nature genetics

Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin

Sean Whalen[1–3], Rebecca M Truty[4] & Katherine S Pollard[1–3]

Discriminating the gene target of a distal regulatory element from other nearby transcribed genes is a challenging problem with the potential to illuminate the causal underpinnings of complex diseases. We present TargetFinder, a computational method that reconstructs regulatory landscapes from diverse features along the genome. The resulting models accurately predict individual enhancer–promoter interactions across multiple cell lines with a false discovery rate up to 15 times smaller than that obtained using the closest gene. By evaluating the genomic features driving this accuracy, we uncover interactions between structural proteins, transcription factors, epigenetic modifications, and transcription that together distinguish interacting from non-interacting enhancer–promoter pairs. Most of this signature is not proximal to the enhancers and promoters but instead decorates the looping DNA. We conclude that complex but consistent combinations of marks on the one-dimensional genome encode the three-dimensional structure of fine-scale regulatory interactions.

Genotyping, exome sequencing, and whole-genome sequencing have linked thousands of noncoding variants to traits in humans and other eukaryotes[1–6]. Noncoding variants are more likely to cause common disease than are nonsynonymous coding variants[7], and they can account for the vast majority of heritability[8]. Yet few noncoding mutations have been functionally characterized or mechanistically linked to human phenotypes[7,9]. Comparative[10] and functional[11–13] genomics, together 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]. However, this new information will only improve understanding of disease and other phenotypes if functional noncoding elements can be accurately linked to the genes, pathways, and cellular processes they regulate. This is a difficult problem because vertebrate promoters and their regulatory elements can be separated by thousands or millions of base pairs[19]. The closest promoter is usually not the true target of enhancers in humans[20], although this varies by species[21], but remains a common heuristic for mapping target genes. Incorrectly mapping regulatory variants to genes prevents meaningful downstream studies.

Until recently, very few validated distal regulatory interactions were known. Hence, previous studies defined interactions indirectly via genomic proximity coupled with genetic associations (for example, expression quantitative trait loci (eQTLs)[22]), gene expression[14,23–25], or promoter chromatin state[26,27]. High-throughput methods for assaying chromatin interactions now exist, including paired-end tag sequencing (ChIA-PET)[28] and extensions of the chromosome conformation capture (3C) assay[29] (5C and Hi-C)[30,31]. When resolution is high enough to measure individual enhancer–promoter interactions[32–35], Hi-C provides an opportunity to examine the genomic features that distinguish the true target of an enhancer from other nearby expressed genes. We hypothesized that modeling relationships between DNA sequences, structural proteins, transcription factors, and epigenetic modifications that together predict looping chromatin might identify new protein functions and molecular mechanisms of distal gene regulation that are not immediately obvious from the Hi-C data itself.

We implemented an algorithm called TargetFinder that integrates hundreds of genomics data sets to identify the minimal subset of features necessary to accurately predict individual enhancer–promoter interactions across the genome. We focused on enhancers because of their large impact on gene regulation[36] and our ability to predict their locations across the genome, although our approach works with other classes of regulatory elements. Our goal was to build a fine-scale model capable of distinguishing individual enhancer–promoter pairs from among the many possible interactions within a topologically associating domain (TAD) or contact domain. Applying TargetFinder to six human Encyclopedia of DNA Elements (ENCODE) cell lines[11] with high-resolution Hi-C data[32], we discovered that interacting enhancer–promoter pairs can be distinguished from non-interacting pairs within the same locus with extremely high accuracy. These analyses also 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 identified specific proteins and chemical modifications on the chromatin loop that bring an enhancer in contact with its target promoter and not with nearby 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

[1]Gladstone Institutes, San Francisco, California, USA. [2]Division of Biostatistics, Institute for Human Genetics, University of California, San Francisco, San Francisco, California, USA. [3]Institute for Computational Health Sciences, University of California, San Francisco, San Francisco, California, USA. [4]Invitae Corporation, San Francisco, California, USA. Correspondence should be addressed to S.W. (sean.whalen@gladstone.ucsf.edu) or K.S.P. (kpollard@gladstone.ucsf.edu).

Received 17 December 2015; accepted 7 March 2016; published online 4 April 2016; doi:10.1038/ng.3539

experimental data to shed new light on the mechanisms of distal gene regulation.

RESULTS

Annotating the genomic features of regulatory interactions

We annotated enhancer–promoter interactions in six human ENCODE cell lines that have rich functional genomics data as well as high-resolution interaction data generated by Rao *et al.*[32]: K562 (mesoderm-lineage cells from a patient with leukemia), GM12878 (lymphoblastoid cells), HeLa-S3 (ectoderm-lineage cells from a patient with cervical cancer), HUVEC (umbilical vein endothelial cells), IMR90 (fetal lung fibroblasts), and NHEK (epidermal keratinocytes). We identified active promoters and enhancers in each cell line using segmentation-based annotations from ENCODE and Roadmap Epigenomics, as well as gene expression data from ENCODE (**Supplementary Table 1**). Enhancers are typically a few hundred base pairs long, whereas promoters are mostly 1–2 kb (**Supplementary Figs. 1**–**3**). Alternative enhancer and promoter definitions gave qualitatively similar results (**Supplementary Note**).

We annotated all enhancer–promoter pairs as interacting or non-interacting using high-resolution genome-wide measurements of chromatin contacts in each cell line[32], the majority of which were also detected by capture Hi-C[35]. Non-interacting pairs were sampled (20 per interacting pair) to have enhancer–promoter distances similar to those of interacting pairs, all of which were less than 2 Mb. To focus on distal regulatory enhancers, any promoter–enhancer pair separated by less than 10 kb was dropped. We did not remove interactions crossing TAD boundaries, but most enhancer–promoter pairs occurred within the same TAD (88% in GM12878 cells and 77% in K562 cells; ref. 37). It is important to emphasize that, by design, all enhancers and promoters in our study, including those in non-interacting pairs, had marks of activation and open chromatin. The challenging question we address is whether interacting pairs have any distinguishing characteristics.

We generated lists of features for all enhancer–promoter pairs in each cell line using functional genomics data such as measures of open chromatin, DNA methylation, gene expression, and ChIP-seq peaks for transcription factors, architectural proteins, and modified histones (**Supplementary Table 2**). We quantified signal at the promoter, at the enhancer, and in the genomic window between them. We also computed features for conserved synteny of the enhancer and promoter, as well as the similarity of transcription factor and target gene annotations, which are associated with experimentally validated interactions[25].

Finally, we created a 'combined' data set by pooling the enhancer–promoter pairs and features from four cell lines (K562, GM12878, HeLa-S3, and IMR90), which we used to discover features of looping chromatin that generalize across lines. Only features measured in all four lines were retained to avoid problems with missing data. The NHEK and HUVEC lines had only ~20 data sets each (versus >50 for the other cell lines; **Supplementary Table 2**) and were therefore excluded from the combined data set.

No single feature distinguishes true enhancer targets

The signal profiles at enhancers and promoters showed many expected differences between interacting and non-interacting pairs (**Fig. 1**). These included higher RNA polymerase II (Pol II) signal at the transcription start site (TSS) of interacting promoters (**Fig. 1a**) and enrichment of acetylation of histone H3 at lysine 27 (H3K27ac) and trimethylation of histone H3 at lysine 4 (H3K4me3) with depletion of monomethylation of histone H3 at lysine 4 (H3K4me1) in regions flanking the TSS of interacting promoters (**Fig. 1b**–**d**). Across cell types, CTCF and RAD21 were enriched near interacting promoters (**Fig. 1e**,**f**). Structural proteins and their cofactors were also enriched near interacting enhancers (**Fig. 2**).

