Clustered, information-dense transcription factor binding sites identify genes with similar tissue-wide expression profiles

Ruipeng Lu[1], Peter K. Rogan[1,2,3*]

[1] Department of Computer Science, Western University, London, Canada, N6A 5B7

[2] Department of Biochemistry, Western University, London, Canada, N6A 5C1

[3] Cytognomix Inc., London, Canada, N5X 3X5

* Correspondence: Dr. Peter K. Rogan (progan@uwo.ca), 1 (519) 661-4255

Running title: Binding site clusters identify transcription targets

ABSTRACT

Background: The distribution and composition of *cis*-regulatory modules (e.g. transcription factor binding site (TFBS) clusters) in promoters substantially determine gene expression patterns and TF targets, whose expression levels are significantly regulated by TF binding. TF knockdown experiments have revealed correlations between TF binding profiles and gene expression levels. We present a general framework capable of predicting genes with similar tissue-wide expression patterns from activated or repressed TF targets using machine learning to combine TF binding and epigenetic features.

Methods: Genes with correlated expression patterns across 53 tissues were identified according to their Bray-Curtis similarity. DNase I HyperSensitive region (DHS) -accessible promoter intervals of direct TF target genes were scanned with previously derived information theory-based position weight matrices (iPWMs) of 82 TFs. Features from information density-based TFBS clusters were used to predict target genes with machine learning classifiers. The accuracy, specificity and sensitivity of the classifiers were determined for different feature sets. Mutations in TFBSs were also introduced to examine their impact on cluster densities and the regulatory states of predicted target genes.

Results: We initially chose the glucocorticoid receptor gene (*NR3C1*), whose regulation has been extensively studied, to test this approach. *SLC25A32* and *TANK* were found to exhibit the most similar expression patterns to this gene across 53 tissues. Prediction of other genes with similar expression profiles was significantly improved by eliminating inaccessible promoter intervals based on DHSs. A Random Forest classifier exhibited the best performance in detecting such coordinately regulated genes (accuracy was 0.972 for training, 0.976 for testing). Target gene prediction was confirmed using CRISPR knockdown data of TFs, which was more accurate than siRNA inactivation. Mutation analyses of TFBSs also revealed that one or more information-dense TFBS clusters in promoters are required for accurate target gene prediction.

Conclusions: Machine learning based on TFBS information density, organization, and chromatin accessibility accurately identifies gene targets with comparable tissue-wide expression patterns. Multiple, information-dense TFBS clusters in promoters appear to protect promoters from the effects of deleterious binding site mutations in a single TFBS that would effectively alter the expression state of these genes.

KEYWORDS

Information theory, transcription factors, DNA binding sites, gene expression, mutation analysis, machine learning

BACKGROUND

The distinctive organization and combination of transcription factor binding sites (TFBSs) and regulatory modules in promoters dictates specific expression patterns within a set of genes [1]. Clustering of multiple adjacent binding sites for the same TF (homotypic clusters) and for different TFs (heterotypic clusters) defines *cis*-regulatory modules (CRMs) in human gene promoters and can amplify the influence of individual TFBSs on gene expression through increasing binding affinities, facilitated diffusion mechanisms and funnel effects [2]. Because tissue-specific TF-TF interactions in TFBS clusters are prevalent, these features can assist in identifying correct target genes by altering binding specificities of individual TFs [3]. Previously, we derived iPWMs from ChIP-seq data that can accurately detect TFBSs and quantify their strengths by computing associated R_i values (Rate of Shannon information transmission for an individual sequence [4]), with $R_{sequence}$ being the average of R_i values of all binding site sequences and representing the average binding strength of the TF [3]. Furthermore, information density-based clustering (IDBC) can effectively identify functional TF clusters by taking into account both the spatial organization (i.e. intersite distances) and information density (i.e. R_i values) of TFBSs [5].

TF binding profiles, either derived from in vivo ChIP-seq peaks [6–8] or computationally detected binding sites and CRMs [9], have been shown to be predictive of absolute gene expression levels using a variety of tissue-specific machine learning classifiers and regression models. Because signal strengths of ChIP-seq peaks are not strictly proportional to TFBS strengths [3], representing TF binding strengths by ChIP-seq signals may not be appropriate; nevertheless, both achieved similar accuracy [10]. CRMs have been formed by combining two or three adjacent TFBSs [9], which is inflexible, as it arbitrarily limits the number of binding sites contained in a module, and does not consider differences between information densities of different CRMs. Chromatin structure (e.g. histone modification (HM) and DNase I hypersensitivity) were also found to be highly redundant with TF binding in explaining tissue-specific mRNA transcript abundance at a genome-wide level [7,8,11,12], which was attributed to the heterogeneous distribution of HMs across chromatin domains [8]. Combining these two types of data explained the largest fraction of variance in gene expression levels in multiple cell lines [7,8], suggesting that either contributes unique information to gene expression that cannot be compensated for by the other.

The number of genes directly bound by a TF significantly exceeds the number of genes whose expression levels significantly change upon knockdown of the TF. Only a small subset of genes whose promoters overlap ChIP-seq peaks were differentially expressed (DE) after individually knocking 59 TFs down using small interfering RNAs (siRNAs) in the GM19238 cell line [13]. Correlation between TFBS counts and gene expression levels across 10 different cell lines among 8,872 genes from these knockdown data were more predictive of DE targets than setting a minimum threshold on TFBS counts [14]. Their TFBS counts were defined as the number of ChIP-seq peaks overlapping the promoter, though it was unknown how many binding sites were present in these peaks; true positives might not be direct targets in the TF regulatory cascade, as the promoters of these targets were not intersected with ChIP-seq peaks. By

perturbing gene expression with CAS9-directed clustered regularly interspaced short palindromic repeats (CRISPR) of 10 different TF genes in K562 cells, the regulatory effects of each TF on 22,046 genes were dissected by single cell RNA sequencing with a regularized linear computational model [15]; this accurately revealed DE targets and new functions of individual TFs, some of which were likely regulated through direct interactions at TFBS in their corresponding promoters. Machine learning classifiers have also been applied in a small number of gene instances to predict targets of a single TF using features extracted from *n*-grams derived from consensus binding sequences [16], or from TFBSs and homotypic binding site clusters [5].

To investigate whether the distribution and composition of information theory-based CRMs in promoters substantially determines gene expression profiles of direct TF targets, we developed a general machine learning framework that predicts which genes have similar expression profiles to a given gene and DE direct TF targets by combining information theory-based TF binding profiles with DHSs. Upon filtering for accessible promoter intervals with DHSs, features designed to capture the spatial distribution and information composition of CRMs were extracted from clusters identified by the IDBC algorithm from iPWM-detected TFBSs. Though not all direct targets regulated by multiple TFs share a common tissue-wide expression profile, this framework provides insight into the transcriptional program of genes with similar profiles by dissecting their *cis*-regulatory element organization and strengths. We identify genes with comparable tissue-wide expression profiles by application of Bray-Curtis similarity [17]. Using transcriptome data generated by CRISPR [15] and siRNA-based [13] TF knockdowns, we verified predicted direct TF targets whose promoters overlap tissue-specific ChIP-seq peaks, in contrast with correlation-based approaches [14].

