Comprehensive enhancer-target gene assignments improve gene set level interpretation of genome-wide regulatory data

Tingting Qin[1], Christopher Lee[1,2], Raymond Cavalcante[3], Peter Orchard[1], Heming Yao[1], Hanrui Zhang[1], Shuze Wang[1], Snehal Patil[1], Alan P Boyle[1,4] and Maureen A Sartor[1,2*]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, Michigan, USA.

[2]Department of Biostatistics, School of Public Health, University of Michigan Medical School, Ann Arbor, Michigan, USA.

[3]Biomedical Research Core Facilities, Epigenomics Core, University of Michigan Medical School, Ann Arbor, Michigan, USA.

[4]Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan, USA

*** Correspondence:**

Maureen A Sartor

sartorma@umich.edu

Keywords: Enhancer, Target gene, Gene Ontology, Gene set enrichment test, Genomic regions, ChIP-seq

Abstract

Revealing the gene targets of distal regulatory elements is challenging yet critical for interpreting regulome data. Experiment-derived enhancer-gene links are restricted to a small set of enhancers and/or cell types, while the accuracy of genome-wide approaches remains elusive due to the lack of a systematic evaluation. We combined multiple spatial and *in silico* approaches for defining enhancer locations and linking them to their target genes aggregated across >500 cell types, generating 1,860 human genome-wide distal **En**hancer to **T**arget gene **Def**initions (**EnTDefs**). To evaluate performance, we used gene set enrichment testing on 87 independent ENCODE ChIP-seq datasets of 34 transcription factors (TFs) and assessed concordance of results with known TF Gene Ontology (GO) annotations., assuming that greater concordance with TF-GO annotation signifies better enrichment results and thus more accurate enhancer-to-gene assignments. Notably, the top ranked 741 (40%) EnTDefs significantly outperformed the common, naïve approach of linking distal regions to the nearest genes (FDR < 0.05), and the top 10 ranked EnTDefs performed well when applied to ChIP-seq data of other cell types. These general EnTDefs also showed comparable performance to EnTDefs generated using cell-type-specific data. Our findings illustrate the power of our approach to provide genome-wide interpretation regardless of cell type.

Background

Enhancers, silencers and insulators are key genomic cis-regulatory elements that play pivotal roles in spatiotemporal control of gene expression by physical contact with the promoters of

target genes they control [1-3]. Promoters are located immediately upstream of the transcription start sties (TSSs), facilitating the recruitment of transcription factors and RNA polymerase II (RNAPII) to instruct the initiation and direction of gene transcription, whereas enhancers and silencers can be located anywhere in the genome and often at distal regions, such as upstream, downstream or in introns of target genes or unrelated genes. Via interaction with promoters of the target genes, enhancers are bound by activator proteins and stimulate the rate of transcription, while silencers were bound by repressor proteins and decrease the rate. In certain cases when the interactions between enhancers/silencers and promoters are unwanted, insulators can block their interactions[4]. Bound by tissue-specific transcription factors and cofactors, such as p300 and Mediator, the cis-regulatory elements and promoter connections direct what, when and how the genome is transcribed so as to control cell fate decisions during development and differentiation [5-7]. For simplicity, we will refer to these distal cis-regulatory elements as general “enhancers” (>5kb from a transcription start site [TSS]) hereinafter.

Perturbation of enhancer activities and/or functions induced by genomic variants, epigenomic dysregulation, and/or aberrant chromosomal rearrangements can underlie disease susceptibility and developmental malformations [8, 9]. A prototypic example of this is the point mutation in the *Shh* enhancer, ZRS (*zone of polarizing activity regulatory sequence*), which can lead to limb malformations such as polydactyly in humans [10]. Recently, genome-wide association studies (GWAS) identified that >88% of disease-linked variants occur within non-coding regulatory DNA [11], especially enriched in enhancers [12]. These findings confirm the importance of enhancers

63 in orchestrating transcriptional regulation and reveal that the dysregulation of enhancer function
64 contributes to the pathogenesis of a variety of diseases, referred to as “enhanceropathies” [13].
65
66 A challenge in enhancer biology is to decipher their target genes and the mechanisms underlying
67 the precise enhancer-gene interactions, which is reviewed in Pennacchio LA *et al* [14]. The
68 enhancer to target gene specificity is essential to understand how gene expression is
69 programmed during normal development and differentiation, and how the ectopic enhancer
70 and/or non-target gene interactions can lead to diseases. However, interpreting genome-wide
71 regulatory data is significantly hampered by our limited knowledge of enhancers and their target
72 genes for multiple reasons. First, enhancers are commonly located distal to their target genes
73 with multiple intervening genes in between, and greatly varying distances. One enhancer can act
74 on multiple genes and one gene can be regulated by multiple enhancers [15]. Second, enhancers
75 act in a dynamic and often cell type-specific manner, which further complicates the definition of
76 a comprehensive set of enhancers and their target genes. Third, enhancers and promoters share
77 various characteristics and functions [16, 17], thus making it challenging to disentangle the two
78 elements based on functional genomic data.
79
80 With the breathtaking progress in technologies such as massive parallel sequencing and high-
81 resolution chromosome conformation capture, our knowledge of cis-regulatory elements’
82 function and spatial organization have grown considerably over the past decade [18-23]. In most
83 cases, enhancers are located at regions distal of their target genes up to hundreds of kilobases,
84 and they can bypass more proximally located genes to bind to the promoters of the genes they

control through long-range 3D chromosomal interactions [19, 24, 25]. The 3D genome is organized in hierarchical layers, from bottom to top including chromatin loops (or insulated neighborhoods), topological associating domains (TADs), and compartments [26]. The chromatin loops are the fundamental structural and functional building blocks of genome organization, which form between two convergent CTCF (CCCTC binding factor) binding sites bound by the cohesin protein complex [27].

Large epigenomics consortia like ENCODE [28-30] and Roadmap Epigenomics [31], have generated a tremendous amount of regulatory data across various tissue and cell types, including genome-wide transcription factor (TF) binding by ChIP-seq [32], chromatin accessibility assays (e.g. DNase-seq [33], ATAC-seq [34]), genome-wide chromatin mark profiles, and 3D chromosome organization. However, enhancer-promoter interactions are still highly restricted to a small number of cell types, which is probed by Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET [35]), and the genome-wide interaction map is still limited due to the high cost of Hi-C experiments [36]. Other enhancer-promoter interaction datasets have been generated by mathematical and/or bioinformatic approaches. The FANTOM5 [37] dataset is based on the gene expression correlation between enhancer and promoter regions, and Thurman et al. exploited DNase signal correlation between enhancers and promoters using DNase-seq data [38]. However, the reliability and generalization of these approaches remains elusive due to the lack of a systematic evaluation.

Gene set enrichment (GSE) testing is widely applied to infer the regulatory networks embedded in the abundant high-throughput gene regulation data, including ChIP-seq, Bisulfite sequencing, DNase-seq, and ATAC-seq. The first step in this analysis is to assign the genomic regions identified by the assays to their target genes, and most methods simply do the assignment using the nearest genes regardless of the actual regulatory targets [39-42]. Since enhancers and their target genes have long-range chromosomal contact, adjacent gene assignments tend to link enhancers to non-target genes, leading to incorrect interpretation for distal enhancer regulation. In this study, we aimed to determine the best sets of human "enhancers" (enhancers, silencers and insulators) and their gene targets. By all possible combinations of existing experimental and/or computationally-derived datasets, we generated 1,860 **En**hancer to **T**arget gene **Def**initions, referred to as EnTDefs, and systematically evaluated their performance based on the concordance of GSE results of 87 ENCODE ChIP-seq datasets with known TF biological processes, resulting in a handful of best-performing EnTDefs. We also showed that as opposed to being random, target genes that are often missed or often falsely identified using adjacent gene assignments are biased to specific Gene Ontology terms. In addition, we compared cell-type-specific EnTDefs (CT-EnTDefs) with non-cell-type-specific ones (general EnTDefs) and found that general EnTDefs were more favorable. Our findings demonstrate that the novel, top-performing EnTDefs significantly enhance the biological interpretation for genomic region data regardless of cell type.

RESULTS

Creation and ranking of genome-wide Enhancer-to-Target gene Definitions (EnTDefs)

Several approaches to define human enhancer locations and their target genes have been proposed in the literature, but no systematic study has been performed to evaluate their performance separately or in combination on a genome-wide scale. To determine the best sets of human enhancers and their distal gene targets, we generated a total of 1,860 genome-wide Enhancer-Target gene Definitions (EnTDefs) using existing experiments and/or literature-derived data, and systematically evaluated their performance. This was done by applying all possible combinations of methods for defining *1) enhancer region locations*, identified from four data sources (ChromHMM[43], DNase-seq[38], FANTOM5[37, 44, 45] and Thurman[38]), and *2) enhancer-target gene links,* defined by four different methods (ChIA-pet data ["ChIA"][46, 47], DNAase-signal correlation ["Thurman"][38], gene expression correlation ["FANTOM5"][45] and loop boundaries with convergent CTCF motif ["L"][48]), including combinations using multiple of each (see Methods for details). Overall, these included a total of 1,768,201 possible individual enhancer-target links across >500 cell types by integrating all of the 4 enhancer-defining datasets and all of the 4 enhancer-gene link datasets. These enhancer-target links were defined from 685,921 enhancers and 21,094 linked target genes. **Figure 1** demonstrates the workflow for the creation and evaluation of these 1,860 EnTDefs. For the "L" enhancer-gene linking method, we evaluated the loops with up to 3 genes (L1: one gene, L2: $\leq$ two genes, or L3: $\leq$ three genes), allowing the links between the enhancer to any of the included genes within the loop. Because current knowledge of enhancers is far from complete and the experimental data that assay enhancers to target genes is limited, the genome coverage of EnTDefs defined by the experimentally and/or computationally derived methods (**Figure 1A:** four enhancer-defining methods and four enhancer-target gene linking methods) was expected to be low. Therefore, we

extended the enhancer regions up to 1kb and/or assigned regions outside of enhancers and promoters (within 5kb of a transcription start site (TSS)) to the gene with the nearest TSS (**Figure 1A:** Extension and Additional link), resulting in 100% coverage of distal genomic regions (>5kb of TSS). All of the 1,860 EnTDefs were evaluated and ranked based on how well they performed in gene set enrichment (GSE) testing with genes' distal ChIP-seq peaks. Specifically, the Gene Ontology biological process (GO BP) enrichment results from 87 ENCODE ChIP-seq datasets for 34 distinct transcription factors (TFs) were compared with the curated GO BP terms annotated to the same tested TFs (GO annotation by GO database) using F1 scores (*see Methods*). EnTDefs demonstrating higher concordance ranked higher, as they were better able to identify the known functions of the TFs based on their distal binding regions (non-promoters).

Overview of the EnTDef characteristics

We first investigated the characteristics of the 1,860 EnTDefs by comparing them to simply assigning distal genomic regions (i.e. >5kb from a TSS) to the genes with the nearest TSS (>5kb **Loc**us **Def**inition [LocDef]) (**Figure 2A, Supplementary Figure S1**). The EnTDefs were ranked in decreasing order by their average F1 score across 34 TFs, and the top 741 EnTDefs (~40%) were found to significantly outperform the >5kb LocDef (Wilcox signed-rank test, FDR < 0.05). The best performing EnTDef (No. 1 ranked) was defined by DNase-seq plus FANTOM5 enhancers and ChIA, Thurman and FANTOM5 enhancer-target gene link methods with the "nearest_All" addition. For the top 741 EnTDefs, the percentage of genome covered and percent of distal peaks caught (outside of 5kb regions around TSSs) was as high as 100% (89% - 100%), the median number of genes assigned to each enhancer was 2 (range of 1 - 2), and the median number of enhancers

assigned to each gene was 20 (range of 2 - 98). Out of the 741 EnTDefs, those ranked 2 through 19 were not significantly worse than the best performing EnTDef (Wilcoxon signed-rank test, p > 0.01. **Supplementary Table S1**), suggesting that these 19 EnTDefs performed equally well. This finding was robust to the specific set of GOBP annotations used (i.e. with or without IEA-based GO to gene annotations; see *Methods*, data not shown).

By examining the types of methods used to generate the top 741 EnTDefs (**Figure 2B**), we found that: *i)* over half of them included FANTOM5 (54%), Thurman (53%) and DNase-seq (51%) enhancer regions, while chromHMM defined enhancers were least used (~45%); *ii)* approximately 60% of them were generated without enhancer regions extension; *iii)* all of them used the "nearest_All" addition to assign the distal regions that were not in enhancers to the nearest gene's TSS; and *iv)* the "ChIA" method applying ChIA-PET data to assign enhancers to target genes was included most frequently (~60%), followed by FANTOM5 (~51%) and Thurman (42%), whereas the "L" method assigning genes to enhancers within the same loop boundaries with convergent CTCF motifs was least used (~26%). It is not surprising that all of the top 741 EnTDefs included the "nearest_all" addition, because this significantly increased genome coverage by assigning all regions outside enhancers and promoters to the nearest distance gene (>5kb LocDef), leading to improved sensitivity and thus F1 score (**Figure 2A**). On the other hand, the fact that these 741 EnTDefs outperformed the >5kb LocDef suggests that the "smart" enhancer to target gene assignments more accurately capture real biological regulatory elements for distal enhancer regions when compared to the simplistic assignment to nearest genes. In addition, all of the top performing EnTDefs were generated using combinations of at least two different datasets/methods for enhancer definitions (ChromHMM, DNase-seq, FANTOM5,

and/or Thurman) and enhancer-gene assignments (ChIA, FANTOM5, L, and/or Thurman), illustrating the importance of high genomic coverage and that the integration of multiple data sources and methods indeed improves the performance of enhancer to target gene assignments.

EnTDefs plus promoter regions outperforms the nearest TSS method

Our analyses thus far have focused on the assessment of distal gene regulation. However, often the goal is to assess the functional regulation from anywhere in the genome, including binding both distal and proximal to TSSs. One commonly used method for ChIP-seq GSE testing is to link all peaks to the nearest gene, hereinafter referred to as the "nearest TSS" method (**Supplementary Figure S1**: "nearest TSS" LocDef), resulting in all peaks having at least one assigned gene. EnTDefs were generated for distal regions (outside the 5kb windows around TSSs) and any regions within 5kb of a TSS were ignored, whereas the "nearest TSS" method includes all genomic regions. Thus, to compare fairly with the "nearest TSS" method, we added promoter regions to the top 10 ranked EnTDefs, referred to as "EnTDef_plus5kb". That is, peaks within 5kb of a TSS were assigned to the nearest gene (**Supplementary Figure S1**: "5kb" LocDef), while distal peaks were assigned according to the EnTDef. All ten of the EnTDef_plus5kbs significantly outperformed the "nearest TSS" method (~0.05 increase in average F1 score, Wilcoxon signed-rank test, $p < 0.0001$) (**Figure 2C**), using the same evaluation method based on F1 scores as used above (see *Methods*).

We next determined if our 'smart' EntDefs using only distal binding events could even outperform the use of all peaks (promoter and enhancer) with naïve assignments to the genes with the nearest TSS. When compared with the "nearest TSS" method, the top 10 best performing EnTDefs showed slightly lower F1 scores (~0.03 lower), but the difference among the

top half of them were not significantly different from "nearest TSS" (Wilcoxon signed-rank test, $p > 0.05$). Thus, although they did not outperform it, the best were not significantly worse. This illustrates the great importance of regulation from promoters in GSE testing.

Two other commonly used GSE methods for genomic regions, GREAT[39] and Fisher's exact test (FET) using peaks within 5kb of a TSS (**Supplementary Figure S1**: "5kb" LocDef), were also evaluated using the same scheme. Notably, the three GSE testing methods (Poly-Enrich, GREAT and FET using 5kb LocDef to assign peak to gene) performed equally well (Friedman test, $p = 0.91$), but significantly worse than the top 10 EnTDefs (Figure 2C, average F1 = 0.45 vs 0.47, Wilcoxon rank-sum test, $p < 0.007$). In addition, both top 10 EnTDefs and 5kb LocDef (i.e. assigning promoters to the nearest gene) significantly outperformed >5kb LocDef (i.e. the naïve approach of assigning distal regions to the nearest gene) (average F1 = 0.47, 0.45 vs 0.27, Wilcoxon signed-rank test, $p = 2.37\times10^{-14}$ and 1.32×10^{-8} respectively). In summary, although the naïve approach of linking distal regions to the nearest gene (>5kb LocDef) did not outperform the use of promoter data only (5kb LocDef), the use of distal binding events with 'smart' gene assignments (EnTDefs) did outperform the use of promoter data only. These findings illustrate the importance of accurately modeling regulation from enhancers, and that when done well, enhancers have the potential to provide more regulatory information than promoters. We conclude that GSE testing using our top EnTDefs exceeds the commonly used nearest distance-based and promoter-only based GSE approaches.

Our EnTDefs are generalizable to different cell lines

Next, we sought to investigate whether the EnTDefs (which were selected based on their performance in GM12878, H1-HESC and K562 cell lines) can perform equally well on testing ChIP-seq data from different cell lines (A549, HEPG2, HUVEC and NB4). Surprisingly, the average F1 score of the top 10 EnTDefs across the *test* ChIP-seq datasets (different cell lines) was significantly higher than that from the *evaluation* ChIP-seq datasets (average F1 = 0.60 vs. 0.56, Wilcoxon sum-rank test, p = 0.0059) (**Figure 2D, and Supplementary Figure S2**). This may be due to the test ChIP-seq datasets containing more peaks than the evaluation datasets (**Supplementary Figure S3A**, Wilcoxon sum-rank test, p = 0.092), and indeed we found that the *F1 scores* were significantly corelated with the number of peaks (**Supplementary Figure S3B**, Pearson's correlation r = 0.65, p = 4.57×10^{-6}). The findings indicate that the performance of the top selected EnTDefs are independent of the cell types of ChIP-seq datasets, but likely strongly influenced by the quality of the datasets themselves. We reasoned that the EnTDefs were created based on the combinations of diverse data sources stemming from >500 different cell types, resulting in a consensus set of enhancer and gene assignments across various cell types, and therefore representative of the background interactions between enhancer and target genes across many cell types. The high generalizability of our top EnTDef makes it feasible to integrate with GSE testing in a cell type-independent manner.