However, any given interaction had a complex combination of genomic features, some of which also occurred at non-interacting pairs in the same locus. For example, *LPIN3* had an enhancer that looped over approximately 400 kb of intervening DNA containing the active promoters of *TOP1*, *PLCG1*, and *ZHX3* in K562 cells (**Fig. 3**). No single mark distinguished *LPIN3* from these alternate targets, although their gene bodies were covered by broad repressive marks (heterochromatin-associated monomethylation of histone H4 at lysine 20 (H4K20me1)) and by broad activating marks (elongation-associated trimethylation of histone H3 at lysine 36 (H3K36me3)). Notably, the alternate promoters lacked binding of RAD21, whereas *ZHX3* and *PLCG1* lacked binding of CUX1, which has been linked to both activation and repression. In GM12878 cells, an intronic

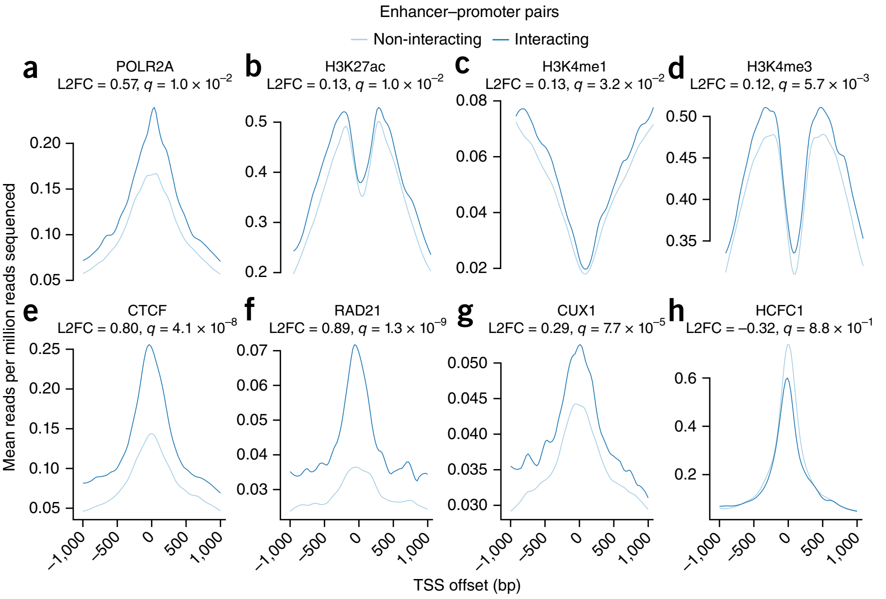


Figure 1 Predictive power of promoter-proximal genomic features. (**a**–**h**) Ratio of various ChIP-seq signals, including Pol II (POLR2A) (**a**), enhancer- and promoter-associated histone modifications (**b**–**d**), known looping factors (**e**,**f**), and selected transcription factors (**g**,**h**), anchored at the TSS of interacting versus non-interacting promoters in K562 cells, along with the $\log_2$-transformed fold change (L2FC) and *P* value corrected for multiple testing (*q* value). All promoters have active chromatin marks and show transcription. The top row shows expected patterns for promoter-associated marks at the TSS, such as a high ratio of H3K4me3 to H3K4me1. Some of these marks are enriched in interacting promoters, whereas others such as lysine 4 methylation patterns are not. The bottom row shows TSS-proximal patterns for several proteins associated with chromatin looping. CTCF and RAD21 are enriched at interacting promoters, whereas the transcription factors CUX1 and HCFC1 are enriched and depleted, respectively.

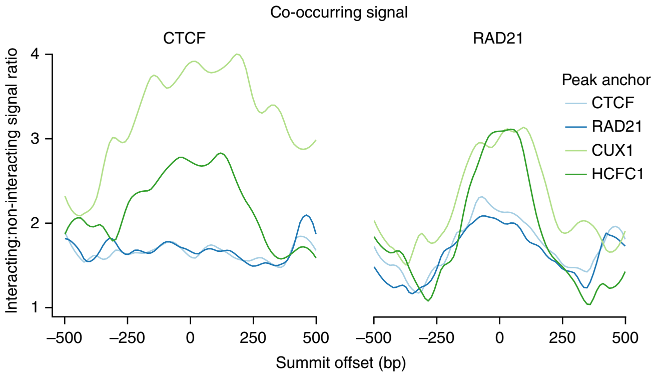


Figure 2 Ratio of the CTCF and RAD21 ChIP-seq signals occurring within interacting enhancers and non-interacting enhancers, anchored at peaks for CTCF, RAD21, and the transcription factors CUX1 and HCFC1 for the K562 cell line. CUX1 and HCFC1 are highly enriched at loop-associated enhancers when co-occurring with CTCF and RAD21. The context dependence of protein binding is demonstrated by RAD21, which is not enriched at interacting promoters (**Fig. 1**). Note that CTCF and RAD21 are already enriched at their respective peaks within interacting enhancers but are further enriched when anchored at CUX1 or HCFC1 peaks.

enhancer targeting *CUTC* looped over the promoter of *ENTPD7*, which had many activation marks but lacked RAD21 (**Supplementary Fig. 4**). This complexity motivated us to model enhancer–promoter interactions as a function of diverse genomic signatures.

Ensemble learning predicts enhancer–promoter pairs with high accuracy

To quantitatively model the interaction status of enhancer–promoter pairs as a function of their genomic features, we built a machine learning pipeline called TargetFinder (**Fig. 4**). The inputs are pairs of enhancers and promoters, annotated as interacting or non-interacting, and genomic features associated with each pair. The algorithm finds an optimal combination of features to distinguish interacting from non-interacting pairs. Multiple machine learning techniques are implemented in the pipeline in a modular way so that performance can be optimized and conclusions can be tested for robustness to the prediction method. The outputs are a model for predicting whether new enhancer–promoter pairs interact, assessments of model performance on held-out data, and estimates of each feature's individual importance to the model as well as its importance in combination with other features. The predictive contribution of different genomic regions and data types is explored by varying the feature set and quantifying predictive performance. By building models for many cell types, their shared and unique characteristics of looping chromatin can be discovered. The method is easily extended to other types of regulatory elements or interactions, such as promoter–promoter or enhancer–enhancer interactions.

We hypothesized that ensemble learning algorithms would have the highest precision and recall on held-out data because they are robust to overfitting and account for nonlinear feature interactions that could encode complex patterns of histone modifications and transcription factor binding. In particular, a technique called boosting is used to iteratively train models that place increasing emphasis on misclassified samples. Indeed, ensembles of boosted decision trees performed better than other methods and a random guessing null model on all cell lines and the combined data set (**Fig. 5** and **Supplementary Table 3**). Accuracy was high by all measures, especially given the noise in functional genomics data and the fact that some non-interacting pairs might be weakly interacting but fall below the significance cutoff (false discovery rate (FDR) = 10%; ref. 32). TargetFinder with boosted trees achieved a balance of precision and recall (F_1) of 77–90% (mean = 83%) and an FDR of 8–15% (mean = 12%). By comparison, all commonly used bioinformatics methods had much higher FDR values and lower recall. For example, using the closest actively transcribed gene results in an FDR of 53–77% (refs. 20,38,39). The gain in predictive accuracy provided by ensemble learning was consistent across cell lines and in the combined data set (**Supplementary Fig. 5**). This predictive accuracy demonstrates that there is rich information about chromatin looping in one-dimensional genomic data sets that are easier and less costly to collect than high-resolution Hi-C data.

