Protein binding and methylation on looping chromatin accurately predict distal regulatory interactions

Sean Whalen Rebecca M. Truty Katherine S. Pollard

May 8, 2015

Abstract

Identifying the gene targets of distal regulatory sequences is a challenging problem with the potential to illuminate the causal underpinnings of complex diseases. However, current experimental methods to map enhancer-promoter interactions genome-wide are limited by their cost and complexity. We present *TargetFinder*, a computational method that reconstructs a cell's three-dimensional regulatory landscape from two-dimensional genomic features. *TargetFinder* achieves outstanding predictive accuracy across diverse cell lines with a false discovery rate up to fifteen times smaller than common heuristics, and reveals that distal regulatory interactions are characterized by distinct signatures of protein interactions and epigenetic marks on the DNA loop between an active enhancer and targeted promoter. Much of this signature is shared across cell types, shedding light on the role of chromatin organization in gene regulation and establishing *TargetFinder* as a method to accurately map long-range regulatory interactions using a small number of easily acquired datasets.

Genotyping, exome sequencing, and whole-genome sequencing have linked thousands of non-coding variants to traits in humans and other eukaryotes [1–6] via genome-wide association studies, family studies, and other approaches. Non-coding variants are more likely to cause common disease than are non-synonymous coding variants [7], and they can account for the vast majority of heritability [8]. Yet few non-coding mutations have been functionally characterized or mechanistically linked to human phenotypes [7, 9]. Comparative [10] and functional [11–13] genomics, coupled with bioinformatics, are generating annotations of regulatory elements in many organisms and cell types [14], as well as tools for exploring or predicting the impact of mutations in regulatory DNA [15–18]. Massively parallel reporter assays [19] provide a way to test and refine some of these predictions. However, this new information will only improve our understanding of disease and other phenotypes if we can accurately link functional non-coding elements to the genes, pathways, and cellular processes they regulate. This is a difficult problem because promoters and their regulatory elements can be separated by thousands—even greater than one million [20]—base pairs (bp), and the closest promoter is usually not the true target in humans [21] (though this varies by species [22]). Incorrectly mapping regulatory variants to genes prevents meaningful downstream studies. Our goal is to address this problem by developing a computational method to accurately predict distal regulatory interactions with relatively easy-to-collect data.

A computational method is desirable for two reasons. First, experimental mapping of chromatin interactions at the resolution of individual promoters and regulatory elements on a genome-wide scale in many cell types and developmental stages is prohibitively expensive and technically challenging. Until recently, very few validated distal regulatory interactions were known. Hence, previous studies defined interactions indirectly via genomic proximity coupled with genetic associations (e.g., eQTLs [23]), gene expression [14, 24–26], or promoter chromatin state [27, 28]. High-throughput methods for assaying chromatin interactions now exist, including paired-end tag sequencing (ChIA-PET) [29] and extensions of the chromosome conformation capture (3C) assay [30] (5C, Hi-C) [31, 32]. Recent improvements to the resolution of genome-wide Hi-C give an unprecedented look at chromatin structure [33, 34], including studies that utilize sequence capture to enrich for interactions with annotated promoters [35]. But Hi-C at the resolution of individual regulatory elements is still prohibitively expensive for most labs. We therefore saw the first high-resolution Hi-C experiments as an opportunity to test the hypothesis that the spatial proximity of promoters with distal regulatory elements can be inferred from data that is routinely measured genome-wide, including DNA sequences, epigenetic marks, and/or protein-DNA binding events. An accurate and generalizable model would enable high-resolution *in silico* Hi-C for many cell types using data that already exists or can be collected rapidly. A second motivation to computationally model regulatory interactions using functional genomics data is to learn relationships between DNA sequences, structural proteins, transcription factors and modified histones in the context of chromatin looping and gene activation. Learning which combinations of features best predict looping might reveal novel protein functions and molecular mechanisms of distal gene regulation.

We implemented an algorithm called *TargetFinder* that integrates hundreds of functional genomics and sequence datasets to identify the minimal subset necessary for accurately predicting enhancer-promoter interactions across the genome. We focused on enhancers due to their large impact on gene regulation [36] and our ability to predict their locations genome-wide, though our approach could easily be extended to other classes of regulatory elements. Applying *TargetFinder* to three human ENCODE cell lines [11] with high resolution Hi-C data [33], we discovered that enhancer-promoter interactions can be predicted with extremely high accuracy. Interestingly, these analyses showed that functional genomics data marking the window *between* the enhancer and promoter are more useful for identifying true interactions than are proximal marks at the enhancer and promoter. Exploration of this phenomenon revealed specific proteins and chemical modifications on the chromatin loop that bring an enhancer in contact with its target promoter and not with nearby repressed or active but non-targeted promoters. Thus, *TargetFinder* provides a framework for accurately assaying three-dimensional genomic interactions, as well as techniques for mining massive collections of experimental data to shed new light on the mechanisms of distal gene regulation.

Results

Ensemble learning of regulatory interactions from genomic data

The core component of *TargetFinder* is a machine learning pipeline that builds and evaluates ensemble models of distal regulatory interactions from genomic features such as epigenetic marks, protein binding events, gene annotations, and evolutionary signatures (Figure 1). Ensemble learning methods have excellent performance, account for non-linear interactions between features, and estimate the predictive importance of each feature. We applied multiple ensemble methods, including random forests and gradient boosted trees, to ensure our conclusions are robust. The inputs to *TargetFinder* are pairs of enhancers and promoters, annotated as interacting or not in a given cell type, and features (i.e., summaries of genomic datasets) associated with each pair. The outputs are a model for predicting interactions in new enhancer-promoter pairs, assessments of model performance on held-out data, and measurements of how predictive each feature is alone and in combination with other features. The predictive contribution of different genomic regions is explored by separately quantifying the importance of features marking enhancers, promoters, and the genomic window between them, and we also examine the importance of each feature alone versus in combination with other features. We specifically implemented *TargetFinder* to address the challenging problem of distinguishing validated enhancer-promoter interactions from all non-interacting pairs of transcribed gene promoters and active distal enhancers ($>$ 10 kilobases (Kb) from the transcription start site (TSS)) within any 2 megabase (Mb) locus. The method is easily extended to include other types of regulatory elements, such as inactive promoters and enhancers, but we found that including inactive elements resulted in less informative models.

TargetFinder predicts enhancer-promoter pairs with high accuracy

We first identified active promoters and enhancers in three ENCODE cell lines that have rich functional genomics data as well as interaction data produced by Rao et al. [33]. These included K562 (mesoderm lineage cells from a leukemia patient), GM12878 (lymphoblastoid cells), and HeLa-S3 (ectoderm lineage cells from a cervical cancer patient). Datasets included measures of open chromatin, DNA methylation, and chromatin immunoprecipitation followed by sequencing (ChIP-seq) for transcription factors (TFs), architectural proteins, and modified histones (Supplemental Table S2). We also included features representing conserved synteny (occurring nearby across species) and co-annotation of the target gene and TFs with motifs in the enhancer (Methods). Using high resolution genome-wide measurements of chromatin interactions [33] in these lines, we annotated interacting and non-interacting enhancer-promoter pairs and generated their corresponding features. For each individual cell line as well as their combination, we repeatedly built models using a random subset of the data and then quantified predictive accuracy on the held-out data using various metrics (Methods), including a balance of precision and recall (power) called $F_{\max}$. Due to the large number of non-interacting pairs, precision and recall are less biased than the commonly reported area under the receiver operating characteristic curve (AUC) that de-emphasizes false positives.

TargetFinder performed well on held-out data in all cell lines, achieving $F_{\max}$ between 83-88% corresponding to 5-12% false discovery rate (FDR) or approximately 87% power at a 10% false positive rate (Table 1). Performance was similar across lines, with *TargetFinder* performing slightly worse in GM12878 compared to K562 and HeLa-S3. This variability is due in part to differences in the number of training samples as well as the quality and quantity of functional genomics data. Performance was nearly identical using random forests and gradient boosting, and both performed significantly better than non-ensemble methods. Since gradient boosting provides more options for estimating feature importance, subsequent results are from this algorithm. Interestingly, we found *TargetFinder* has very high precision and recall largely independent of enhancer-promoter interaction distances in the range of 10 Kb to 2 Mb (Supplementary Figures S1 and S2). By comparison, all commonly used bioinformatics methods have much higher FDR. For example, using the closest active gene has an estimated FDR of 53-77% [21, 37, 38].

Variable importance highlights key datasets for predicting interactions

TargetFinder quantifies the importance of each feature and annotates whether it is associated with interacting or non-interacting enhancer-promoter pairs. This enabled us to deeply explore the genomic data associated with chromatin loops and revealed several interesting patterns. First, among hundreds of diverse features, the most predictive were functional genomics experiments including specific ChIP-seq, DNase-seq, and DNA methylation experiments (Figure 2). We observed that features differed in importance across cell lines for many reasons, including real functional differences (e.g., different co-factors), lack of expression (e.g., tissue-specific TFs), data quality, and the experimental design of ENCODE (including data processing methods and availability of replicates). Combining cell lines, the most predictive features were DNA methylation and binding of structural proteins and histones. Proteins related to activation (in particular, activator protein 1 (AP-1) complex [39]) and repression (in particular, polycomb repressive complex 2 (PRC2) [40]) also boost performance. These results demonstrate the rich information about chromatin looping that is present in datasets that are easier and less costly to collect than high-resolution Hi-C.

The most surprising discovery is that, contrary to our prior belief that *TargetFinder* would mostly utilize proximal marks at the enhancer and promoter, the most informative features were instead protein binding events and epigenetic modifications in the genomic window between an enhancer and its target (Figure 3). This is true despite the fact that average signal (e.g., ChIP-seq peak density) was higher at enhancers and promoters for most features.

Window features are highly informative for several reasons. Some window features are directly involved in chromatin looping including CTCF, the cohesin complex (SMC3/RAD21), and zinc finger protein ZNF143. The latter interacts with CTCF to provide sequence specificity for chromatin interactions [41], and also interacts with lineage-specific TFs that ZNF143 binds at interacting promoters (e.g., HCFC1 in HeLa-S3 [42]). Other window features impact the likelihood that additional promoters in the locus are the true targets of an enhancer. For example, RNA polymerase II (Pol II) at a promoter is not informative alone, because it can indicate either active transcription or a gene that is poised for rapid activation. In the paused state, Pol II blocks premature activation by acting as an insulator [43]. *TargetFinder* learned that non-targets can be easily distinguished by a lack of activators or coactivators [44] as well as histone marks, such as H3K36me3 and H3K79me2, that are associated with elongation. When these features occur in the window between an enhancer and a promoter, they indicate that an intervening promoter may be the true target. On the other hand, the presence of heterochromatin, PRC2 silencing [40], and various insulators (including existing looping interactions [45] marked by cohesin) in the window suggest that intervening genes are unavailable for binding. Window-associated marks may also be proxies for relevant but un-assayed histone modifications [46]. Thus, functional genomics data in the window between an enhancer and promoter carries rich information about the chromatin conformation of the locus that *TargetFinder* utilizes to predict if the enhancer and promoter physically interact.

Finally, we observed that many of the most important features mark non-interacting enhancer-promoter pairs. In other words, the absence of one feature may be more predictive of looping chromatin than the presence of a different feature. One example is the absence of activators in the window, which is more informative than the presence of Pol II, which might be paused. We also found that the association of a feature with interacting versus non-interacting pairs may be different at promoters, enhancers, and windows between these (Figure 2). For instance, SMC3 at promoters and enhancers is positively associated with interactions, while SMC3 binding in the window region is associated with non-interacting enhancer-promoter pairs because it increases the likelihood that an intervening promoter is the true target. The same holds for histones associated with activation, elongation, and repression.

TargetFinder identifies complex interactions between DNA-binding proteins and epigenetic marks

There are many top ranked features with similar predictive power. This is due in part to strong correlations between feature values, for example, due to multi-protein complexes or groups of related histone modifications with similar binding patterns across enhancer-promoter pairs (Figure 4). Correlated blocks of top ranked features fall into several broad categories including architectural proteins (CTCF, RAD21, SMC3, ZNF143), DNA methylation, and several types of histone modifications related to elongation (H3K36me3, H3K79me2), heterochromatin (H4K20me1) and activation (H2AZ, H3K4me1/2/3, H3K9ac, H3K27ac). These clusters of features often divide into sub-blocks such that one is associated with interactions and the other with non-interactions, suggesting different roles in chromatin organization despite correlated genomic distributions.

One consequence of correlated features is that proteins performing multiple distinct functions tend not to be highly predictive on their own. Instead, *TargetFinder* preferentially uses their co-factors that specify function. For example, the histone acetyltransferase EP300 is not always a top ranked feature, despite being strongly associated with active enhancers due to its ability to acetylate H3K27 [47]. The importance of EP300 is reduced in some cell lines because it is highly correlated with more predictive co-factors. One such co-factor is C/EBPβ that has been shown to phosphorylate and modulate the activity of EP300, as well as translocate it to specific gene regions [48]. The predictive importance of ZNF143 is also sometimes reduced due to its correlation with CTCF and the cohesin complex. Considered independent of these other features, however, ZNF143 is among the top 10 predictors for each cell line. Similarly, JUN typically ranks below its co-factors, but has higher importance when *TargetFinder* is trained using all cell lines. This is because the combined model can only utilize features present in all cell lines, and the more predictive co-factors of JUN are not uniformly assayed by ENCODE.