General EnTDefs perform comparably to cell-type-specific EnTDefs

To contrast with the EnTDefs generated by integrating data for many cell types, hereafter called "general EnTDefs", we created "cell-type-specific EnTDefs" (CT-EnTDef) using ChIA-PET datasets of a particular cell type. Since many enhancers and regulatory relationships between enhancer and target genes are considered to be tissue and cell-type-specific, we sought to examine how

the general EnTDefs perform when compared with CT-EnTDefs. For each tested TF (the average number of TFs tested in each cell type is ~60 [6 - 104], see **Supplementary Table S2**), F1 scores were calculated (see *Methods*) and compared between a pair of EnTDefs which were created using the same methods (the same combinations of enhancer and enhancer-gene link methods) but based on data from different cell types: *i)* general EnTDef vs. CT-EnTDef using the same cell type (same-CT-EnTDefs), *ii)* general EnTDef vs. CT-EnTDef using a different cell type (diff-CT-EnTDefs), and *iii)* same CT-EnTDefs vs. different CT-EnTDefs. Notably, there was no significant difference in F1 scores among the three comparative EnTDefs for cell types GM12878, H1HESC or K562 (**Figure 3A**, Kruskal-Wallis test, $p \geq 0.5$). For MCF7, the same-CT-EnTDef performed near significantly better than the general CT-EnTDef (Wilcoxon sum-rank test, $p = 0.03$; three groups: Kruskal-Wallis test, $p = 0.059$). It is worth noting that only four TFs were tested in cell type MCF7, whereas ≥46 TFs were tested in the other three cell types, so this exception might not be visible for a larger set of TFs.

We also observed that the average F1 scores of TFs across all possible EnTDefs were significantly correlated between the same-CT-EnTDef and diff-CT-EnTDef for all four cell lines (**Figure 3B and Supplementary Figure S4**, Pearson's correlation > 0.9, $p < 0.0001$). This implies that the performance of the results is driven more by the quality and quantity of input data than by whether a general or CT-EnTDef is used. The trend of correlation still held at the individual TF and EnTDef level (F1 score per TF per EnTDef, **Supplementary Figure S5**). As shown in **Figure 3C**, regardless of the type of EnTDef (general EnTDefs, same-CT-EnTDefs and diff-CT-EnTDefs) used for evaluation, the average F1 score across all TFs and EnTDefs were similar, with the difference ranging from 0 to 0.14. Taken together, these findings suggest that CT-EnTDefs are overall

comparable to general EnTDefs, and the benefit of using CT-EnTDefs is minor and depends on the quality and quantity of data for a particular cell type (e.g. MCF7 in **Figure 3C**). This is good news since it is costly and difficult to generate cell-type-specific ChIA-PET experiments, which are required to create the corresponding CT-EnTDef. In contrast, the general EnTDefs, which capture real enhancer and target gene interactions in a similar way to CT-EnTDef, are more practically and economically favorable for GSE testing.

Incorrect gene assignments by nearest distance method are not random

Since enhancers are known to be located up to 1Mbp away from their regulatory genes [14, 49], several interceding genes can reside between a TF binding site (peak) in an enhancer and its target gene(s), as modeled by our EnTDefs (**Supplementary Figure S6**). In contrast, the nearest distance method simply links a peak to the gene with the nearest TSS without accounting for interceding genes. By ranking the genes based on the average number of interceding genes across the enhancers that target them, we investigated whether the number of interceding genes is randomly distributed across genes and GO terms, or if there are GO terms significantly enriched with genes having more or fewer interceding genes[50]. We investigated the best performing EnTDef excluding the "nearest_all" addition, in order to assess the 'smart' enhancer-target links only. The genes least likely to have interceding genes were found to be significantly enriched in *G protein-coupled receptor activity* (FDR = 1.41×10^{-14}), *olfactory receptor activity* (FDR = 6.21×10^{-12}), *detection of chemical stimulus* (FDR = 3.23×10^{-11}), *phenol-containing compound metabolic process* (FDR = 1.91×10^{-4}), *GABA-ergic synapse* (FDR = 2.35×10^{-4}), *RISC complex* (FDR = 2.39×10^{-4}), *postsynaptic membrane* (FDR = 4.13×10^{-4}) and *behavior* (FDR =4.71×10^{-4}) (**Figure 4A**). These

GO terms enriched with genes least likely to have interceding genes (lower ranked genes) are most likely to be correctly assigned by the nearest distance method (**Supplementary Figure S1**: >5kb LocDef), and thus most easily detectable by current GSE testing. Conversely, the GO terms enriched with higher numbers of interceding genes (upper ranked genes) were *mRNA metabolic process* (FDR = 8.09×10^{-8}), *regulation of catabolic process* (FDR = 8.40×10^{-8}), *chromatin organization* (FDR = 2.53×10^{-7}), *kinase binding* (FDR = 1.75×10^{-6}), *heterocycle catabolic process* (FDR = 3.22×10^{-6}), *chromatin* (FDR = 7.25×10^{-6}), *hemopoiesis* (FDR = 9.47×10^{-6}) and *RNA processing* (FDR = 2.27×10^{-5}) (Figure 4A). Those GO terms are least likely to be assigned by the nearest distance method, and most likely missed using current methods for GSE testing.

To determine if this observation is robust to different EnTDefs, we performed the same analysis on all top 10 best performing EnTDefs without the "nearest_all" addition, and combined the results by calculating FDR-adjusted harmonic mean p-values, followed by removing redundant terms (see *Methods*). Consistently, *G protein-coupled receptor activity*, *olfactory receptor activity*, *RISC complex* and *postsynaptic membrane* were still the top 5 enriched terms for the genes with fewer interceding genes, and similarly, *regulation of catabolic process*, *chromatin organization*, *kinase binding* and *heterocycle catabolic process* were the top 5 enriched terms in upper ranked genes with more interceding genes (**Figure 4B**). These findings indicate that both the genes with the most and fewest interceding genes are not random: chemical stimulus and neuron-related genes can be easily assigned with the nearest distance method, whereas metabolic processing and chromatin organization genes may be frequently missed. It is concordant with the knowledge that enhancers regulate genes via long-range chromatin interactions, which are able to be captured by our EnTDefs.

Nearest distance assignment method leads to false positive and false negative GSE results

Our finding that the nearest distance method tends to incorrectly assign TF binding sites in some biological functions more than others triggered us to investigate how this bias affects the results of GSE testing. By comparing the GSE results from assigning distal regions to the nearest genes (i.e. using >5kb LocDef) to those using the best-performing EnTDef, and evaluating using TF GO annotations as above (see Methods), we identified all false positive (FP) and false negative (FN) enriched GO terms by the >5kb LocDef, that were correctly called by our EnTDef. First, we ranked the FP GO terms in descending order by significance across the 34 tested TFs. These FPs represent GO terms with genes that tend to be in between enhancers and their targets (i.e. the interceding genes), and they are not annotated to the TFs of interest. For instance, the GO term *blood vessel morphogenesis* is annotated with *PROK1*, an interceding gene between an enhancer bound by the TF *YY1* (peak:1503) and its target gene *SLC16A4* (**Figure 5A left panel**). Using the nearest distance method, the *YY1* binding site would be falsely assigned to the nearest gene *PROK1,* rather than to the real target gene *SLC16A4*, leading to the enrichment of *blood vessel morphogenesis* that is not regulated by the TF *YY1* (i.e. a FP term by >5kb LocDef).

The top ranked FP biological processes (**Figure 5B**) included those related to development and regulation (*gland development, skeletal muscle organ development, regulation of DNA biding, and regulation of neuron death*), metabolic process or response to different stimuli (*alcohol metabolic process*, *response to calcium ion*, *response to drug*), and cell differentiation processes (*Stem cell differentiation*, *lymphocyte differentiation* and *T cell activation*). Among them, *blood vessel morphogenesis* was identified to be the most often falsely enriched term by >5kb LocDef (Fisher's FDR = 9.48×10^{-14}) with five out of seven TFs falsely called as positives (FP rate = 71%). It

344 is noteworthy that the Fisher's p values of FP terms calculated based on >5kb LocDef GSE results
345 were overall significantly lower than those based on EnTDef GSE results (Wilcoxon rank-sum test,
346 p= 3.95×10^{-09}), indicating that those FP terms were much less likely identified by our EnTDef
347 method.

348 On the other hand, common FN GO terms failed to be identified by >5kb LocDef while being
349 successfully identified by the top EnTDef. In contrast to the FP GO terms, FN terms tend to be
350 missed by assignment to the nearest distance gene, but correctly identified using the top EnTDefs.
351 As shown in **Figure 5A** (right panel), the GO term *histone modification* contains the gene *CHD3*,
352 the gene target of an enhancer bound by the TF *CTCF* (the peak:6206). Using the nearest distance
353 method, the *CTCF* binding site would be falsely assigned to the nearest gene *ALOX15B,* rather
354 than the target gene *CHD3*, failing to identify *histone modification* as regulated by the TF *CTCF*
355 (i.e. a FN term by >5kb LocDef). These FN GO terms consistently point to chromosome
356 organization and modification processes, including *Protein-DNA complex subunit organization*
357 (FN rate = 50%), *histone modification* (FN rate = 40%), *Peptidyl-lysine modification* (FN rate = 38%),
358 and *regulation of chromosome organization* (FN rate = 25%) (**Figure 5C**). Their Fisher's p values
359 based on the >5kb LocDef GSE results were significantly higher than those based on the EnTDef
360 GSE results (Wilcoxon rank-sum test, p= 6.24×10^{-15}), in line with the finding that chromosome
361 organization-related terms are very often missed by the nearest distance method. Together,
362 these results demonstrate that GSE analysis using the nearest distance gene assignment method
363 cannot always identify biological processes induced by long-range chromosome organization;
364 rather, they tend to favor development and cell differentiation functions which are not related
365 to the TFs. In contrast, the EnTDef method can successfully detect distal enhancer and target

366 gene interactions even for biological processes with complex long-range interactions such as
367 chromosomal organization-related terms, and avert potential false positive results.

368 **Guidance for selecting a peak-to-gene assignment method in GSE analysis**

369 The first step in GSE testing of cis-regulome data, such as TF binding sites or chromatin marks
370 from ChIP-seq, is to assign the genomic regions or peaks to their target genes. The different
371 assignment methods can lead to variable enrichment results and FP and/or FN findings, as
372 discussed above (nearest distance method vs. EnTDef). To avoid misinterpretation of genome-
373 wide regulatory data, we need to select an appropriate LocDef method with care, which should
374 be specific to the particular research question and the genomic regions of interest. **Figure 6**
375 summarizes three general categories of research questions and the corresponding regions of
376 interest: *i)* the 5kb or 1kb LocDef should be selected when interested in how a TF and/or
377 chromatin mark regulates gene expression from promoters; *ii)* the EnTDef (enhancer) should be
378 selected when interested in how a TF and/or chromatin mark regulates gene expression from
379 distal regions; and *iii)* when the comprehensive regulatory signature is of interest, including both
380 promoter and distal regions, our EnTDef plus 5kb LocDef (enhancer.5kb) should be selected. The
381 promoter LocDef has the lowest genome coverage (10% for <5kb LocDef and 2% for <1kb LocDef),
382 while the EnTDef plus 5kb has 100% genome coverage, and the EnTDef has intermediate genome
383 coverage (90%). We incorporated our top performing EnTDef and EnTDef.plus5kb into the
384 Bioconductor package *chipenrich* [42] and the ChIP-Enrich website (https://chip-
385 enrich.med.umich.edu), allowing users to select the most suitable genomic regions-gene
386 assignment methods, gene sets and GSE method to correctly interpret their genome-wide
387 regulatory data. In addition, we provide a peak-to-gene assignment functionality in our GSE Suite

388 (http://gsesuite.dcmb.med.umich.edu), by which users can select any possible combination of
389 enhancer location and enhancer-to-gene target methods (as described in this study) and obtain
390 the gene assignments for a user uploaded list of genomic regions, based on the selected EnTDef,
391 or other method (e.g. promoters, exons, introns or anywhere in the genome).

392

393 **DISCUSSION**

394 A greater appreciation of the central role that distal regulatory elements play in genetic diseases
395 and cancers has motivated a multitude of enhancer studies. As a result of the increasing
396 availability of functional genomics data, growing attention has been paid to matching Enhancer-
397 Target Gene pairs (ETG) in the field of computational biology and genomics. Over the past decade,
398 a variety of algorithms and tools have been developed by leveraging multiple genomic features
399 and functional data, as recently reviewed in [51]. Briefly, they can be categorized into four groups:
400 *1)* correlation-based (e.g. Thurman *et al* [38], PreSTIGE [52], ELMER [53, 54], etc.); *2)* supervised
401 learning-based (e.g. IM-PET [55], TargetFinder[56], McEnhancer[57], etc.); *3)* regression-based
402 (e.g. RIPPLE[58], JEME[59], FOCS[60], etc.); and *4)* score-based methods (e.g. EpiTensor[61],
403 GeneHancer[62] and PEGASUS [63, 64]). Although these algorithms have significantly advanced
404 our knowledge of ETGs, they are affected by one or more of the following issues: *1)* the lack of a
405 genome-wide exhaustive reference list of enhancers; *2)* the lack of a large gold standard which is
406 required for supervised learning algorithms, i.e. experimentally-validated true positive and true
407 negative enhancer-target gene pairs, and *3)* the lack of a systematic evaluation of their reliability
408 and generalization in various cell types. To overcome these issues, we developed a gold standard-

free approach to generate and prioritize comprehensive sets of ETGs based on their performance in the interpretation of regulome data.

In this study, we identified a best set of **En**hancer to **T**arget gene **Def**initions (**EnTDefs**) by investigating and evaluating all possible combinations of existing reliable sources for human enhancer location definitions and enhancer-target gene pair definitions across various cell types. Purposely, we coupled EnTDefs with GSE testing to systematically evaluate their performance when interpreting regulome data. By carefully selecting datasets of high quality and resolution, we explored ENCODE ChromHMM, DNase-seq, FANTOM5 and Thurman datasets for enhancer regions and ENCODE ChIA-PET interactions, Thurman DHS correlation-based, FANTOM5 and ENCODE ChIA-PET CTCF loops-based enhancer-target gene interactions. We also systematically evaluated the performance of all possible combinations of datasets when applied on ENCODE TF ChIP-seq data in GO GSE testing and compared the enriched GO terms with the curated TF GO annotations (TF-annotated GO BP terms by the GO database). In contrast to the statistical model-based or machine learning-based algorithms as described above, our approach integrates various data sources and directly couples the EnTDefs with GSE testing for a systematic evaluation, resulting in an EnTDef with maximally balanced sensitivity and specificity (assessed by F1-score). Our approach to generating EnTDefs is assumption-free and independent of true positive/negative pairs, but based on a systematic evaluation using GSE testing. The results demonstrate that the DNase-seq and FANTOM5 enhancers with the integrated enhancer-target gene pairs from ChIA-PET, Thurman and FANTOM5 interactions performed best, suggesting that both chromosome accessibility and conformation, as well as transcriptional correlation, are beneficial for identifying enhancer-target regulatory relationships.

The nearest distance method, which naively assigns genomic regions of interest to the nearest gene, is commonly used in GSE for regulome data. Our analysis showed that this naïve approach commonly fails to identify certain functions, such as those related to chromosome organization, which lead to false negatives, whereas our top EnTDefs can successfully identify these long-range enhancer regulatory functions. These findings re-confirmed that universally assigning cis-regulatory elements to the gene with the nearest TSS is problematic, resulting in misleading and/or incomplete functional interpretation.

Our EnTDefs were generated by leveraging different genomic data across >500 cell types and can be applied to different cell types, demonstrating performance comparable to their cell-type-specific counterparts. Our top integrated EnTDef based on many cell types represents a comprehensive set of enhancer regions (only a subset of which will be active in any one cell type); our data indicate this performs well because current cell-type-specific enhancer-target genes (ETGs) are not yet sufficiently comprehensive (except for a few cell types such as GM12878). Research performed on cancer samples, less commonly used cell lines, and other complex tissue samples will greatly benefit from this integrated EnTDef. While cell-type-specific ETGs are important for studying regulation at specific locations, our results demonstrate that for genome-wide approaches such as GSE, the comprehensiveness outweighs the need for specificity.

Besides DNase-seq, ChIA-PET, CAGE-seq, and RNA-seq data, Hi-C and eQTL data are also used to infer ETG [62, 65]. However, we found that current Hi-C data often have insufficient resolution, with genomic windows being a few to several kb wide due to low coverage, and high quality Hi-C data is not available for nearly as many cell types as the other approaches. Although eQTL data is available for many tissues and cell types, it is similarly restricted by limited population diversity

453 and low resolution. The tissue-specific eQTL data from the GTEx project [66] is widely used,
454 however, it was generated for only 49 tissues from <1000 donors with the majority being
455 Caucasian (84.6%), making it difficult to apply to other tissues/populations. In addition, eQTL data
456 is highly correlated with the linkage disequilibrium, and thus its resolution is associated with the
457 size of haplotype blocks, which is highly variable across populations (on average ~10kb) [67],
458 whereas enhancers are usually short genomic regions (50-1,500bp). Due to this low resolution of
459 Hi-C and eQTL data, we excluded them from our analysis. On the other hand, another data-
460 integration method, HACER, was recently developed [68], which utilized the nascent eRNA
461 information from GRO-seq and PRO-seq, along with the FANTOM5 CAGE-seq data, to identify
462 cell-type-specific ETGs. In future work, we will incorporate the GRO-seq and PRO-seq data
463 deposited in HACER and evaluate if the new datasets can further boost our EnTDef performance
464 when coupled with GSE.