Variable importance highlights predictive data sets

We next asked whether the ability of TargetFinder to predict enhancer–promoter interactions depends on a particular subset of the features. By omitting different categories of features and evaluating performance with cross-validation, we learned that synteny and gene annotations contribute little to predictive accuracy. We therefore proceeded to evaluate models using only functional genomics features.

To derive mechanistic insights from the model, TargetFinder estimates feature importance for each genomics data set within enhancer, promoter, and window regions (Online Methods). Decision trees inherently estimate predictive importance when deciding which features to split; importance is estimated per feature per tree and then averaged across

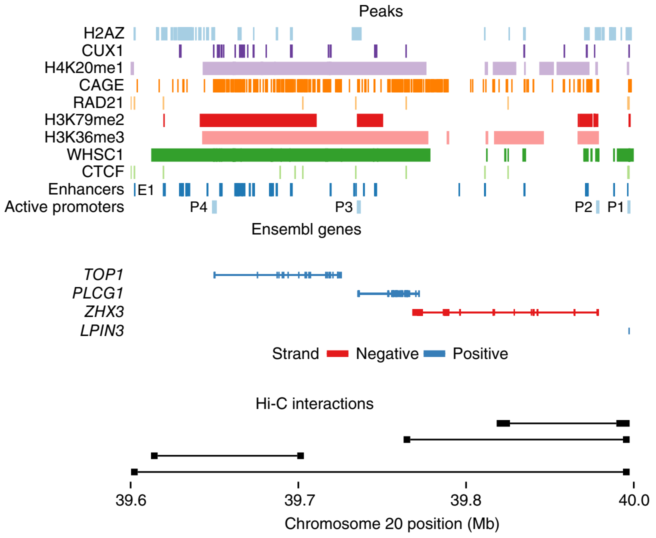
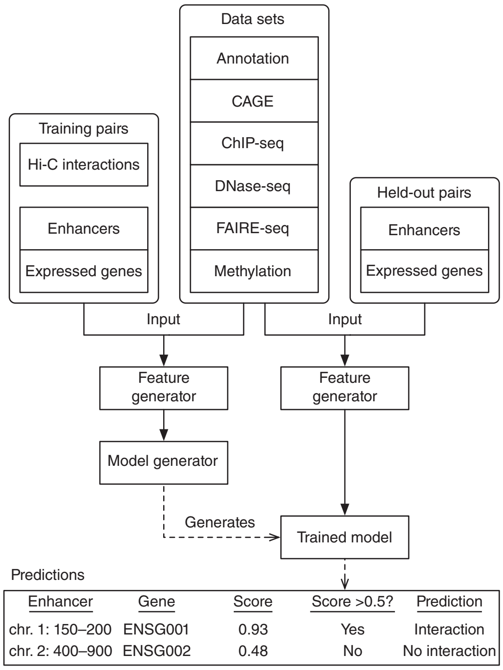


Figure 3 Predicting a chromatin loop that skips over multiple active promoters in K562 cells. ENCODE-called peaks are shown for the top nine predictive data sets of an interacting promoter (P1) and enhancer (E1) in K562 cells, separated by other active promoters and enhancers. Active enhancers are segments marked 'E' by combined ChromHMM and Segway annotations, and active promoters are segments marked 'TSS' and expressed in K562 cells with RPKM >0.3. Ensembl genes are also displayed, with introns denoted as thin lines and exons denoted as rectangles. The left and right fragments of the Hi-C assay are also shown to visually confirm that E1 interacts with P1. This figure shows a straightforward example of an enhancer (E1) looping over multiple active promoters (P2–P4) to reach its true target (P1). Existing interactions in the window between E1 and P1 do not block looping, and P1 is the target of other distal regulatory elements within the window. P2–P4 are each missing a looping-associated RAD21 mark that has elevated predictive importance in this cell line. In addition, P2 and P3 are missing the highly predictive CUX1 transcription factor (**Fig. 2**). Interpreting loops often depends on a more complex interaction of marks (**Supplementary Fig. 4**).



Enhancer	Gene	Score	Score >0.5?	Prediction
chr. 1: 150–200	ENSG001	0.93	Yes	Interaction
chr. 2: 400–900	ENSG002	0.48	No	No interaction

Figure 4 The TargetFinder pipeline. Features are generated from hundreds of diverse data sets 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 predicts whether enhancer–promoter pairs from new or held-out samples interact, as well as estimates the importance of each feature for accurate prediction. Classifier predictions are probabilities, and a decision threshold (commonly 0.5 but with the possibility of adjustment) converts these to positive or negative prediction labels. This figure excludes selection of minimal predictor sets and evaluation of the accuracy of output predictions using held-out Hi-C interaction data.

all trees in the ensemble (Online Methods). This approach enabled us to deeply explore the genomic data associated with chromatin loops and identified several interesting patterns.

The most predictive features that were robust across cell lines were DNA methylation, activation- and elongation-associated histone marks, binding of structural proteins, open chromatin, binding of proteins related to repression (MXI1, MAZ, and MAFK), and cap analysis of gene expression (CAGE) data (**Fig. 6**). Other trends emerged across many but not all cell lines, including importance of the activator protein 1 (AP-1) complex[40]. Features differ in importance across cell lines for many reasons, including real functional differences (for example, due to different co-factors), lack of expression (for example, due to tissue-specific transcription factors), and differences in laboratory protocols and antibody quality (**Supplementary Fig. 6**). Interestingly, although there was some overlap with known looping factors such as CTCF and cohesin, the features predictive of individual enhancer–promoter interactions were largely different than those used to identify TAD boundaries

