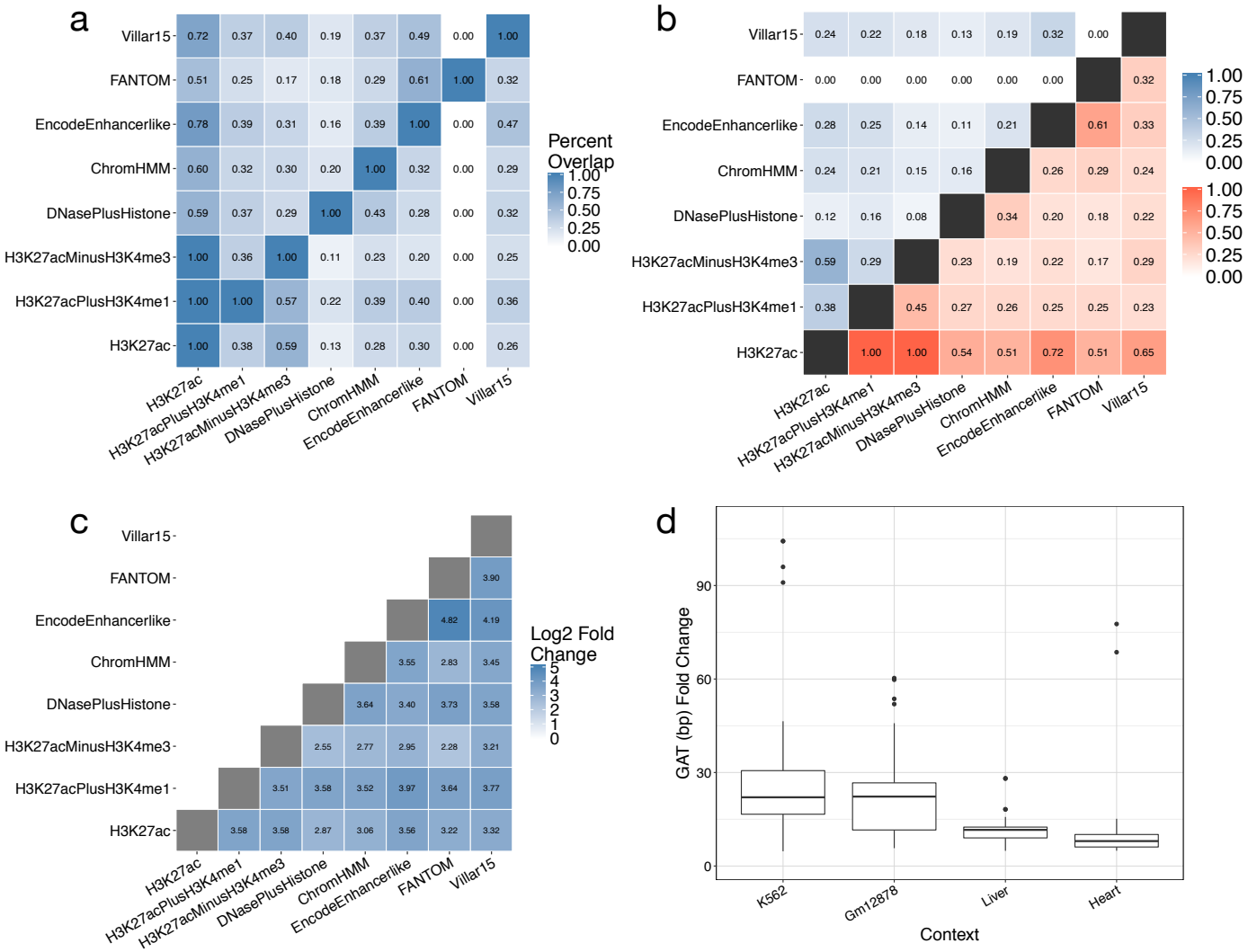