465 In conclusion, we identified a best set of enhancer-target gene pairs (EnTDef) by leveraging
466 existing data sources of chromosome accessibility and/or conformation and transcriptome data
467 across numerous cell types, which significantly improved the biological interpretation of distal
468 regulation in GSE compared to assigning genomic regions to the nearest gene. Our approach
469 performs well across a wide range of cell types, making it feasible to apply on extensive genomic
470 data sets. The limitations of our EnTDef are inherited from the existing data sources, including
471 low genome coverage, low resolution, and small number of cell types with good quality ChIA-PET
472 data. With the continued growth in volume of functional genomics data and advances in data
473 quality and resolution, we expect further improvement of our EnTDef in the future.

474

Conclusions

In summary, we provide an optimized enhancer-to-target gene assignment approach, which is critical for interpreting genome-wide regulatory data. This study has important implications for which type of enhancer-target gene methods are most accurate, and the relative importance of comprehensiveness versus cell-type specific accuracy. To the best of our knowledge, there is currently no such a comprehensive resource of distal regulatory region-to-target gene links which are feasible to apply on various types of regulome data (eg.ChIP-seq, ATAC-seq) regardless of cell types.

METHODS

Generation of general enhancer-target gene definitions

We generated genome-wide definitions of human distal enhancer locations and their target gene assignments for the hg19 genome using all possible combinations of the below enhancer location methods and enhancer-gene linking data (**Figure 1A**: Enhancer, Extension, Enhancer-target gene link and Additional links). These are based on enhancers from: *1)* "ChromHMM": ENCODE ChromHMM UCSC tracks (9 cell types) [43], *2)* "DNase-seq": DNase hypersensitive sites (DHSs) from 125 cell types processed by ENCODE [38], *3)* "FANTOM5": Cap Analysis Gene Expression (CAGE) experiment-derived enhancers across 421 distinct cell lines/tissue/primary cells from FANTOM5 project [37, 44, 45], and/or *4)* "Thurman": distal and non-promoter DHS within 500 kb of the correlated promoter DHSs from 79 cell types, referred to as the first author of the publication [38]. Since our motivation was to identify the target genes of distal regulatory

496 elements that do not have clear target genes based on close proximity to a TSS, we constrained
497 the enhancer regions to be outside of 5 kb from a transcription factor start site (TSS) by trimming
498 the bases from the above defined enhancers overlapping with the 5kb windows of TSSs. The hg19
499 TSS locations were obtained from the Bioconductor *chipenrich* package version 3.5.0 [42]. To
500 identify target genes, we used: *1)* "ChIA" method: enhancer and gene interactions identified by
501 ChIA-PET2 using default parameters [69] from 10 ChIA-PET datasets of 5 cell types
502 (**Supplementary Table S3**) [46, 47], *2)* "Thurman" method: the enhancer and promoter
503 interactions identified by Thurman *et al*, which were defined by high correlation ($r > 0.7$) between
504 cross-cell-type DNase I signal at each DHS position and all promoters within ±500 kb [38], *3)*
505 "FANTOM5" method: the regulatory targets of enhancers predicted by correlation tests using the
506 expression profiles of all enhancer-promoter pairs within 500kb[45], and *4)* "Loop" method: any
507 possible interactions between enhancers and genes that are encompassed within in a RAD21,
508 cohesin and/or CTCF ChIA-PET loop with convergent CTCF motifs [48], and depending on the
509 number of genes included in the loop, this method was referred to as "L1" (one gene), "L2" (≤
510 two gene) or "L3" (≤ three genes) (**Figure 1B**).

511 All possible combinations of the above, allowing multiple at a time, defined 465 of the **En**hancer
512 to **T**arget gene **Def**initions (EnTDefs) (**Figure 1A, B)**. In addition, to increase the genome coverage,
513 we tested extending the enhancer regions to 1kb (i.e. "enhancer extension", 500 bp extension at
514 both sides of the midpoint), and assigning regions outside of enhancers and promoters (within
515 5kb of a TSS) to the gene with the nearest TSS (i.e. "nearest_all" additional links). The additional
516 combinations using these options brought us to a total of 1,860 distinct EnTDefs.

Evaluation of enhancer-target gene definitions

To evaluate the performance of each individual EnTDef, we performed Gene Ontology (GO Biological Processes [GOBP]) enrichment testing using Poly-Enrich [70] in the *chipenrich* Bioconductor package[42] on 87 ChIP-seq datasets of 34 TFs selected from the tier 1 ENCODE cell lines (**Supplementary Table S4**). We then compared the significantly enriched GOBP terms with the GO BP annotations of each TF (i.e. the GOBP terms assigned to the 34 TFs by the GO database, excluding the terms with <15 or >2000 assigned genes) (**Figure 1C**: Evaluation of the Enhancer-Target gene Definition), to identify the EnTDefs with greatest concordance. The assumption of this approach, used previously in[70, 71], is that TFs tend to the regulate genes in the biological processes to which they belong, and thus greater overlap with TF GO BP annotation indicates more accurate enrichment results, and thus more accurate peak-to-gene assignments. For a full justification of this, see Supplementary Methods. To minimize runtime for the initial pass analysis, we used the PE.Approx method (an approximate version of Poly-Enrich[70], see Supplementary Methods and **Supplementary Figure S7**). To alleviate the bias caused by the unbalanced number of positive and negative assignments, we generated the same number of true negative assignments for each TF as there were positive by randomly selecting GOBP terms from the set that were not assigned to the particular TF, and excluding the offspring terms and their siblings of the assigned terms (hereafter called “true negative” terms, depicted in **Supplementary Figure S8A**). In order to control for the confounding of GOBP size (i.e. the number of assigned genes to each GOBP term), random sampling was performed among the negative terms of comparable size to the corresponding true positive term (bin size = 20). In each sampling, the PE results were assessed by the number of true positive (TP), false positive (FP), true negative (TN) and false

539 negative (FN) GOBP terms according to the following definitions: *1)* TP: the number of GOBP
540 terms that were significantly enriched (FDR < 0.05) and assigned to the TF by the GO database;
541 *2)* FP: the number of GOBP terms that were significantly enriched (FDR < 0.05), but not assigned
542 to the TF by the GO database; 3) TN: the number of GOBP terms that were not significantly
543 enriched (FDR > 0.05 or "depleted") and also not assigned to the TF; and 4) FN: the number of
544 GOBP terms that were not significantly enriched (FDR > 0.05, or "depleted"), but assigned to the
545 TF. The *F1 score* ($F1\ score = 2 \times \frac{Precision \times Recall}{Precision+Recall}$) was calculated to measure the overall
546 performance of an EnTDef for a TF. We repeated the sampling process 10 times, and took the
547 average *F1 score* for each EnTDef and TF. The average of these F1 scores across TFs provided the
548 final ranking for each EnTDef.

549 To assess the robustness of our approach, we also evaluated the performance of EnTDefs using
550 more conservative GO annotations, in which GOBP assignments based on "automatically
551 assigned, inferred from Electronic Annotation" (IEA) were excluded, thus minimizing false
552 annotations in GOBP. For the positive GO annotations we used only the leaf GO terms (the lowest
553 level in the GO hierarchical tree) and their parent and grandparent GO terms, while the negative
554 terms were sampled from all other terms, excluding positive terms, and ancestors of positive
555 terms, siblings of ancestors of positives, as well as offspring of positives (depicted in
556 **Supplementary Figure S8B**).

557 After ranking all EnTDefs in descending order by their average F1 scores (**Figure 1C**: Rank of
558 EnTDef), we identified the set of best EnTDefs. Paired Wilcox signed-rank tests were performed
559 to compare the F1 score of the 1st ranked EnTDef with each of the sequential ones, and the rank

at which the EnTDef showed significantly lower F1 score than the 1st ranked one ($p < 0.01$) was selected as the cutoff. The EnTDefs ranking above the cutoff were defined as the best set of EnTDefs. In addition, we performed the same F1 score evaluation of previously-defined methods for genomic region-to-gene assignments, termed gene **Loc**us **Def**initions (LocDefs, see **Supplementary Figure S1** for details) that do not use "smart" enhancer-target links (i.e., ">5kb": distal regions assigned to the gene with the nearest TSS; "<5kb": regions within 5kb of a TSS assigned to the gene with that TSS; and "nearest TSS": all regions assigned to the gene with the nearest TSS). These LocDefs are used by Poly-Enrich in the *chipenrich* R Bioconductor package [42], and represent the current standard practice for enhancer-to-gene assignments for gene set analysis. The F1 scores were compared between each EnTDef and the distal nearest distance (">5kb") LocDef by Wilcoxon signed-rank tests. We also evaluated and compared two commonly used Gene Set Enrichment (GSE) testing methods, Fisher's Exact Test (FET) and GREAT [39], which were implemented by the R *chipenrich* package using the FET and binomial method respectively, coupled with the "5kb" LocDef. To obtain a final assessment, a second round of GSE testing using Poly-Enrich (*PE.Exact* method; see Supplementary methods for details) was applied on the subset of EnTDefs which significantly outperformed the nearest distance assignments (>5kb LocDef), and the average F1 scores were calculated and used to refine the final ranking of EnTDefs.

Validation of the EnTDefs with ChIP-seq data from different cell types

To further evaluate the performance of EnTDefs in different cell lines, we selected 13 additional ENCODE ChIP-seq datasets from four non-tier 1 ENCODE cell lines (A549, HEPG2, HUVEC and NB4), which contain ChIP-seq experiments for at least three TFs in each cell line (**Supplementary**

581 **Table S5).** In comparison to the 87 *evaluation* ChIP-seq peak sets from ENCODE tier 1 cell lines
582 (GM12878, H1HESC and K562), these 13 datasets are the *test* datasets. The six TFs (*C-JUN, C-MYC,*
583 *CEBPB, CTCF*, *MAX* and *NRSF*) assayed by the 13 *test* datasets were also included in the 87
584 *evaluation* datasets. The top 10 best EnTDefs were evaluated using the *PE.Exact* method as
585 described above and these 13 test ChIP-seq datasets. For each EnTDef, the average F1 score
586 across the 13 ChIP-seq datasets was calculated and compared with the average F1 score
587 generated using the evaluation ChIP-seq datasets (n=16) of the corresponding TFs.

588 **Generation of cell-type-specific EnTDefs**

589 We used "ChIA" and/or "L"-derived enhancer-to-gene assignment methods (**Figure 1B**) to
590 generate cell-type-specific enhancer-target gene definitions, hereafter called CT-EnTdefs. Since
591 the enhancer-gene linking data defined by Thurman and FANTOM5 datasets were non-cell-type
592 specific, we did not include these. The cell types were selected based on the availability and
593 quality of cell-type specific ChIP-seq and ChIA-PET data in ENCODE. As shown in **Supplementary**
594 **Table S5**, four cell types were selected: GM12878 (tier 1), H1-hESC (tier 1), K562 (tier 1) and MCF7
595 (tier 2). The multiple ChIA-PET datasets were combined for each cell type. All combinations of
596 enhancer location definitions, along with ChIA, L1 (or L2 or L3) enhancer-gene assignment
597 methods, with or without enhancer location extension and "nearest_all" addition, were used to
598 generate the CT-EnTDefs, resulting in a total of 630 CT-EnTDefs for each of the 4 cell types.

599 **Evaluation of CT-EnTDefs**

600 To evaluate the performance of the CT-EnTDefs, we performed GSE testing of Gene Ontology
601 (GO Biological Processes [BOBP]) using Poly-Enrich [70] on the TF ChIP-seq peak sets of the same

cell type from which each CT-EnTDef was generated (**Supplementary Table S5**. See details as described above). For comparison, we also applied GSE testing on the same TF ChIP-seq peak sets using the corresponding general EnTDefs (i.e. not cell-type specific, using the same enhancer regions and target gene link methods as those of the comparative CT-EnTDef), as well as the CT-EnTDef from a different cell type (Figure 3C, i.e. MCF7 CT-EnTDefs were applied on GM12878 TF ChIP-seq peaks, GM12878 CT-EnTDefs on H1hESC peaks, H1hESC EnTDefs on K562 peaks, K562 EnTDefs on MCF7 peaks). For each TF, the average *F1* scores across all evaluated EnTDefs were calculated and compared between using the respective CT-EnTDef, general EnTDef and different cell type CT-EnTDef. For each cell type, Pearson's correlation test was used to evaluate the pair-wise correlation among the F1 scores, and Wilcoxon Rank-Sum test was used to compare their differences. Finally, the overall performance of CT-EnTDefs, general EnTDefs and different cell type CT-EnTDefs were assessed using the average F1 scores across all evaluated EnTDefs and TFs in each cell type.

Testing for functions that have significantly more or fewer interceding genes between enhancers and their target genes

We investigated the number of interceding genes between an enhancer and its target gene(s) (i.e. genes between the entire region of an enhancer and the target gene in an EnTDef, depicted in **supplementary Figure S6**), and ranked all target genes based on their average number of interceding genes. By definition of nearest distance enhancer-target gene assignment (e.g. >5kb LocDef), the bottom genes with low numbers of interceding genes are most likely to be correctly assigned to their enhancers, while the top ranked genes with high numbers of interceding genes

are least likely assigned to their true enhancers. We used the best performing EnTDef without "nearest_all" addition, defined by DNase-seq plus FANTOM5 enhancers and ChIA, Thurman and FANTOM5 enhancer-target gene link methods, as an example for this analysis. Gene Ontology (GO) enrichment testing was performed by *LRpath*[50] using GO Cellular Component (CC), Biological Process (GOBP), and Molecular Function (MF) terms of size ranging from 10 to 1000 genes. The rank-based inverse normal transformation (INT) implemented by the *rankNorm* function in R package *RNOmni*[72] was applied to the average number of interceding genes to have approximately normally distributed scores. *LRpath* took the genes that were linked to at least one enhancer and their exponential transformed INT scores as the input (the input scores were log transformed internally by *LRpath* program) and performed logistic regression-based enrichment testing on each GO term. The significant GO terms (FDR < 0.05) with positive coefficients indicate the functions enriched in genes with less interceding genes (lower ranked), while those with negative coefficients are functions enriched in genes with more interceding genes (higher ranked). For reporting purposes, we filtered out closely related GO terms, using the *GO.db* R package [73] to determine relationships among significant terms. A GO term was filtered if one or more of its parents, children or siblings had a higher rank in the list [74].

To determine the robustness of the results, we performed the same analysis for all top 10 best performing EnTDefs without "nearest_all" addition. The enrichment results were combined across EnTDefs for each GO term by taking the Harmonic Mean (HM) p-values[75]. The significant terms were extracted using FDR-adjusted HM p-values (HM FDR < 0.05), followed by redundant term filtering as described above.

Identification of common false positive and false negative enrichment results by nearest distance enhancer-target gene assignment

Next, we examined the false positive (FP) and false negative (FN) GO terms identified using the nearest distance assignment (>5kb LocDef), that were correctly identified using the 'smart' enhancer-target gene assignments. To do this, we compared the Poly-Enrich [70] GSE results of the 87 evaluation ChIP-seq datasets using >5kb LocDef for nearest distance to those of the best performing EnTDef (defined by DNase-seq plus FANTOM5 enhancers and ChIA, Thurman and FANTOM5 enhancer-target gene link methods with "nearest_All" addition) using the known TF-GO BP annotated terms (i.e. GO annotation). The FP and FN GO BP terms were defined according to the following criteria: *i)* FPs: GO terms identified as significantly enriched (FDR < 0.05) by the >5kb LocDef, but not significant (FDR $\geq$ 0.05) by EnTdef and not annotated by the GO database; *ii)* FNs: GO terms identified as insignificant (FDR $\geq$ 0.05) by the >5kb LocDef, but significantly enriched (FDR < 0.05) by EnTDef and also annotated by the GO database. GO terms were removed if they were annotated to < 4 (<10%) or > 30 (>90%) of the 34 used TFs to ensure the possibility of having both true and false annotations present among the TFs, and only terms with < 1000 total annotated genes were used in this analysis. For each GO term, the number of FPs and FNs among the 34 TFs, detected using '>5kb', were summarized, and the proportion of FP and FN were calculated using the total number of positive TFs (FDR < 0.05) or negative TFs (FDR $\geq$ 0.05) by >5kb LocDef as the denominator, respectively. To summarize the significance level of each FP or FN term in GSE testing, a meta-p-value was generated using Fisher's method [76] across the annotated TFs using the raw p-values from the >5kb LocDef or EnTDef approach, respectively. Finally, the FP (or FN) GO terms were ranked by the proportion of FP (or FN)

666 decreasingly and meta-p value increasingly, followed by redundant GO term filtering as described
667 above.

668

669 **DATA AVAILABILITY**

670 The top performing EnTDef and EnTDef.plus5kb have been included in the Bioconductor package
671 *chipenrich* (42) and the ChIP-Enrich website (https://chip-enrich.med.umich.edu). The peak-to-
672 gene assignment functionality provided by our GSE Suite (http://gsesuite.dcmb.med.umich.edu)
673 allows users to select all possible combinations of enhancer and/or enhancer-to-gene link
674 methods (as described in this study) and obtain the gene assignments for user uploaded genomic
675 regions based on the selected sources and methods. Genomic regions can also be assigned to
676 target genes based on other approaches (e.g. promoters, exons or anywhere in the genome using
677 the nearest distance method).