METHODS

To identify genes with similar tissue-wide expression patterns, we formally define gene expression profiles and pairwise similarity measures between profiles of different genes. A general machine learning framework relates features extracted from the organization of TFBSs in these genes to their tissue-wide expression patterns. True positives (TPs) and negatives (TNs) for predicting direct DE TF targets were validated using CRISPR- and siRNA-generated knockdown data (see below).

Similarity between gene expression profiles

The median RPKM (Reads Per Kilobase of transcript per Million mapped reads) of 56,238 genes across 53 tissues were obtained from the Genotype-Tissue Expression (GTEx) project [18]. To capture the tissue-wide overall expression pattern of a gene instead of within a single tissue, the expression profile of a gene was defined as its median RPKM across the 53 tissues, which forms a vector of size 53 and does not distinguish between different isoforms whose expression patterns may significantly differ from each other. To obtain ground-truth genes that have similar expression profiles to a given gene, the Bray-Curtis Similarity (Equation 1) was used to compute the similarity value between the expression profiles of two genes, because it takes both the directions and lengths of the vectors into account while maintaining strict bounds of 0 and 1.

$$sim_{Bray-Curtis}(EP^A, EP^B) = \begin{cases} 1, & if\ \sum_{i=1}^{53} EP_i^A = \sum_{i=1}^{53} EP_i^B = 0 \\ 1 - \frac{\sum_{i=1}^{53} |EP_i^A - EP_i^B|}{\sum_{i=1}^{53} (EP_i^A + EP_i^B)}, & otherwise \end{cases} \quad (1)$$

where EP^A and EP^B are respectively the expression profiles of genes A and B, EP_i^A and EP_i^B are respectively the median RPKM of genes A and B in the ith tissue. If $EP^A = EP^B$, then $sim_{Bray-Curtis}(EP^A, EP^B) = 1$.

Prediction of genes with similar expression profiles

The framework for identifying genes that have similar expression profiles to a specific gene is shown in Figure 1A and 1B. All DHSs in 95 cell types generated by the ENCODE project [18; hg38 assembly] were intersected with known promoters [20], then 94 iPWMs exhibiting primary binding motifs for 82 TFs [3] were used to detect TFBSs in overlapping intervals. When detecting heterotypic TFBS clusters with the IDBC algorithm, a minimum threshold $0.1 * R_{sequence}$ was set for R_i values of TFBSs, in order to remove weak binding sites that were likely to be false positive TFBSs.

The information density-related features derived from each TFBS cluster include: 1) The distance between this cluster and the transcription start site (TSS); 2) The length of this cluster; 3) The information content of this cluster (i.e. the sum of R_i values of all TFBSs in this cluster); 4) The number of binding sites of each TF within this cluster; 5) The number of strong binding sites ($R_i > R_{sequence}$) of each TF within this cluster; 6) The sum of R_i values of binding sites of each TF within this cluster; 7) The sum of R_i values of strong binding sites ($R_i > R_{sequence}$) of each TF within this cluster.

For a gene instance, each of Features 1-3 is defined as a vector whose size equals the number of clusters in the promoter; thus, the entire vector could be input into a classifier. If two instances contained different numbers of clusters, the maximum number of clusters among all instances was determined, and null clusters are added at the 5' end of promoters with fewer clusters, enabling all instances to have the same cluster count. Machine learning classifiers in Weka [21] were implemented for training and testing.

Prediction of differentially expressed direct targets of TFs

Using gene expression in the CRISPR-based perturbations

Dixit et al. performed CRISPR-based perturbation experiments using multiple guide RNAs for each of ten TFs in K562 cells, resulting in a regulatory matrix of coefficients that indicate the

effect of each guide RNA on each of 22,046 genes [15]. The coefficient of a guide RNA on a TF gene target is defined as the $\log_{10}$(fold change in gene expression level) [15]. Among these ten TFs, we have previously derived iPWMs exhibiting primary binding motifs for seven (EGR1, ELF1, ELK1, ETS1, GABPA, IRF1, YY1) [3]. Therefore, the framework for predicting direct TF targets in the K562 cell line (Figure 1A and 1C) was applied to these TFs. The criteria for defining a TP (i.e. a DE direct target), of a TF was:

1) The fold change in the expression level of this gene for each guide RNA of the TF was > (or <) 1, consistent with the possibility that the gene was regulated by the TF, and

2) The average fold change in the expression level of this gene for all guide RNAs of the TF was > threshold ε (or < $1/\varepsilon$), and

3) The promoter interval (10 kb) upstream of a TSS of this gene overlaps a ChIP-seq peak of the TF in the K562 cell line.

If the coefficients of all guide RNAs of the TF for a gene are zero, the gene was defined as a TN. As the threshold $\boldsymbol{\varepsilon}$ increases, the number of TPs strictly decreases; as $\boldsymbol{\varepsilon}$ decreases, we have increasingly lower confidence in the fact that the TPs were indeed differentially expressed because of the TF perturbation. To achieve a balance between sensitivity and specificity, we evaluated three different values (i.e. 1.01, 1.05 and 1.1) for $\boldsymbol{\varepsilon}$. For each TF, all ENCODE ChIP-seq peak datasets from the K562 cell line were merged to determine TPs. To make the numbers of TNs and TPs equal, the Bray-Curtis function was applied to compute the similarity values in the expression profile between all TNs and the TP with the largest average coefficient, then the TNs with the smallest values were selected (Figure 1C).

Because TFs act upon different sets of target genes in different tissues [3], the iPWMs of EGR1, ELK1, ELF1, GABPA, IRF1, YY1 from the K562 cell line were used to detect binding sites; for ETS1, we used the only available iPWM from the GM12878 cell line [3]. Six features

were derived from each homotypic cluster (i.e. Features 3 and 6 converged to the same value, because only binding sites from a single TF were used).

Using gene expression in the siRNA-based knockdown

In the GM19238 cell line, 59 TFs were individually knocked down using siRNAs, and significant changes in the expression levels of 8,872 genes were indicated according to their corresponding P-values [13]. In these cases, the P-value of a gene for a TF is the probability of observing the change in the expression level of this gene under the null hypothesis of no differential expression after TF knockdown; thus the larger the change in the expression level, the smaller the P-value and the more likely this gene is differentially expressed. They also indicated whether the promoters of these genes display evidence of binding to TFs by intersecting with ChIP-seq peaks in the GM12838 cell line. Among these 59 TFs, we have previously derived accurate iPWMs exhibiting primary binding motifs for 11 (BATF, JUND, NFE2L1, PAX5, POU2F2, RELA, RXRA, SP1, TCF12, USF1, YY1) [3]. Therefore, the framework for predicting direct TF targets in the GM19238 cell line (Figure 1A and 1D) was applied to these 11 TFs.