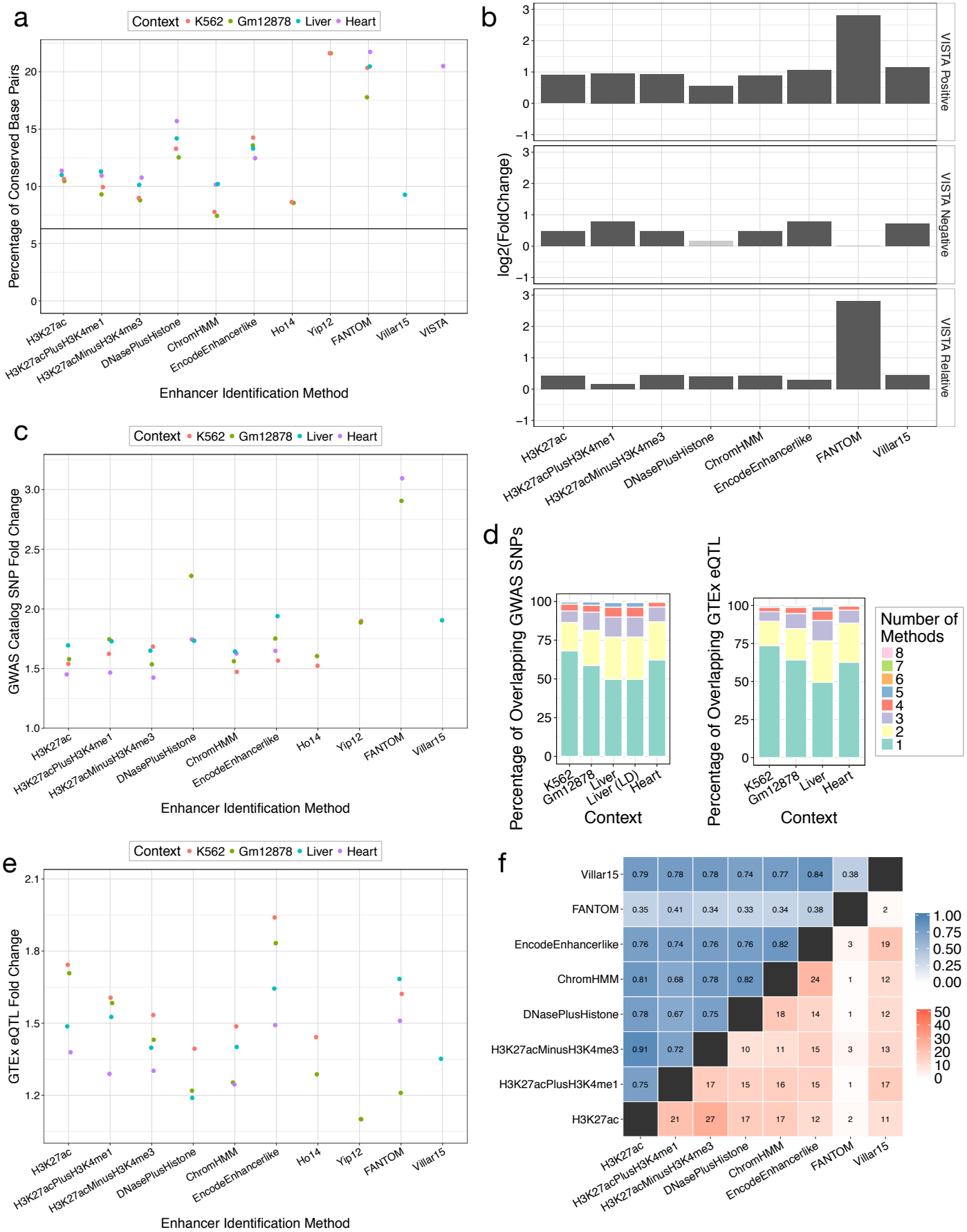