678 **ACKNOWLEDGEMENTS**

679 T.Q. performed the bioinformatics analysis and wrote the manuscript. C.L. contributed the GSE
680 testing. R.C. and P.O. contributed the data collection and EnTDef generation. H.Y., H.Z. and S.W.
681 contributed specific bioinformatics analysis. S.P. developed the GSE Suite. A.P.B. contributed to
682 manuscript writing and provided analysis suggestions. M.A.S. conceived, supervised the study,
683 participated in data interpretation and manuscript writing.

684 **FUNDING**

685 This work was supported by National Cancer Institute P01 grant CA240239 and P30 grant
686 CA046592, and by National Institute of Environmental Health Science P30 grant ES017885.

FIGURE LEGENDS

Figure 1. Workflow for generating and evaluating EnTDefs. (A) Enhancers were defined by ENCODE ChromHMM UCSC tracks, ENCODE DNase-seq hypersensitive sites (DHSs), Cap Analysis Gene Expression (CAGE) experiment-derived enhancers from the FANTOM5 project, and/or distal and non-promoter DHS within 500kb of the correlated promoter DHSs from Thurman *et al*. (B) The enhancer-target gene links were defined by ChIA-PET interactions from ENCODE ChIA-PET data (ChIA), DNase-signal correlation-based links from Thurman *et al*, expression correlation-based interactions from FANTOM5, and/or interactions between enhancers and genes within loop (L) boundaries of ChIA-PET with convergent CTCF motifs (L1 [one gene], L2 [≤ two gene] or L3 [≤ three genes] were allowed). An enhancer can be assigned to multiple genes. To increase the genome coverage, we allowed the extension of enhancers to 1kb (i.e. enhancer extension), and assigned other regions outside of 5kb from a TSS to the nearest gene (i.e. "nearest_All" additional links). All combinations of the above, allowing multiple at a time, defined the possible **En**hancer to **T**arget gene **Def**initions (EnTDefs). (C) Left: 1,860 EnTDefs were generated and GOBP GSE testing was performed on 87 ENCODE TF ChIP-seq datasets using each of the EnTDefs. By comparing the significant GOBP terms identified by GSE with each EntDef to those assigned to the TF by the GO database ("GO annotation"), the F1-score was calculated for each EnTDef-TF pair. Right: the EnTDefs were ranked by average F1-score across TFs in descending order. TF paired Wilcoxon sum-rank test was performed between the top ranked EnTDef and each of the sequential ones to identify the set of best EnTDefs (top 1 until the rank with p value < 0.01).

Figure 2. Characteristics of EnTDefs. (A) Overview of the characteristics of 1,860 EnTDefs and >5kb LocDef ranked by F1-score in descending order. F1-score, sensitivity, specificity, number of enhancers, average number of genes per enhancer, average number of enhancers per gene, average proportion of caught TF peaks, average proportion of caught TF peaks outside of 5kb of TSSs, proportion of genome coverage, and whether the EnTDef was significantly better than the >5kb LocDef are shown. (B) The frequency of each method (four enhancer definition methods, with or without enhancer extension, seven enhancer-target gene link methods) among the 741 EnTDefs that significantly outperformed the >5kb LocDef. For simplification, the "nearest_all" additional link method was grouped in enhancer-target gene link method. (C) Bar plot of average F1-scores for the top 10 EnTDefs plus 5kb LocDef (blue bars), top 10 EnTDefs (purple bars), nearest TSS method (pink bar), >5kb LocDef (mustard bar) and 5kb LocDef used by PE.Approx (yellow bar), by GREAT (dark green bar) and by Fisher's exact test (green bar). (D) Distribution of average F1-scores of the top 10 EnTDefs used on *evaluation* ChIP-seq datasets or *testing* ChIP-seq datasets. The dashed lines link the same EnTDefs used in the two different ChIP-seq datasets. The p value of Wilcoxon signed-rank test was shown in the figure.

Figure 3. Evaluation of cell type-specific (CT)-EnTDefs and general EnTDefs. (A) Distribution of the average F1-scores of same-CT EnTDefs, diff-CT EnTDefs and general EnTDefs that were applied on the same TF ChIP-seq data. (B) Correlation between average F1-scores calculated on a TF in a particular cell type using CT-EnTDefs of the matched cell type (x-axis) and the ones calculated on the same TF using CT-EnTDefs of a different cell type (y-axis). Each dot represents an average F1-score of a TF across EnTDefs, and each panel is one of four cell types (GM12878, H1HESC, K562 and MCF7) for which the CT-EnTDefs were created and evaluated, respectively. (C)

730 Evaluation summary of different types of EnTDefs in four different cell types. Comparative
731 average F1-scores associated with a particular cell type are grouped in a grey box: blue refers to
732 using the CT-EnTDef on a TF ChIP-seq from the same cell line, green refers to using the general
733 CT-EnTDef on that TF ChIP-seq, and red refers to using the diff-CT-EnTDef on TF ChIP-seqs from
734 that TF ChIP-seq.

735 **Figure 4. GO terms often missed or falsely identified by the nearest distance method of**
736 **assigning genomic regions to target genes.** (A) Distribution of the rank-based inverse normal
737 transformation (INT) of average interceding gene numbers for the best EnTDef without the
738 "nearest_all" addition. The top ranked enriched GO terms most likely or less likely to be identified
739 by nearest distance method were listed. (B) The enriched GO terms in the genes with fewest
740 interceding genes and the ones with the most interceding genes across the top 10 EnTDefs and
741 their associated -log10 Harmonic Mean (HM) FDR.

742 **Figure 5. False positive and false negative GSE results by nearest distance assignment method.**
743 (A) Examples of false positive (left) and false negative results. The false positive term *blood vessel*
744 *morphogenesis* was annotated with gene *PROK1*, the interceding genes between an enhancer
745 bound by a TF *YY1* (peak:1503) and its target gene *SLC16A4*. The false negative term *histone*
746 *modification* was annotated with gene *CHD3*, a target gene of the enhancer bound by the TF *CTCF*
747 (peak:6206). (B) The false positive GO terms identified as enriched by the nearest distance
748 method, but not by the best performing EnTDef and not defined by the TFs' GO annotation. (C)
749 The false negative GO terms which were failed to be identified by the nearest distance method,
750 but correctly identified by the best performing EnTDef and also defined by the TFs' GO annotation.
751 The significance levels (-log10 Fisher's FDR) are shown in dark blue for >5kb LocDef and dark red

752 for the EnTDef. The numbers listed at the end of each paired bars are false positive rate (B) and
753 false negative rate (C).

754 **Figure 6. User guidelines for selecting an appropriate enhancer-to-gene assignment method**
755 **(LocDef) for GSE testing.** Depending on the specific research questions, three types of LocDefs
756 can be selected for GSE testing from the *chipenrich* R package: 1) "5kb" or "1kb" for promoter
757 regulation, 2) "enhancer" for distal regulation, and 3) "enhancer.5kb" for whole genome
758 regulation. Different LocDefs have different genome coverages as shown in the last column.
759 Options in other GSE testing software for genomic regions will differ. We no longer recommend
760 using nearest TSS method for Poly-Enrich analysis.

761 **REFERENCES**

762

763 1. Long HK, Prescott SL, Wysocka J: **Ever-Changing Landscapes: Transcriptional Enhancers**
764 **in Development and Evolution.** *Cell* 2016, **167:**1170-1187.
765 2. Schoenfelder S, Fraser P: **Long-range enhancer-promoter contacts in gene expression**
766 **control.** *Nat Rev Genet* 2019, **20:**437-455.
767 3. Riethoven JJ: **Regulatory regions in DNA: promoters, enhancers, silencers, and**
768 **insulators.** *Methods Mol Biol* 2010, **674:**33-42.
769 4. Kolovos P, Knoch TA, Grosveld FG, Cook PR, Papantonis A: **Enhancers and silencers: an**
770 **integrated and simple model for their function.** *Epigenetics Chromatin* 2012, **5:**1.
771 5. Plank JL, Dean A: **Enhancer function: mechanistic and genome-wide insights come**
772 **together.** *Mol Cell* 2014, **55:**5-14.
773 6. Pombo A, Dillon N: **Three-dimensional genome architecture: players and mechanisms.**
774 *Nat Rev Mol Cell Biol* 2015, **16:**245-257.
775 7. Furlong EEM, Levine M: **Developmental enhancers and chromosome topology.** *Science*
776 2018, **361:**1341-1345.
777 8. Sakabe NJ, Savic D, Nobrega MA: **Transcriptional enhancers in development and**
778 **disease.** *Genome Biol* 2012, **13:**238.
779 9. Rickels R, Shilatifard A: **Enhancer Logic and Mechanics in Development and Disease.**
780 *Trends Cell Biol* 2018, **28:**608-630.
781 10. Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de
782 Graaff E: **A long-range Shh enhancer regulates expression in the developing limb and**
783 **fin and is associated with preaxial polydactyly.** *Hum Mol Genet* 2003, **12:**1725-1735.

11. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A* 2009, **106:**9362-9367.
12. Corradin O, Scacheri PC: **Enhancer variants: evaluating functions in common disease.** *Genome Med* 2014, **6:**85.
13. Smith E, Shilatifard A: **Enhancer biology and enhanceropathies.** *Nat Struct Mol Biol* 2014, **21:**210-219.
14. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G: **Enhancers: five essential questions.** *Nat Rev Genet* 2013, **14:**288-295.
15. Peng Y, Zhang Y: **Enhancer and super-enhancer: Positive regulators in gene transcription.** *Animal Model Exp Med* 2018, **1:**169-179.
16. Andersson R, Sandelin A, Danko CG: **A unified architecture of transcriptional regulatory elements.** *Trends Genet* 2015, **31:**426-433.
17. Dao LTM, Galindo-Albarran AO, Castro-Mondragon JA, Andrieu-Soler C, Medina-Rivera A, Souaid C, Charbonnier G, Griffon A, Vanhille L, Stephen T, et al: **Genome-wide characterization of mammalian promoters with distal enhancer functions.** *Nat Genet* 2017, **49:**1073-1081.
18. Schoenfelder S, Furlan-Magaril M, Mifsud B, Tavares-Cadete F, Sugar R, Javierre BM, Nagano T, Katsman Y, Sakthidevi M, Wingett SW, et al: **The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements.** *Genome Res* 2015, **25:**582-597.
19. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, et al: **Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C.** *Nat Genet* 2015, **47:**598-606.
20. Siersbaek R, Madsen JGS, Javierre BM, Nielsen R, Bagge EK, Cairns J, Wingett SW, Traynor S, Spivakov M, Fraser P, Mandrup S: **Dynamic Rewiring of Promoter-Anchored Chromatin Loops during Adipocyte Differentiation.** *Mol Cell* 2017, **66:**420-435 e425.
21. Rubin AJ, Barajas BC, Furlan-Magaril M, Lopez-Pajares V, Mumbach MR, Howard I, Kim DS, Boxer LD, Cairns J, Spivakov M, et al: **Lineage-specific dynamic and pre-established enhancer-promoter contacts cooperate in terminal differentiation.** *Nat Genet* 2017, **49:**1522-1528.
22. Sanjana NE, Wright J, Zheng K, Shalem O, Fontanillas P, Joung J, Cheng C, Regev A, Zhang F: **High-resolution interrogation of functional elements in the noncoding genome.** *Science* 2016, **353:**1545-1549.
23. Rowley MJ, Corces VG: **Organizational principles of 3D genome architecture.** *Nat Rev Genet* 2018, **19:**789-800.
24. Sanyal A, Lajoie BR, Jain G, Dekker J: **The long-range interaction landscape of gene promoters.** *Nature* 2012, **489:**109-113.
25. Spitz F: **Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles.** *Semin Cell Dev Biol* 2016, **57:**57-67.
26. Gibcus JH, Dekker J: **The hierarchy of the 3D genome.** *Mol Cell* 2013, **49:**773-782.
27. Ghirlando R, Felsenfeld G: **CTCF: making the right connections.** *Genes Dev* 2016, **30:**881-891.

28. Consortium EP, Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, et al: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.
29. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, Hilton JA, Jain K, Baymuradov UK, Narayanan AK, et al: **The Encyclopedia of DNA elements (ENCODE): data portal update.** *Nucleic Acids Res* 2018, **46:**D794-D801.
30. Consortium EP, Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, et al: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447:**799-816.
31. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, et al: **Integrative analysis of 111 reference human epigenomes.** *Nature* 2015, **518:**317-330.
32. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al: **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.** *Genome Res* 2012, **22:**1813-1831.
33. Song L, Crawford GE: **DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells.** *Cold Spring Harb Protoc* 2010, **2010:**pdb prot5384.
34. Buenrostro JD, Wu B, Chang HY, Greenleaf WJ: **ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide.** *Curr Protoc Mol Biol* 2015, **109:**21 29 21-21 29 29.
35. Li G, Fullwood MJ, Xu H, Mulawadi FH, Velkov S, Vega V, Ariyaratne PN, Mohamed YB, Ooi HS, Tennakoon C, et al: **ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing.** *Genome Biol* 2010, **11:**R22.
36. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J: **Hi-C: a comprehensive technique to capture the conformation of genomes.** *Methods* 2012, **58:**268-276.
37. Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, Abugessaisa I, Fukuda S, Hori F, Ishikawa-Kato S, et al: **Gateways to the FANTOM5 promoter level mammalian expression atlas.** *Genome Biol* 2015, **16:**22.
38. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al: **The accessible chromatin landscape of the human genome.** *Nature* 2012, **489:**75-82.
39. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G: **GREAT improves functional interpretation of cis-regulatory regions.** *Nat Biotechnol* 2010, **28:**495-501.
40. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al: **Enrichr: a comprehensive gene set enrichment analysis web server 2016 update.** *Nucleic Acids Res* 2016, **44:**W90-97.
41. Cavalcante RG, Lee C, Welch RP, Patil S, Weymouth T, Scott LJ, Sartor MA: **Broad-Enrich: functional interpretation of large sets of broad genomic regions.** *Bioinformatics* 2014, **30:**i393-400.

42. Welch RP, Lee C, Imbriano PM, Patil S, Weymouth TE, Smith RA, Scott LJ, Sartor MA: **ChIP-Enrich: gene set enrichment testing for ChIP-seq data.** *Nucleic Acids Res* 2014, **42:**e105.
43. Ernst J, Kellis M: **ChromHMM: automating chromatin-state discovery and characterization.** *Nat Methods* 2012, **9:**215-216.
44. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, et al: **An atlas of active enhancers across human cell types and tissues.** *Nature* 2014, **507:**455-461.
45. Consortium F, the RP, Clst, Forrest AR, Kawaji H, Rehli M, Baillie JK, de Hoon MJ, Haberle V, Lassmann T, et al: **A promoter-level mammalian expression atlas.** *Nature* 2014, **507:**462-470.
46. Dowen JM, Fan ZP, Hnisz D, Ren G, Abraham BJ, Zhang LN, Weintraub AS, Schujiers J, Lee TI, Zhao K, Young RA: **Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes.** *Cell* 2014, **159:**374-387.
47. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, Trzaskoma P, Magalska A, Wlodarczyk J, Ruszczycki B, et al: **CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription.** *Cell* 2015, **163:**1611-1627.
48. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell* 2014, **159:**1665-1680.
49. Maston GA, Evans SK, Green MR: **Transcriptional regulatory elements in the human genome.** *Annu Rev Genomics Hum Genet* 2006, **7:**29-59.
50. Kim JH, Karnovsky A, Mahavisno V, Weymouth T, Pande M, Dolinoy DC, Rozek LS, Sartor MA: **LRpath analysis reveals common pathways dysregulated via DNA methylation across cancer types.** *BMC Genomics* 2012, **13:**526.
51. Hariprakash JM, Ferrari F: **Computational Biology Solutions to Identify Enhancers-target Gene Pairs.** *Comput Struct Biotechnol J* 2019, **17:**821-831.
52. Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sal lari R, Lupien M, Markowitz S, Scacheri PC: **Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits.** *Genome Res* 2014, **24:**1-13.
53. Yao L, Shen H, Laird PW, Farnham PJ, Berman BP: **Inferring regulatory element landscapes and transcription factor networks from cancer methylomes.** *Genome Biol* 2015, **16:**105.
54. Silva TC, Coetzee SG, Gull N, Yao L, Hazelett DJ, Noushmehr H, Lin DC, Berman BP: **ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles.** *Bioinformatics* 2019, **35:**1974-1977.
55. He B, Chen C, Teng L, Tan K: **Global view of enhancer-promoter interactome in human cells.** *Proc Natl Acad Sci U S A* 2014, **111:**E2191-2199.
56. Whalen S, Truty RM, Pollard KS: **Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin.** *Nat Genet* 2016, **48:**488-496.
57. Hafez D, Karabacak A, Krueger S, Hwang YC, Wang LS, Zinzen RP, Ohler U: **McEnhancer: predicting gene expression via semi-supervised assignment of enhancers to target genes.** *Genome Biol* 2017, **18:**199.