To further explore correlations between co-localizing proteins, we systematically compared how well each feature predicts enhancer-promoter interactions in isolation against its performance in combination with other features (Supplementary Tables S3 to S6). In K562, large changes in predictive rank included SPI1 (PU.1) that has been linked to chromatin looping [49, 50], NCOR1 which represses transcription by inhibiting Pol II elongation [51], TBP that has been linked with long range interactions [52] and whose TAF3 subunit is recruited by CTCF to distal promoters [53], and SRF which regulates FOS [54] and interacts with C/EBPβ [55]. Large HeLa-S3 rank changes included TBP (see previous), TCF7L2 that is known to form DNA loops by bending [56], SMARCC1 (BAF155) that is part of the SWI/SNF complex implicated in long range looping [57], MAFK that is a subunit of NF-E2 linked to long range interactions and β-globin activation [58], and STAT1 linked to chromatin remodeling of the MHC locus [59]. GM12878 had large rank changes for FOXM1 and EBF1 having known roles in B cell fate [60], as well as NFATC1 involved in enhancer-promoter communication [61]. Other large rank changes commonly included activating histone marks such as H2AZ and H3K9ac that may help distinguish active enhancers and promoters, including non-targets within window regions that cannot be discriminated solely by activation marks at their promoters. The elevated importance of H2AZ might also be explained by the link between H2A ubiquitination and polycomb silencing [62]. Thus, changes in predictive rank recapitulate known protein interactions and can identify under-appreciated or novel biological interactions. Lineage-specific proteins without large rank changes also have strong evidence of their relevance to enhancer-promoter interactions, such as RFX5 that can interact with SMC3 to tether looping enhancers to their targets [63].

Next, we combined the correlations shown in Figure 4 with protein-protein interaction data to derive a network of highly predictive features for each cell line that sheds light on the biological interplay between different clusters of regulatory proteins (Figure 5). This provided further support for the observation that histone modifications shared between diverse cell lines are highly predictive when combined with tissue-specific TFs. For example, *TargetFinder* learned that H4K20me1 interacts with H3K9me1/2/3, and this combination is known to mark different types of silent chromatin [64]. It also identified the known interaction between PHF8 and H4K20, which is demethylated by PHF8 during cell cycle progression [65]. These observations are consistent with the finding that histone modifications alone cannot drive transcription, even when Pol II is successfully recruited [66]. Instead, our results underscore that the interaction of histone-modifying complexes such as deacetylases, acetyltransferases, demethylases, and methyltransferases are essential to predicting and understanding the mechanisms of distal gene regulation [67]. A powerful advantage of the tree-based ensemble classifiers used in *TargetFinder* is their ability to detect and utilize these complex, non-linear interactions between features. No single dataset or simple rule captures the genomic signature of all interactions, but *TargetFinder* can learn the more intricate rules needed for distinguishing the true target of an active enhancer (Figure 6, Supplemental Figure S3).

TargetFinder generalizes across cell types

We developed and validated *TargetFinder* on the data-rich cell lines provided by ENCODE with an eye towards its application to cell lines or tissues with more limited data, including those without validated enhancer-promoter interactions necessary for training a new model. To explore how well *TargetFinder* generalizes to other cell lines, we used models trained on K562, GM12878, and HeLa-S3 to make predictions on each of the other lines using only the top 16 features from the training cell line that were also available in the test cell line. We also made predictions on the HUVEC line (human umbilical epithelial cells), which has only 11 available features (Supplemental Table S2). Despite the significant differences between cell lines noted above, we found that training *TargetFinder* on one cell line and predicting on another identified a large portion of validated enhancer-promoter interactions (Supplementary Table S1). Estimates of $F_{\max}$ ranged between 38 and 46%, including on HUVEC, which indicates reasonable precision and recall (e.g., an $F_{\max}$ of 44% corresponded to 36% precision and 55% recall). Such performance may be sufficient to discover novel enhancer targets without the considerable expense or labor of high-resolution Hi-C experiments.

Since functional genomics data is both highly predictive and costly to obtain, we also evaluated the cross-validation performance of *TargetFinder* using a small number of features. For all cell lines, *TargetFinder* needed $\sim$16 features to achieve nearly optimal performance ($F_{\max}$ 0.76-0.81, Figure 7), and comparable performance was achieved with just 8 features ($F_{\max}$ 0.70-0.77). Thus, regulatory interactions in a new cell type could be predicted by generating less than ten ChIP-seq datasets and using a *TargetFinder* trained on cell types where validated interactions have already been obtained.

Discussion

Through precise chromatin looping, regulatory elements physically interact with promoters of their target genes over long genomic distances, while avoiding other nearby active and inactive promoters. How do they do this? We hypothesized that transcription factors, histones, and architectural proteins might combine to drive—or at least mark—distal regulatory interactions. If so, then we ought to be able to computationally model known interactions from functional genomics data, and the most important genomic datasets in the model might shed light on the mechanisms of gene regulation in three dimensions. To test this hypothesis, we built *TargetFinder*, a machine learning algorithm that finds an optimal combination of genomic features for predicting experimentally validated enhancer-promoter pairs. The resulting models of distal regulatory interactions achieved outstanding performance, with a balance of precision and recall across the K562, GM12878, and HeLa-S3 ENCODE cell lines ranging from 83-88%and an FDR up to fifteen times smaller than using the closest gene. Our findings were robust across multiple enhancer and promoter definitions, and were reproduced using multiple algorithms and programming languages.

Which functional genomics experiments are most informative about chromatin interactions?

A unique feature of our approach is that we combine high resolution genome-wide Hi-C interaction data [33] with the vast functional genomics datasets provided by the ENCODE project for predicting distal enhancer targets. By integrating these diverse datasets and examining their relevance to enhancer-promoter interactions, we discovered the most predictive datasets and highlighted the complex interplay between regulatory proteins and DNA in the three-dimensional genome. All of the top ranking features were functional genomics experiments, rather than conserved synteny or similarity of TF and target gene annotations. We identified EBF1, FOXM1, NCOR1, PML, RFX5, SMARCC1, SRF, STAT1, TBP, and TCF7L2 as combinatorially predictive proteins whose role in distal enhancer-promoter interactions may be under-appreciated. Other predictive proteins that were independently predictive included CDS1, C/EBPβ, GABPA, GATA2, HCFC1, JUN/JUNB/JUND, MEF2A, NFATC1, NFKB1, PHF8, REST, SAP30, SP2, SPI1, and THAP1. Many of these features interacted with the cohesin complex and ZNF143, which was recently shown to provide sequence specificity to cohesin-assisted chromatin looping [41]. Members of the cohesin complex (SMC3/RAD21), CTCF, and ZNF143 were also highly ranked and have greater potential to generalize across cell types than lineage-specific TFs. Such activators and repressors nonetheless boost performance in individual cell lines, particularly those related to AP-1 and PRC2 complexes. Well-known histone marks necessary for ChromHMM/Segway annotations of promoters and enhancers are also necessary, though we found activating marks H2AZ/H3K9ac and elongation marks H3K36me3/H3K79me2 were especially predictive.

DNA between interacting enhancers and promoters carries a distinct genomic signature

The knowledge gained in this study depended critically on our decision to include genomic data from the window between each enhancer and promoter in the analyses. We discovered these window features dominated those encoding chromatin states at the promoter and enhancer themselves. The genomic signature of looping DNA had several components. First, interacting pairs tended not to have cohesin complex bound to the window, although it was prevalent near the enhancer and promoter. Long-range loops in the window between a candidate enhancer and promoter greatly reduced their interaction probability, suggesting pre-existing loops act as a kind of insulator between flanking elements. Secondly, DNA between interacting enhancers and promoters tended not to contain activating TFs and epigenetic marks of elongation and active transcription, all of which could indicate the presence of an alternative promoter target. On the other hand, windows did contain epigenetic marks associated with heterochromatin, polycomb-associated proteins, and co-factors of CTCF associated with its insulator function. Given this, our predictive features are more relevant to looping models of interaction than alternatives such as facilitated tracking [68]. Polycomb complexes appear to play several roles in distinguishing nearby targets. For example, PRC2-targeted CpG islands are enriched for REST and CUX1 binding motifs, both transcriptional repressors [69] with high predictive importance. In Drosophila, cohesin co-localizes with PRC1 at promoters and interacts to control gene silencing [70]. Given the conservation of PRC between flies and humans [71], this has implications for the interaction of cohesin and PRC for mammalian gene silencing and thus discrimination of target promoters. Also, distal enhancers may sometimes serve to clear PRC from CpG islands [72]. Finally, recent work shows that cohesin spatially clusters enhancers [73] and is consistent with our observation that the presence of active marks at alternate nearby enhancers often increase the likelihood of interaction. These are several of many possible explanations for the ability of window-based features to predict distal enhancer-promoter interactions with high precision and recall—explanations that may be refined by analysis of new functional genomics datasets.

How does TargetFinder distinguish targets from non-target promoters in the same locus?

Careful examination of many enhancer-promoter pairs across cell lines suggests several broad rules influence *TargetFinder's* score of an enhancer-promoter interaction: 1) do the enhancer and other nearby enhancers look active? 2) does the target transcript look like it is actively elongating? 3) is the target promoter cell type-specific? 4) do other promoters near the target have repressive marks or marks of paused polymerase? 5) is another pair interacting within the window? and 6) are there marks of chromatin remodelers or architectural proteins in the window, plus cohesion complex adjacent to the promoter and enhancer, that might facilitate looping interactions?

Figure 6 illustrates how these rules are combined to learn that an enhancer loops over the promoters of intervening genes (INTS6, WDFY2) to interact with the promoter of DHRS12 roughly 400 kilobases away. No single mark distinguishes the target. All active promoters have a repressive H4K20me1 mark and an activating (via Pol II elogination) H3K36me3 mark. Furthermore, PHF8 is present at every promoter in the region, while SP2 is present at none. Thus two highly predictive features do not separate targets from non-targets in this locus. Instead, the CTCF mark lacking cohesin complex marks suggests the WDFY2 promoter is not tethered to a distal enhancer via chromatin looping. However, DHRS12 has a cohesin complex mark (RAD21) at its promoter, and both RAD21 and SMC3 nearby. This interaction may also be defined by more complex interactions, including FOS and JUN binding on the looping chromatin, which is associated with changes in conformation [74, 75] that could possibly be relevant to the angle of the loop and not necessarily limited to their presence or absence at competing promoters. The situation appears even more complicated in loci with multiple active enhancers, including physically associated enhancers targeting the same promoter (Supplemental Figure S3). But *TargetFinder* can still predict enhancer-promoter pairs with high accuracy in such loci, indicating a degree of modularity in the genomic signature of interactions across loci, regardless of their architectures.

Prospects for predicting regulatory interactions in many cell types

In addition to better understanding the mechanisms behind distal enhancer-promoter interactions within a specific cell line, we aimed to train *TargetFinder* on data-rich cell lines such as those provided by ENCODE, identify a minimal subset of easy to collect datasets needed for prediction, and make accurate predictions on new cell lines. Cross-cell line prediction is a difficult task as enhancers and promoters vary, functional genomics assays are noisy and may have different peak strengths due to numerous factors, and as few as 55% of interactions were shared between cell lines [33]. This condition is sometimes termed *covariate shift* and violates the assumptions of most machine learning methods. Despite these challenges, we discovered that accurate prediction requires only 8 ChIP-seq datasets, and nearly optimal prediction requires only ∼16. Importantly, many of these proteins are not routinely interrogated, and several frequently studied histones and TFs are redundant with or less predictive than proteins they interact with. Additionally, our analyses highlighted proteins that are predictive either in isolation or in combination with others. This impacts the probability that a dataset will generalize across cell types.

We therefore conclude that a researcher seeking to collect data for enhancer-promoter prediction in a new system might prioritize experiments that are in the top ∼16 features (Figure 2) for the most similar well-characterized cell type or use features that score well across multiple cell lines. This will direct researchers towards predictive co-factors rather than multi-functional proteins that may be better known but less predictive. A *TargetFinder* trained on this reduced feature set from the well-characterized line(s) with validated interactions could then be applied to the new system by plugging in the values of the features for enhancer-promoter pairs in the new cell type, without the need for generating validated interactions. A subset of the resulting candidate interactions could then be tested using low-throughput assays to validate the predictions. In addition to reducing the burden on experimentalists, a model learned from fewer features is less likely to overfit and thus more likely to perform well on new cell lines. Our study demonstrates that this approach has the potential to be much more accurate than simply mapping enhancers to the closest promoter. Thus, *TargetFinder* is not only a tool for predicting the interactions of distal regulatory elements, but also a screening tool for estimating the relevance of unassayed DNA-binding proteins and epigenetic marks in disparate cell lines—and potentially disparate organisms.

Materials and Methods

All code was implemented in Python using the `scikit-learn` machine learning library [76] and the `pandas` analytics library [77] in combination with `bedtools` [78]. Results were verified using a comparable pipeline implemented in R using the `caret` [79], `randomForest` [80], `gbm` [81], and `glmnet` [82] packages. Genome-wide data was obtained from the UCSC Genome Browser for the ENCODE Project [11] (`http://genome.ucsc.edu/ENCODE/`) for the K562 (tier-1), GM12878 (tier-1), and HeLa-S3 (tier-2) cell lines. GENCODE [83] version 19 annotations and expression data were obtained directly from the ENCODE portal (`https://www.encodeproject.org/data/annotations`). Chromatin interaction data generated by Rao et. al [33] was obtained from the Gene Expression Omnibus (GEO). Training data was generated using the same methods and parameters for each cell line. When possible, separate features were generated for enhancer, promoter, and window regions defined as all base pairs between the proximal edges of the enhancer and the promoter.

Promoter Identification

In each cell line, we identified actively transcribed protein coding genes with mean FPKM > 0.3 [84] and irreproducible discovery rate < 0.1 [85]. Corresponding promoters were regions labeled “TSS” (predicted promoter region including transcription start site) by the combined ChromHMM [86] and Segway [87] annotations available from the UCSC Genome Browser. This resulted in 9863, 10092, and 9303 active promoters for the above cell lines out of 20345 annotated protein coding genes. We also evaluated performance using GENCODE version 7 annotations and expression data, as well as promoter regions defined as a GENCODE TSS ± 2 kilobases (Supplementary).