We defined a TP (i.e. a DE direct target) for a TF, if the P-value of this gene for the TF was ≤ 0.01, and the promoter interval (10kb) upstream of a TSS of this gene overlapped a ChIP-seq peak of the TF in the GM12878 cell line. A TN for a TF exhibited the following properties: a P-value > 0.01 for the TF, and this gene was annotated to exhibit a single promoter and one constitutive transcript. Because different transcripts can display different tissue-specific expression [22], the use of genes with one single transcript guaranteed that the analyzed promoters functionally induce their expression in the GM12878 cell line. TPs and TNs were ranked according to their Bray-Curtis similarity values prior to being separated into training and test sets (Figure 1D).

The DHSs in the GM19238 cell line mapped from the hg19 genome assembly were first remapped to the hg38 assembly using liftOver (available at genome.ucsc.edu) [23]. Aside from RELA and NFE2L1, the iPWMs of TFs from the GM12878 cell line were used to detect binding sites. For RELA, the iPWM from the GM19099 cell line was used; for NFE2L1, the only available iPWM was derived from K562 cells and was applied. Although the knockdown was performed in GM19238, GM12878 and GM19099 are also lymphoblastic cell lines, with GM19099 and GM19238 both being derived from Yorubans. For this analysis, the iPWMs derived in GM12878 and GM19099 were more appropriate than the iPWM from K562, since GM12878 and GM19099 are of the same tissue type and are thus more likely comparable to GM19238 than to K562.

Mutation analyses on promoters of differentially expressed direct targets

To better understand the significance of individual binding sites for information-dense clusters and the regulatory state of direct targets, we evaluated the effects of sequence changes that altered the R_i values of these sites on cluster formation and whether a gene was predicted to be a TF target. Mutations were sequentially introduced into the strongest binding sites in TFBS clusters of the EGR1 target gene, *MCM7*, to determine the threshold for cluster formation after disappearing clusters disabled induction of *MCM7* expression. For one target gene of each TF from the CRISPR-generated perturbation data, effects of naturally occurring TFBS variants present in dbSNP (https://www.ncbi.nlm.nih.gov/projects/SNP/) [24] were also evaluated to explore aspects of TFBS organization that enabled both clusters and promoter activity to be resilient to binding site mutations. This was done by analyzing whether the occurrence of individual or multiple single nucleotide polymorphisms (SNPs) lead to the loss of binding sites and the clusters that contain them, and result in changes in the predictions of these targets.

RESULTS

Similarity between gene expression profiles

To confirm that the Bray-Curtis Similarity can indeed effectively measure how akin the expression profiles of two genes are to each other, it was applied to compute the similarity values between the expression profiles of the glucocorticoid receptor (*GR* or *NR3C1*) gene and all other 56,237 genes. *NR3C1* is an extensively characterized TF with many known direct target genes [22]. As a constitutively expressed TF activated by glucocorticoid ligands, it can mediate the up-regulation of anti-inflammatory genes by binding of homodimers to glucocorticoid response elements and down-regulation of proinflammatory genes by complexing with other activating TFs (e.g. NFKB and AP1) and eliminating their ability to bind targets [22]. NR3C1 can bind its own promoter forming an auto-regulatory loop, which also contains functional binding sites of 11 other TFs (e.g. SP1, YY1, IRF1, NFKB) whose iPWMs have been developed and/or mutual interactions have been described in Lu et al. [3,22]. However, the expression profile of *NR3C1* integrates all different splicing and translational isoforms (e.g. *GRα-A* to *GRα-D*, *GRβ*, *GRγ*, *GRδ*), whereas these isoforms have tissue-specific expression patterns (e.g. levels of the *GRα-C* isoforms are significantly higher in the pancreas and colon, whereas levels of *GRα-D* are highest in spleen and lungs) [22]. *SLC25A32* and *TANK* have the greatest similarity values to *NR3C1* (0.880 and 0.877 respectively), which is evident intuitively based on their overall similar expression patterns across the 53 tissues (Figure 2).

Prediction of genes with similar expression profiles

The framework for predicting genes with similar expression profiles was based on promoter scans with each TFBS, followed by the derivation of the spatial density- and information density-related features from clusters in each promoter for genes with an *NR3C1*-like expression pattern (as shown in Figure 1A and 1B). We investigated two versions of this framework, depending on whether promoter sequences were first intersected with DHSs. Under both scenarios, all classifiers (Naïve Bayes, two types of Decision trees and three types of Support vector

machines (SVM)) were applied to both the training and test sets, successfully distinguishing similar from dissimilar genes in terms of expression profiles (i.e. accuracy, sensitivity and specificity all > 0.5) (Table 1, Additional file 5). We found, however, that generally all TFBSs in a DHS formed a binding site cluster, and the performance of all classifiers were significantly improved by inclusion of DHS information (i.e. accuracy, sensitivity and specificity were all increased) (Table 1, Additional file 5). The SVM classifier with the RBF kernel and the Random Forest classifier were the only two classifiers with accuracies exceeding 0.97, and each performed equally well on both the training and test sets (Table 1).

Prediction of differentially expressed direct TF targets

Between the two classifiers with the best performance in distinguishing genes with similar expression profiles to *NR3C1* from others (i.e. SVM with RBF kernel and Random Forest), we used a Random Forest (RF) classifier to predict direct TF targets respectively based on the CRISPR- [15] and siRNA-generated [13] perturbation data, because the SVM classifier with the RBF kernel did not perform as well (Additional file 5).

After eliminating TFBSs in inaccessible promoter intervals, i.e. those excluded from tissue-specific DHSs, the RF classifier predicted direct targets with greater accuracy and specificity (Table 2 and 3, Additional file 5). Specifically, predictions based on CRISPR-generated knockdown data for TFs: EGR1, ELK1, ELF1, ETS1, GABPA, and IRF1 were more accurate than for YY1, which itself represses or activates a wide range of promoters by binding to sites overlapping the TSS (Table 2, Additional file 5). Accordingly, the perturbation data indicated that YY1 has ~3-23 times more targets in the K562 cell line than the other TFs (ε = 1.05), and its binding has a more significant impact on the expression levels of target genes (for YY1, the ratio of the target counts at ε = 1.1 vs ε = 1.01 was 0.328, which significantly exceeded those of the other TFs (0.019-0.081); Additional file 3). This is concordant with our previous finding that YY1 extensively interacts with 11 cofactors (e.g. DNA-binding IRF9 and TEAD2; non-DNA-binding

287 DDX20 and PYGO2) in K562 cells, consistent with a central role in specifying erythroid-specific
288 lineage development [3].