Figure 3. Enhancer sets have low genomic overlap. (a) The percent base pairs (bp) overlap between all pairs of liver enhancer sets. Percent overlap in each cell was calculated by dividing the number of shared base pairs among the two sets by the total number of base pairs of the set on the y-axis. High overlap is observed only for pairs based on very similar data, e.g., H3K27ac, H3K27acPlusH3K4me1, and H3K27acMinusH3K4me3. (b) The Jaccard similarity between all pairs of liver enhancer sets. The upper triangle gives the Jaccard similarity, and the lower triangle gives the relative Jaccard similarity in which the observed similarity is divided by the maximum possible similarity for the pair of sets. (c) Pairwise bp enrichment values (log2 fold change) for overlap between each liver enhancer set, compared to the expected overlap between randomly distributed, length-matched regions. (d) The enrichment for bp 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 between contexts).


a
Context • K562 • Gm12878 • Liver • Heart
Percentage of Conserved Base Pairs
H3K27ac
H3K27acPlusH3K4me1
H3K27acMinusH3K4me3
DNasePlusHistone
ChromHMM
EncodeEnhancerlike
Ho14
Yip12
FANTOM
Villar15
VISTA
Enhancer Identification Method
b
log2(FoldChange)
VISTA Positive
VISTA Negative
VISTA Relative
c
GWAS Catalog SNP Fold Change
d
Percentage of Overlapping GWAS SNPs
Percentage of Overlapping GTEx eQTL
K562
Gm12878
Liver
Liver (LD)
Heart
Context
Number of Methods
e
GTEx eQTL Fold Change
f
Villar15 0.79 0.78 0.78 0.74 0.77 0.84 0.38
FANTOM 0.35 0.41 0.34 0.33 0.34 0.38 2
EncodeEnhancerlike 0.76 0.74 0.76 0.76 0.82 3 19
ChromHMM 0.81 0.68 0.78 0.82 24 1 12
DNasePlusHistone 0.78 0.67 0.75 18 14 1 12
H3K27acMinusH3K4me3 0.91 0.72 10 11 15 3 13
H3K27acPlusH3K4me1 0.75 17 15 16 15 1 17
H3K27ac 21 27 17 17 12 2 11

1

Figure 4. Enhancers identified by different methods differ in functional attributes. (a) Enhancer sets vary in their degree of evolutionary conservation. Each point represents the percentage of base pairs (bp) in the enhancer set that overlap a primate or vertebrate PhastCons conserved element. FANTOM, Yip12, and VISTA are the most conserved (~20% of bp, while sets based on histone modification data alone are among the least conserved (~10% of bp). The black line at 6.4% denotes the average conservation level for the genomic background. (b) Plot of the element-wise enrichment for positive (upper panel) and negative (middle panel) VISTA enhancers. The bottom panel is the logarithm of the relative enrichment ratio for liver enhancers with VISTA positives compared to VISTA negatives. A positive ratio indicates higher enrichment for positive VISTA enhancers and a negative value indicates more enrichment for non-enhancers tested by VISTA. Equal enrichment in both sets yields a score of 0. Most enhancer sets have enrichment for positive VISTA enhancers. Overall, FANTOM enhancers have the highest enrichment for experimentally validated enhancers relative to the negative set. (c) GWAS SNP enrichment (fold change compared to randomly shuffled regions) among all enhancer sets for each biological context. All sets except FANTOM in K562 and liver contexts (not shown due to a small number of overlaps) are significantly enriched. (d) Among all GWAS SNPs (left) and GTEx eQTL (right) that overlap at least one enhancer in a context enhancer, the colored bars represent the number of methods that identified the region as an enhancer. The majority of these functional variants are supported by a single method; very few functional variants are shared among all methods. The percentage is similar for SNPs in high LD ($r^2 > 0.9$) with the GWAS Catalog SNPs in liver (Liver LD). (e) eQTL enrichment among all enhancer sets for each biological context. All sets were enriched except for DNasePlusHistone in heart (not shown). (f) Enhancer sets from the same biological context have different functional associations. We identified Gene Ontology (GO) functional annotations enriched among genes likely to be regulated by each enhancer set. The upper triangle represents the pairwise semantic similarity for significant MF GO terms associated with predicted liver enhancers. The lower triangle shows the number of shared MF GO terms in the top 50 significant hits for liver enhancer sets.