Enhancer Identification

Enhancers were segments labeled “E” (strong enhancer) by the combined ChromHMM [86] and Segway [87] annotations available from the UCSC Genome Browser. To focus our models on distal interactions, enhancers closer than 10 kilobases to the nearest promoter were discarded. This resulted in 44227, 51631, and 41734 active enhancers for the above cell lines. We also evaluated performance using clustered TF binding sites (Supplementary).

Chromatin Interactions

Hi-C is an unbiased method for genome-wide identification of chromatin interactions [88]. Recent work [33] applied Hi-C with improved resolution to 9 cell types, 3 of which also had extensive ENCODE data. Hi-C interaction data obtained from GEO lists statistically significant interactions at 10% FDR, which we further filtered down to 1% FDR. Positive training samples were interactions with at least one active enhancer and at least one active promoter intersecting with forward and reverse Hi-C fragments. This resulted in 1100, 1368, and 855 interacting enhancer-promoter pairs for the above cell lines. We also evaluated performance using ENCODE 5C data (Supplementary).

Negatives were random pairs of active enhancers and promoters without a statistically significant Hi-C interaction. To select negatives matching the distribution of positive interaction distances, positives were assigned to one of 5 bins using quantile discretization of the distance between enhancer and promoter. For each bin, 5 negatives per positive were randomly selected for the training set. This number was chosen for computational efficiency, but cross-validated performance was similar using the complete set of negative enhancer-promoter pairs ($\approx$ 700-900k depending on the cell line).

Features

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) identifies where and how strongly TFs, architectural proteins, and modified histones bind along the genome. ChIP-seq assays for 141, 98, and 71 different proteins were performed genome-wide by the ENCODE consortium for the above cell lines. Peaks were called by ENCODE using a uniform pipeline and biological replicates where possible, then provided as BED files that specify peak locations and strengths. These were intersected with promoter, enhancer, and window regions. The average peak strength per region was used as a feature (computed as the sum of peak strengths divided by the region length in base pairs), resulting in 3 features per ChIP-seq dataset.

DNase I hypersensitive sites sequencing (DNase-seq) and Formaldehyde-Assisted Isolation of Regulatory Elements followed by sequencing (FAIRE-seq) are similar assays for identifying regulatory regions in the genome. These assays were converted to features using the above ChIP-seq methodology.

Reduced representation bisulphite sequencing (RRBS) identifies methylated DNA regions and was performed genome-wide by the ENCODE consortium for the above cell lines. These regions and their methylated base counts were intersected with our promoter, enhancer, and window regions. The percent of methylated bases within each region was used as a feature, resulting in 3 features.

Annotation-based features were derived from STRING [89], IMP [90], and GeneMANIA [91] by summing the interaction scores between the gene and all TFs predicted by CENTIPEDE to bind the enhancer [92]. Features for each tool were derived separately. A synteny-based feature was derived using the phylogenetic distance between enhancers and promoters covered by the same syntenic nets [93], summed over the 23 mammals in the UCSC Genome Browser 46-way multi-species alignment (`http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/`). Annotation and synteny features were excluded from our final trained version of *TargetFinder* due to their low predictive importance relative to their high computational cost.

Machine Learning

Supervised ensemble learning algorithms [94] were used to predict enhancer-promoter interactions. Ensemble learning is a subfield of machine learning that trains multiple diverse models and combines their predictions to achieve performance greater than the best individual model. Supervised algorithms require labeled training data; enhancer-promoter pairs are labeled as *positive* if the regions interact according to Hi-C data (see above) and *negative* otherwise. We used both random forests [95] and gradient boosted trees [96] to ensure consistent results. The former constructs independent decision trees in parallel, while the latter iteratively constructs decision trees and places increasing emphasis on high-error samples.

For boosting, we used the `gbm` R package [81] and `GradientBoostingClassifier` in the `scikit-learn` Python package [76]. Nested cross-validation achieved optimal performance using 4096 iterations (trees), shrinkage (learning rate) 0.1, and interaction depth (maximum tree depth) 9. Similar performance was achieved with slightly more conservative parameters.

For random forests, we used the `randomForest` R package [80] and `RandomForestClassifier` in the `scikit-learn` Python package [76]. We used 1500 trees and left all other parameters at defaults. A much smaller forest achieved similar performance; the larger forest was used solely to stabilize estimates of feature importance.

We verified that the cross-validated performance and feature importances of boosting and random forests were similar, though differences are expected by design. In addition, we evaluated logistic regression (using `scikit-learn` [76]) and elastic nets tuned via nested cross-validation (using `caret` [79] and `glmnet` [82]). The resulting performance drop was substantial and emphasizes the importance of capturing non-additive feature interactions for predicting enhancer-promoter interactions. Baseline performance was estimated using random training labels.

For linear classifiers, features were first mean-centered and scaled to unit variance. For all classifiers, training samples within each CV fold were assigned weights inversely proportional to their class prevalence in order to compensate for severe class imbalance.

Feature importances given in the paper were estimated using only gradient boosting, for simplicity.

Performance Evaluation

True positives (tp), false negatives (fn), false positives (fp), and true negatives (tn) are defined by the following contingency table comparing actual and predicted labels:

Our chosen classifiers generate scores representing confidence that a sample belongs to the positive class. To evaluate performance, scores above a threshold are given a positive label and otherwise are labeled negative. Raising this threshold results in fewer but more confident positive predictions. As a result, we used two metrics that summarize performance over all possible thresholds using the following base metrics:

	Predicted	
Actual	Positive	Negative
Positive	tp	fn
Negative	fp	tn

$$
\begin{aligned}
\text{true positive rate} &= \frac{\text{tp}}{\text{tp} + \text{fn}} \\
\text{false positive rate} &= \frac{\text{fp}}{\text{fp} + \text{tn}} \\
\text{precision} &= \frac{\text{tp}}{\text{tp} + \text{fp}} \\
\text{recall} &= \text{true positive rate} \\
F_\beta &= (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}
\end{aligned}
$$

Area under the receiver operating characteristic curve (auROC or more commonly AUC) measures the area under the curve formed by the true positive and false positive rate of the classifier over all possible thresholds for the positive class. F_β is the weighted harmonic mean of precision and recall. We used the common F_1 score (F_β where $\beta = 1$) to equally weight precision and recall. $F_{\max}$ is the maximum F_1 score over all possible thresholds for the positive class.

AUC and $F_{\max}$ were estimated with 10-fold cross validation where data is split into 10 non-overlapping training and test sets. Samples were weighted with the inverse of their class counts to compensate for the imbalance between negative and positive samples. Classifiers were constructed for each training set and predictions were generated for the corresponding test set. Performance was evaluated independently for each test set and averaged to produce a single estimate per metric. For example, a 3-fold cross validation might result in an AUC of 0.7, 0.8, and 0.9 for folds 1, 2, and 3, resulting in an average AUC of 0.8.

Ensemble Feature Selection

Recursive Feature Elimination (RFE) is an embedded multivariate feature selection technique [97] that has recently been adapted to random forests [98, 99]. The selection process is similar to parameter tuning where the best number of features is considered a parameter of the model, and requires nested cross-validation as detailed by Ambroise and McLachlan [100] to obtain unbiased performance estimates. Our analysis used `caret`'s implementation of RFE.

For each cross-validation fold, the ensemble is first trained using the complete feature set. Importances are then estimated by permuting the values of each feature over all out-of-bag samples and measuring the average loss in accuracy. Out-of-bag samples are those excluded at each tree during training as a result of resampling with replacement. The performance of feature subsets is then evaluated using inner cross-validation. The performance of the best subset from inner cross-validation is then re-estimated using only the outer test fold to avoid bias. The smallest subset size having average performance within 1.5% of the best performance across all folds was selected as optimal. To reduce computation time, we evaluated subset sizes from 1 up to the maximum number of features counting by powers of 2.

RFE performance was estimated using random forests, but not boosting, due to limitations in `caret`.

Interaction Networks

A network of feature interactions was created using the top 10 predictive datasets per cell line. Nodes in the network were connected with an orange edge if they had Pearson correlation above 0.3 at enhancer, promoter, or window regions. Purple edges connected nodes with known protein interactions according to BioGrid 3.3.122 [101]. The the most central node (part of the most shortest paths between all nodes) were shown in bold. Central nodes often correspond to datasets with large rank changes in univariate versus multivariate performance. A limited number of nodes were shown to conserve space.

Competing interests

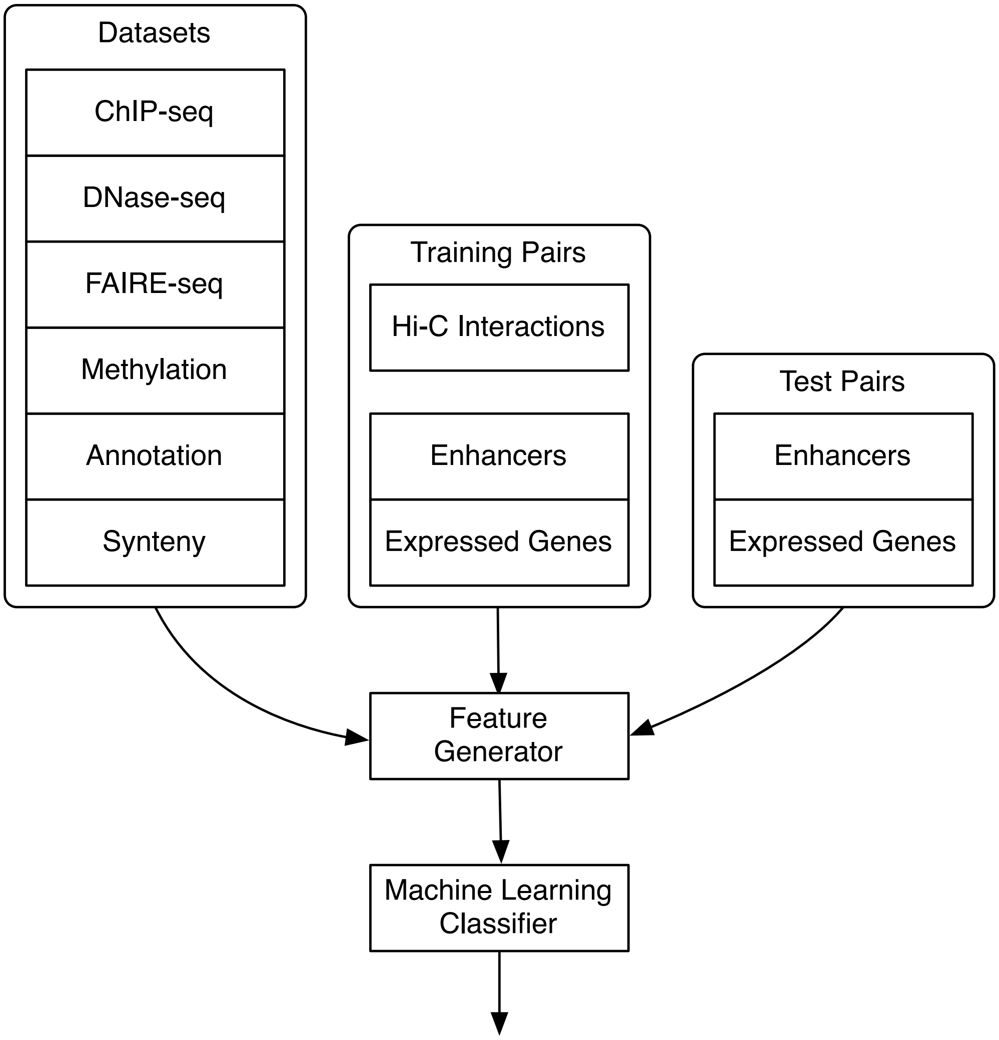
The authors declare that they have no competing interests.

Acknowledgments

This project was supported by the Bench to Bassinet Program of the NHLBI (U01HL098179), NIH/NHLBI (HL089707), the San Simeon Fund, and the Gladstone Institutes.