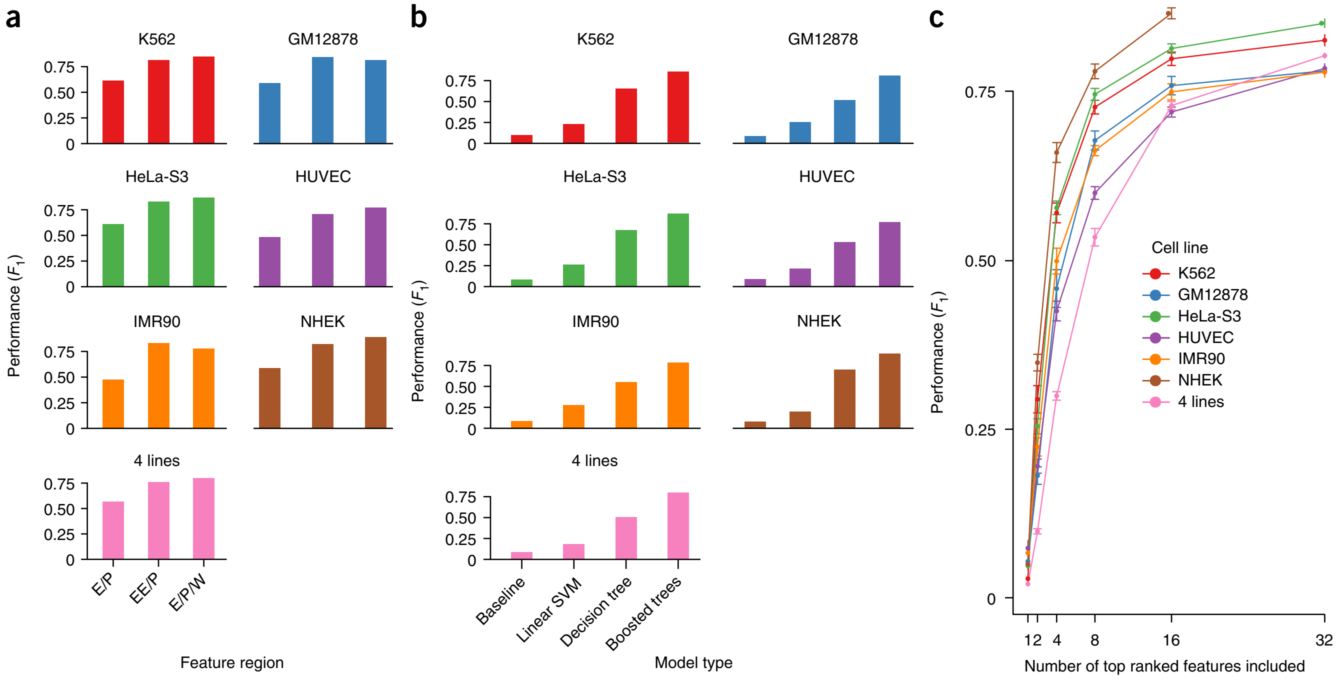


Figure 5 TargetFinder performance by cell line, model type, and number of features. (**a**) Performance of boosted trees using features for enhancers and promoters only (E/P), extended enhancers and promoters (EE/P), and enhancers and promoters plus the windows between them (E/P/W). (**b**) Cross-validated performance of TargetFinder predictions for a baseline (random guessing null) model, a linear support vector machine (SVM), a single decision tree, and a boosted ensemble of decision trees. Performance is given as the balance of precision and recall (F_1), averaging 83% across cell lines and corresponding to a mean FDR of 12%. Ensemble methods use complex interactions between features to greatly increase the accuracy of predicted interactions. Performance is also high for a combined cell line set comprising the K562, GM12878, HeLa-S3, and IMR90 data sets, with features restricted to data sets shared by all cell lines. (**c**) Recursive feature elimination (Online Methods) evaluates predictor subsets of size 1 up to the maximum for each cell line, increasing by powers of 2 for computational efficiency. Near-optimal performance was achieved using ~16 predictors for lineage-specific models as well as the combined model, whereas lower but acceptable performance required 8 predictors. The maximum size for the feature subset shown is 32 to enhance the visibility of smaller feature subsets. NHEK lacks a measurement at a subset size of 32 because its data set included fewer than 32 total features. There were ten runs per cell line; error bars, s.e.m.

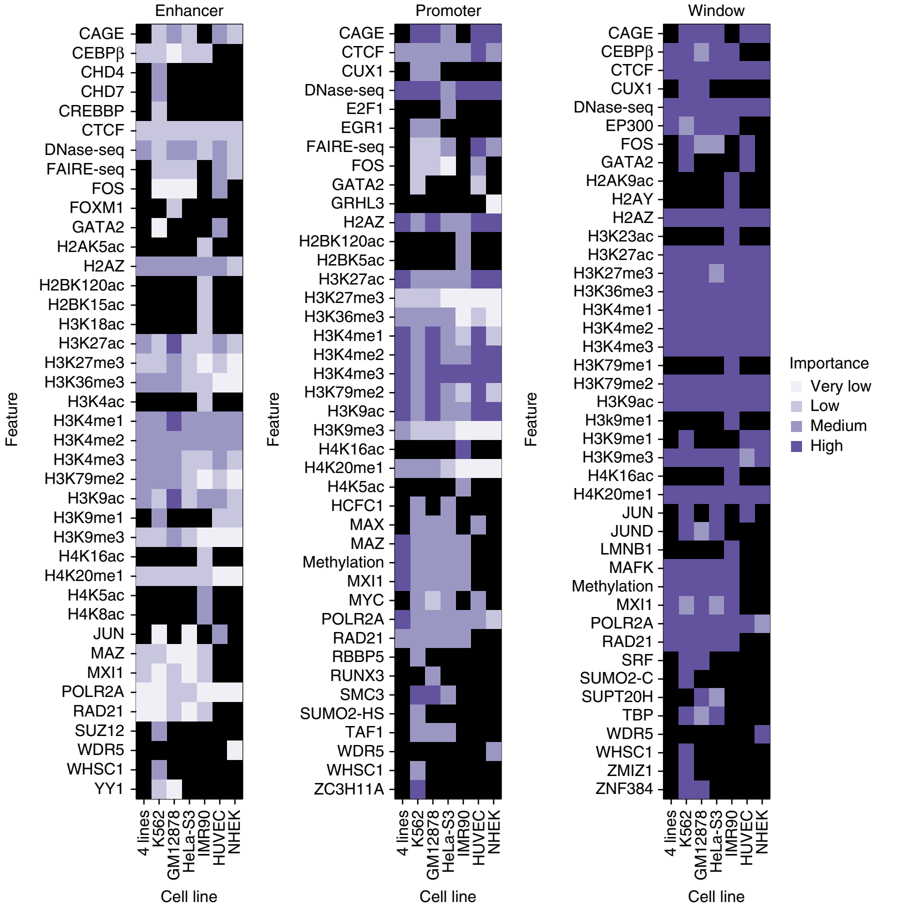


Figure 6 Predictive importance of genomic features across cell lines and regions. Importance (Online Methods) is discretized by quartiles; grid entries are filled in black when a data set was unavailable in a cell line. The highest average importance is assigned to features in the window region, followed by those in promoters. Promoter methylation and Pol II occupancy are more important in the combined '4 lines' classifier (K562, GM12878, HeLa-S3, and IMR90) than in individual cell lines. Data on highly predictive features such as CAGE were available in most but not all cell lines needed for inclusion in the combined model. Data for certain transcription factors were available in multiple cell lines but are not universally predictive, such as FOS in the window region. Data for other transcription factors were only available in a single cell line but are highly predictive, such as WHSC1 and ZMIZ1 in the window region of K562 cells and RUNX3 in the window region of GM12878 cells.

and large-scale chromatin organization[37]. This points to different molecular mechanisms operating across these scales.

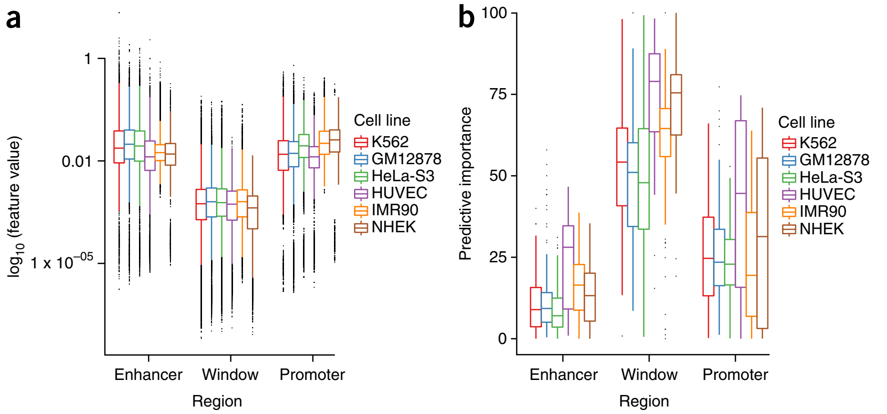
Proteins between enhancers and promoters are predictive

TargetFinder mines a diverse collection of hundreds of genomic features to build its models. To determine whether such a large feature set is needed, we applied recursive feature elimination (Online Methods). Near-optimal performance required only ~16 features (**Fig. 5c**), with performance varying by cell line owing to differences in the number of enhancer–promoter pairs as well as the quality and quantity of functional genomics data (**Supplementary Fig. 6**).