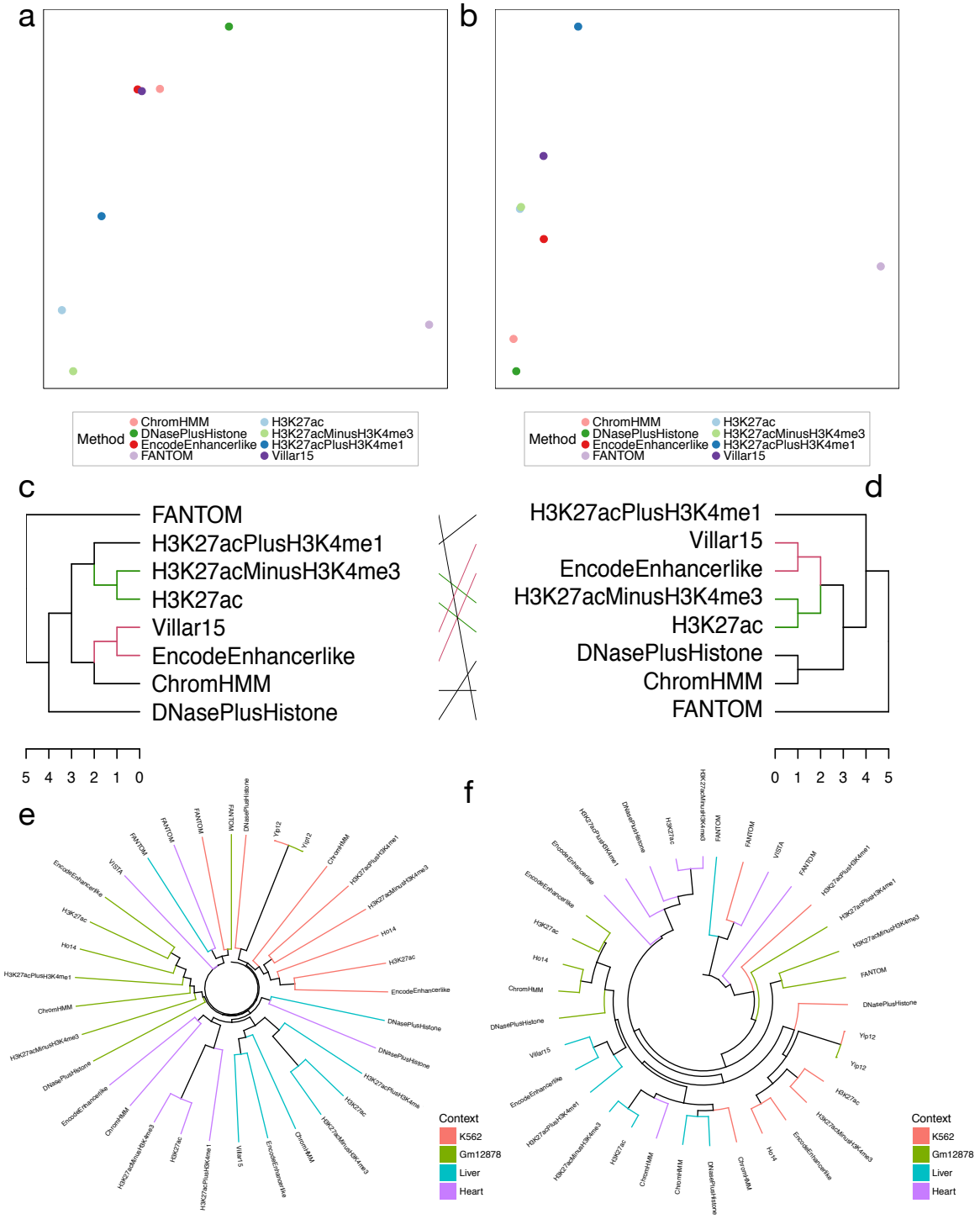


Figure 5. The genomic and functional similarity between enhancer sets is not consistent. (a) Multidimensional scaling (MDS) plot of liver enhancer sets based on the Jaccard similarity of the genomic distributions (Figure 3B). (b) MDS plot for liver enhancers based on distances calculated from molecular function (MF) Gene Ontology (GO) term semantic similarity values (Figure 4F). (c,d) Hierarchical clustering based on the Jaccard similarities of the genomic distributions (c) of all liver enhancer sets compared to clustering based on GO semantic similarity (d). Colored branches denote identical subtrees within the hierarchy. (e) Hierarchical clustering based on genomic Jaccard distances for all methods and all contexts. (f) Hierarchical clustering of all available enhancer sets based on GO term distances. Terminal branches in E and F 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.