289 Despite a high accuracy of target recognition, sensitivity was consistently higher than
290 specificity (Table 2, Additional file 5), implying that the classifier more effectively identified direct
291 targets compared to non-targets. This is attributable to the fact that the promoters of false
292 positive target genes also contain accessible, but non-functional TFBSs. In vivo co-regulation
293 mediated by interacting cofactors, which was excluded by the classifier, assisted in
294 distinguishing these non-functional sites that do not significantly affect gene expression [3,13].

295 As the threshold ε increased, the accuracy of the classifier monotonically increased on the
296 training sets of all the TFs (Figure 3) as expected. For a gene to be defined as a DE target of a
297 TF, the average fold change in its expression level for all guide RNAs that downregulated the
298 TF were required to reach the minimum threshold ε. Upon TF knockdown, higher ε is inversely
299 correlated with the number of target genes, but positively correlated with larger fold changes in
300 their corresponding expression levels. In general, more significantly DE genes have been
301 associated with a higher number of TFBSs in their promoters [13]. Thus, at greater ε, there are
302 larger differences in the values of machine learning features derived from TFBS clusters
303 between direct targets and non-targets (Additional File 1). Note that this inference holds valid
304 only when taking all direct targets and non-targets of a TF into account; it may not be true for a
305 specific pair of genes (i.e. the promoter of a gene that is not a DE target may contain a greater
306 number of accessible, but non-functional TFBSs) (Additional File 1). We noted this trend on the
307 test sets of only ELF1 and IRF1 (Figure 3); for the other five TFs (EGR1, ELK1, ETS1, GABPA,
308 YY1), differences in the clustered TFBS counts between targets and non-targets did not
309 necessary increase with larger values of ε, since the test set consists of both targets and non-
310 targets in equal proportions (Additional File 1). However, the classifier performed well in each

311 instance, because the count differences were still sufficiently large to discriminate between
312 targets and non-targets (Figure 3).

313 With the siRNA-generated knockdown data, the performance of the RF classifier was
314 compared to an approach inferring DE targets by correlating TF binding with gene expression
315 levels across ten cell types [14]. In this correlation-based approach, three measures (i.e. the
316 absolute Pearson correlation coefficient (PC), the absolute Spearman correlation coefficient
317 (SC), and the absolute combined angle ratio statistic (CARS)), whose performance was
318 evaluated with precision-recall curves, were alternatively used to compute a correlation score
319 between the number of ChIP-seq peaks overlapping the promoter and gene expression values.
320 Genes predicted to be DE targets had scores above the threshold resulting in a 1.5-fold
321 increase compared to the background precision. For example, in the case of the TF YY1, which
322 was analyzed by both approaches, the performance of the RF classifier on the training set was
323 0.66 (precision) and 0.456 (recall), and the test set was 0.672 and 0.396 (Table 3). This
324 classifier outperformed all three correlation measures (PC: 0.467 and 0.003; SC: 0.467 and
325 0.006; CARS: 0.467 and 0.003), even though the correlation approach used a less stringent P-
326 value threshold (0.05) for defining differential expression of likely non-direct targets, and
327 intersected ChIP-seq peaks over shorter 5kb promoter intervals upstream of the TSS.

328 **Intersection of genes with similar expression profiles and direct targets**

329 To determine how many direct targets have similar tissue-wide expression profiles, we
330 intersected the set of targets with the set of 500 genes with the most similar expression profiles
331 for each TF (Table 4, Additional file 6). The TFs PAX5 and POU2F2 are primarily expressed in
332 B cells, and their respective targets *IL21R* and *CD86* are also B cell-specific, which accounts for
333 the high similarity in the expression profile between them. There are respectively 21 and 7
334 nuclear mitochondrial genes (e.g. *MRPL9* and *MRPS10*, which are subunits of mitochondrial
335 ribosomes) in the intersections for YY1 in the K562 and GM19238 cell lines [25]. Previous

studies reported that YY1 upregulates a large number of mitochondrial genes by complexing with PGC-1α in C2C12 cells [26], and genes involved in the mitochondrial respiratory chain in K562 cells [15], which is consistent with the idea that YY1 may broadly regulate mitochondrial function (within all 53 tissues in addition to the erythrocyte, lymphocyte and skeletal muscle cell lines).

Between 0.4%-25% of genes with similar expression profiles to the TFs are actually direct targets (Table 4); the majority are non-targets whose promoters contain non-functional binding sites that are distinguished from targets by their lack of coregulation by corresponding cofactors. For YY1 and EGR1, we validated this hypothesis by contrasting the flanking cofactor binding site distributions and strengths in the promoters of the most similarly expressed target genes (YY1: *MRPL9*, *BAZ1B*; EGR1: *CANX*, *NPM1*) and non-target genes (YY1: *ADNP*, *RNF25*; EGR1: *AC142293.3*, *AP000705.7*). Strong and intermediate recognition sites for TFs: SP1, KLF1, CEBPB formed heterotypic clusters with adjacent YY1 sites; as well TFBSs of SP1, KLF1, and NFY were frequently present adjacent to EGR1 binding sites. These patterns contrasted with the enrichment of CTCF and ETS binding sites in gene promoters of YY1 and EGR1 non-targets (Additional file 7). Previous studies have reported that KLF1 is essential for terminal erythroid differentiation and maturation [27], direct physical interactions between YY1 and the constitutive activator SP1 synergistically induce transcription [28], the activating CEBPB promotes differentiation and suppresses proliferation of K562 cells by binding the promoter of the G-CSFR gene encoding a hematopoietin receptor [29], EGR1 and SP1 synergistically cooperate at adjacent non-overlapping sites on EGR1 promoter but compete binding at overlapping sites [30]; whereas CTCF functions as an insulator blocking the effects of *cis*-acting elements and preventing gene activation [31], and ETV6, a member of the ETS family, is a transcriptional repressor required for bone marrow hematopoiesis and associated with leukemia development [32].

Mutation analyses on promoters of direct targets

Because the promoters of most direct targets contain multiple binding site clusters, we anticipate that this enables these genes' expression to be naturally robust against binding site mutations; in other words, the other clusters can compensate for the loss of a cluster destroyed by mutations in binding sites, so that the mutated promoters are still capable of effectively inducing gene transcription upon TF binding. First, we validated this hypothesis by examining whether introducing artificial variants into binding sites in the promoter of the target gene *MCM7* in the test set of EGR1 changes the classifier output (Figure 4). Specifically, in the K562 cell line, *MCM7* is upregulated by EGR1. Knockdown of *MCM7* has an anti-proliferative and pro-apoptotic effect on K562 cells [33] and the loss of EGR1 increases leukemia initiating cells [34], which suggests that EGR1 may act as a tumor suppressor in K562 cells through the *MCM7* pathway.