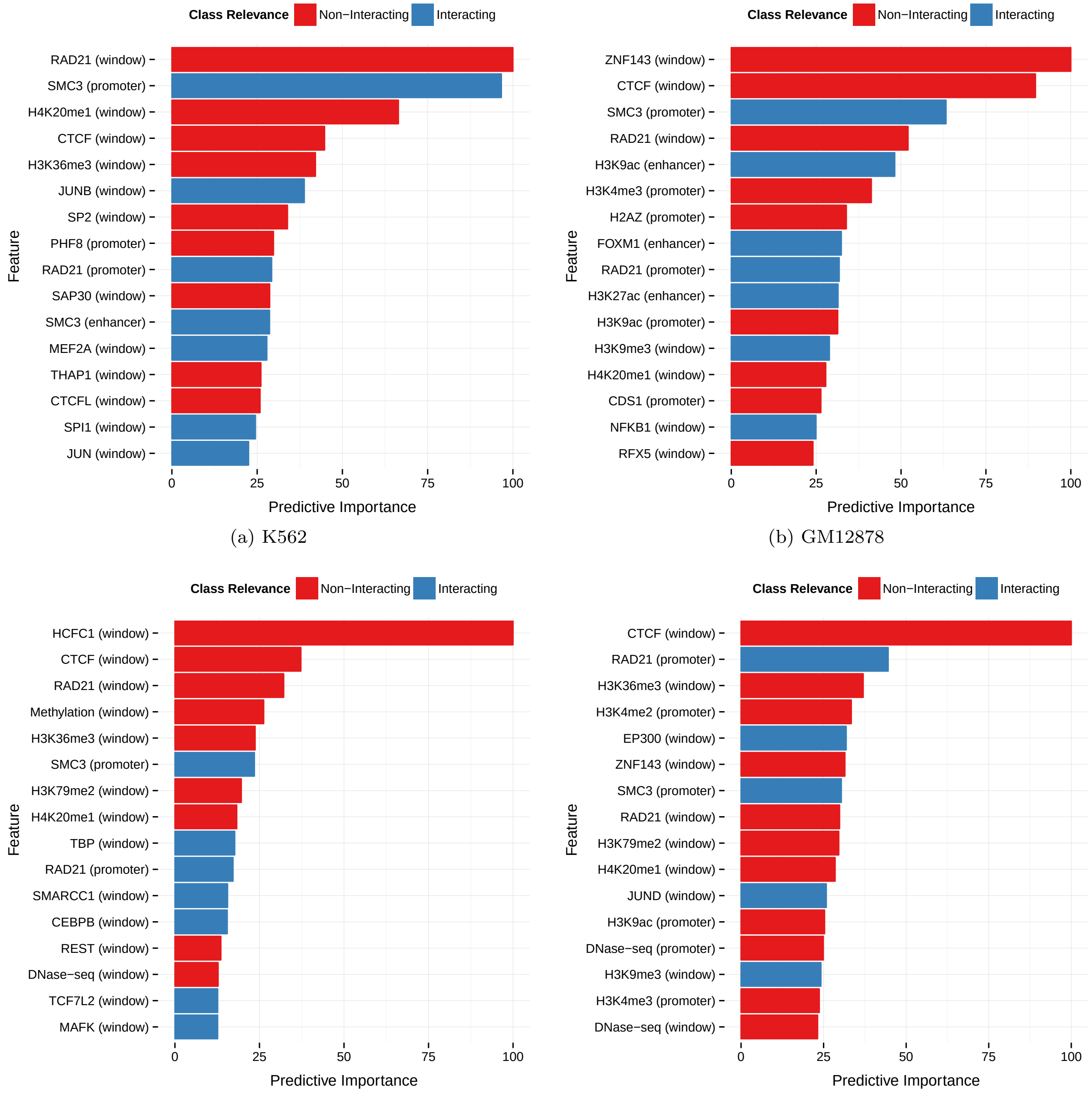
	Model	AUC	TPR	Precision	Recall	$F_{\max}$	MCC
Cell Line							
K562	Baseline	0.50	0.00	0.17	1.00	0.29	0.01
K562	Logistic Regression	0.79	0.48	0.48	0.56	0.51	0.42
K562	Gradient Boosting	0.95	0.87	0.88	0.80	0.84	0.82
K562	Random Forest	0.95	0.87	0.92	0.80	0.85	0.83
GM12878	Baseline	0.50	0.00	0.17	1.00	0.29	0.00
GM12878	Logistic Regression	0.78	0.43	0.42	0.60	0.49	0.39
GM12878	Gradient Boosting	0.94	0.85	0.87	0.76	0.81	0.79
GM12878	Random Forest	0.94	0.85	0.90	0.76	0.83	0.79
HeLa-S3	Baseline	0.49	0.00	0.17	1.00	0.29	-0.02
HeLa-S3	Logistic Regression	0.81	0.51	0.50	0.58	0.53	0.46
HeLa-S3	Gradient Boosting	0.96	0.89	0.93	0.81	0.87	0.84
HeLa-S3	Random Forest	0.96	0.89	0.95	0.83	0.88	0.85
Combined	Baseline	0.50	0.00	0.17	1.00	0.29	-0.00
Combined	Logistic Regression	0.76	0.41	0.39	0.59	0.47	0.36
Combined	Gradient Boosting	0.94	0.86	0.86	0.77	0.81	0.80
Combined	Random Forest	0.95	0.87	0.91	0.77	0.83	0.80

Table 1: **TargetFinder performance on held out data.** Metrics include precision, recall, the maximum harmonic mean of precision and recall over all scoring thresholds ($F_{\max}$), Matthews correlation coefficient (ϕ), area under the ROC curve, and power (true positive rate) at a 10% false positive rate. Ensemble methods (random forests and gradient boosting) have similarly high precision and recall compared to linear models due to their ability to capture non-linear feature interactions. The gap between AUC and precision/recall measures demonstrates how the former is biased due to de-emphasis of false positives. Metrics were computed on predictions generated for the test split of each cross-validation fold and then averaged. Precision and recall were computed using the $F_{\max}$ threshold. Baseline performance was estimated using random training labels.



Enhancer	Gene	Score	Score > 0.5?	Prediction
chr1:150-200	ENSG001	0.93	Yes	Interaction
chr2:400-900	ENSG002	0.48	No	No Interaction

Figure 1: **The *TargetFinder* pipeline.** Features are generated from hundreds of diverse datasets for pairs of enhancers and promoters of expressed genes found to have significant Hi-C interactions (positives), as well as random pairs of enhancers and promoters without significant interactions (negatives). These labeled samples are used to train an ensemble classifier that is used to examine the predictive importance of each feature or predict whether new enhancer-promoter pairs interact. Classifier predictions are probabilities, and a decision threshold (commonly 0.5 but may be adjusted) converts these to positive or negative prediction labels. Though any classifier can be used, we selected two popular ensemble methods (random forests and gradient boosted trees; Methods) for their predictive accuracy and interpretability. This figure excludes the selection of minimal predictor sets for simplicity.



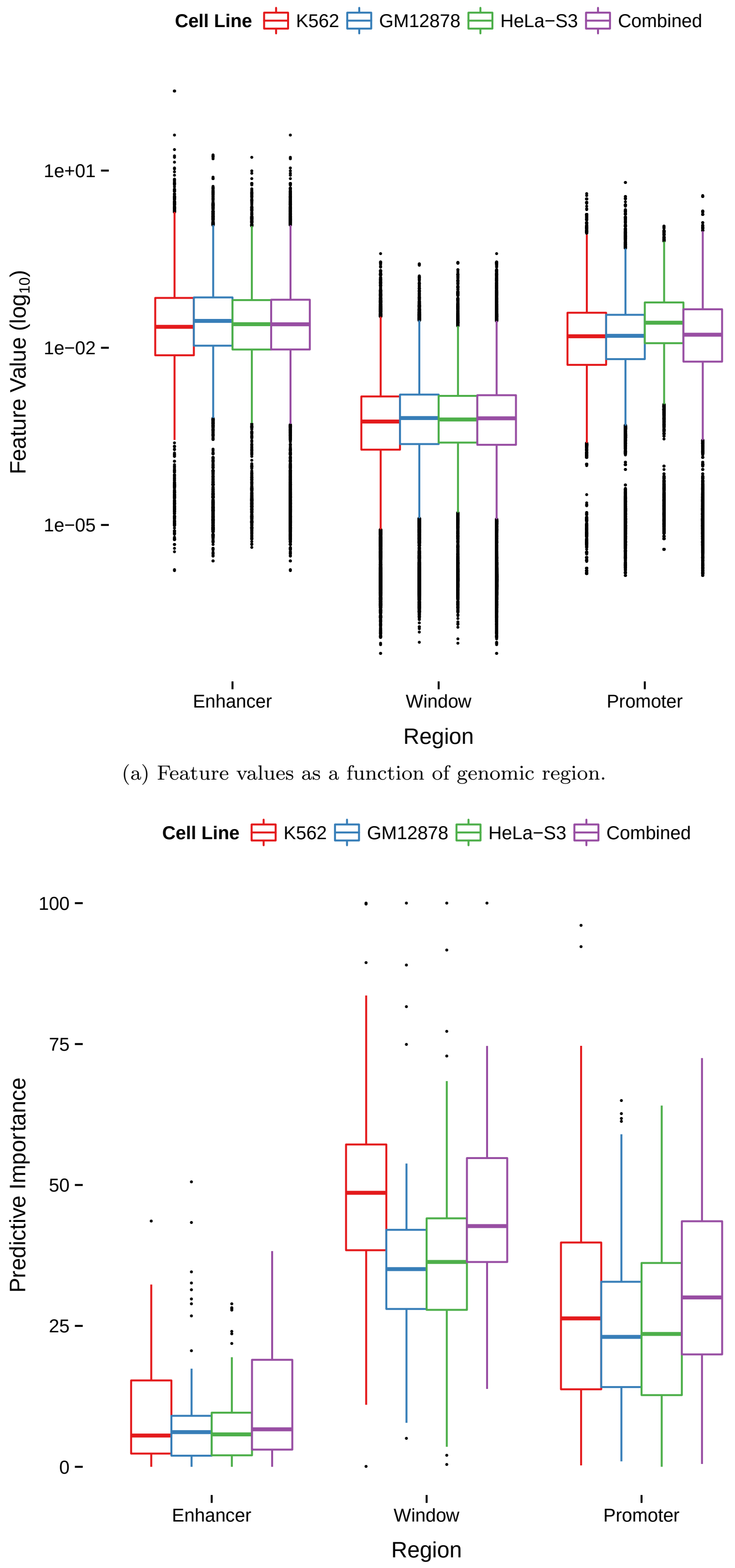
(a) K562

(b) GM12878

(c) HeLa-S3

(d) Combined

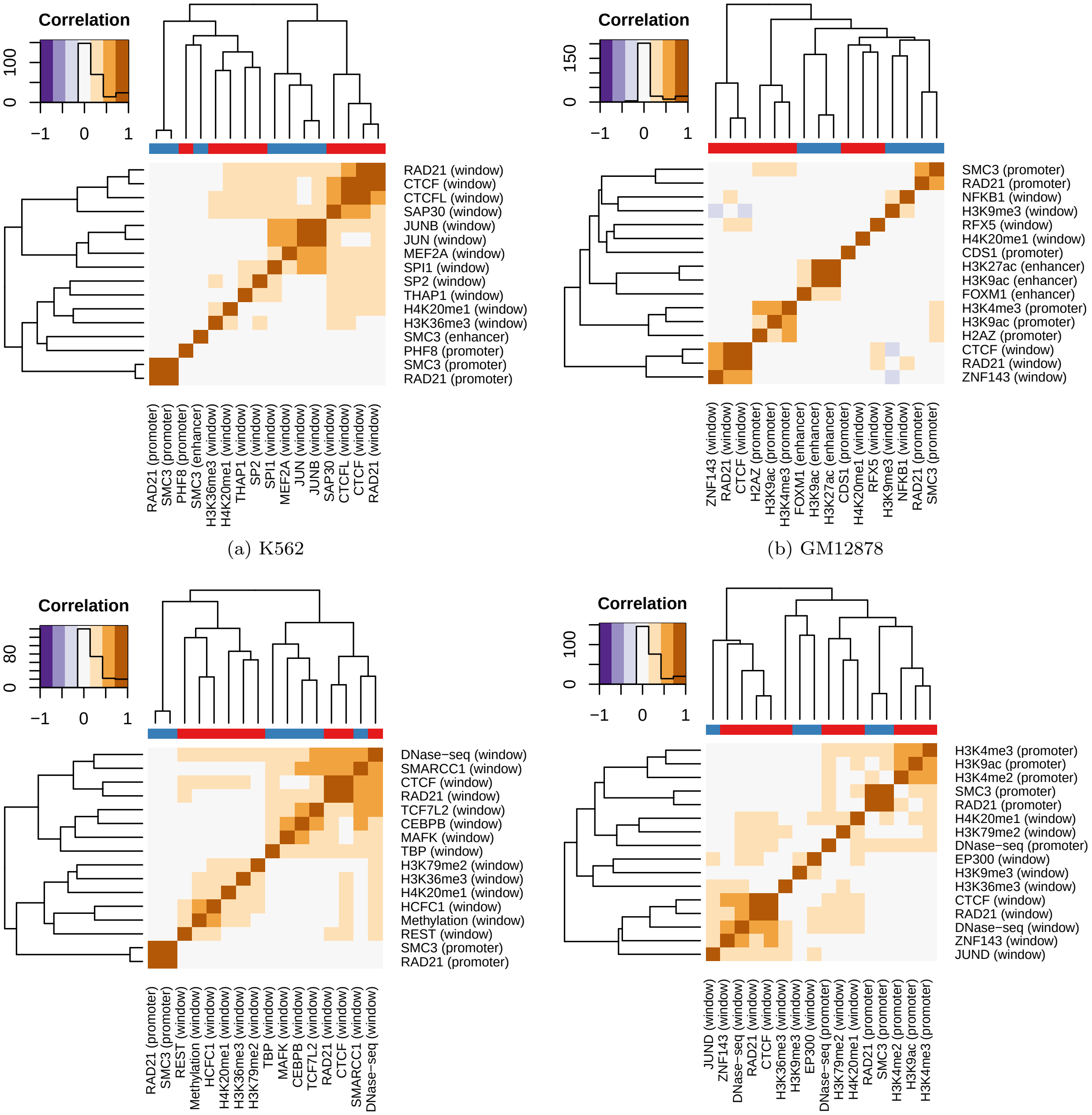
Figure 2: **Top 16 predictive features per cell line.** Color indicates relevance to interacting (blue) or non-interacting (red) enhancer-promoter pairs, estimated by a gradient boosting classifier. Features at the region between the enhancer and promoter (the *window*) are more prevalent than those directly at the enhancer and promoter. The same feature (e.g., CTCF, cohesin complex) may be relevant to either interacting or non-interacting pairs depending on whether it binds in the window versus at enhancers or promoters.



(a) Feature values as a function of genomic region.

(b) Predictive importance as a function of genomic region.

Figure 3: **Feature values and predictive performance for enhancer, promoter, and window regions.** Despite having the lowest feature values, the predictive importance of the window dominates that of enhancer and promoter regions.