58. Roy S, Siahpirani AF, Chasman D, Knaack S, Ay F, Stewart R, Wilson M, Sridharan R: **A predictive modeling approach for cell line-specific long-range regulatory interactions.** *Nucleic Acids Res* 2015, **43:**8694-8712.
59. Cao Q, Anyansi C, Hu X, Xu L, Xiong L, Tang W, Mok MTS, Cheng C, Fan X, Gerstein M, et al: **Reconstruction of enhancer-target networks in 935 samples of human primary cells, tissues and cell lines.** *Nat Genet* 2017, **49:**1428-1436.
60. Hait TA, Amar D, Shamir R, Elkon R: **FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map.** *Genome Biol* 2018, **19:**56.
61. Zhu Y, Chen Z, Zhang K, Wang M, Medovoy D, Whitaker JW, Ding B, Li N, Zheng L, Wang W: **Constructing 3D interaction maps from 1D epigenomes.** *Nat Commun* 2016, **7:**10812.
62. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, Rosen N, Kohn A, Twik M, Safran M, et al: **GeneHancer: genome-wide integration of enhancers and target genes in GeneCards.** *Database (Oxford)* 2017, **2017**.
63. Clement Y, Torbey P, Gilardi-Hebenstreit P, Crollius HR: **Enhancer-gene maps in the human and zebrafish genomes using evolutionary linkage conservation.** *Nucleic Acids Res* 2020.
64. Naville M, Ishibashi M, Ferg M, Bengani H, Rinkwitz S, Krecsmarik M, Hawkins TA, Wilson SW, Manning E, Chilamakuri CS, et al: **Long-range evolutionary constraints reveal cis-regulatory interactions on the human X chromosome.** *Nat Commun* 2015, **6:**6904.
65. Ron G, Globerson Y, Moran D, Kaplan T: **Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains.** *Nat Commun* 2017, **8:**2237.
66. Consortium GT: **The Genotype-Tissue Expression (GTEx) project.** *Nat Genet* 2013, **45:**580-585.
67. Kim SA, Yoo YJ: **Effects of Single Nucleotide Polymorphism Marker Density on Haplotype Block Partition.** *Genomics Inform* 2016, **14:**196-204.
68. Wang J, Dai X, Berry LD, Cogan JD, Liu Q, Shyr Y: **HACER: an atlas of human active enhancers to interpret regulatory variants.** *Nucleic Acids Res* 2019, **47:**D106-D112.
69. Li G, Chen Y, Snyder MP, Zhang MQ: **ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis.** *Nucleic Acids Res* 2017, **45:**e4.
70. Lee CT, Cavalcante RG, Lee C, Qin T, Patil S, Wang S, Tsai ZTY, Boyle AP, Sartor MA: **Poly-Enrich: count-based methods for gene set enrichment testing with genomic regions.** *NAR Genom Bioinform* 2020, **2:**lqaa006.
71. Lee C, Wang K, Qin T, Sartor MA: **Testing Proximity of Genomic Regions to Transcription Start Sites and Enhancers Complements Gene Set Enrichment Testing.** *Front Genet* 2020, **11:**199.
72. Z. M: **Rank normal transformation omnibus test. https://cran.r-project.org/web/packages/RNOmni/RNOmni.pdf.** 2019.
73. M C: **GO.db: A set of annotation maps describing the entire Gene Ontology.** *package version 382* 2019.

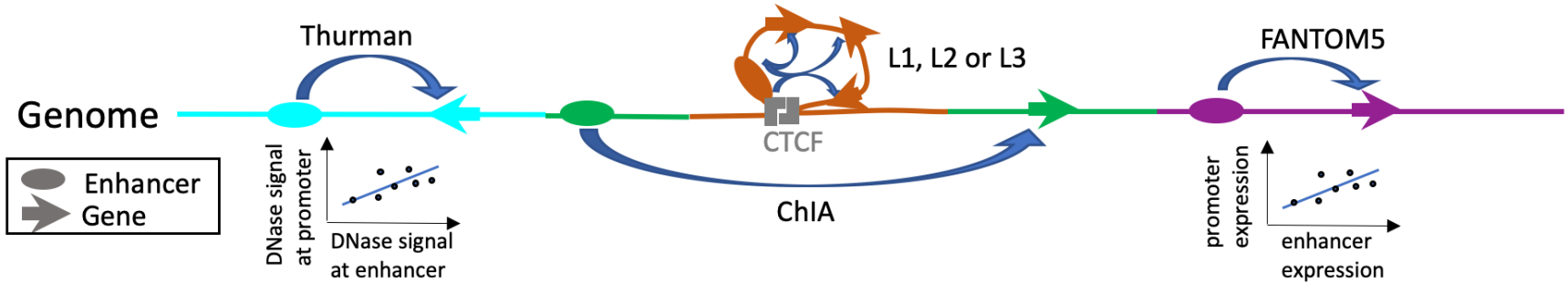
74. Koneva LA, Zhang Y, Virani S, Hall PB, McHugh JB, Chepeha DB, Wolf GT, Carey TE, Rozek LS, Sartor MA: **HPV Integration in HNSCC Correlates with Survival Outcomes, Immune Response Signatures, and Candidate Drivers.** *Mol Cancer Res* 2018, **16:**90-102.
75. Wilson DJ: **The harmonic mean p-value for combining dependent tests.** *Proc Natl Acad Sci U S A* 2019, **116:**1195-1200.
76. Fisher R: *Statistical methods for research workers.* Oliver and Boyd: Edinburgh; 1925.

Figure 1

(A)

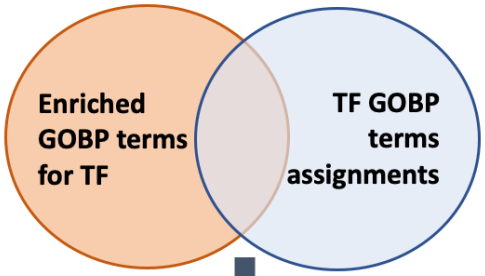
Enhancer	Extension	Enhancer-target gene link	Additional link
1. ChromHMM 2. DNase-seq 3. FANTOM5 4. Thurman	• + 0 kb • + 1 kb	1. ChIA-PET interaction (ChIA) 2. DNase-signal correlation (Thurman) 3. Expression correlation (FANTOM5) 4. Loop boundaries of ChIA-PET with CTCF motifs (L1, L2 or L3)	1. No addition 2. Assign non-enhancer distal regions to nearest gene (nearest_all)

(B)



(C) Evaluation of Enhancer-Target gene Definition (EnTDef) | Rank of EnTDef

- Data: 87 ENCODE ChIP-seq datasets for 34 transcription factors (TFs)
- Gold Standard: the GO biological processes (GOBPs) that each TF is assigned to
- Assumption: TFs regulate genes in the biological processes that they play a role in



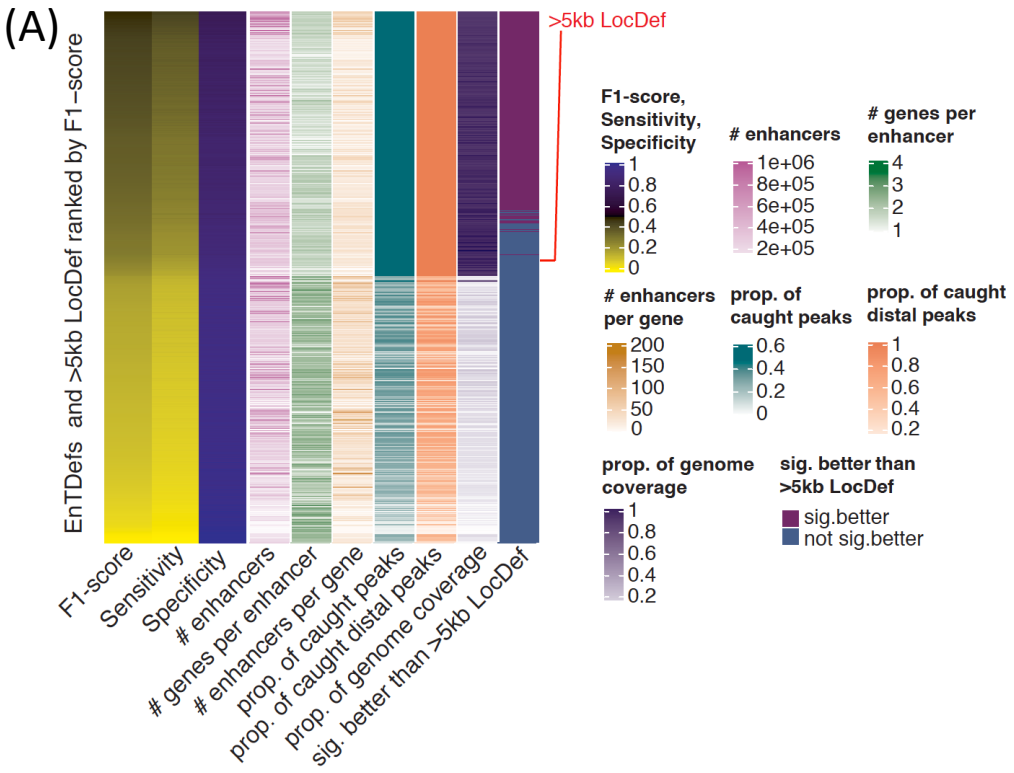
Repeat: 87 datasets & 1,860 EnTDefs

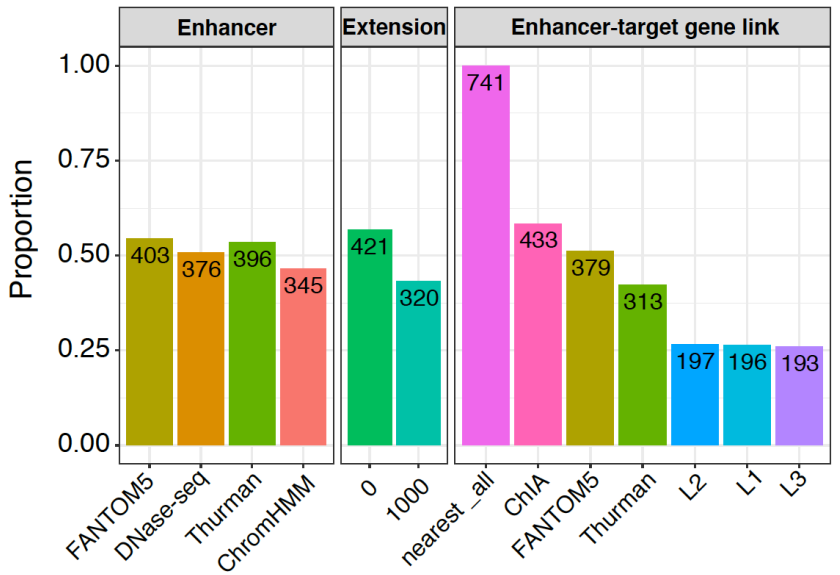
	TF1	TF2	...
EnTDef A			
EnTDef B			
...			

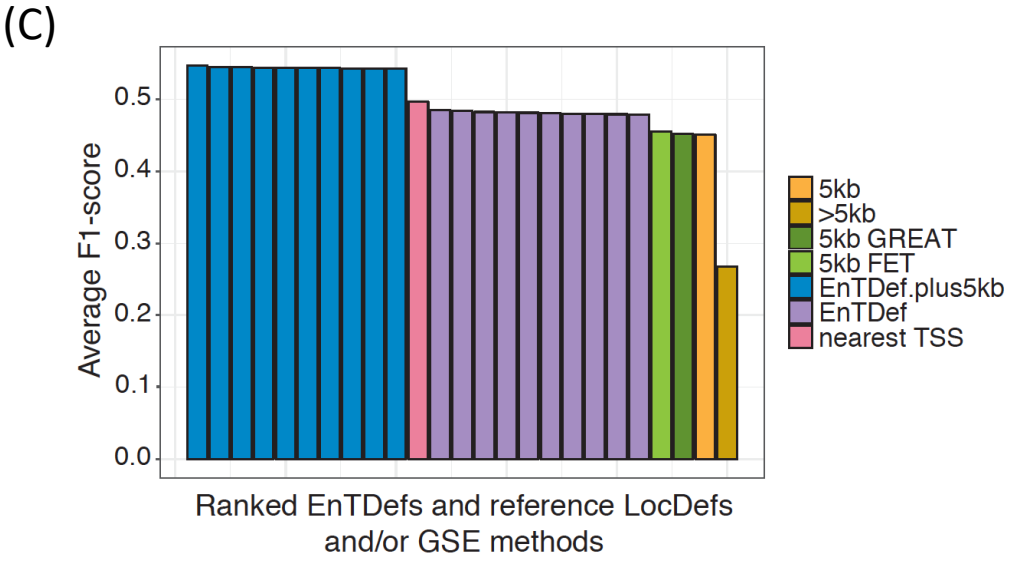
- Rank EnTDefs by average F1-score
- Paired test: any EnTDef vs. the 1st ranked one
- Best EnTDef set: until the rank with p value < 0.01

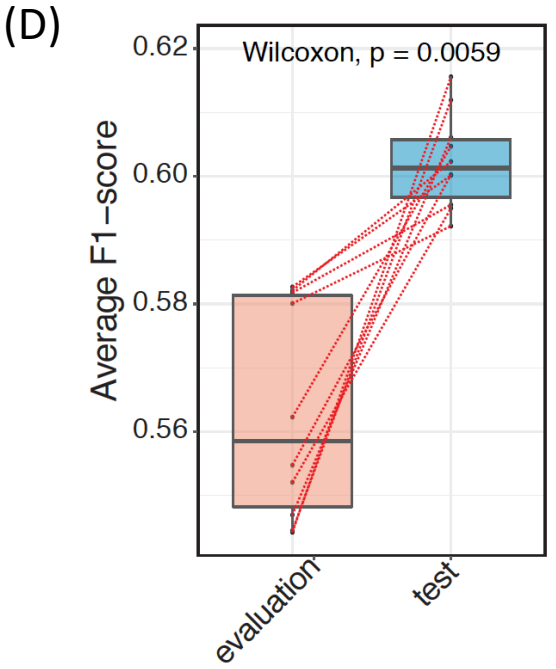
EnTDef	rank	F1 mean	P value (w.r.t. 1st ranked EnTDef)
A	1	0.67	NA
B	2	0.54	0.18
C	3	0.47	0.51
...	...	...	...
L	20	0.46	0.01
...		...	...
N	3	0.20	8.13E-16

Figure 2

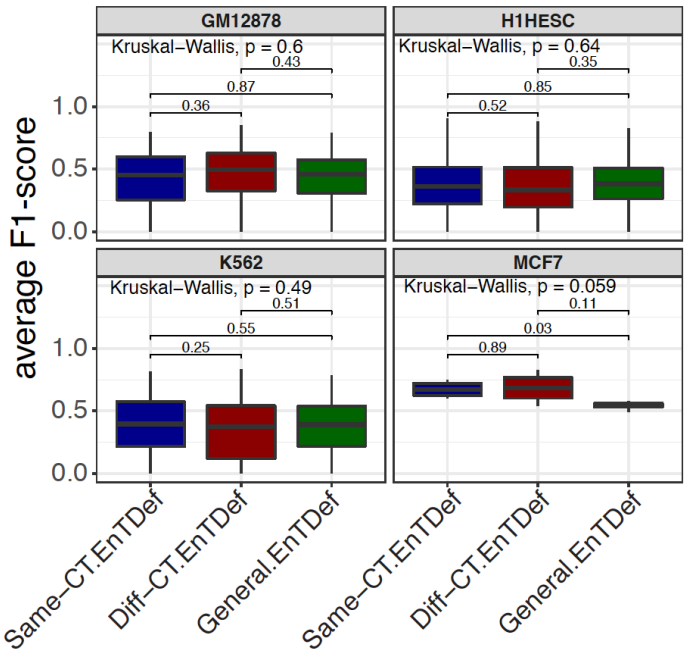

(A)
EnTDefs and >5kb LocDef ranked by F1−score
>5kb LocDef
F1-score
Sensitivity
Specificity
enhancers
genes per enhancer
enhancers per gene
prop. of caught peaks
prop. of caught distal peaks
prop. of genome coverage
sig. better than >5kb LocDef
F1-score, Sensitivity, Specificity
1
0.8
0.6
0.4
0.2
0
enhancers
1e+06
8e+05
6e+05
4e+05
2e+05
genes per enhancer
4
3
2
1
enhancers per gene
200
150
100
50
0
prop. of caught peaks
0.6
0.4
0.2
0
prop. of caught distal peaks
1
0.8
0.6
0.4
0.2
prop. of genome coverage
1
0.8
0.6
0.4
0.2
sig. better than >5kb LocDef
sig.better
not sig.better


(B)
Enhancer
Extension
Enhancer-target gene link
Proportion
1.00
0.75
0.50
0.25
0.00
403
376
396
345
421
320
741
433
379
313
197
196
193
FANTOM5
DNase-seq
Thurman
ChromHMM
0
1000
nearest
_all
ChIA
FANTOM5
Thurman
L2
L1
L3

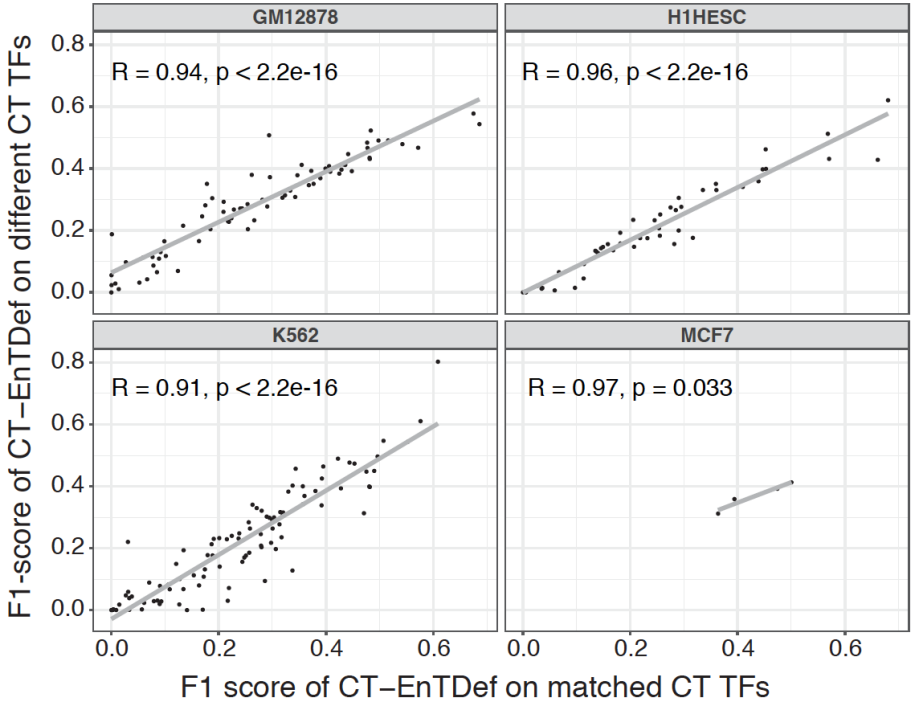

(C)
Average F1-score
0.5
0.4
0.3
0.2
0.1
0.0
Ranked EnTDefs and reference LocDefs
and/or GSE methods
5kb
>5kb
5kb GREAT
5kb FET
EnTDef.plus5kb
EnTDef
nearest TSS