First, the strongest binding site (at position chr7:100103347 [hg38], - strand, R_i = 12.0 bits) in the promoter was eliminated by a G->A mutation, resulting in the disappearance of Cluster 1, which consists of two sites (the other site at chr7:100103339, -, 4.3 bits). EGR1 was still predicted to compensate for this mutation, due to the presence of the other two clusters comprising weaker binding sites of intermediate strength (chr7:100102252, +, 7.0 bits; chr7:100102244, +, 1.3 bits; chr7:100101980, +, 8.9 bits; chr7:100101977, +, 2.2 bits; chr7:100101984, +, 1.9 bits), enabling the promoter to maintain capability of inducing *MCM7* expression (Figure 4). These adjacent clustered sites, which may not be strong enough to bind TFs and individually activate transcription, can stabilize each other's binding [2]. The weaker sites flanking a strong binding site in a cluster can direct the TF molecule to the strong site and extend the period of the molecule physically associating with the strong site, which is termed, the funnel effect [2]. Further, Clusters 2 and Cluster 3 were respectively removed by G->T and C->G mutations abolishing the strongest site in either cluster, which altered the prediction, that

386 is, EGR1 lost the capability to induce *MCM7* transcription (Figure 4). The remaining four sparse
387 weak sites do not form a cluster and cannot completely supplant the disrupted strong sites.

388 Further, we examined the impacts of known natural SNPs on binding site strengths, clusters
389 and the regulatory state of the promoter for a direct target of each of the seven TFs from the
390 CRISPR-generated perturbation data (Table 5). Often a single SNP (e.g. rs996639427 of EGR1)
391 can affect the strengths of multiple binding sites (Table 5). Apart from SNPs that are predicted
392 to abolish binding (Figure 4), leaky variants that merely weaken TF binding are common (Table
393 5). Binding stabilization between adjacent sites and the funnel effect enable the CRMs
394 comprised of information-dense clusters to be robust to mutations in individual binding sites. In
395 this way, neither mutations that abolish TFBSs nor leaky SNPs in flanking weak sites can
396 destroy functional clusters (e.g. rs1030185383 and rs5874306 of IRF1), whereas SNPs with
397 large reductions in R_i values of central strong sites are more likely to abolish clusters (e.g.
398 rs865922947, rs946037930, rs917218063 and rs928017336 of YY1) (Table 5). More generally,
399 the presence of multiple clusters enables promoters to be effectively resilient to the effects
400 binding site mutations; only the complete abolishment of all clusters resulting from the
401 simultaneous occurrence of multiple SNPs can transform the promoter to be unresponsive to TF
402 binding to residual weak sites (e.g. rs997328042, rs1020720126 and rs185306857 of GABPA)
403 (Table 5). Furthermore, a relatively small number of SNPs that strengthen TF binding and
404 eventually amplify the regulatory effect of the TF on the gene expression level are also present
405 (e.g. rs887888062 of EGR1 and rs751263172 of ELF1) (Table 5), suggesting that, in addition to
406 deleterious mutations, benign variants may also be found in promoters, consistent with the
407 expectations of neutral theory [35].

408 **DISCUSSION**

409 In this study, the Bray-Curtis Similarity function was initially shown (for the *NR3C1* gene) to
410 measure the relatedness of overall expression patterns between genes across a diverse set of

411 tissues. The resulting machine learning framework distinguished similar from dissimilar genes
412 based on the distribution, strengths and compositions of TFBS clusters in accessible promoters,
413 which can substantially account for the corresponding gene expression patterns. Using
414 knockdown data as the gold standard, the combinatorial use of TF binding profiles and
415 chromatin accessibility was also demonstrated to be predictive of DE direct TF targets. A
416 binding site comparison confirmed that coregulatory cofactors are responsible for distinguishing
417 between functional sites in targets and non-functional ones in non-targets. Furthermore,
418 mutation analyses on binding sites of targets demonstrated that the existence of both multiple
419 TFBSs in a cluster and multiple information-dense clusters in a promoter enables both the
420 cluster and the promoter to be resilient to binding site mutations.

421 The Random Forest classifier improved after intersecting promoters with DHSs in both
422 prediction of genes with similar expression profiles to *NR3C1* and prediction of direct TF targets
423 (Table 1, 2 and 3, Additional file 5). This intersection eliminated noisy binding sites that are
424 inaccessible to TF proteins in promoters; specifically, it widened discrepancies in feature vectors
425 between TPs and TNs. If the 10kb promoter of a gene instance does not overlap DHSs, its
426 feature vector will only consist of 0; the percentages of TNs whose promoters do not overlap
427 DHSs considerably exceeded those of TPs (Additional file 8), which led to an excess of TN
428 feature vectors containing only 0 after intersection. This explains why these TNs are not
429 functional targets of the TFs in the K562 and GM19238 cell lines, because their entire
430 promoters are not open to TF molecules; other regulatory regions besides the proximal
431 promoters (e.g. distal enhancers) still enable the TFs to effectively control the expression of the
432 TPs with inaccessible promoters.

433 The relatively poor performance of the classifier on YY1 (Table 2) is attributable to its smaller
434 percentage of TNs with inaccessible promoters (Additional file 8). Additionally, the Random
435 Forest classifier was more predictive of functional TF binding on the CRISPR-generated

436 knockdown data than the siRNA-generated ones (Table 2 and 3). This larger discrepancy in
437 feature vectors between TPs and TNs from CRISPR-based perturbations is also attributable to
438 the greater differences in the percentages between TPs and TNs with inaccessible promoters
439 (Additional file 8). Among the 22,046 genes whose expression levels were measured in the
440 CRISPR-based perturbations, most of the TNs with inaccessible promoters merely have one
441 transcript and specific functions (e.g. *VENTXP1* for the TF, EGR1), whereas many such TNs
442 were excluded from the 8,872 genes whose knockdown data were generated by siRNA
443 inactivation.

444 Our mutation analyses revealed that some deleterious TFBS mutations could be
445 compensated for by other information-dense clusters in a promoter; thus predicting the effects
446 of mutations in individual binding sites would not be sufficient to interpretation of downstream
447 effects. Though compensatory clusters may maintain gene expression, the promoter will provide
448 lower levels of activity than the wild-type promoter could, which is a recipe for achieving natural
449 phenotypic diversity. Few published studies in molecular diagnostics have specifically examined
450 the effects of naturally occurring variants within clustered TFBSs; thus IDBC-based machine
451 learning provided an alternative computational approach to predict deleterious mutations that
452 actually impact (i.e. repress or abolish) transcription of target genes and result in abnormal
453 phenotypes, and to simultaneously minimize false positive calls of TFBS mutations that
454 individually have little or no impact.