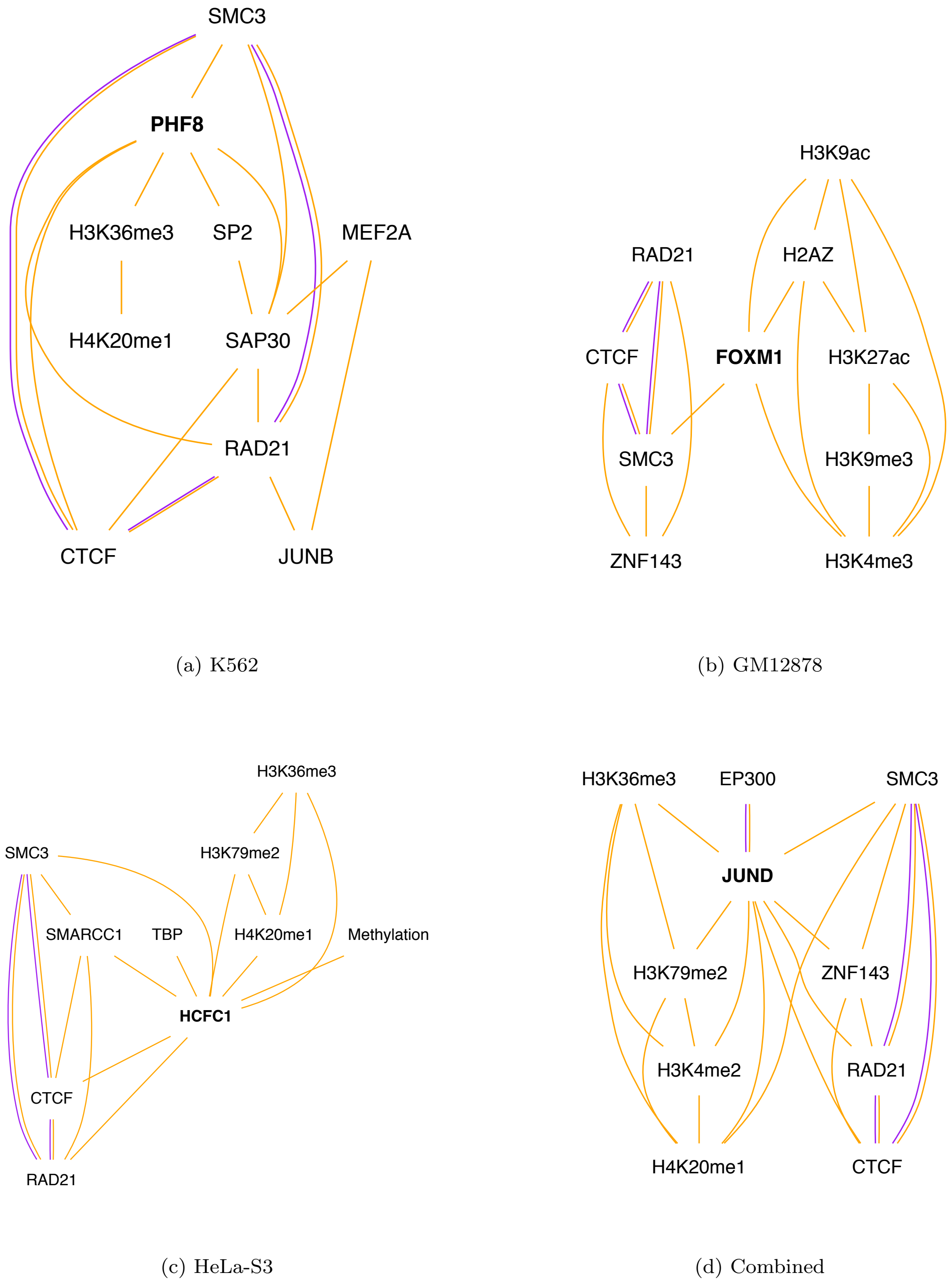
(a) K562

(b) GM12878

(c) HeLa-S3

(d) Combined

Figure 4: **Predictive features co-occur on the genome.** Clustered correlation heatmap of the 16 most predictive features per cell line. The color bar along the top of the heatmap indicates feature relevance to interacting (blue) or non-interacting (red) enhancer-promoter pairs. Blocks of correlated predictors along the diagonal often indicate proteins found in the same complex, or histones interacting with a chemical modifier such as an acetyltransferase or deacetylase.



(a) K562

(b) GM12878

(c) HeLa-S3

(d) Combined

Figure 5: **Predictive features physically interact in protein complexes.** Interaction network for the top 10 most predictive features per cell line. Purple edges indicate known protein interactions according to BioGrid (Methods). Other features co-occur with these known protein interactions (orange edges indicating moderate to high correlation of peak locations along the genome) and may form complexes with them. The node that is part of the most shortest paths between any two nodes (the most *central* node) is shown in bold and is often a lineage-specific TF.

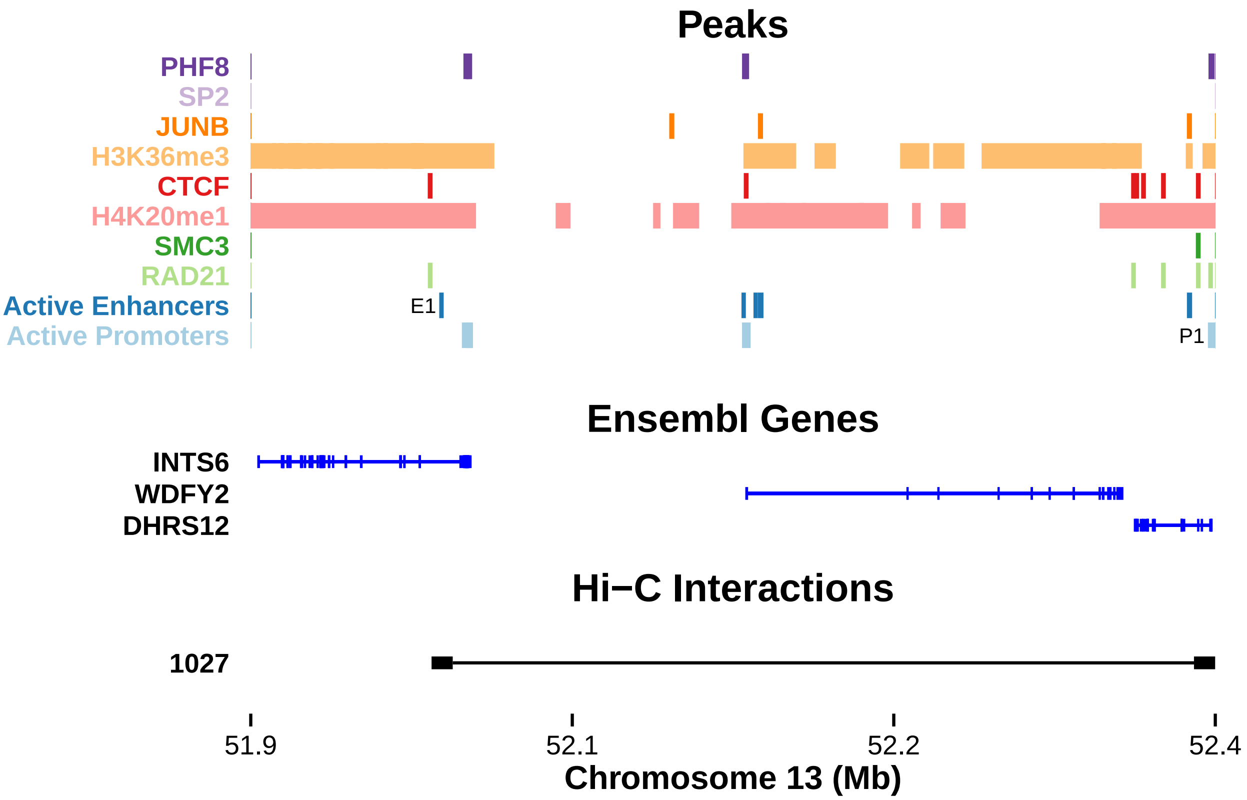
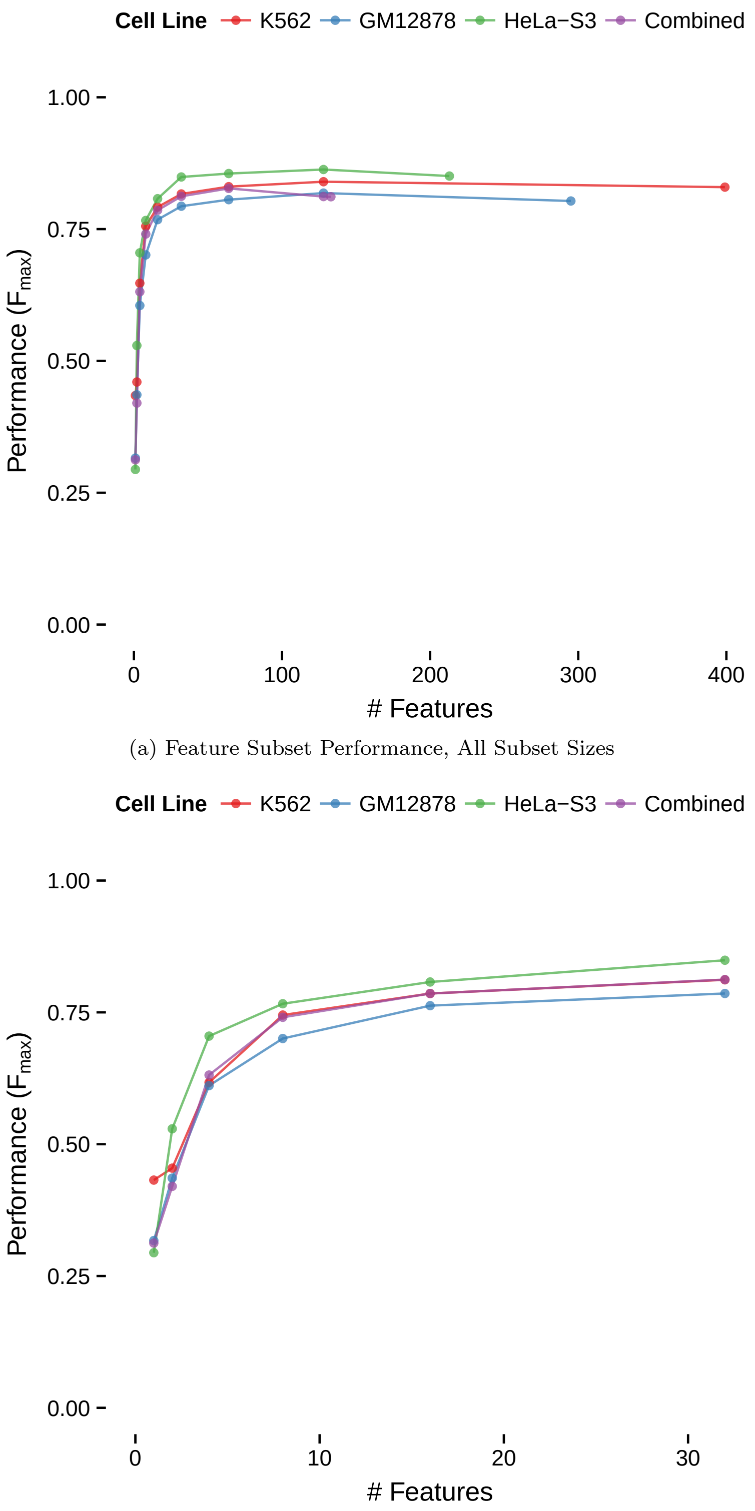


Figure 6: **Predicting a chromatin loop that skips over two active promoters in K562 cells.** Browser-like tracks display the top 8 predictive datasets (significant peak values), ChromHMM enhancers and promoters, Ensembl genes (introns = thin lines, exons = squares), and Hi-C interactions (left and right fragments connected by a line). An enhancer (E1) interacts with a promoter of DHRS12 (P1) nearly 400 kilo bases away and not with the intervening active promoters of INTS6 and WDFY2. *TargetFinder* integrates multiple datasets (including ones beyond the top 8 displayed) to learn the architecture of this locus. It appears likely that WDFY2 is not targeted due to the presence of the insulator CTCF without interaction-associated cohesin complex marks (SMC3/RAD21), while DHRS12 is marked by RAD21 at the promoter and SMC3/RAD21 nearby. Interestingly, P1 has marks for both H4K20me1 (repression) and H3K36me3 (elongation leading to activation), and the highly predictive mark SP2 is not present in this particular region. These characteristics underscore the need for a machine-learning approach to integrate complex genomic signatures for accurate target prediction.



(a) Feature Subset Performance, All Subset Sizes

(b) Feature Subset Performance, Subset Sizes ≤ 32

Figure 7: **Performance as a function of the number of features.** Performance of recursive feature elimination (Methods) that evaluates predictor subsets of size 1 up to the maximum per cell line and increasing by powers of 2 for computational efficiency. Near optimal performance was achieved using only ∼16 predictors for lineage-specific models as well as the combined model, while lower but acceptable performance required only 8 predictors. Performance slowly degrades past a certain number of predictors due to the noise introduced by uninformative features. For extra detail, the second figure shows performance using the top 1, 2, 4, 8, 16, and 32 predictors.

Supplemental Material

Alternate promoter and enhancer definitions

Before transitioning to annotation-based enhancers, we defined candidate enhancers as transcription factor binding sites (TFBS) identified by CENTIPEDE [92], lifted these over from the hg18 to hg19 assembly, and clustered them [102] using the DBSCAN algorithm [103] with `eps = 300` and `min_samples = 1`. Finally, we intersected the resulting TFBS clusters with p300, H3K27ac, and H3K4me1 ChIP-seq peaks from the same cell line and retained all clusters that overlapped at least one of these ChIP-seq marks. Clusters closer than 10kb to the nearest promoter were discarded. This approach had comparable performance for 5C-assayed interactions but was outperformed by annotation-defined enhancers for Hi-C-assayed interactions.