Many of the top features for each line and the combined model were from the genomic window between the enhancer and the promoter, rather than proximal signals at regulatory elements (**Fig. 7a**). This was true despite the fact that average signal (for example, ChIP-seq peak density) was higher at enhancers and promoters for most features (**Fig. 7b**). To further validate the importance of features marking the looping chromatin, we retrained TargetFinder with two alternative sets of features per cell line. The first set included features for the enhancer and promoter only (E/P) and the second set included features for an extended enhancer (using 3 kb of flanking sequence) and a non-extended promoter (EE/P), to test the hypothesis that only the enhancer-proximal part of the window is important for prediction of looping. We found a large performance gap when using only the enhancer and promoter, without marks flanking the enhancer or in the window (**Fig. 5a**). This indicates that there is substantial information relevant to looping interactions outside the enhancers and promoters themselves, and we observed this consistently across cell lines (**Supplementary Fig. 5**). Performance was better when using enhancers and promoters plus the window between them (E/P/W) than with the EE/P set, especially after accounting for the lower dimensionality of the EE/P set (two regions versus three per genomics data set), which generally improved the performance of the machine learning models. Using smaller windows around the enhancers for the EE/P set resulted in lower performance, showing that the relevant signal is not immediately next to the enhancer. Thus, signals relevant to looping are located throughout the genomic window between an enhancer and a promoter but especially within 3 kb of the enhancer.

The surprising discovery that the interaction status of an enhancer–promoter pair can be predicted with high accuracy using protein binding and epigenetic marks on DNA between them, plus a few proximal marks, made sense when we examined the specific window features that the model ranked most important. Some window features are directly involved in chromatin looping, including CTCF, the cohesin complex (SMC3–RAD21), and zinc-finger proteins such as ZNF384 and ZNF143. The zinc-finger proteins interact with CTCF to provide sequence specificity for chromatin interactions[41] by binding lineage-specific transcription factors at interacting promoters (for example, HCFC1 in HeLa-S3 cells[42]). Other window features influence the likelihood that additional promoters in the locus are the true targets of an enhancer. For example, Pol II occupancy at a promoter is not predictive by itself because it can indicate either active transcription or a gene that is poised for rapid activation. Such non-targets are distinguished by a lack of activators or co-activators[43] as well as the elongation-associated histone marks H3K36me3 and dimethylation of histone H3 at lysine 79 (H3K79me2). When these features occur in the window between an enhancer and a promoter, they increase the likelihood that an intervening promoter may be the true target. In contrast, the presence of heterochromatin, PRC2 silencing[44], and various insulators in the window suggests that intervening genes are unavailable for binding and was therefore associated with non-interacting pairs in our analyses (**Supplementary Figs. 7** and **8**). However, note that many interacting pairs had different architectures and were exceptions to this trend, including the distal enhancer of *LPIN3*

Figure 7 Influence of features by region. (**a**,**b**) Feature values (**a**) and predictive importance (**b**) for features in promoter, enhancer, and window regions. Despite having the lowest feature values, the window dominates had higher predictive importance than the enhancer and promoter regions. There were ten runs per cell line. The middle line in each plot represents the median; error bars represent 1.5 times the interquartile range.



shown in **Figure 3**. These results emphasize that TargetFinder accurately predicts interactions by learning complex genomic signatures across loci.

Window features do not directly encode distance between the enhancer and promoter, although they may serve as a kind of proxy for active chromatin or domain boundaries. To offset this possibility, we matched the distance distributions for interacting and non-interacting pairs and normalized features by the length of the region. TargetFinder has high precision and recall largely independent of enhancer–promoter interaction distances in the range of 10 kb to 2 Mb (**Supplementary Fig. 9**). In fact, performance often improved with interaction distance, which is consistent with window features encoding information about contact domain boundaries. Indeed, domain boundaries were significantly enriched in non-interacting pairs as compared to interacting pairs separated by similar distances (**Supplementary Fig. 10**). Other interaction patterns are shown in **Supplementary Figures 11–14**. Window-associated marks may also be proxies for relevant but unassayed histone modifications marking alternate targets[45].

DNA looping has a complex genomic signature

The complex patterns of co-occurrence for DNA-binding proteins and known looping factors provide mechanistic insights into the looping process itself. For example, we found that CUX1 and HCFC1 interact with CTCF and RAD21 within enhancers to increase the likelihood of looping interactions in K562 cells (**Fig. 2**). Interestingly, CUX1 was also significantly enriched at interacting promoters relative to non-interacting promoters (**Fig. 1g**), whereas HCFC1 was not (**Fig. 1h**). The importance of co-factors extends beyond this example. TargetFinder identified numerous cell-type-specific transcription factors with high feature importance that increased the probability of an enhancer being involved in an interaction when they co-occurred near the enhancer with CTCF and/or RAD21. This pattern emerged only because we quantified features separately in enhancer, promoter, and window regions.

We also learned that proteins performing multiple functions are rarely predictive on their own. Instead, TargetFinder learns to use co-factors that determine their function. For example, the histone acetyltransferase EP300 was rarely a top ranked feature, despite its strong association with active enhancers because of its ability to acetylate H3K27 (ref. 46). However, EP300 was correlated with highly predictive co-factors such as C/EBPβ which phosphorylates and modulates the activity of EP300, as well as translocates it to specific gene regions[47]. The high predictive importance of C/EBPβ may thus be due to its ability to determine the localized activity of EP300.

To further explore such context dependence, we plotted the predictive rank of an individual feature against its predictive rank when

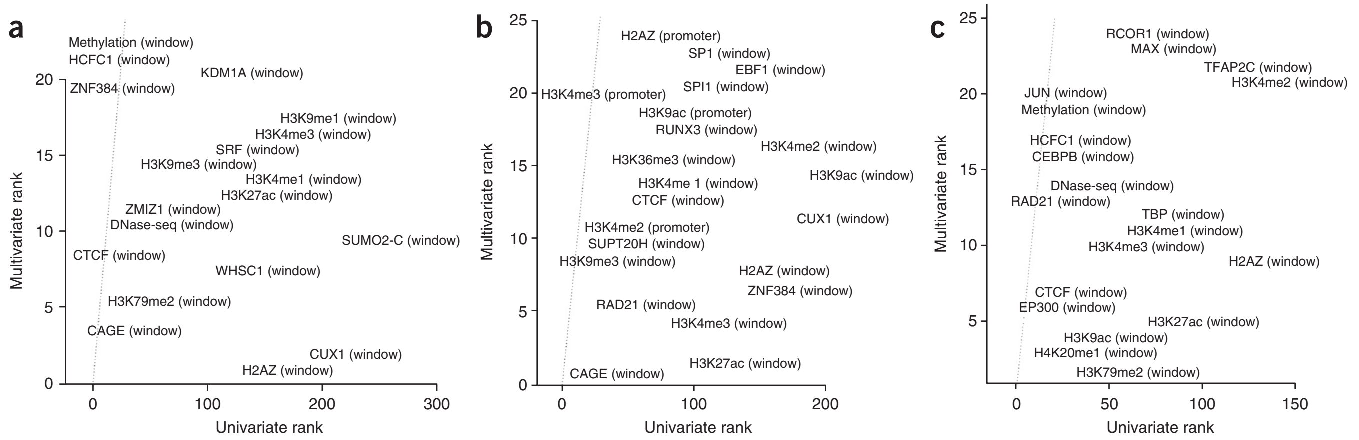


Figure 8 Identification of complex interactions between DNA-binding proteins and epigenetic marks. (**a**–**c**) Results for three cell lines are shown: K562 (**a**), GM12878 (**b**), and HeLa-S3 (**c**). Scatterplots show univariate feature significance (two-sample Kolmogorov–Smirnov test) versus multivariate feature importance (estimated via a boosted trees classifier) for the three cell lines. To highlight data sets that are predictive in combination with other features (multivariate) but are not predictive individually (univariate), only features with a multivariate rank less than 25 and a univariate rank greater than 25 are shown. For example, the lower right corner of the K562 plot in **a** shows that H2AZ, WHSC1, CUX1, and SUMO2 are among the top ten predictive features when the colocalization of other proteins is known. H2AZ has similar context-dependent importance in GM12878 (**b**) and HeLa-S3 (**c**) cells. Many features predictive in one or more cell lines were not assayed uniformly and thus could not be included in the combined model (for example, HCFC1, CUX1, and SUMO2).