455 Apart from these TFs, the Bray-Curtis Similarity can be directly applied to identify the ground-
456 truth genes with overall similar tissue-wide expression patterns to any other gene whose
457 expression profile is known. Further studies could investigate the biological significance
458 underlying the phenomenon that all these genes share a common expression pattern, including
459 the similarity between other regulatory regions besides proximal promoters in terms of TFBSs
460 and epigenetic markers. This machine learning framework can also be applied to predict direct

DE targets for other TFs and in other cell lines, depending on the availability of corresponding knockdown data.

There are a number of limitations of our approach. The Bray-Curtis function seems unable to accurately measure the similarity between gene expression profiles of a ubiquitously expressed gene (e.g. *NR3C1*) and a tissue-specific gene (e.g. stomach-specific *PGA3*), which exhibit quite different tissue-wide expression patterns (i.e. $sim_{Bray-Curtis}(NR3C1, PGA3) = 0.007$). Intuitively, in terms of expression patterns *PGA3* is more similar to a gene (e.g. *MIR23A*) without any detectable mRNA in any of the 53 tissues analyzed than *NR3C1*; however, the Bray-Curtis similarity values indicate that both *PGA3* and *NR3C1* bear no similarity to *MIR23A* (i.e. $sim_{Bray-Curtis}(NR3C1, MIR23A) = sim_{Bray-Curtis}(PGA3, MIR23A) = 0$). Another possible limitation in classifier performance in the prediction of genes with similar tissue-wide expression profiles is that only binding sites of 82 TFs were analyzed due to a lack of available iPWMs for other TFs, given that 2000-3000 sequence-specific DNA-binding TFs are estimated to be encoded in the human genome [36]. For example, four TFs (CREB, MYB, NF1, GRF1) were previously reported to bind the promoter of the *NR3C1* gene to activate or repress its expression, however their iPWMs exhibiting known primary motifs could not be successfully derived from ChIP-seq data [3,22]. Regarding the CRISPR-generated knockdown data used here, TPs were inferred to be direct targets by intersecting promoters with their corresponding ChIP-seq peaks, which may not be completely accurate, due to the presence of noise peaks that do not contain true TFBSs [3,37]. In instances where small fold changes in the expression levels of DE targets were evident, these peaks could arise from compromised efficiency of knockdowns as a result of suboptimal guide RNAs or the limitations of perturbing only a single allele of the TF. Finally, the framework developed here only takes into account the 10kb interval proximal to the TSS, and would not therefore capture long range enhancer effects beyond this

485 distance; by contrast, correlation based approaches have successfully incorporated multiple
486 definitions of promoter length [14].

487 CONCLUSIONS

488 The Bray-Curtis similarity measure is able to effectively identify genes with similar tissue-
489 wide expression profiles. By analysis of promoter information theory-based TF binding profiles
490 that captured the spatial distribution and information contents of TFBS clusters, ChIP-seq and
491 chromatin accessibility data, we described a machine learning framework that distinguished
492 tissue-wide expression profiles of similar vs dissimilar genes and identified direct DE targets of
493 TFs. Functional binding sites in target genes that significantly alter expression levels upon direct
494 binding are also distinguished by TF-cofactor coregulation from non-functional sites in non-
495 targets. Finally, depending on how multiple TFBSs are organized in information-dense clusters
496 in target gene promoters, sequence variations in these binding sites may be protective, i.e.
497 resilient to dysregulation or, if deleterious, abrogate their normal transcriptional programs.

498 LIST OF ABBREVIATIONS

499 TF: transcription factor, TFBS: transcription factor binding site, CRM: *cis*-regulatory modules,
500 iPWM: information theory-based position weight matrix, IDBC: information density-based
501 clustering, ChIP-seq: chromatin immunoprecipitation with massively parallel DNA sequencing,
502 HM: histone modification, mRNA: messenger RNA, siRNA: small interfering RNA, CRISPR:
503 clustered regularly interspaced short palindromic repeats, DHS: deoxyribonuclease I
504 hypersensitive region, TP: true positive, TN: true negative, RPKM: reads per kilobase of
505 transcript per million mapped reads, GTEx: genotype-tissue expression, ENCODE:
506 encyclopedia of DNA elements, TSS: transcription start site, SVM: support vector machine, RBF:
507 radial basis function, PC: absolute Pearson correlation coefficient, SC: the absolute Spearman

508 correlation coefficient, CARS: the absolute combined angle ratio statistic, SNP: single
509 nucleotide polymorphism.

510 **ADDITIONAL FILES**

511 **Additional file 1:** The workflow of the IDBC algorithm, the mathematical definitions of five
512 statistical variables to measure classifier performance, and the correlation between ε values and
513 the RF classifier accuracy

514 Format: .docx

515 **Additional file 2:** The lists of TPs and TNs in the machine learning classifiers to predict genes
516 with similar tissue-wide expression profiles

517 Format: .xlsx

518 **Additional file 3:** The lists of TPs and TNs in the Random Forest classifier to predict DE direct
519 targets based on the CRISPR-generated knockdown data

520 Format: .xlsx

521 **Additional file 4:** The lists of TPs and TNs in the Random Forest classifier to predict DE direct
522 targets based on the siRNA-generated knockdown data

523 Format: .xlsx

524 **Additional file 5:** The classifier native performance leaving out intersecting promoters with
525 DHSs, and the SVM classifier performance on knockdown data

526 Format: .xlsx

527 **Additional file 6**: The list of the most similar 500 genes to each TF in terms of expression
528 profiles, and the intersection of these 500 genes and DE direct targets of the TF

Format: .xlsx

Additional file 7: Cofactor binding sites adjacent to YY1 and EGR1 sites in the promoters of their targets and non-targets

Format: .docx

Additional file 8: The percentages of TPs and TNs whose promoters do not overlap DHSs

Format: .xlsx

DECLARATIONS

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The median RPKM, TSS coordinate, DNase I hypersensitivity and ChIP-seq data are respectively available from the GTEx Analysis V6p release (www.gtexportal.org), Ensembl Biomart (www.ensembl.org) and ENCODE (www.encodeproject.org). The CRISPR- and siRNA-generated knockdown data are available from the supplementary information files of Dixit et al. [15] and Cusanovich et al. [13]. The code implementing this machine learning framework is available in Zenodo (https://doi.org/10.5281/zenodo.1145458). All other data supporting the findings of this study are available within the article and its supplementary information files.

Competing interests

549 PKR is the inventor of US Patent 5,867,402 and other patents pending, which apply iPWMs to
550 the prediction and validation of mutations. He cofounded Cytognomix, Inc., which is developing
551 software based on this technology for complete genome or exome mutation analysis.