Alternate interaction data

Before transitioning to Hi-C-assayed interactions, we used chromosome conformation capture carbon copy (5C) data from ENCODE that also identifies physically interacting segments of the genome [31]. Enhancers were intersected with forward 5C fragments and promoters with reverse 5C fragments. Following the ENCODE standard for interaction significance, enhancer-promoter pairs with fragments found to interact across both 5C biological replicates were given positive labels in our training data. To select negatives matching the distribution of interaction distances, positives were first assigned a bin number using quantile discretization of the distance between enhancers and promoters. For each positive distance bin, 200 negatives were generated by randomly selecting non-interacting enhancer-promoter pairs within the ENCODE pilot regions. The number of negatives per bin was limited by the number of active promoters covered by reverse 5C fragments. Due to the limited number of positives, we transitioned to Hi-C data when it became available with sufficient resolution.

Supplemental Tables

Test Cell Line	GM12878	HUVEC	HeLa-S3	K562
Training Cell Line				
GM12878		0.39	0.43	0.40
HUVEC	0.39		0.38	0.40
HeLa-S3	0.43	0.41		0.38
K562	0.46	0.44	0.45	

Table S1: *TargetFinder* performance ($F_{\max}$) when trained on one cell line and tested against another. The top 16 features in the training cell line that were also present in the test cell line were used. An additional cell line not present in other evaluations, HUVEC, was used to test performance on a cell line where few (11) datasets were available.

Table S2: Datasets available for each cell line, driven primarily by availability from ENCODE.

	Cell Line			
	K562	GM12878	HeLa-S3	HUVEC
ChIP-seq (ARID3A)	X			
ChIP-seq (ATF1)	X			
ChIP-seq (ATF2)		X		
ChIP-seq (ATF3)	X	X		
ChIP-seq (BACH1)	X			
ChIP-seq (BATF)		X		
ChIP-seq (BCL11A)		X		

Table S2: Datasets available for each cell line, driven primarily by availability from ENCODE.

	Cell Line			
	K562	GM12878	HeLa-S3	HUVEC
ChIP-seq (BCL3)	X	X		
ChIP-seq (BCLAF1)	X	X		
ChIP-seq (BDP1)	X		X	
ChIP-seq (BHLHE40)	X	X		
ChIP-seq (BRCA1)		X	X	
ChIP-seq (BRF1)	X		X	
ChIP-seq (BRF2)	X		X	
ChIP-seq (CBX2)	X			
ChIP-seq (CBX3)	X			
ChIP-seq (CBX8)	X			
ChIP-seq (CCNT2)	X			
ChIP-seq (CDS1)		X		
ChIP-seq (CEBPB)	X	X	X	
ChIP-seq (CEBPD)	X			
ChIP-seq (CHD1)	X	X		
ChIP-seq (CHD2)	X	X	X	
ChIP-seq (CHD4)	X			
ChIP-seq (CHD7)	X			
ChIP-seq (CREB1)	X	X		
ChIP-seq (CREBBP)	X			
ChIP-seq (CTCF)	X	X	X	X
ChIP-seq (CTCFL)	X			
ChIP-seq (CUX1)	X			
ChIP-seq (E2F1)			X	
ChIP-seq (E2F4)	X	X	X	
ChIP-seq (E2F6)	X		X	
ChIP-seq (EBF1)		X		
ChIP-seq (EGR1)	X	X		
ChIP-seq (ELF1)	X	X		
ChIP-seq (ELK1)	X	X	X	
ChIP-seq (ELK4)			X	
ChIP-seq (EP300)	X	X	X	
ChIP-seq (ETS1)	X	X		
ChIP-seq (EZH2)	X	X	X	X
ChIP-seq (FOS)	X	X	X	X
ChIP-seq (FOSL1)	X			
ChIP-seq (FOXM1)		X		
ChIP-seq (GABPA)	X	X	X	
ChIP-seq (GATA1)	X			
ChIP-seq (GATA2)	X			X
ChIP-seq (GTF2B)	X			
ChIP-seq (GTF2F1)	X		X	
ChIP-seq (GTF3C2)	X		X	
ChIP-seq (H2AZ)	X	X	X	
ChIP-seq (H3K27ac)	X	X	X	
ChIP-seq (H3K27me3)	X	X	X	
ChIP-seq (H3K36me3)	X	X	X	
ChIP-seq (H3K4me1)	X	X	X	
ChIP-seq (H3K4me2)	X	X	X	

Table S2: Datasets available for each cell line, driven primarily by availability from ENCODE.

	Cell Line			
	K562	GM12878	HeLa-S3	HUVEC
ChIP-seq (H3K4me3)	X	X	X	
ChIP-seq (H3K79me2)	X	X	X	
ChIP-seq (H3K9ac)	X	X	X	
ChIP-seq (H3K9me1)	X			
ChIP-seq (H3K9me3)	X	X	X	
ChIP-seq (H4K20me1)	X	X	X	
ChIP-seq (HA-E2F1)			X	
ChIP-seq (HCFC1)	X		X	
ChIP-seq (HDAC1)	X			
ChIP-seq (HDAC2)	X			
ChIP-seq (HDAC6)	X			
ChIP-seq (HDAC8)	X			
ChIP-seq (HMGN3)	X			
ChIP-seq (IKZF1)		X		
ChIP-seq (IRF1)	X			
ChIP-seq (IRF3)		X	X	
ChIP-seq (IRF4)		X		
ChIP-seq (JUN)	X		X	X
ChIP-seq (JUNB)	X			
ChIP-seq (JUND)	X	X	X	
ChIP-seq (KAT2A)		X	X	
ChIP-seq (KAT2B)	X			
ChIP-seq (KDM1A)	X			
ChIP-seq (KDM5B)	X			
ChIP-seq (MAFF)	X			
ChIP-seq (MAFK)	X	X	X	
ChIP-seq (MAX)	X	X	X	X
ChIP-seq (MAZ)	X	X	X	
ChIP-seq (MEF2A)	X	X		
ChIP-seq (MEF2C)		X		
ChIP-seq (MTA3)		X		
ChIP-seq (MXI1)	X	X	X	
ChIP-seq (MYC)	X	X	X	X
ChIP-seq (NCOR1)	X			
ChIP-seq (NELFE)	X			
ChIP-seq (NFATC1)		X		
ChIP-seq (NFE2)	X	X		
ChIP-seq (NFIC)		X		
ChIP-seq (NFKB1)		X		
ChIP-seq (NFYA)	X	X	X	
ChIP-seq (NFYB)	X	X	X	
ChIP-seq (NR2C2)	X	X	X	
ChIP-seq (NR2F2)	X			
ChIP-seq (NR4A1)	X			
ChIP-seq (NRF1)	X	X	X	
ChIP-seq (PAX5)		X		
ChIP-seq (PBX3)		X		
ChIP-seq (PHF8)	X			
ChIP-seq (PML)	X	X		

Table S2: Datasets available for each cell line, driven primarily by availability from ENCODE.

	Cell Line			
	K562	GM12878	HeLa-S3	HUVEC
ChIP-seq (POL2)				X
ChIP-seq (POL2A)				X
ChIP-seq (POLR2A)	X	X	X	
ChIP-seq (POLR3A)			X	
ChIP-seq (POLR3G)	X	X		
ChIP-seq (POU2F2)		X		
ChIP-seq (PRDM1)			X	
ChIP-seq (RAD21)	X	X	X	
ChIP-seq (RBBP5)	X			
ChIP-seq (RCOR1)	X	X	X	
ChIP-seq (REST)	X	X	X	
ChIP-seq (RFX5)	X	X	X	
ChIP-seq (RNF2)	X			
ChIP-seq (RUNX3)		X		
ChIP-seq (RXRA)		X		
ChIP-seq (SAP30)	X			
ChIP-seq (SETDB1)	X			
ChIP-seq (SIN3A)		X		
ChIP-seq (SIN3AK20)	X			
ChIP-seq (SIRT6)	X			
ChIP-seq (SIX5)	X	X		
ChIP-seq (SMARCA4)	X		X	
ChIP-seq (SMARCB1)	X		X	
ChIP-seq (SMARCC1)			X	
ChIP-seq (SMARCC2)			X	
ChIP-seq (SMC3)	X	X	X	
ChIP-seq (SP1)	X	X		
ChIP-seq (SP2)	X			
ChIP-seq (SPI1)	X	X		
ChIP-seq (SREBF1)		X		
ChIP-seq (SREBF2)		X		
ChIP-seq (SRF)	X	X		
ChIP-seq (STAT1)	X	X	X	
ChIP-seq (STAT2)	X			
ChIP-seq (STAT3)		X	X	
ChIP-seq (STAT5A)	X	X		
ChIP-seq (SUPT20H)		X	X	
ChIP-seq (SUZ12)	X			
ChIP-seq (TAF1)	X	X	X	
ChIP-seq (TAF7)	X			
ChIP-seq (TAL1)	X			
ChIP-seq (TBL1XR1)	X	X		
ChIP-seq (TBP)	X	X	X	
ChIP-seq (TCF12)		X		
ChIP-seq (TCF3)		X		
ChIP-seq (TCF7L2)			X	
ChIP-seq (TEAD4)	X			
ChIP-seq (TFAP2A)			X	
ChIP-seq (TFAP2C)			X	

Table S2: Datasets available for each cell line, driven primarily by availability from ENCODE.

	Cell Line			
	K562	GM12878	HeLa-S3	HUVEC
ChIP-seq (THAP1)	X			
ChIP-seq (TRIM28)	X			
ChIP-seq (UBTF)	X			
ChIP-seq (USF1)	X	X		
ChIP-seq (USF2)	X	X	X	
ChIP-seq (WHSC1)	X			
ChIP-seq (WRNIP1)		X		
ChIP-seq (XRCC4)	X			
ChIP-seq (YY1)	X	X		
ChIP-seq (ZBTB33)	X	X		
ChIP-seq (ZBTB7A)	X			
ChIP-seq (ZC3H11A)	X			
ChIP-seq (ZEB1)		X		
ChIP-seq (ZKSCAN1)			X	
ChIP-seq (ZMIZ1)	X			
ChIP-seq (ZNF143)	X	X	X	
ChIP-seq (ZNF263)	X			
ChIP-seq (ZNF274)	X	X	X	
ChIP-seq (ZNF384)	X	X		
ChIP-seq (ZZZ3)		X	X	
DNase-seq	X	X	X	X
FAIRE-seq	X	X	X	X
Methyl-RRBS	X	X	X	X
GeneMANIA	X	X	X	X
IMP	X	X	X	X
STRING	X	X	X	X
Synteny	X	X	X	X

	Multivariate Rank	Univariate Rank	Absolute Rank Difference
RAD21 (window)	1	1	0
SMC3 (promoter)	2	11	9
H4K20me1 (window)	3	3	0
CTCF (window)	4	2	2
H3K36me3 (window)	5	4	0
JUNB (window)	6	22	16
SP2 (window)	7	6	1
PHF8 (promoter)	8	29	21
RAD21 (promoter)	9	20	11
SAP30 (window)	10	17	7
SMC3 (enhancer)	11	113	102
MEF2A (window)	12	40	28
THAP1 (window)	13	33	20
CTCFL (window)	14	8	6
SPI1 (window)	15	63	48
JUN (window)	16	13	3
SIX5 (window)	17	7	10
NCOR1 (window)	18	128	110
DNase-seq (promoter)	19	47	28
SRF (window)	20	140	120
TBP (promoter)	21	250	229
H3K4me2 (promoter)	22	105	83
JUND (window)	23	70	47
GTF2F1 (window)	24	42	18
H2AZ (window)	25	165	140
MAZ (promoter)	26	162	136
ZNF143 (window)	27	10	17
HCFC1 (promoter)	28	26	2
TEAD4 (window)	29	14	15
CHD4 (window)	30	154	124
FAIRE-seq (window)	31	49	18
H3K79me2 (window)	32	61	29

Table S3: Univariate and multivariate feature ranks for the top 32 predictors in K562. Large differences between multivariate and univariate ranks indicate features that are not predictive on their own but become highly predictive in combination with other features. Such outliers may identify novel biological interactions or resolve ambiguities caused by noisy assays.

	Multivariate Rank	Univariate Rank	Absolute Rank Difference
ZNF143 (window)	1	1	0
CTCF (window)	2	2	0
SMC3 (promoter)	3	8	5
RAD21 (window)	4	3	1
H3K9ac (enhancer)	5	29	24
H3K4me3 (promoter)	6	19	13
H2AZ (promoter)	7	44	37
FOXM1 (enhancer)	8	82	74
RAD21 (promoter)	9	6	3
H3K27ac (enhancer)	10	46	36
H3K9ac (promoter)	11	57	46
H3K9me3 (window)	12	27	15
H4K20me1 (window)	13	53	40
CDS1 (promoter)	14	13	1
NFKB1 (window)	15	42	27
RFX5 (window)	16	37	21
H3K4me3 (window)	17	126	109
H3K79me2 (window)	18	60	42
ATF2 (promoter)	19	12	7
DNase-seq (promoter)	20	62	42
BATF (window)	21	28	7
NFATC1 (window)	22	69	47
H3K4me1 (window)	23	92	69
H3K9ac (window)	24	221	197
NRF1 (window)	25	5	20
H3K36me3 (window)	26	59	33
H3K4me2 (promoter)	27	39	12
H3K27me3 (window)	28	16	12
EBF1 (window)	29	158	129
H3K79me2 (promoter)	30	98	68
H2AZ (enhancer)	31	111	80
REST (window)	32	55	23

Table S4: Univariate and multivariate feature ranks for the top 32 predictors in GM12878. Large differences between multivariate and univariate ranks indicate features that are not predictive on their own but become highly predictive in combination with other features. Such outliers may identify novel biological interactions or resolve ambiguities caused by noisy assays.