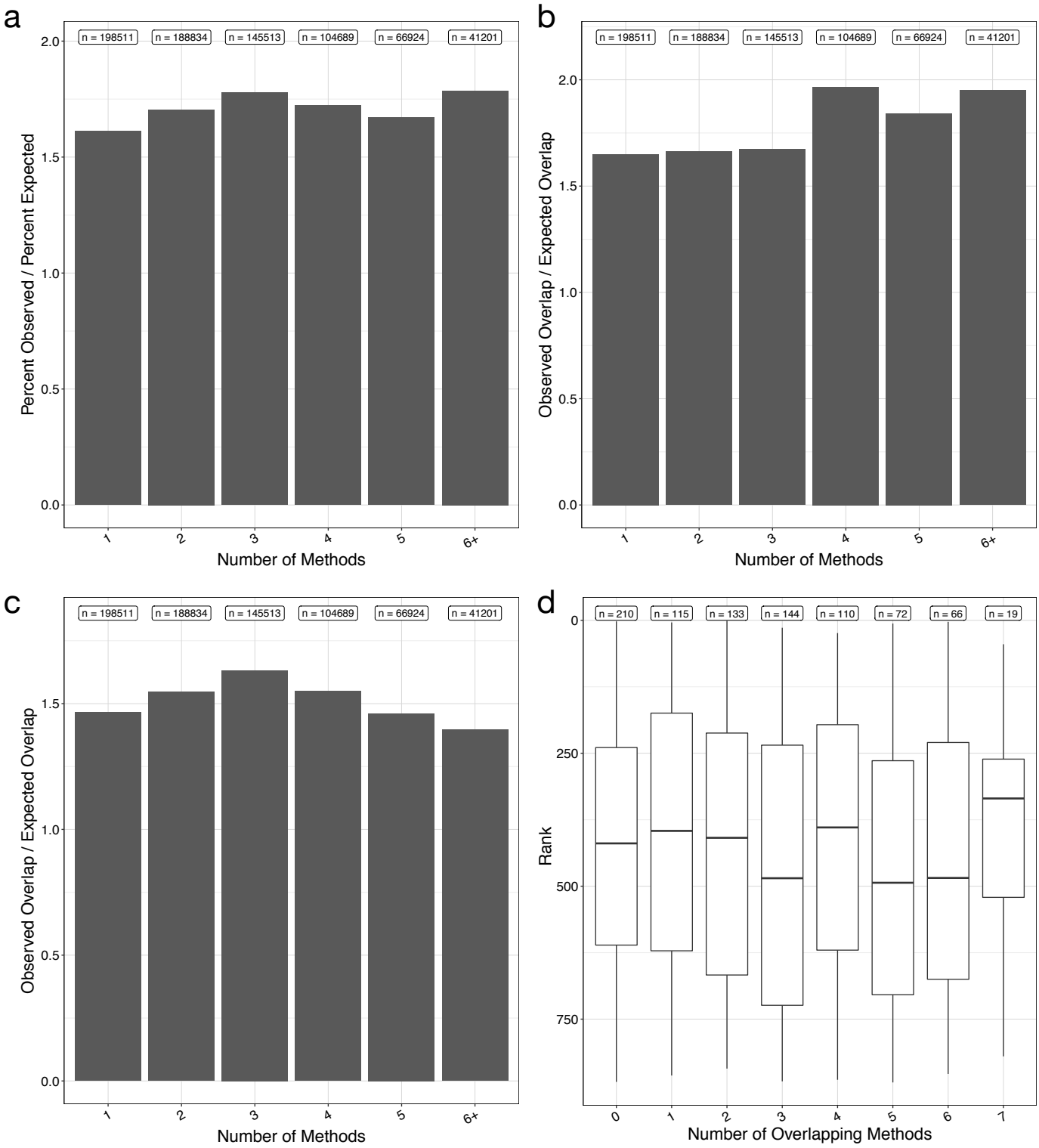


Figure 6. Enhancers identified by multiple methods are not more likely to have evidence of regulatory function. (a) Bar plot of the percentage of base pairs conserved for liver enhancers identified by different numbers of strategies compared to the expected percentage. Regions identified as enhancers by multiple methods are not more conserved. (b) Bar plot of the number of overlapping GWAS SNPs for all combinations of liver enhancer sets over the number of expected overlapping GWAS SNPs. On average, the number of SNPs is similar for all levels of overlap between enhancer sets. (c) Bar plot of the number of overlapping GTEx eQTL for all combinations of liver enhancer sets. GTEx eQTL show similar trends to the percentage of conservation and GWAS SNP overlap. In (A–C), the number of enhancers in each bin is provided. (d) Boxplots showing the distribution of ranked confidence scores for FANTOM enhancers in liver partitioned into bins based on the number of other methods that also identify the region as an enhancer. Lower ranks indicate higher

1 confidence. Enhancers identified by many methods have similar confidence distributions to those identified by a single
2 method (Supplementary Figure 7).

1 **TABLES**
2
3 **Table 1.** Summary of all enhancer sets analyzed in this study.
4

Context	Enhancer Set	Number of Base Pairs (kb)	Number of Enhancers	Median Enhancer Length	Fraction of Genome
K562	H3K27ac	120,073	45,230	783	0.0421
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	H3K27ac	117,212	45,105	1,156	0.0411
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	H3K27ac	231,900	113,221	1,218	0.0814
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	H3K27ac	211,406	171,819	701	0.0742
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

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

Enhancer Set	GO Terms
H3K27ac	cytoskeletal adaptor activity
	14-3-3 protein binding
	nucleobase-containing compound transmembrane transporter activity
	high-density lipoprotein particle binding
	JUN kinase binding
H3K27acPlusH3K4me1	cytoskeletal adaptor activity
	14-3-3 protein binding
	leukotriene-C4 synthase activity
	nucleobase-containing compound transmembrane transporter activity
	FAD binding
H3K27acMinusH3K4me3	14-3-3 protein binding
	cytoskeletal adaptor activity
	thyroid hormone receptor binding
	ARF guanyl-nucleotide exchange factor activity
	high-density lipoprotein particle binding
DNasePlusHistone	cytoskeletal adaptor activity
	glucocorticoid receptor binding
	nucleobase-containing compound transmembrane transporter activity
	high-density lipoprotein particle binding
	14-3-3- protein binding
ChromHMM	high-density lipoprotein particle binding
	nucleobase-containing compound transmembrane transporter activity
	cytoskeletal adaptor activity
	14-3-3 protein binding
	retinoid X receptor binding
EncodeEnhancerlike	cytoskeletal adaptor activity
	14-3-3 protein binding
	nucleobase-containing compound transmembrane transporter activity
	apolipoprotein A-I binding
	high-density lipoprotein particle binding