it was combined with other features (**Fig. 8**). We observed many off-diagonal features that were not useful on their own (larger rank) but were extremely predictive (lower rank) in combination with additional features. In K562 cells, for example, these features included WHSC1, SUMO2, CUX1, and H2AZ. The latter two were assayed in other cell lines and showed a similar pattern. Across cell lines, large changes in rank commonly occurred for activating histone marks, such as acetylation of histone H3 at lysine 9 (H3K9ac) and the histone 2A variant H2AZ, that may help distinguish active from poised enhancers and promoters within window regions that cannot be discriminated by single activation marks. The elevated importance of H2AZ might also be explained by the link between H2A ubiquitination and Polycomb silencing[48]. Chromatin modifiers such as methyltransferases and acetyltransferases also appeared to disambiguate the state of enhancers and alternate promoter targets.

Efficient screening for relevance to chromatin looping

Motivated by results showing that histone modifications can be predicted by transcription factor binding[45], we sought to determine whether predictive transcription factors were proxies for important but unassayed post-translational modifications such as ubiquitination or sumoylation. We analyzed genome-wide SUMO ChIP-seq data for heat-shocked and non-shocked K562 cells[49] to evaluate the usefulness of sumoylation for predicting enhancer–promoter interactions. SUMO proteins are involved in protein stability and transcriptional regulation[50], and CTCF post-translationally modified by SUMO proteins 1–3 organizes repressive chromatin domains[51]. When added to the TargetFinder K562 model, sumoylation in the window between an enhancer and a promoter was a top predictor of interactions—nearly as important as CTCF. Thus, increased accuracy and insight into mechanisms of chromatin looping will be gained as additional genomic features are measured across many cell lines.

DISCUSSION

Our ability to accurately predict interactions up to 2 Mb apart at high resolution and the identification of minimal sets of predictive features quantified by genomic region, as well as a focus on high-resolution intra- rather than inter-TAD interactions, distinguish TargetFinder from previous work. Machine learning has been shown to accurately identify TADs and other larger chromatin structures (for example, A and B compartments) from two-dimensional genomic data[37], but it has not yet been applied to such fine-scale interactions within TADs.

Although some of the predictive accuracy of TargetFinder derives from genomic features whose measurement was limited to one or a few cell types, many of the top ranked features are similar in predictive importance across cell types and in the combined model. For example, members of the cohesin complex (SMC3–RAD21) and CTCF are highly predictive, as is CAGE when it is assayed. DNA methylation and Pol II have elevated importance in the combined cell line set where the model was trained on fewer data sets that excluded some transcription factors and other features measured in only a subset of the cell lines. Marks of heterochromatin and elongation are also consistently important. These robust, general features of looping chromatin promise to be useful assays for predicting regulatory interactions in new cell types, perhaps in combination with data on cell-type-specific regulators. They also suggest that, as these predictive features are assayed in more cell types, we may be able to develop a generic TargetFinder model that could perform accurate *in silico* Hi-C on independent cell types that do not have genome-wide high-resolution chromatin interaction data. To do so will require rigorous normalization, as TargetFinder relies on numeric values of genomic data being comparable across data sets.

We identified numerous features whose role in distal enhancer–promoter interactions may be underappreciated. These include the DNA-binding proteins CUX1, ZNF384, SUPT20H, RUNX3, SPI1, SP1, EBF1, RCOR1, MAX, TFAP2C, HCFC1, C/EBPβ, JUND, TBP, SRF, ZMIZ1, and WHSC1 (**Fig. 8**). Most of these are predictive only in combination with other features, some of which have roles in chromatin structure. For instance, several interact with the cohesin complex, and ZNF143 was recently shown to provide sequence specificity to cohesin-associated chromatin looping[41]. Predictive transcription factors often belong to activating or repressive complexes such as AP-1, AP-2γ, or PRC2 or are chromatin modifiers such as methyltransferases or acetyltransferases that help determine whether enhancers or promoters are in an active or poised state. These general trends are consistent across cell types, but the particular transcription factors that provide a predictive boost are often specific to a small number of cell lines. In addition, we identified several more general predictors of looping chromatin. Sumoylation is a combinatorially predictive post-translational modification not assayed by ENCODE or Roadmap Epigenomics. The activating marks H2AZ and H3K9ac and elongation marks H3K36me3 and H3K79me2 were also especially useful for chromatin loop prediction, more so than many of the well-known histone marks necessary for ChromHMM and Segway annotations of promoters and enhancers. CAGE is also a consistently top ranked feature, providing information on the activation state of annotated enhancers and alternate targets in the window that is complementary to data from ChIP-seq assays.

Many of the top features used by TargetFinder are not predictive on their own. To our knowledge, several of our most predictive features have received little to no study in the context of chromatin looping, although others have well-known biological relevance either to regulatory elements or their interactions. Examples include SRF, which regulates FOS[52] and interacts with C/EBPβ (ref. 53); TFAP2C (AP-2γ), which is a pioneer factor associated with estrogen-receptor-binding events and FOXA1 expression[54]; ZMIZ1 (hZimp10), which promotes expression and sumoylation of the androgen receptor[55]; and KDM1A, which interacts with RCOR1 to demethylate H3K4 (ref. 56). We identified several other proteins with poor univariate importance that nonetheless have known roles in chromatin looping and were highly ranked by TargetFinder. These include SP1 (refs. 57,58), SPI1 (PU.1)[59,60], HCFC1, which colocalizes with looping factor ZNF143 (ref. 42), and TBP, whose TAF3 subunit is recruited by CTCF to distal promoters[61] and which is linked with long-range interactions[62]. Finally, WHSC1 (NSD2) is a histone methyltransferase of H3K36me3 and therefore is associated with predictive marks of elongation[63]. Thus, changes in univariate versus multivariate predictive rank recapitulate known protein interactions as well as identify underappreciated or potentially new biological interactions, often involving cell-line-specific transcription factors.