	Multivariate Rank	Univariate Rank	Absolute Rank Difference
HCFC1 (window)	1	1	0
CTCF (window)	2	2	0
RAD21 (window)	3	28	25
Methylation (window)	4	3	1
H3K36me3 (window)	5	7	2
SMC3 (promoter)	6	14	8
H3K79me2 (window)	7	51	44
H4K20me1 (window)	8	15	7
TBP (window)	9	99	90
RAD21 (promoter)	10	8	2
SMARCC1 (window)	11	59	48
CEBPB (window)	12	9	3
REST (window)	13	26	13
DNase-seq (window)	14	49	35
TCF7L2 (window)	15	41	26
MAFK (window)	16	75	59
H3K4me1 (promoter)	17	56	39
H3K4me2 (promoter)	18	78	60
JUN (window)	19	5	14
H2AZ (window)	20	146	126
NR2C2 (window)	21	37	16
H3K27ac (promoter)	22	90	68
H3K4me1 (window)	23	68	45
JUND (window)	24	12	12
H3K27ac (window)	25	87	62
Methylation (promoter)	26	67	41
CTCF (enhancer)	27	73	46
DNase-seq (promoter)	28	72	44
NFYB (window)	29	19	10
RCOR1 (window)	30	88	58
ZNF143 (window)	31	11	20
STAT1 (window)	32	145	113

Table S5: Univariate and multivariate feature ranks for the top 32 predictors in HeLa-S3. Large differences between multivariate and univariate ranks indicate features that are not predictive on their own but become highly predictive in combination with other features. Such outliers may identify novel biological interactions or resolve ambiguities caused by noisy assays.

	Multivariate Rank	Univariate Rank	Absolute Rank Difference
CTCF (window)	1	1	0
RAD21 (promoter)	2	4	2
H3K36me3 (window)	3	8	5
H3K4me2 (promoter)	4	32	28
EP300 (window)	5	13	8
ZNF143 (window)	6	2	4
SMC3 (promoter)	7	6	1
RAD21 (window)	8	3	5
H3K79me2 (window)	9	26	17
H4K20me1 (window)	10	23	13
JUND (window)	11	42	31
H3K9ac (promoter)	12	45	33
DNase-seq (promoter)	13	34	21
H3K9me3 (window)	14	21	7
H3K4me3 (promoter)	15	36	21
DNase-seq (window)	16	22	6
H2AZ (promoter)	17	24	7
H2AZ (window)	18	89	71
H3K9ac (enhancer)	19	37	18
REST (window)	20	30	10
H3K4me1 (window)	21	49	28
H3K27ac (promoter)	22	100	78
H3K4me3 (window)	23	62	39
H3K4me1 (promoter)	24	41	17
FAIRE-seq (window)	25	80	55
Methylation (window)	26	5	21
TBP (promoter)	27	38	11
H3K9ac (window)	28	81	53
Methylation (promoter)	29	67	38
H3K36me3 (promoter)	30	51	21
H3K27ac (window)	31	88	57
POLR2A (promoter)	32	91	59

Table S6: Univariate and multivariate feature ranks for the top 32 predictors across all cell lines. Large differences between multivariate and univariate ranks indicate features that are not predictive on their own but become highly predictive in combination with other features. Such outliers may identify novel biological interactions or resolve ambiguities caused by noisy assays.

Supplemental Figures

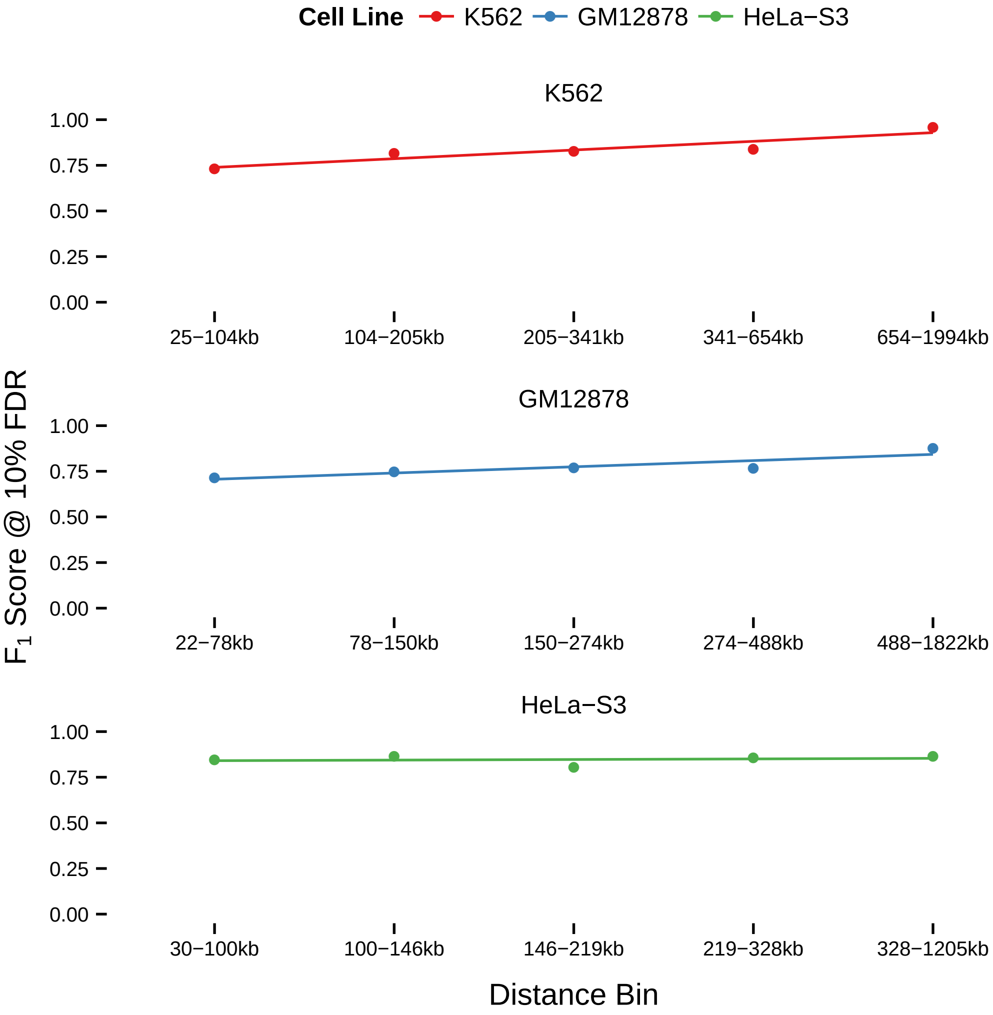


Figure S1: Performance as a function of distance between enhancer and promoter. Samples are first grouped into equal-count bins via distance quantiles, then predictions within each bin are evaluated at 10% FDR.

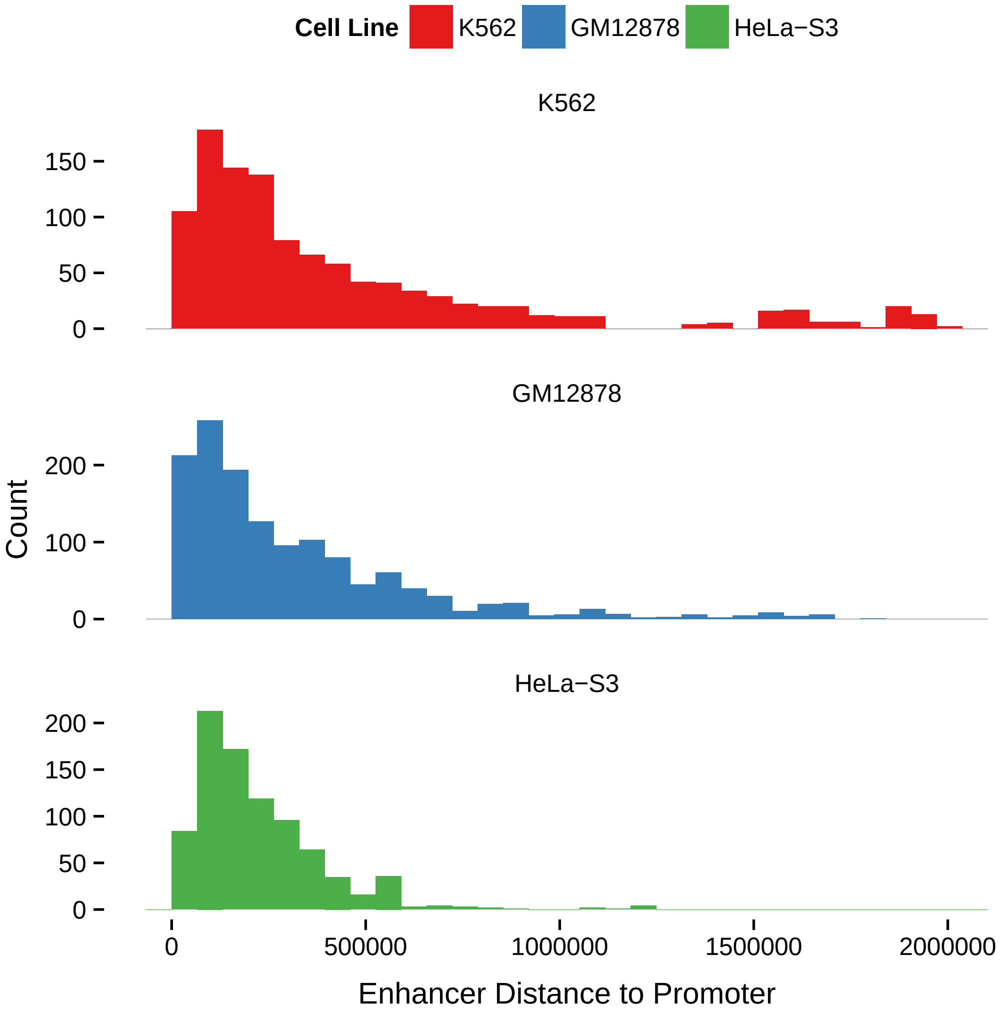


Figure S2: Distribution of genomic distance between enhancers and promoters in the training data for each cell line.

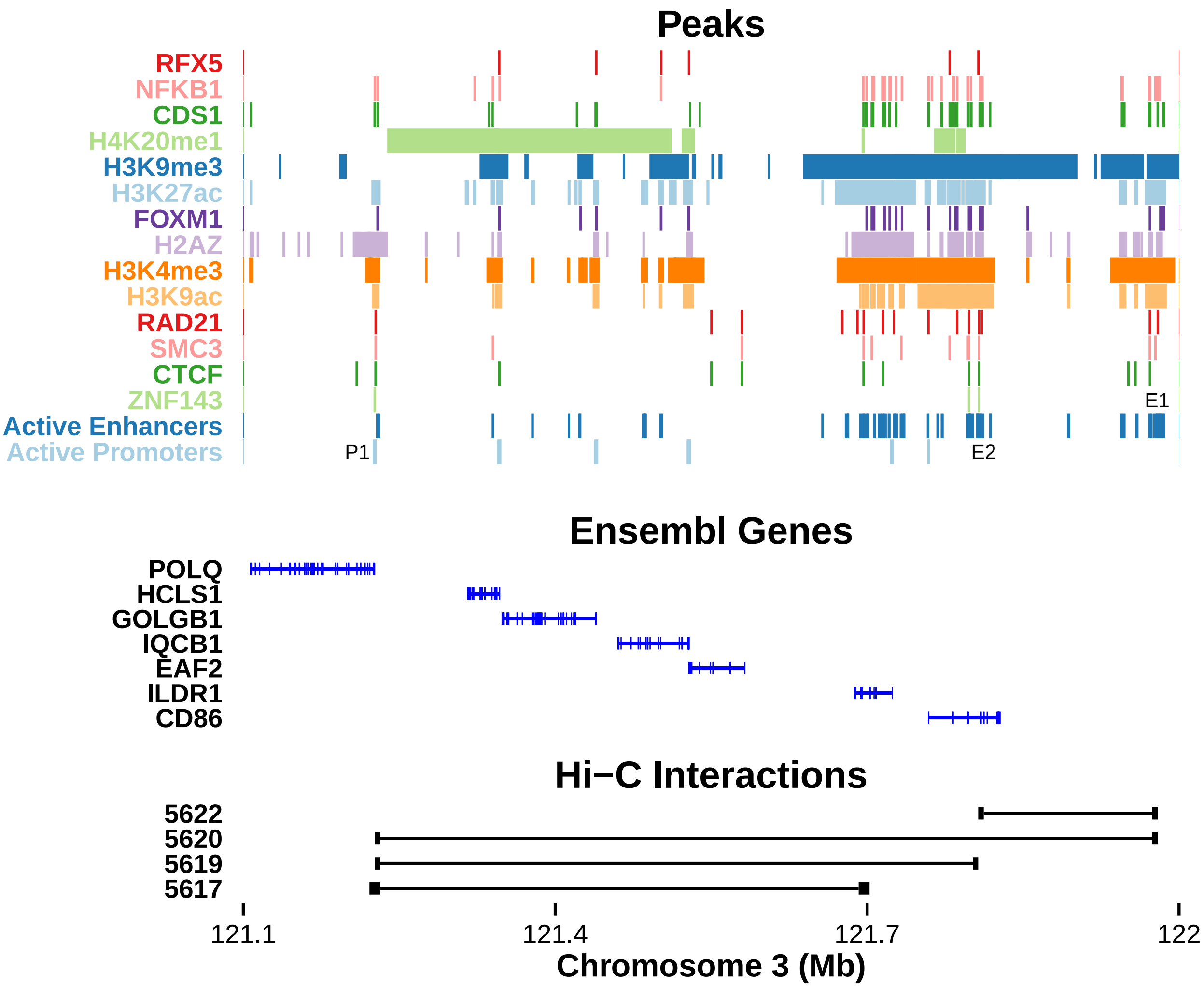


Figure S3: Significant peak values of the top 14 predictive datasets for an interacting promoter (P1) and enhancer (E1) in GM12878, separated by other active promoters and enhancers. In contrast to Figure 6, enhancer E1 interacts not only with P1 but also with E2, and P1 is the target of multiple enhancers. Active enhancers are segments marked “E” by combined ChromHMM/Segway annotations, and active promoters are segments marked “TSS” and expressed in GM12878 (determined by RNA-seq with expression threshold 0.3). Ensembl genes are also displayed, with introns denoted as thin lines and exons as squares. Left and right fragments of the Hi-C assay are also shown to visually confirm E1 interacts with P1 and other targets in the window. Note that P1 has all expected loop-associated marks including CTCF, cohesin, and ZNF143, as well as all other activation-associated marks. Also note spans of the repressive H4K20me1 and H3K9me3 marks that may rule out several alternate targets. As in the K562 example, the presence or absence of a single mark does not rule out a potential target and should instead be considered in combination with other marks.