FANTOM	glucocorticoid receptor binding
	protein kinase binding
	kinase binding
	methylglutaconyl-CoA hydratase activity
	vitamin D response element binding
Villar15	protease binding
	phosphatidylinositol 3-kinase binding
	14-3-3 protein binding
	cytoskeletal adaptor activity
	glucocorticoid receptor binding

3

REFERENCES

1. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15,** 272–86 (2014).
2. Andersson, R., Sandelin, A. & Danko, C. G. A unified architecture of transcriptional regulatory elements. *Trends in Genetics* **31,** 426–433 (2015).
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. 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).
5. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science (80-.).* **337,** 1190–1195 (2012).
6. Corradin, O. & Scacheri, P. C. Enhancer variants: evaluating functions in common disease. *Genome Med.* **6,** 85 (2014).
7. Sholtis, S. J. & Noonan, J. P. Gene regulation and the origins of human biological uniqueness. *Trends Genet.* **26,** 110–118 (2010).
8. 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
9. Mack, K. L. & Nachman, M. W. Gene Regulation and Speciation. *Trends Genet.* **33,** 68–80 (2016).
10. Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A. & Bejerano, G. Enhancers: five essential questions. *Nat. Rev. Genet.* **14,** 288–95 (2013).
11. Kleftogiannis, D., Kalnis, P. & Bajic, V. B. Progress and challenges in bioinformatics approaches for enhancer identification. *Brief. Bioinform.* **17,** 967–979 (2016).
12. Inoue, F. *et al.* A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *bioRxiv* 61606 (2016). doi:10.1101/061606
13. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106,** 159–164 (2015).
14. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39,** 311–8 (2007).
15. Hoffman, M. M. *et al.* Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.* **41,** 827–841 (2013).
16. 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).
17. Crawford, G. E. *et al.* Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Res.* **16,** 123–131 (2006).
18. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).
19. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 (2009).
20. Woolfe, A. *et al.* Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3,** (2005).
21. Pennacchio, L. a *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444,** 499–502 (2006).
22. Visel, A. *et al.* Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nat Genet* **40,** 158–160 (2008).
23. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457,** 854–8 (2009).
24. Blow, M. J. *et al.* ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* **42,** 806–810 (2010).
25. Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507,** 455–61 (2014).
26. 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).
27. Arner, E. *et al.* Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. *Science (80-.).* **347,** 1010 LP-1014 (2015).
28. Erwin, G. D. *et al.* Integrating Diverse Datasets Improves Developmental Enhancer Prediction. *PLoS Comput. Biol.* **10,** (2014).
29. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–9 (2011).