(D)
Wilcoxon, p = 0.0059
Average F1−score
0.62
0.60
0.58
0.56
evaluation
test

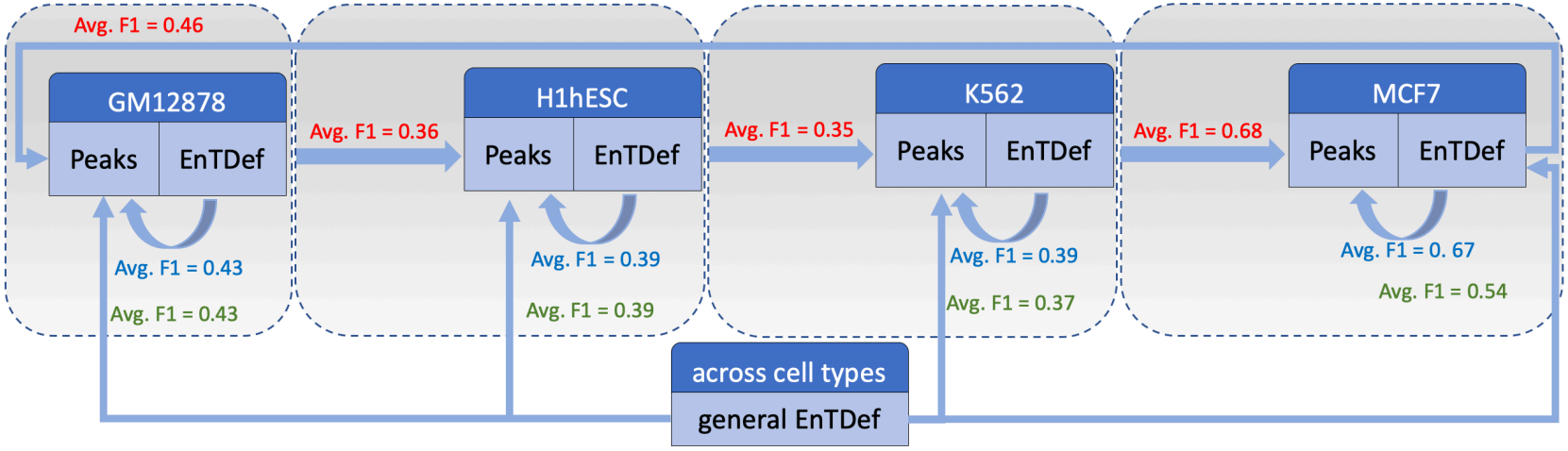
(A)


GM12878
Kruskal–Wallis, p = 0.6
H1HESC
Kruskal–Wallis, p = 0.64
K562
Kruskal–Wallis, p = 0.49
MCF7
Kruskal–Wallis, p = 0.059
average F1-score
Same–CT.EnTDef
Diff–CT.EnTDef
General.EnTDef

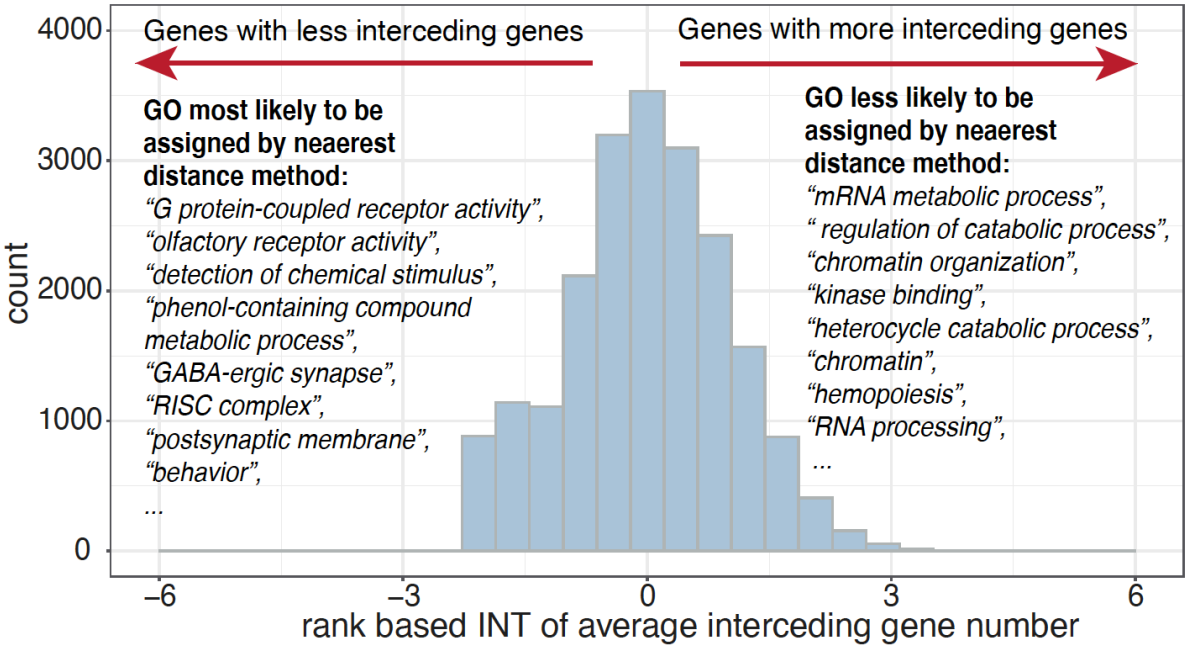
(B)


GM12878
R = 0.94, p < 2.2e-16
H1HESC
R = 0.96, p < 2.2e-16
K562
R = 0.91, p < 2.2e-16
MCF7
R = 0.97, p = 0.033
F1-score of CT–EnTDef on different CT TFs
F1 score of CT–EnTDef on matched CT TFs

(C)


Avg. F1 = 0.46
GM12878
Peaks
EnTDef
Avg. F1 = 0.43
Avg. F1 = 0.43
Avg. F1 = 0.36
H1hESC
Peaks
EnTDef
Avg. F1 = 0.39
Avg. F1 = 0.39
Avg. F1 = 0.35
K562
Peaks
EnTDef
Avg. F1 = 0.39
Avg. F1 = 0.37
Avg. F1 = 0.68
MCF7
Peaks
EnTDef
Avg. F1 = 0. 67
Avg. F1 = 0.54
across cell types
general EnTDef

(A)


Genes with less interceding genes
Genes with more interceding genes
GO most likely to be assigned by neaerest distance method:
"G protein-coupled receptor activity",
"olfactory receptor activity",
"detection of chemical stimulus",
"phenol-containing compound metabolic process",
"GABA-ergic synapse",
"RISC complex",
"postsynaptic membrane",
"behavior",
...
GO less likely to be assigned by neaerest distance method:
"mRNA metabolic process",
" regulation of catabolic process",
"chromatin organization",
"kinase binding",
"heterocycle catabolic process",
"chromatin",
"hemopoiesis",
"RNA processing",
...
count
4000
3000
2000
1000
0
−6
−3
0
3
6
rank based INT of average interceding gene number

(B)

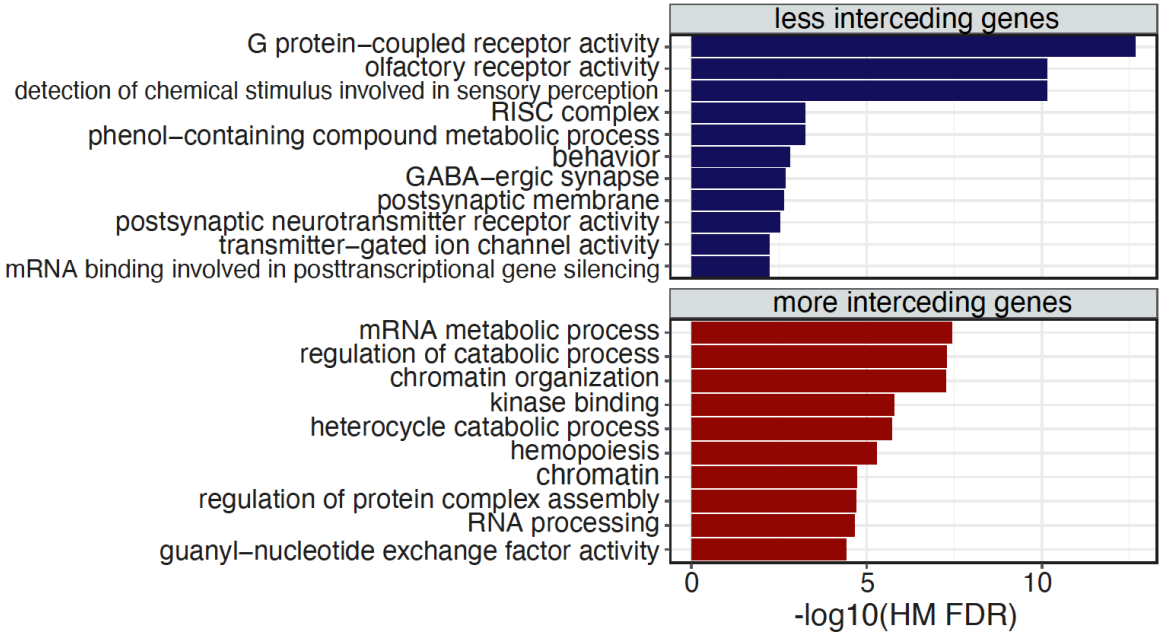
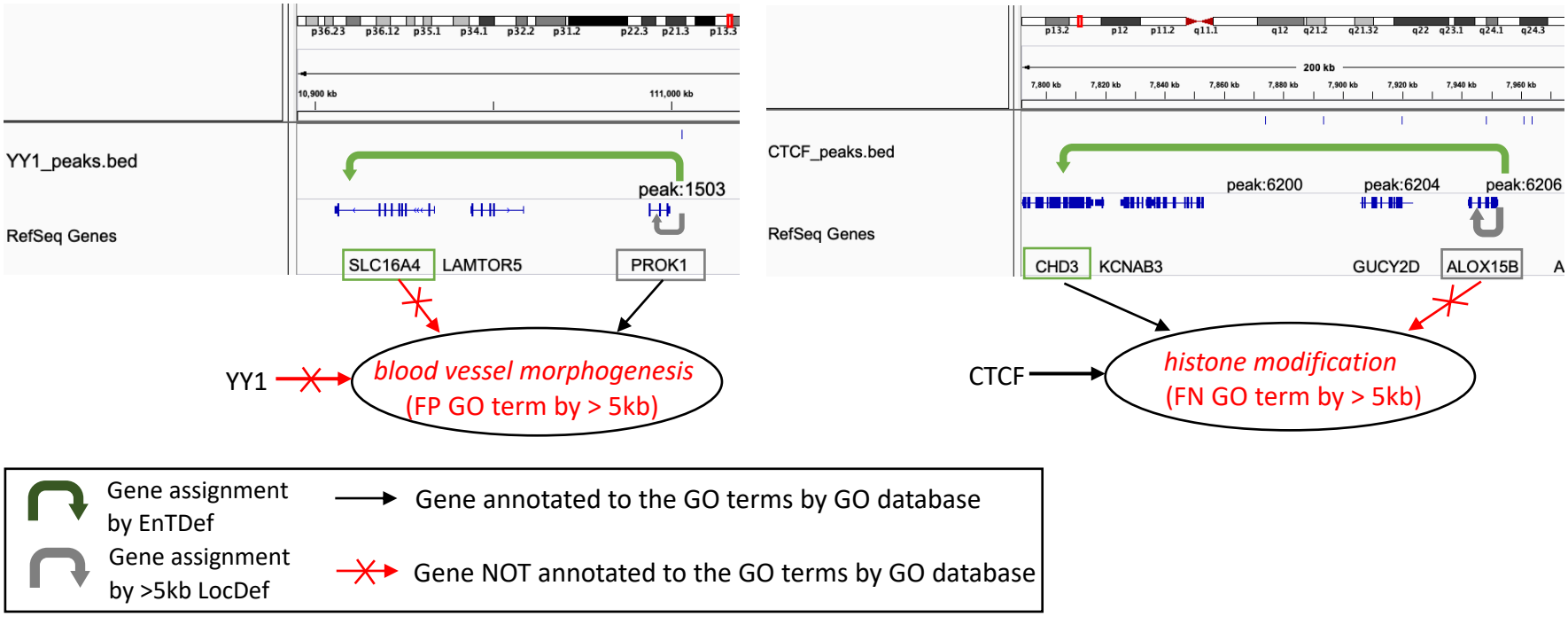
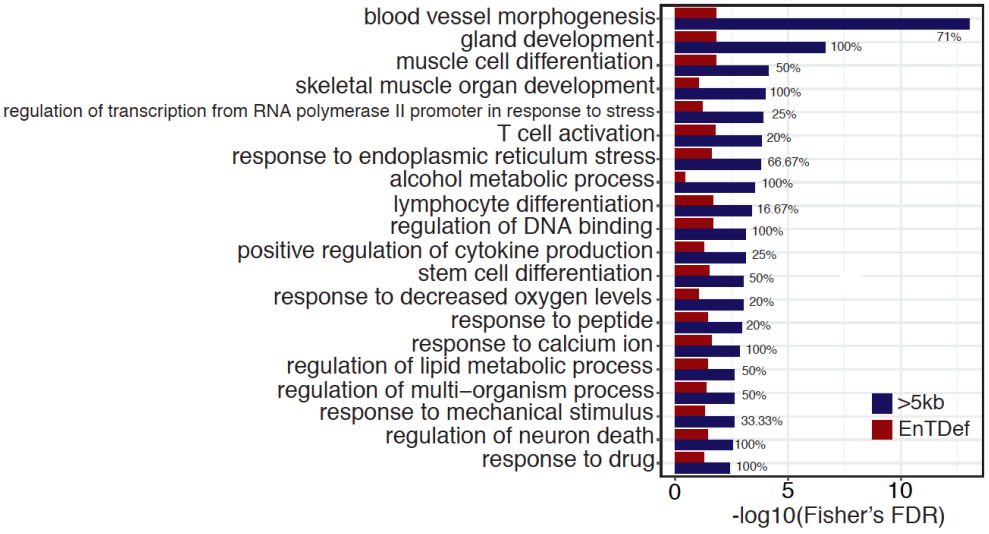

less interceding genes
G protein−coupled receptor activity
olfactory receptor activity
detection of chemical stimulus involved in sensory perception
RISC complex
phenol−containing compound metabolic process
behavior
GABA−ergic synapse
postsynaptic membrane
postsynaptic neurotransmitter receptor activity
transmitter−gated ion channel activity
mRNA binding involved in posttranscriptional gene silencing
more interceding genes
mRNA metabolic process
regulation of catabolic process
chromatin organization
kinase binding
heterocycle catabolic process
hemopoiesis
chromatin
regulation of protein complex assembly
RNA processing
guanyl−nucleotide exchange factor activity
0
5
10
-log10(HM FDR)

Figure 5

(A)


YY1_peaks.bed
peak:1503
RefSeq Genes
SLC16A4
LAMTOR5
PROK1
YY1
blood vessel morphogenesis
(FP GO term by > 5kb)
CTCF_peaks.bed
peak:6200
peak:6204
peak:6206
RefSeq Genes
CHD3
KCNAB3
GUCY2D
ALOX15B
CTCF
histone modification
(FN GO term by > 5kb)
Gene assignment by EnTDef
Gene assignment by >5kb LocDef
Gene annotated to the GO terms by GO database
Gene NOT annotated to the GO terms by GO database

(B)


False positive GO terms by nearest distance
blood vessel morphogenesis
gland development
muscle cell differentiation
skeletal muscle organ development
regulation of transcription from RNA polymerase II promoter in response to stress
T cell activation
response to endoplasmic reticulum stress
alcohol metabolic process
lymphocyte differentiation
regulation of DNA binding
positive regulation of cytokine production
stem cell differentiation
response to decreased oxygen levels
response to peptide
response to calcium ion
regulation of lipid metabolic process
regulation of multi−organism process
response to mechanical stimulus
regulation of neuron death
response to drug
71%
100%
50%
100%
25%
20%
66.67%
100%
16.67%
100%
25%
50%
20%
20%
100%
50%
50%
33.33%
100%
100%
>5kb
EnTDef
0
5
10
-log10(Fisher's FDR)

(C)