References

1. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Research* **22,** 1748–59 (Sept. 2012).
2. Lomelin, D., Jorgenson, E. & Risch, N. Human genetic variation recognizes functional elements in noncoding sequence. *Genome Research* **20,** 311–9 (Mar. 2010).
3. Alexandrov, N. N. *et al.* Features of Arabidopsis genes and genome discovered using full-length cDNAs. *Plant Molecular Biology* **60,** 69–85 (Jan. 2006).
4. Hillier, L. W. *et al.* Whole-genome sequencing and variant discovery in C. elegans. *Nature Methods* **5,** 183–8 (Feb. 2008).
5. Massouras, A. *et al.* Genomic variation and its impact on gene expression in Drosophila melanogaster. *PLOS Genetics* **8,** e1003055 (Jan. 2012).
6. Tang, R. *et al.* Candidate genes and functional noncoding variants identified in a canine model of obsessive-compulsive disorder. *Genome Biology* **15,** R25 (Jan. 2014).
7. Manolio, T. A., Brooks, L. D. & Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *Journal of Clinical Investigation* **118,** 1590–605 (May 2008).
8. Gusev, A. *et al.* Regulatory variants explain much more heritability than coding variants across 11 common diseases. *American Journal of Human Genetics* **95,** 535–552 (2014).
9. Frazer, K. A., Murray, S. S., Schork, N. J. & Topol, E. J. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics* **10,** 241–51 (Apr. 2009).
10. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–82 (Oct. 2011).
11. Bernstein, B. E. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (Sept. 2012).
12. Celniker, S. E. *et al.* Unlocking the secrets of the genome. *Nature* **459,** 927–30 (June 2009).
13. Bernstein, B. E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology* **28,** 1045–8 (Oct. 2010).
14. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 (2011).
15. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research* **22,** 1790–7 (Sept. 2012).
16. Ward, L. D. & Kellis, M. HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research* **40,** D930–4 (Jan. 2012).
17. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* **46,** 310–5 (Mar. 2014).
18. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nature Genetics* **47,** 276–283 (Jan. 2015).
19. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nature Biotechnology* **30,** 271–7 (Mar. 2012).
20. Lettice, L. A. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics* **12,** 1725–1735 (July 2003).
21. Sanyal, A., Lajoie, B. R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489,** 109–113 (2012).
22. Kvon, E. Z. *et al.* Genome-scale functional characterization of Drosophila developmental enhancers in vivo. en. *Nature* **512,** 91–95 (2014).

23. Wang, D., Rendon, A. & Wernisch, L. Transcription factor and chromatin features predict genes associated with eQTLs. *Nucleic Acids Research* **41,** 1450–1463 (2013).

24. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biology* **13,** R48 (2012).

25. Aran, D., Sabato, S. & Hellman, A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biology* **14,** R21 (2013).

26. Rödelsperger, C. *et al.* Integrative analysis of genomic, functional and protein interaction data predicts long-range enhancer-target gene interactions. *Nucleic Acids Research* **39,** 2492–2502 (2011).

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

28. Wilczynski, B., Liu, Y.-H., Yeo, Z. X. & Furlong, E. E. M. Predicting spatial and temporal gene expression using an integrative model of transcription factor occupancy and chromatin state. *PLOS Computational Biology* **8,** e1002798 (2012).

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

30. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing Chromosome Conformation. *Science* **295,** 1306–1311 (2002).

31. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research* **16,** 1299–1309 (2006).

32. De Wit, E. & de Laat, W. A decade of 3C technologies: Insights into nuclear organization. *Genes & Development* **26,** 11–24 (Jan. 2012).

33. Rao, S. S. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159,** 1665–80 (Dec. 2014).

34. Dixon, J. R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518,** 331–336 (Feb. 2015).

35. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research* (Mar. 2015).

36. Maston, G. A., Evans, S. K. & Green, M. R. Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics* **7,** 29–59 (Sept. 2006).

37. Zhang, Y. *et al.* Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* **504,** 306–10 (Dec. 2013).

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

39. Shaulian, E. & Karin, M. AP-1 as a regulator of cell life and death. *Nature Cell Biology* **4,** E131–6 (May 2002).

40. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469,** 343–9 (Jan. 2011).

41. Bailey, S. D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nature Communications* **2,** 6186 (Jan. 2015).

42. Michaud, J. *et al.* HCFC1 is a common component of active human CpG-island promoters and coincides with ZNF143, THAP11, YY1, and GABP transcription factor occupancy. *Genome Research* **23,** 907–16 (June 2013).

43. Chopra, V. S., Cande, J., Hong, J.-W. & Levine, M. Stalled Hox promoters as chromosomal boundaries. *Genes & Development* **23,** 1505–9 (July 2009).

44. Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: Emerging roles in metazoans. *Nature Reviews Genetics* **13,** 720–31 (Oct. 2012).

45. Doyle, B., Fudenberg, G., Imakaev, M. & Mirny, L. A. Chromatin Loops as Allosteric Modulators of Enhancer-Promoter Interactions. *PLOS Computational Biology* **10,** e1003867 (Oct. 2014).

46. Benveniste, D., Sonntag, H.-J., Sanguinetti, G. & Sproul, D. Transcription factor binding predicts histone modifications in human cell lines. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 13367–72 (Sept. 2014).

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

48. Schwartz, C. *et al.* Recruitment of p300 by C/EBPbeta triggers phosphorylation of p300 and modulates coactivator activity. *The EMBO Journal* **22,** 882–92 (Feb. 2003).

49. Listman, J. A. *et al.* Conserved ETS domain arginines mediate DNA binding, nuclear localization, and a novel mode of bZIP interaction. *The Journal of Biological Chemistry* **280,** 41421–8 (Dec. 2005).

50. Van Riel, B. & Rosenbauer, F. Epigenetic control of hematopoiesis: the PU.1 chromatin connection. *Biological Chemistry* **395,** 1265–74 (Nov. 2014).

51. Zhou, W. *et al.* Histone H2A monoubiquitination represses transcription by inhibiting RNA polymerase II transcriptional elongation. *Molecular Cell* **29,** 69–80 (Jan. 2008).

52. Bertolino, E. & Singh, H. POU/TBP cooperativity: a mechanism for enhancer action from a distance. *Molecular Cell* **10,** 397–407 (Aug. 2002).

53. Liu, Z., Scannell, D. R., Eisen, M. B. & Tjian, R. Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell* **146,** 720–31 (Sept. 2011).

54. Fujioka, S. *et al.* NF-kappaB and AP-1 connection: mechanism of NF-kappaB-dependent regulation of AP-1 activity. *Molecular and Cellular Biology* **24,** 7806–19 (Sept. 2004).

55. Hanlon, M & Sealy, L. Ras regulates the association of serum response factor and CCAAT/enhancer-binding protein beta. *The Journal of Biological Chemistry* **274,** 14224–8 (May 1999).

56. Love, J. J. *et al.* Structural basis for DNA bending by the architectural transcription factor LEF-1. *Nature* **376,** 791–5 (Aug. 1995).

57. Euskirchen, G. M. *et al.* Diverse roles and interactions of the SWI/SNF chromatin remodeling complex revealed using global approaches. *PLOS Genetics* **7,** e1002008 (Mar. 2011).

58. Palstra, R.-J. T. S. Close encounters of the 3C kind: long-range chromatin interactions and transcriptional regulation. *Briefings in Functional Genomics & Proteomics* **8,** 297–309 (July 2009).

59. Christova, R. *et al.* P-STAT1 mediates higher-order chromatin remodelling of the human MHC in response to IFNgamma. *Journal of Cell Science* **120,** 3262–70 (Sept. 2007).

60. Lin, Y. C. *et al.* A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. *Nature Immunology* **11,** 635–43 (July 2010).

61. Cockerill, P. N. NFAT is well placed to direct both enhancer looping and domain-wide models of enhancer function. *Science Signaling* **1,** pe15 (Jan. 2008).

62. Wang, H. *et al.* Role of histone H2A ubiquitination in Polycomb silencing. *Nature* **431,** 873–8 (Oct. 2004).

63. Majumder, P. & Boss, J. M. Cohesin regulates MHC class II genes through interactions with MHC class II insulators. *Journal of Immunology* **187,** 4236–44 (Oct. 2011).

64. Sims, J. K., Houston, S. I., Magazinnik, T. & Rice, J. C. A trans-tail histone code defined by monomethylated H4 Lys-20 and H3 Lys-9 demarcates distinct regions of silent chromatin. *The Journal of Biological Chemistry* **281,** 12760–6 (May 2006).

65. Liu, W. *et al.* PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression. *Nature* **466,** 508–12 (July 2010).

66. Wang, Z. *et al.* Genome-wide mapping of HATs and HDACs reveals distinct functions in active and inactive genes. *Cell* **138,** 1019–31 (Sept. 2009).

67. Lee, J.-S., Smith, E. & Shilatifard, A. The language of histone crosstalk. *Cell* **142,** 682–5 (Sept. 2010).

68. Blackwood, E. M. & Kadonaga, J. T. Going the Distance: A Current View of Enhancer Action. *Science* **281,** 60–63 (July 1998).

69. Islam, A. B., Richter, W. F., Lopez-Bigas, N. & Benevolenskaya, E. V. Selective targeting of histone methylation. *Cell Cycle* **10,** 413–424 (2011).

70. Dorsett, D. & Kassis, J. A. Checks and balances between cohesin and polycomb in gene silencing and transcription. *Current Biology* **24,** R535–9 (June 2014).

71. Levine, S. S. *et al.* The Core of the Polycomb Repressive Complex Is Compositionally and Functionally Conserved in Flies and Humans. *Molecular and Cellular Biology* **22,** 6070–8 (Sept. 2002).

72. Vernimmen, D. *et al.* Polycomb eviction as a new distant enhancer function. *Genes & Development* **25,** 1583–8 (Aug. 2011).

73. Ing-Simmons, E. *et al.* Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. *Genome Research* **25,** gr.184986.114 (Feb. 2015).

74. Kerppola, T. K. & Curran, T. Fos-Jun heterodimers and Jun homodimers bend DNA in opposite orientations: Implications for transcription factor cooperativity. *Cell* **66,** 317–326 (July 1991).

75. Chinenov, Y & Kerppola, T. K. Close encounters of many kinds: Fos-Jun interactions that mediate transcription regulatory specificity. *Oncogene* **20,** 2438–52 (Apr. 2001).

76. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

77. McKinney, W. *Python for Data Analysis* (O'Reilly, Sebastopol, CA, 2012).

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

79. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28,** 1–26 (2008).

80. Law, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2,** 18–22 (2002).

81. Ridgeway, G. *Generalized boosted models: A guide to the gbm package* 2005.

82. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33,** 1–22 (2010).

83. Harrow, J. *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Research* **22,** 1760–74 (2012).

84. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLOS Computational Biology* **5,** e1000598 (Dec. 2009).

85. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics* **5,** 1752–1779 (Sept. 2011).

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

87. Hoffman, M. M. *et al.* Unsupervised Pattern Discovery in Human Chromatin Structure Through Genomic Segmentation. *Nature Methods* **9,** 473–6 (2012).

88. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326,** 289–93 (Oct. 2009).

89. Franceschini, A. *et al.* STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research* **41,** D808–15 (Jan. 2013).

90. Wong, A. K. *et al.* IMP: A multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Research* **40,** W484–90 (July 2012).

91. Zuberi, K. *et al.* GeneMANIA prediction server 2013 update. *Nucleic Acids Research* **41,** W115–22 (July 2013).

92. Pique-Regi, R. *et al.* Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research* **21,** 447–55 (Mar. 2011).

93. Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 11484–9 (Sept. 2003).

94. Yang, P., Yang, Y. H., Zhou, B. B. & Zomaya, A. Y. A Review of Ensemble Methods in Bioinformatics. *Current Bioinformatics* **5,** 296–308 (2010).

95. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32 (2001).

96. Friedman, J. H. Stochastic Gradient Boosting. *Computational Statistics & Data Analysis* **38,** 367–378 (2002).

97. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **46,** 389–422 (2002).

98. Tuv, E., Borisov, A., Runger, G. & Torkkola, K. Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination. *Journal of Machine Learning Research* **10,** 1341–1366 (2009).

99. Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P. & Saeys, Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics* **26,** 392–398 (2010).

100. Ambroise, C. & McLachlan, G. J. Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 6562–6566 (2002).

101. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2015 update. *Nucleic Acids Research* **43,** D470–8 (Jan. 2015).

102. Yan, J. *et al.* Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154,** 801–13 (Aug. 2013).

103. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. *A density-based algorithm for discovering clusters in large spatial databases with noise* in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (1996), 226–231.