30. The ENCODE Project Consortium. An Integrated Encyclopedia of DNA Elements in the Human Genome. *Nature* **489,** 57–74 (2012).
31. Cotney, J. *et al.* Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome Res.* **22,** 1069–1080 (2012).
32. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–47 (2013).
33. Dickel, D. E. *et al.* Genome-wide compendium and functional assessment of in vivo heart enhancers. *Nat. Commun.* **7,** 12923 (2016).
34. Yao, L., Tak, Y. G., Berman, B. P. & Farnham, P. J. Functional annotation of colon cancer risk SNPs. *Nat. Commun.* **5,** 5114 (2014).
35. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–43 (2015).
36. Harismendy, O. *et al.* 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. *Nature* **470,** 264–268 (2011).
37. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470,** 279–83 (2011).
38. Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160,** 554–566 (2015).
39. Hay, D. *et al.* Genetic dissection of the α-globin super-enhancer in vivo. *Nat. Genet.* 1–12 (2016).
40. Zhernakova, D. V *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49,** 139–145 (2017).
41. 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).
42. Parker, S. C. J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 17921–17926 (2013).
43. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95,** 535–552 (2014).
44. 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).
45. 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).
46. Ho, J. W. K. *et al.* Comparative analysis of metazoan chromatin organization. *Nature* **512,** 449–52 (2014).
47. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).
48. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Res.* **35,** (2007).
49. 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).
50. Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nat. Biotechnol.* **30,** 1095–106 (2012).
51. 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).
52. Ashoor, H., Kleftogiannis, D., Radovanovic, A. & Bajic, V. B. DENdb: Database of integrated human enhancers. *Database* **2015,** (2015).
53. 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
54. Pasquali, L. *et al.* Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nat Genet* **46,** 136–143 (2014).
55. Hazelett, D. J. *et al.* Comprehensive Functional Annotation of 77 Prostate Cancer Risk Loci. *PLoS Genet.* **10,** e1004102 (2014).
56. 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).
57. Yu, G. *et al.* GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26,** 976–978 (2010).
58. Weedon, M. N. *et al.* Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat Genet* **46,** 61–64 (2014).
59. 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).

60. 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
61. Pradeepa, M. M. *et al.* Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat. Genet.* **48,** 681–686 (2016).
62. 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).
63. Kim, T. K. & Shiekhattar, R. Architectural and Functional Commonalities between Enhancers and Promoters. *Cell* **162,** 948–959 (2015).
64. Andersson, R. Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays* **37,** 314–323 (2015).
65. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
66. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods* **9,** 215–6 (2012).
67. Kundaje, A. A comprehensive collection of signal artifact blacklist regions in the human genome. ... *Site/Anshulkundaje/Projects/Blacklists (Last Accessed 30 …* (2013).
68. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
69. Wickham, H. *ggplot2*. *Elegant Graphics for Data Analysis* (2009). doi:10.1007/978-0-387-98141-3
70. 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).
71. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** (2014).
72. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
73. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–5 (2013).
74. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28,** 495–501 (2010).
75. 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).
76. Galili, T. dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* **31,** 3718–3720 (2015).