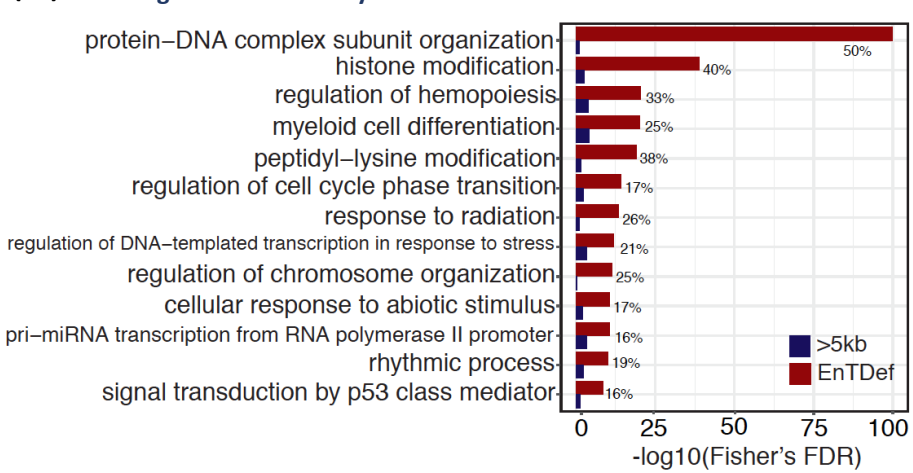
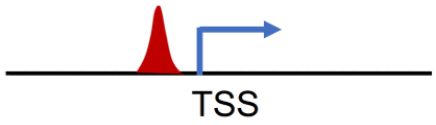
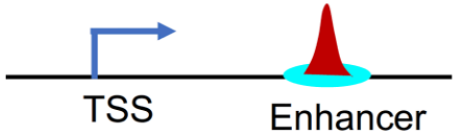
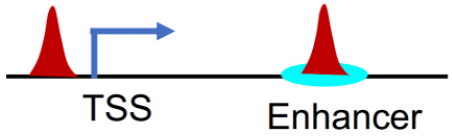
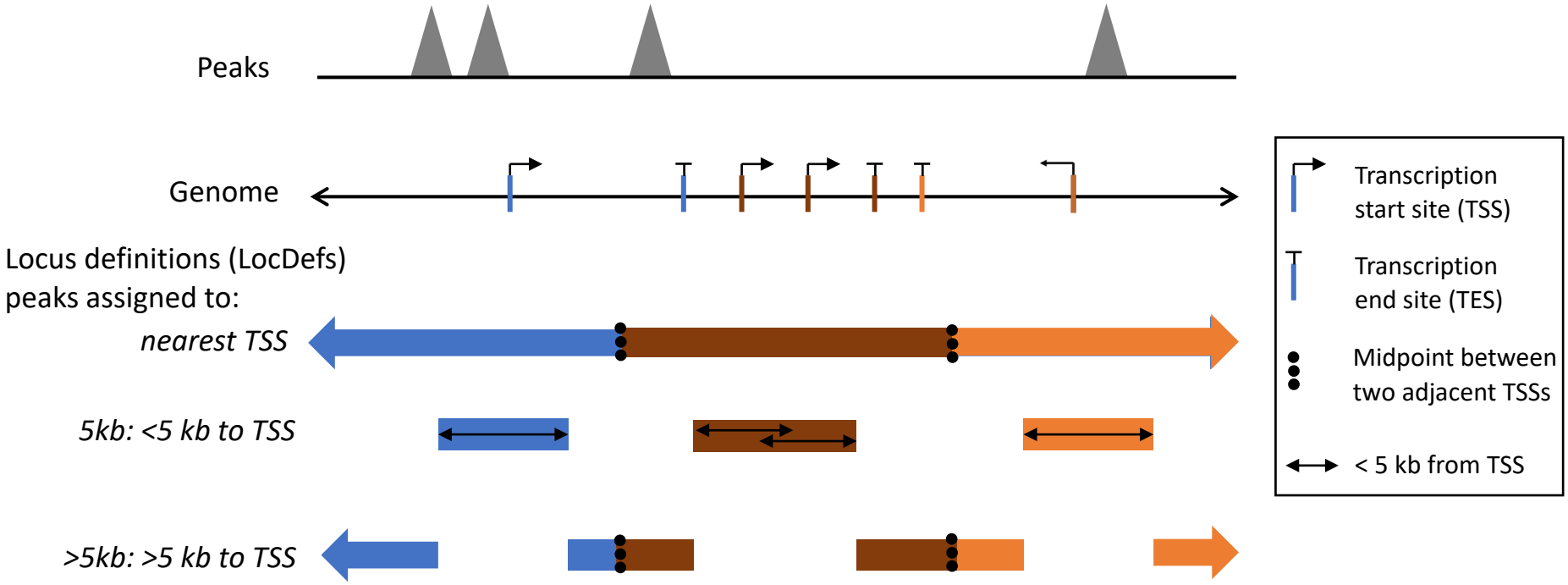

False negative GO terms by nearest distance
protein−DNA complex subunit organization
histone modification
regulation of hemopoiesis
myeloid cell differentiation
peptidyl−lysine modification
regulation of cell cycle phase transition
response to radiation
regulation of DNA−templated transcription in response to stress
regulation of chromosome organization
cellular response to abiotic stimulus
pri−miRNA transcription from RNA polymerase II promoter
rhythmic process
signal transduction by p53 class mediator
50%
40%
33%
25%
38%
17%
26%
21%
25%
17%
16%
19%
16%
>5kb
EnTDef
0
25
50
75
100
-log10(Fisher's FDR)

Figure 6

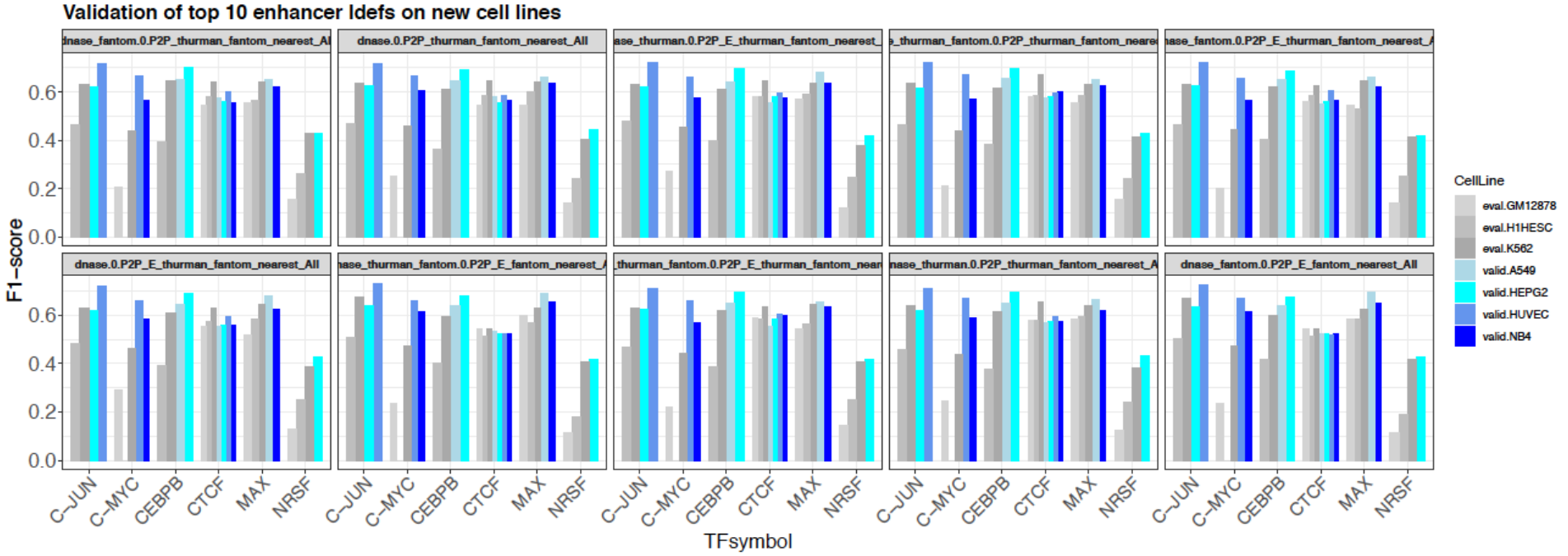
Research Questions	Regions of interest	Options of LocDef	Genome coverage of LocDef
How does the TF and/or chromatin mark regulate gene expression from promoters?	Promoters 	5kb (or 1kb)	10% (2% for 1kb)
How does the TF and/or chromatin mark regulate gene expression from distal regions?	Enhancers 	enhancer	90%
How does the TF and/or chromatin mark regulate gene expression from anywhere on the genome?	Whole genome 	enhancer.5kb	100%

Supplementary figures

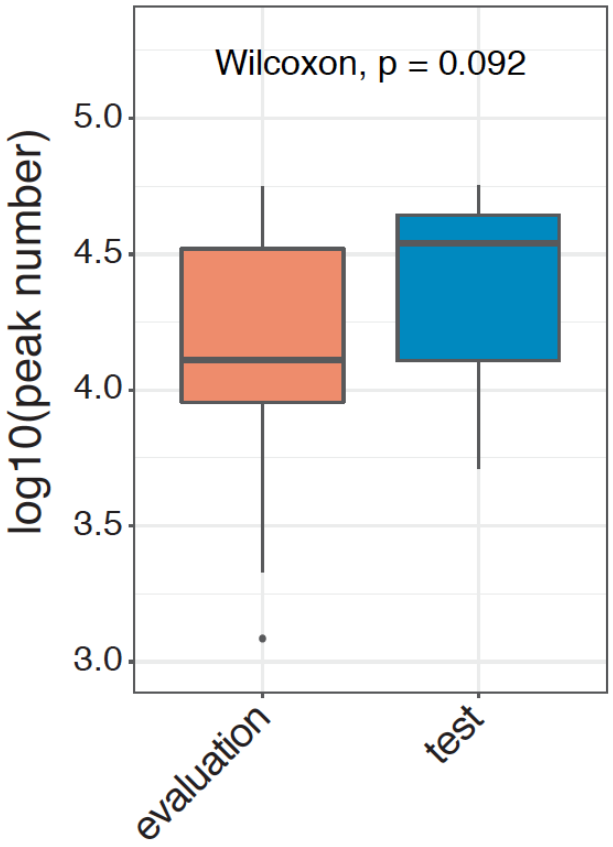
Supp. Figure S1



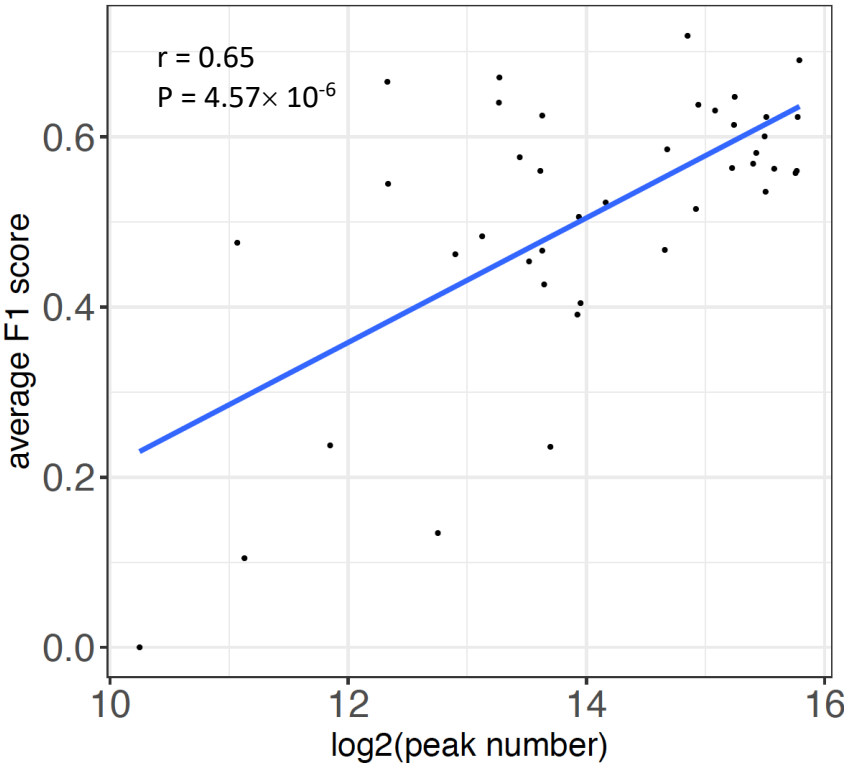
Supp. Figure S2



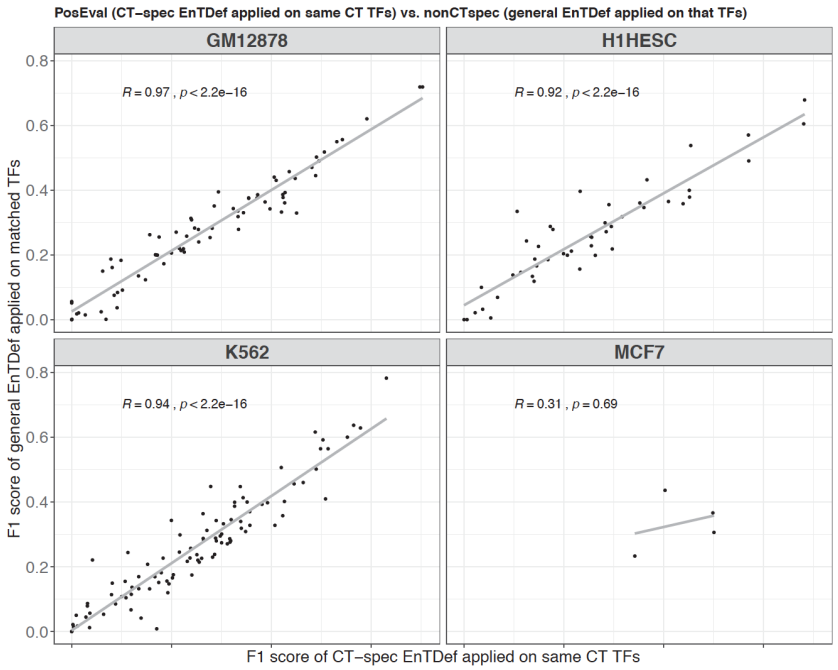
(A)


Wilcoxon, p = 0.092
log10(peak number)
5.0
4.5
4.0
3.5
3.0
evaluation
test

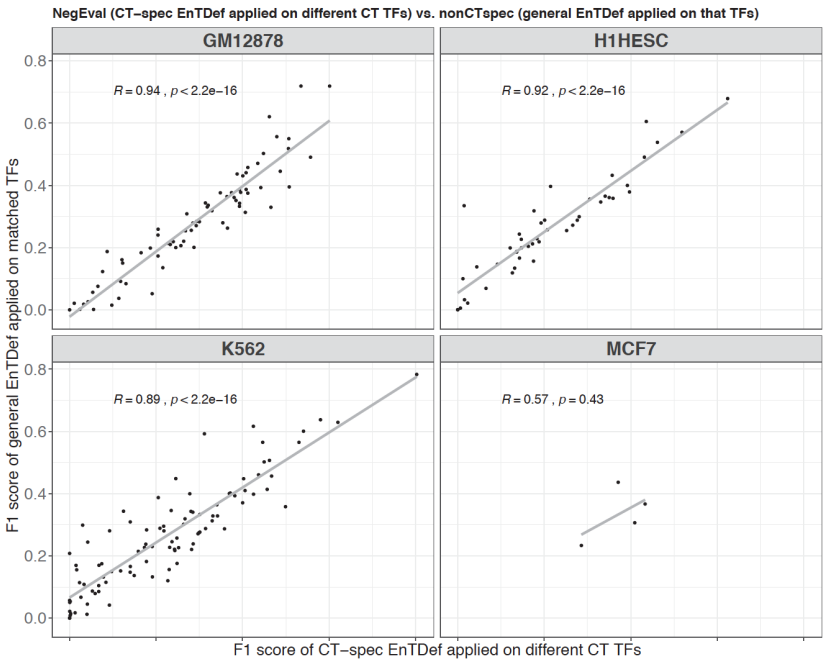
(B)


r = 0.65
P = 4.57× 10-6
average F1 score
0.6
0.4
0.2
0.0
10
12
14
16
log2(peak number)

(A)

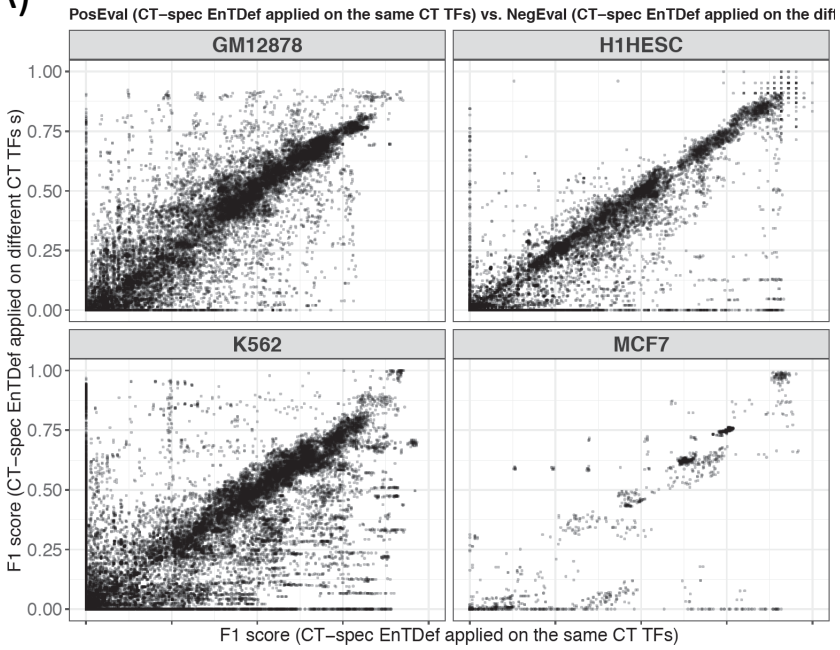

PosEval (CT−spec EnTDef applied on same CT TFs) vs. nonCTspec (general EnTDef applied on that TFs)
GM12878
H1HESC
K562
MCF7
R = 0.97 , p < 2.2e−16
R = 0.92 , p < 2.2e−16
R = 0.94 , p < 2.2e−16
R = 0.31 , p = 0.69
F1 score of general EnTDef applied on matched TFs
F1 score of CT−spec EnTDef applied on same CT TFs
0.0
0.2
0.4
0.6
0.8