552 **Funding**

553 Natural Sciences and Engineering Research Council of Canada Discovery Grant [RGPIN-2015-
554 06290]; Canada Foundation for Innovation; Canada Research Chairs; Cytognomix Inc. Funding
555 for open access charge: University of Western Ontario and the Natural Sciences and
556 Engineering Research Council.

557 **Authors' contributions**

558 PKR defined the objectives and directed the study. RL and PKR devised the general machine
559 learning framework. RL implemented this framework and collected the results. Both RL and
560 PKR interpreted the results and wrote the manuscript.

561 **Acknowledgements**

562 We are grateful to Ben Shirley and Eliseos Mucaki for constructive comments on the paper.

563 **REFERENCES**

564 1. Hosseinpour B, Bakhtiarizadeh MR, Khosravi P, Ebrahimie E. Predicting distinct organization
565 of transcription factor binding sites on the promoter regions: a new genome-based approach to
566 expand human embryonic stem cell regulatory network. Gene. 2013;531:212–9.

567 2. Ezer D, Zabet NR, Adryan B. Homotypic clusters of transcription factor binding sites: A model
568 system for understanding the physical mechanics of gene expression. Comput. Struct.
569 Biotechnol. J. 2014;10:63–9.

570 3. Lu R, Mucaki EJ, Rogan PK. Discovery and validation of information theory-based
571 transcription factor and cofactor binding site motifs. Nucleic Acids Res. 2017;45:e27.

572 4. Bi C, Rogan PK. Bipartite pattern discovery by entropy minimization-based multiple local
573 alignment. Nucleic Acids Res. 2004;32:4979–91.

574 5. Dinakarpandian D, Raheja V, Mehta S, Schuetz EG, Rogan PK. Tandem machine learning
575 for the identification of genes regulated by transcription factors. BMC Bioinformatics. 2005;6:204.

576 6. Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and
577 differential gene expression in embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A.
578 2009;106:21521–6.

579 7. Cheng C, Alexander R, Min R, Leng J, Yip KY, Rozowsky J, et al. Understanding
580 transcriptional regulation by integrative analysis of transcription factor binding data. Genome
581 Res. 2012;22:1658–67.

582 8. Budden DM, Hurley DG, Cursons J, Markham JF, Davis MJ, Crampin EJ. Predicting
583 expression: the complementary power of histone modification and transcription factor binding
584 data. Epigenetics Chromatin. 2014;7:36.

585 9. Smith AD, Sumazin P, Xuan Z, Zhang MQ. DNA motifs in human and mouse proximal
586 promoters predict tissue-specific expression. Proc. Natl. Acad. Sci. U. S. A. 2006;103:6275–80.

587 10. McLeay RC, Lesluyes T, Cuellar Partida G, Bailey TL. Genome-wide in silico prediction of
588 gene expression. Bioinforma. Oxf. Engl. 2012;28:2789–96.

589 11. Karlić R, Chung H-R, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are
590 predictive for gene expression. Proc. Natl. Acad. Sci. U. S. A. 2010;107:2926–31.

591 12. Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, et al. Modeling gene
592 expression using chromatin features in various cellular contexts. Genome Biol. 2012;13:R53.

593 13. Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. The functional consequences of variation
594 in transcription factor binding. PLoS Genet. 2014;10:e1004226.

595 14. Banks CJ, Joshi A, Michoel T. Functional transcription factor target discovery via compendia
596 of binding and expression profiles. Sci. Rep. 2016;6:20649.

597 15. Dixit A, Parnas O, Li B, Chen J, Fulco CP, Jerby-Arnon L, et al. Perturb-Seq: Dissecting
598 Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. Cell.
599 2016;167:1853–1866.e17.

600 16. Cui S, Youn E, Lee J, Maas SJ. An improved systematic approach to predicting transcription
601 factor target genes using support vector machine. PloS One. 2014;9:e94519.

602 17. Bray JR, Curtis JT. An Ordination of the Upland Forest Communities of Southern Wisconsin.
603 Ecol. Monogr. 1957;27:325–349.

604 18. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat. Genet.
605 2013;45:580–5.

606 19. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human
607 genome. Nature. 2012;489:57–74.

608 20. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The
609 accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.

610 21. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining
611 Software: An Update. SIGKDD Explor Newsl. 2009;11:10–18.

612 22. Vandevyver S, Dejager L, Libert C. Comprehensive overview of the structure and regulation
613 of the glucocorticoid receptor. Endocr. Rev. 2014;35:671–93.

614 23. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human
615 genome browser at UCSC. Genome Res. 2002;12:996–1006.

616 24. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI
617 database of genetic variation. Nucleic Acids Res. 2001;29:308–11.

618 25. Calvo SE, Clauser KR, Mootha VK. MitoCarta2.0: an updated inventory of mammalian
619 mitochondrial proteins. Nucleic Acids Res. 2016;44:D1251-1257.

620 26. Cunningham JT, Rodgers JT, Arlow DH, Vazquez F, Mootha VK, Puigserver P. mTOR
621 controls mitochondrial oxidative function through a YY1-PGC-1alpha transcriptional complex.
622 Nature. 2007;450:736–40.

623 27. Tallack MR, Perkins AC. KLF1 directly coordinates almost all aspects of terminal erythroid
624 differentiation. IUBMB Life. 2010;62:886–90.

625 28. Seto E, Lewis B, Shenk T. Interaction between transcription factors Sp1 and YY1. Nature.
626 1993;365:462–4.

627 29. Ferrari-Amorotti G, Mariani SA, Novi C, Cattelani S, Pecorari L, Corradini F, et al. The
628 biological effects of C/EBPalpha in K562 cells depend on the potency of the N-terminal
629 regulatory region, not on specificity of the DNA binding domain. J. Biol. Chem. 2010;285:30837–
630 50.

631 30. Huang RP, Fan Y, Ni Z, Mercola D, Adamson ED. Reciprocal modulation between Sp1 and
632 Egr-1. J. Cell. Biochem. 1997;66:489–99.

633 31. Bell AC, West AG, Felsenfeld G. The protein CTCF is required for the enhancer blocking
634 activity of vertebrate insulators. Cell. 1999;98:387–96.

635 32. Wang LC, Swat W, Fujiwara Y, Davidson L, Visvader J, Kuo F, et al. The TEL/ETV6 gene is
636 required specifically for hematopoiesis in the bone marrow. Genes Dev. 1998;12:2392–402.

637 33. Tian L, Liu J, Xia G-H, Chen B-A. RNAi-mediated knockdown of MCM7 gene on CML cells
638 and its therapeutic potential for leukemia. Med. Oncol. Northwood Lond. Engl. 2017;34:21.