These results are more relevant to looping models of interaction than alternatives such as facilitated tracking[64]. Polycomb complexes appear to have several roles in distinguishing nearby targets. For example, PRC2-targeted CpG islands are enriched for binding motifs for REST and CUX1, both of which are transcriptional repressors[65] with high predictive importance. In *Drosophila melanogaster*, cohesin colocalizes with PRC1 at promoters and interacts with this complex to control gene silencing[66]. Given the conservation of Polycomb complexes between flies and humans[67], this finding has implications for the interaction of cohesin and Polycomb complexes in mammalian gene silencing and thus for the discrimination of target promoters. Also, distal enhancers may sometimes serve to clear Polycomb complexes from CpG islands[68]. Elongation has recently been shown to spatially segregate genes in the HoxD locus present in separate TADs[69], suggesting that its role in inter-TAD gene clusters could contribute to its predictive importance. Finally, recent work shows that cohesin spatially clusters enhancers[70] and is consistent with our observation that the presence of active marks at nearby enhancers often increases 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 data sets. Additional discussion is available in the **Supplementary Note**.

URLs. TargetFinder, https://github.com/shwhalen/targetfinder; ENCODE annotation, http://encodeproject.org/data/annotations.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This project was supported by the Bench to Bassinet Program of the NHLBI (U01HL098179 and UM1HL098179), the NIH/NHLBI (HL089707), the San Simeon Fund, and the Gladstone Institutes.

AUTHOR CONTRIBUTIONS

S.W., R.M.T., and K.S.P. designed the experiments and wrote the manuscript. S.W. and R.M.T. implemented the experiments.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).
2. Lomelin, D., Jorgenson, E. & Risch, N. Human genetic variation recognizes functional elements in noncoding sequence. *Genome Res.* **20**, 311–319 (2010).
3. Alexandrov, N.N. *et al.* Features of *Arabidopsis* genes and genome discovered using full-length cDNAs. *Plant Mol. Biol.* **60**, 69–85 (2006).
4. Hillier, L.W. *et al.* Whole-genome sequencing and variant discovery in *C. elegans*. *Nat. Methods* **5**, 183–188 (2008).
5. Massouras, A. *et al.* Genomic variation and its impact on gene expression in *Drosophila melanogaster*. *PLoS Genet.* **8**, e1003055 (2012).
6. Tang, R. *et al.* Candidate genes and functional noncoding variants identified in a canine model of obsessive-compulsive disorder. *Genome Biol.* **15**, R25 (2014).
7. Manolio, T.A., Brooks, L.D. & Collins, F.S. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118**, 1590–1605 (2008).
8. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
9. Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. Human genetic variation and its contribution to complex traits. *Nat. Rev. Genet.* **10**, 241–251 (2009).
10. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
11. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
12. Celniker, S.E. *et al.* Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
13. Bernstein, B.E. *et al.* The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol.* **28**, 1045–1048 (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 Res.* **22**, 1790–1797 (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 Res.* **40**, D930–D934 (2012).
17. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (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. *Nat. Genet.* **47**, 276–283 (2015).
19. Lettice, L.A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).
20. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
21. Kvon, E.Z. *et al.* Genome-scale functional characterization of *Drosophila* developmental enhancers *in vivo*. *Nature* **512**, 91–95 (2014).
22. Wang, D., Rendon, A. & Wernisch, L. Transcription factor and chromatin features predict genes associated with eQTLs. *Nucleic Acids Res.* **41**, 1450–1463 (2013).
23. Yip, K.Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13**, R48 (2012).
24. Aran, D., Sabato, S. & Hellman, A. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol.* **14**, R21 (2013).
25. Rödelsperger, C. *et al.* Integrative analysis of genomic, functional and protein interaction data predicts long-range enhancer–target gene interactions. *Nucleic Acids Res.* **39**, 2492–2502 (2011).
26. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
27. 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 Comput. Biol.* **8**, e1002798 (2012).
28. Fullwood, M.J. *et al.* An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
29. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
30. Dostie, J. *et al.* Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.* **16**, 1299–1309 (2006).
31. de Wit, E. & de Laat, W. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.* **26**, 11–24 (2012).
32. Rao, S.S.P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
33. Dixon, J.R. *et al.* Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
34. Schoenfelder, S. *et al.* The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.* **25**, 582–597 (2015).
35. Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.* **47**, 598–606 (2015).
36. Maston, G.A., Evans, S.K. & Green, M.R. Transcriptional regulatory elements in the human genome. *Annu. Rev. Genomics Hum. Genet.* **7**, 29–59 (2006).
37. Moore, B.L., Aitken, S. & Semple, C.A. Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization. *Genome Biol.* **16**, 110 (2015).
38. Zhang, Y. *et al.* Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* **504**, 306–310 (2013).
39. Corradin, O. & Scacheri, P.C. Enhancer variants: evaluating functions in common disease. *Genome Med.* **6**, 85 (2014).
40. Shaulian, E. & Karin, M. AP-1 as a regulator of cell life and death. *Nat. Cell Biol.* **4**, E131–E136 (2002).
41. Bailey, S.D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun.* **2**, 6186 (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 Res.* **23**, 907–916 (2013).
43. Adelman, K. & Lis, J.T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13**, 720–731 (2012).
44. Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469**, 343–349 (2011).
45. Benveniste, D., Sonntag, H.-J., Sanguinetti, G. & Sproul, D. Transcription factor binding predicts histone modifications in human cell lines. *Proc. Natl. Acad. Sci. USA* **111**, 13367–13372 (2014).
46. Visel, A. *et al.* ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854–858 (2009).
47. Schwartz, C. *et al.* Recruitment of p300 by C/EBPβ triggers phosphorylation of p300 and modulates coactivator activity. *EMBO J.* **22**, 882–892 (2003).
48. Wang, H. *et al.* Role of histone H2A ubiquitination in Polycomb silencing. *Nature* **431**, 873–878 (2004).
49. Niskanen, E.A. *et al.* Global SUMOylation on active chromatin is an acute heat stress response restricting transcription. *Genome Biol.* **16**, 153 (2015).
50. Hay, R.T. SUMO: a history of modification. *Mol. Cell* **18**, 1–12 (2005).
51. MacPherson, M.J., Beatty, L.G., Zhou, W., Du, M. & Sadowski, P.D. The CTCF insulator protein is posttranslationally modified by SUMO. *Mol. Cell. Biol.* **29**, 714–725 (2009).