(B)

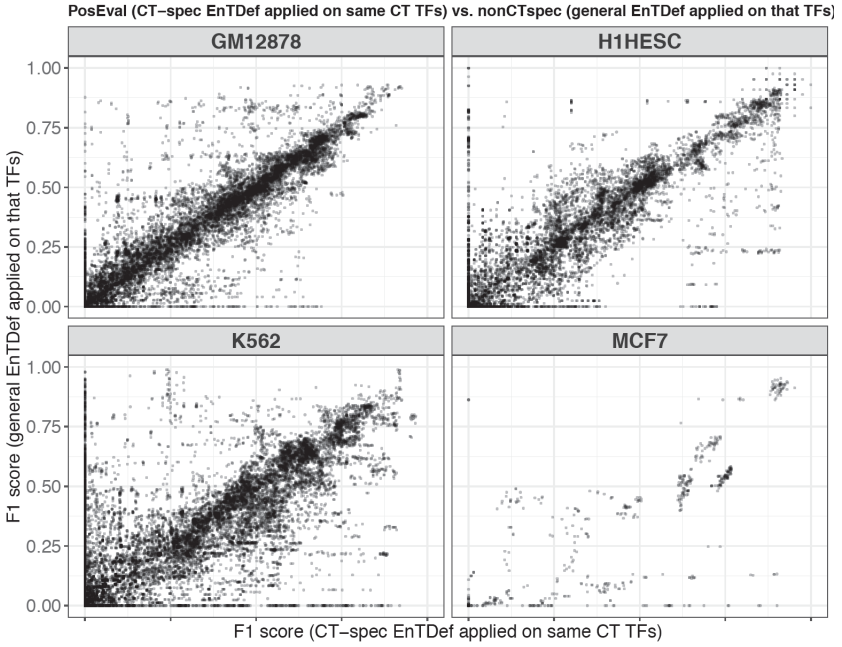

NegEval (CT−spec EnTDef applied on different CT TFs) vs. nonCTspec (general EnTDef applied on that TFs)
GM12878
H1HESC
K562
MCF7
R = 0.94 , p < 2.2e−16
R = 0.92 , p < 2.2e−16
R = 0.89 , p < 2.2e−16
R = 0.57 , p = 0.43
F1 score of general EnTDef applied on matched TFs
F1 score of CT−spec EnTDef applied on different CT TFs
0.0
0.2
0.4
0.6
0.8

Supp. Figure S5

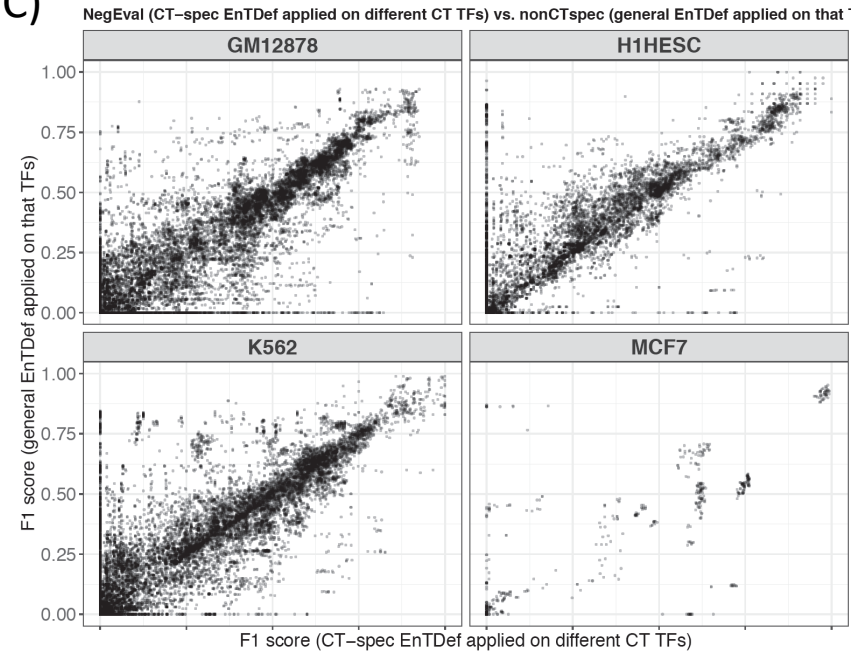
(A)

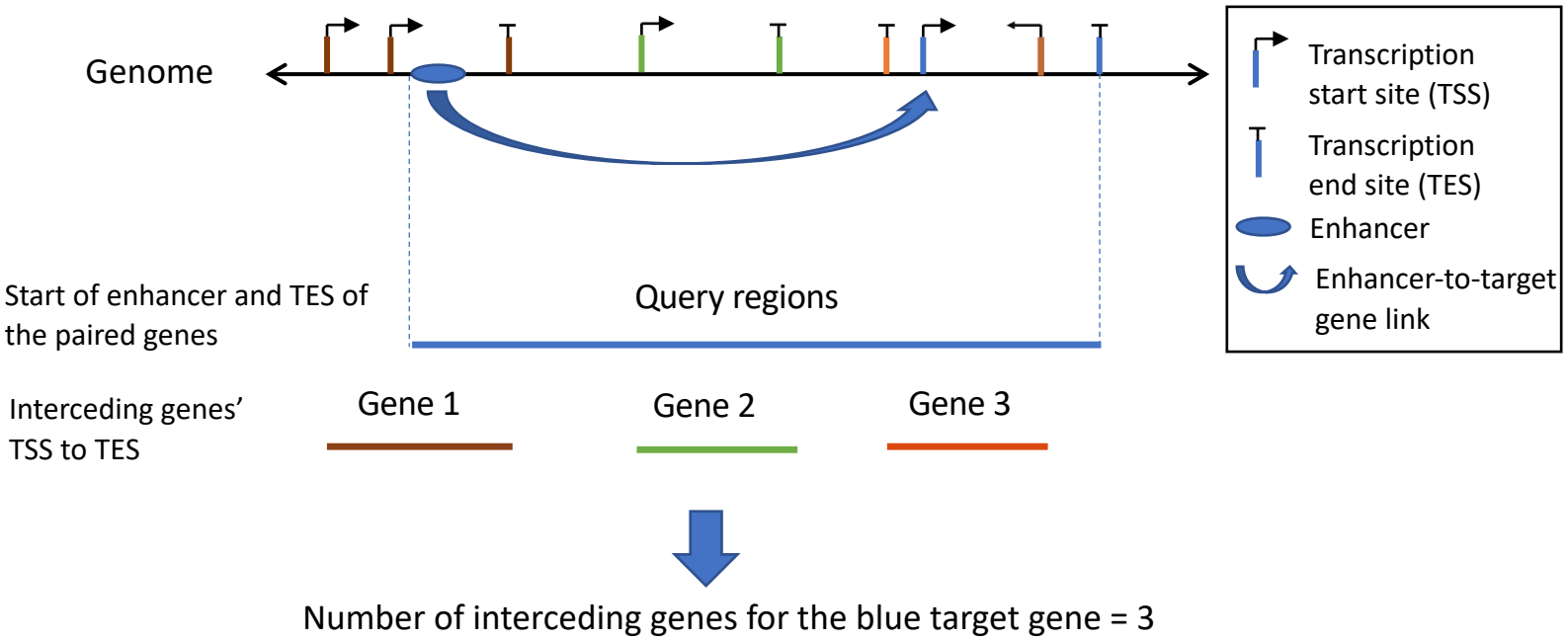


(B)

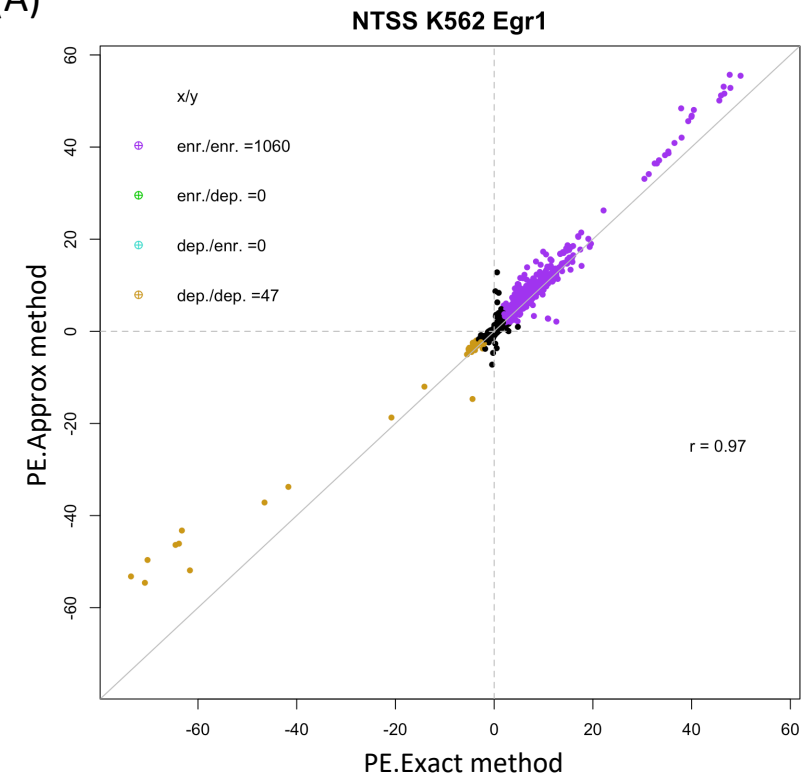


(C)

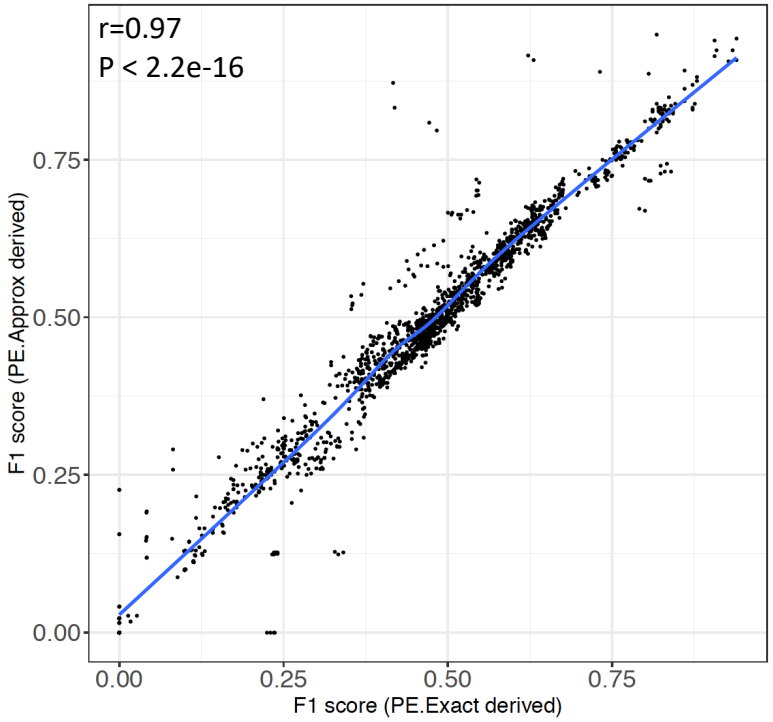



Genome
Transcription start site (TSS)
Transcription end site (TES)
Enhancer
Enhancer-to-target gene link
Start of enhancer and TES of the paired genes
Query regions
Interceding genes' TSS to TES
Gene 1
Gene 2
Gene 3
Number of interceding genes for the blue target gene = 3

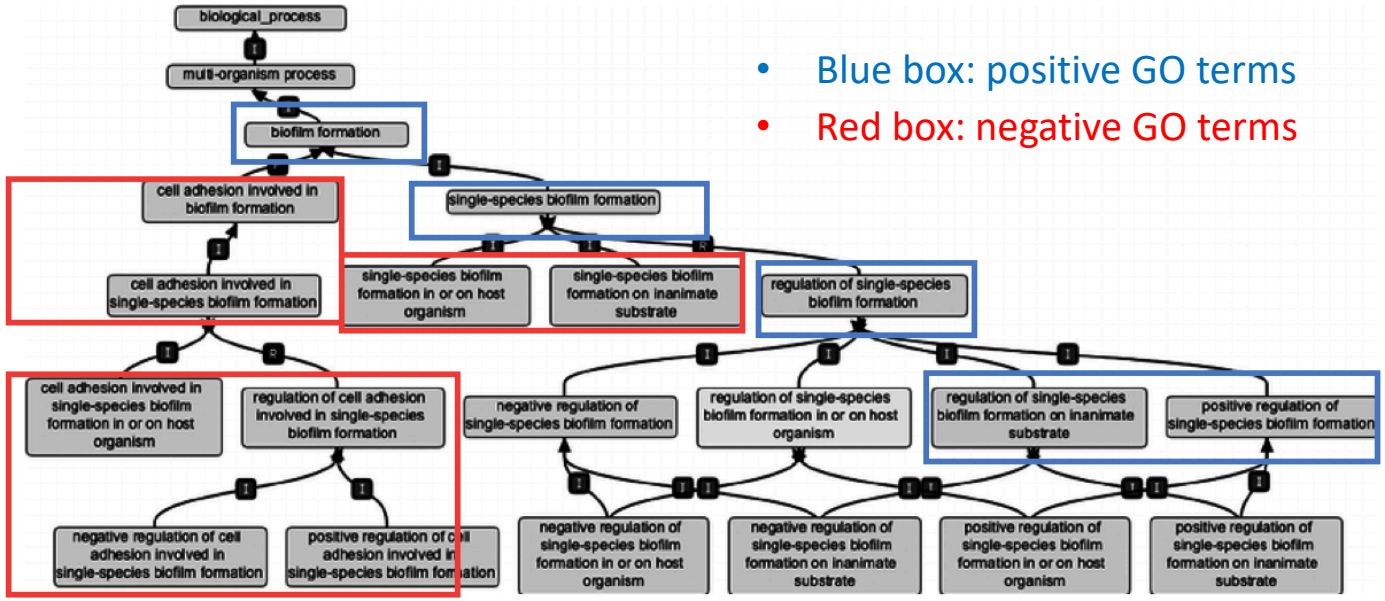
(A)


NTSS K562 Egr1
x/y
enr./enr. =1060
enr./dep. =0
dep./enr. =0
dep./dep. =47
r = 0.97
PE.Approx method
PE.Exact method
60
40
20
0
-20
-40
-60

(B)

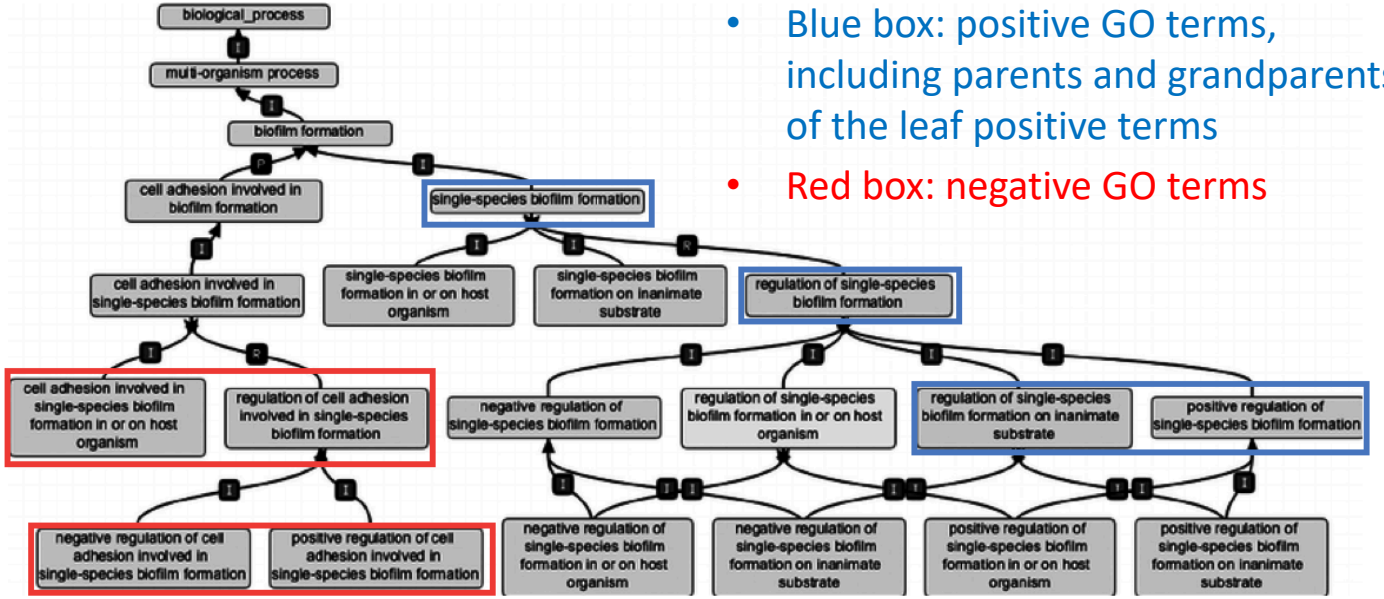

r=0.97
P < 2.2e-16
F1 score (PE.Approx derived)
F1 score (PE.Exact derived)
0.00
0.25
0.50
0.75

(A)



- Blue box: positive GO terms
- Red box: negative GO terms

(B)



- Blue box: positive GO terms, including parents and grandparents of the leaf positive terms
- Red box: negative GO terms