52. Fujioka, S. *et al.* NF-κB and AP-1 connection: mechanism of NF-κB-dependent regulation of AP-1 activity. *Mol. Cell. Biol.* **24**, 7806–7819 (2004).
53. Hanlon, M. & Sealy, L. Ras regulates the association of serum response factor and CCAAT/enhancer-binding protein β. *J. Biol. Chem.* **274**, 14224–14228 (1999).
54. Jozwik, K.M. & Carroll, J.S. Pioneer factors in hormone-dependent cancers. *Nat. Rev. Cancer* **12**, 381–385 (2012).
55. Sharma, M. *et al.* hZimp10 is an androgen receptor co-activator and forms a complex with SUMO-1 at replication foci. *EMBO J.* **22**, 6101–6114 (2003).
56. Upadhyay, G., Chowdhury, A.H., Vaidyanathan, B., Kim, D. & Saleque, S. Antagonistic actions of Rcor proteins regulate LSD1 activity and cellular differentiation. *Proc. Natl. Acad. Sci. USA* **111**, 8071–8076 (2014).
57. Nolis, I.K. *et al.* Transcription factors mediate long-range enhancer-promoter interactions. *Proc. Natl. Acad. Sci. USA* **106**, 20222–20227 (2009).
58. Deshane, J. *et al.* Sp1 regulates chromatin looping between an intronic enhancer and distal promoter of the human heme oxygenase-1 gene in renal cells. *J. Biol. Chem.* **285**, 16476–16486 (2010).
59. Listman, J.A. *et al.* Conserved ETS domain arginines mediate DNA binding, nuclear localization, and a novel mode of bZIP interaction. *J. Biol. Chem.* **280**, 41421–41428 (2005).
60. van Riel, B. & Rosenbauer, F. Epigenetic control of hematopoiesis: the PU.1 chromatin connection. *Biol. Chem.* **395**, 1265–1274 (2014).
61. Liu, Z., Scannell, D.R., Eisen, M.B. & Tjian, R. Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. *Cell* **146**, 720–731 (2011).
62. Bertolino, E. & Singh, H. POU/TBP cooperativity: a mechanism for enhancer action from a distance. *Mol. Cell* **10**, 397–407 (2002).
63. Nimura, K. *et al.* A histone H3 lysine 36 trimethyltransferase links Nkx2-5 to Wolf-Hirschhorn syndrome. *Nature* **460**, 287–291 (2009).
64. Blackwood, E.M. & Kadonaga, J.T. Going the distance: a current view of enhancer action. *Science* **281**, 60–63 (1998).
65. Islam, A.B., Richter, W.F., Lopez-Bigas, N. & Benevolenskaya, E.V. Selective targeting of histone methylation. *Cell Cycle* **10**, 413–424 (2011).
66. Dorsett, D. & Kassis, J.A. Checks and balances between cohesin and polycomb in gene silencing and transcription. *Curr. Biol.* **24**, R535–R539 (2014).
67. Levine, S.S. *et al.* The core of the polycomb repressive complex is compositionally and functionally conserved in flies and humans. *Mol. Cell. Biol.* **22**, 6070–6078 (2002).
68. Vernimmen, D. *et al.* Polycomb eviction as a new distant enhancer function. *Genes Dev.* **25**, 1583–1588 (2011).
69. Fabre, P.J. *et al.* Nanoscale spatial organization of the HoxD gene cluster in distinct transcriptional states. *Proc. Natl. Acad. Sci. USA* **112**, 13964–13969 (2015).
70. Ing-Simmons, E. *et al.* Spatial enhancer clustering and regulation of enhancer-proximal genes by cohesin. *Genome Res.* **25**, 504–513 (2015).

ONLINE METHODS

All code and data are accessible online (see URLs).

Identification of regulatory elements. TSS-containing promoter regions and strong and weak enhancer regions were identified using combined ENCODE Segway[71] and ChromHMM[72] annotations for K562, GM12878, HeLa-S3, and HUVEC cells and Roadmap Epigenomics ChromHMM annotations for NHEK and IMR90 cells. Enhancers closer than 10 kb to the nearest promoter were discarded to focus the model on distal interactions. Promoters were retained if actively transcribed (mean FPKM >0.3 (ref. 73) with irreproducible discovery rate <0.1 (ref. 74)) in each cell line as determined using GENCODE[75] version 19 annotations and RNA-seq data from the ENCODE portal. Promoter and enhancer counts per line are given in **Supplementary Table 1**.

Chromatin interactions. Interacting enhancer–promoter pairs were annotated using high-resolution genome-wide Hi-C data (10% FDR; Gene Expression Omnibus (GEO), GSE63525)[32]. Pairs were assigned to one of five bins on the basis of the distance between the enhancer and the promoter, such that each bin had the same number of interactions. Non-interacting enhancer–promoter pairs were assigned to their corresponding distance bin and then subsampled within each bin, using 20 negatives per positive (**Supplementary Table 1**). Performance was similar without distance matching, with a loss of approximately 1% F_1 per 250,000 additional samples (total loss of 6% F_1 for K562 cells).

Genomic features. Functional genomics data for each cell line were downloaded from ENCODE, Roadmap Epigenomics, or GEO; details and accessions are given in **Supplementary Table 2**. Peak calls for ENCODE data were obtained from GEO; raw reads for the Roadmap Epigenomics and GEO data sets were obtained, quality trimmed using fastq-mcf, aligned to hg19 using Bowtie2 (ref. 76), and subjected to peak calling using MACS2 (ref. 77) with default parameters. Peaks were intersected with promoter, enhancer, extended enhancer, and window regions. The strength of all peaks in a region or the counts of methylated bases in a region were summed and divided by the length of the region in base pairs to generate features. Cluster heat maps of correlations between the top 16 predictive features for each cell line are shown in **Supplementary Figures 15–20**.

Software implementation. TargetFinder was implemented in Python using the scikit-learn machine learning library[78], the pandas analytics library[79], and BEDTools[80]. We used DummyClassifier to measure baseline performance, LinearSVC for a linear SVM[81], DecisionTreeClassifier for a single decision tree[82], and GradientBoostingClassifier for a decision tree ensemble[83]. The linear SVM was fit with parameter class weight = "balanced" as part of a pipeline with a StandardScaler preprocessing step. The boosting classifier was fit with parameters n_estimators = 4,000, learning_rate = 0.1, max_depth = 5, and max_features = "log2". Models were fit with sample weights inversely proportional to class balance to prevent overfitting the negative class. Identical parameters were used for each cell line. Results were consistent with those from an alternative implementation in R (**Supplementary Note**).

All models were evaluated using tenfold cross-validation where data were divided into ten non-overlapping training and test sets. Performance was measured using multiple metrics, and the average overall test sets are reported. Feature importance values were computed by scikit-learn using the method of Hastie *et al.*[84], accessible via the feature_importances_ attribute of eligible models. The following pseudocode summarizes their implementation

```
ensemble_importances = zeros(total_features)
  for each tree in ensemble:

  tree_importances = zeros(total_features)
    for each node in tree:
      if node is not a leaf:
        tree_importances[node.feature_index] +=
        node.sample_count * node.impurity –
        node.left_child.sample_count * node.left_child.impurity –
        node.right_child.sample_count * node.right_child.impurity
  ensemble_importances += tree_importances / total_samples
ensemble_importances /= total_trees
```

where zeros(n) initializes an array of *n* zeros, total features is the total number of features in the data set, node.feature index is the index of the feature used to split samples at a node, node.sample count is the number of samples present at a node before splitting, node.impurity is a measure of error (here, gini impurity), and node.left child and node.right child represent the children of a node. Overall, this method sums the weighted reduction in impurity when splitting on each feature across all trees in the ensemble, normalized by the number of samples per tree and total number of trees. Models were fit ten times, each with a different random seed number, to better estimate the mean and variance of feature importance values.

Recursive feature elimination[85] was used to estimate the optimal number of features via nested cross-validation[86]. Within each training set during 'outer' cross-validation, feature importance values are initially estimated using all features. The performance of the top *n* features is then estimated from 'inner' cross-validation on the training set, with *n* increasing from 1 to the maximum number of features by powers of 2. Finally, the best performing subset identified via inner cross-validation is evaluated against the outer test set to obtain an unbiased performance estimate.

71. Hoffman, M.M. *et al.* Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nat. Methods* **9**, 473–476 (2012).
72. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
73. Ramsköld, D., Wang, E.T., Burge, C.B. & Sandberg, R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput. Biol.* **5**, e1000598 (2009).
74. Li, Q., Brown, J.B., Huang, H. & Bickel, P.J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).
75. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
76. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
77. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
78. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
79. McKinney, W. *Python for Data Analysis* (O'Reilly, 2012).
80. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
81. Burges, C.J.C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **2**, 121–167 (1998).
82. Kingsford, C. & Salzberg, S.L. What are decision trees? *Nat. Biotechnol.* **26**, 1011–1013 (2008).
83. Friedman, J.H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
84. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, 2009).
85. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
86. Ambroise, C. & McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA* **99**, 6562–6566 (2002).