639 34. Maifrede S, Liebermann D, Hoffman B. Egr-1, a Stress Response Transcription Factor and
640 Myeloid Differentiation Primary Response Gene, Behaves As Tumor Suppressor in CML. Blood.
641 2014;124:2211.

642 35. Kimura M. The neutral theory of molecular evolution. Sci. Am. 1979;241:98–100, 102, 108
643 passim.

644 36. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human
645 transcription factors: function, expression and evolution. Nat. Rev. Genet. 2009;10:252–63.

646 37. Kidder BL, Hu G, Zhao K. ChIP-Seq: technical considerations for obtaining high-quality data.
647 Nat. Immunol. 2011;12:918–22.

FIGURE LEGENDS

Figure 1. General framework for predicting genes with similar tissue-wide expression profiles and DE direct TF targets

A) An overview of the machine learning framework. The steps enclosed in the dashed rectangle and for forming training and test sets vary across prediction of genes with similar expression profiles and DE direct TF targets. The step with a dash-dotted border that intersects promoters with DHSs is a variant of the primary approach that provided more accurate results. In the IDBC algorithm (Additional file 1), the parameter *I* is the minimum threshold on the total information contents of TFBS clusters. In prediction of genes with similar expression profiles, the minimum value was 939, which was the sum of mean information contents ($R_{sequence}$ values) of all 94 iPWMs; in prediction of direct targets, this value was the $R_{sequence}$ value of the single iPWM used to detect TFBSs in each promoter. The parameter *d* is the radius of initial clusters in base pairs, whose value, 25, was determined empirically. Eight types of three different classifiers were evaluated with statistics (accuracy, sensitivity and specificity) to measure the classifier performance (Additional file 1). **B)** Formation of the training and test sets for identifying genes with similar expression profiles to a given gene (Additional file 2). **C)** Formation of the training and test sets for predicting direct targets of seven TFs using the CRISPR-generated perturbation data in K562 cells (Additional file 3). **D)** Formation of the training and test sets for predicting direct targets of 11 TFs using the siRNA-generated knockdown data in GM19238 cells (Additional file 4). When genes with single transcripts were more than the TPs, those with the largest P-values were selected as TNs (null hypothesis of differential expression cannot be rejected); when genes with single transcripts were fewer than the TPs, those genes with two transcripts and the largest P-values were also selected. This step was iterated until the number of TNs equaled that of TPs.

673 **Figure 2. Expression profiles of NR3C1, SLC25A32 and TANK**

674 Visualization of the expression values (in RPKM) of these genes across 53 tissues from
675 GTEx. For each gene, the colored rectangle belonging to each tissue indicates the valid RPKM
676 of all samples in the tissue, the black horizontal bar in the rectangle indicates the median RPKM,
677 the hollow circles indicate the RPKM of the samples considered as outliers, and the grey vertical
678 bar indicates the sampling error. By comparing the pictures, the overall expression patterns of
679 the three genes across the 53 tissues resemble each other (e.g. all three genes exhibit the
680 highest expression levels in lymphocytes and the lowest levels in brain tissues).

681 **Figure 3. Accuracy of the Random Forest classifier when using three different values**
682 **for ε**

683 **A)** The accuracy of the classifier on the training sets of the TFs based on 10-fold cross
684 validation. Binding site clusters were derived intersecting promoters with DHSs, for different
685 minimum threshold ε values (i.e. 1.01, 1.05 and 1.1) corresponding to the average fold change
686 in gene expression levels under all guide RNAs of the TF. **B)** The accuracy on the test sets. As
687 ε increased, accuracy on the training sets also increased.

688 **Figure 4. Mutation analyses on the target MCM7 in the test set of EGR1**

689 This figure depicts the effect of a mutation in each EGR1 binding site cluster of the MCM7
690 promoter on the expression level of MCM7, which is a target of the TF EGR1. The strongest
691 binding site in each cluster were abolished by a single nucleotide variant. Upon loss of all three
692 clusters, only weak binding sites remained and EGR1 was predicted to no longer be able to
693 effectively regulate MCM7 expression. Multiple clusters in the promoters of TF targets confers
694 robustness against mutations within individual binding sites that define these clusters.

695 **TABLES**

696 **Table 1. Performance of machine learning classifiers for predicting genes with similar**

697 **expression profiles to NR3C1**

Classifier		After intersecting promoters with DHSs					
		Training set[§]			Test set		
		Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
Naïve Bayes		0.964	0.992	0.936	0.956	0.968	0.944
Decision tree	J48 tree	0.960	0.960	0.960	0.970	0.952	0.988
	Random tree	0.936	0.940	0.932	0.936	0.912	0.960
	Random forest[†]	**0.972**	**0.976**	**0.968**	**0.976**	**0.964**	**0.988**
SVM	**RBF kernel[†]**	**0.972**	**0.976**	**0.968**	**0.976**	**0.964**	**0.988**
	Polynomial kernel of exponent 1	0.964	0.960	0.968	0.968	0.944	0.992
	Polynomial kernel of exponent 2	0.976	0.968	0.984	0.964	0.936	0.992
	Polynomial kernel of exponent 3	0.960	0.932	0.988	0.958	0.920	0.996

698 [†] The two best-performing classifiers were bolded.

699 [§] The results on the training set was obtained using 10-fold cross validation.

700

Table 2. The Random Forest classifier performance for predicting direct TF targets using the CRISPR-generated data

TF[†]	After intersecting promoters with DHSs					
	Training set[§]			Test set		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
EGR1	0.879	0.943	0.816	0.845	0.954	0.736
ELF1	0.846	0.923	0.769	0.863	0.900	0.825
ELK1	0.862	0.897	0.828	0.793	0.948	0.638
ETS1	0.810	0.912	0.708	0.779	0.899	0.659
GABPA	0.819	0.932	0.706	0.770	0.94	0.600
IRF1	0.792	0.860	0.725	0.735	0.860	0.610
YY1	0.595	0.559	0.631	0.587	0.535	0.638

[†] The results for all seven TFs were obtained when setting ε to 1.05, and the transcriptome data generated by CRISPR-based TF knockdowns were obtained from Dixit et al [15].

[§] The results on the training sets was obtained using 10-fold cross validation.

Table 3. The Random Forest classifier performance for predicting direct TF targets using the siRNA-generated data

TF[†]	After intersecting promoters with DHSs					
	Training set[§]			Test set		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
BATF	0.625	0.646	0.604	0.706	0.649	0.763
JUND	0.625	0.646	0.604	0.682	0.682	0.682
NFE2L1	0.633	0.533	0.733	0.75	0.667	0.833
PAX5	0.575	0.614	0.537	0.627	0.563	0.691
POU2F2	0.725	0.818	0.633	0.651	0.796	0.505
RELA	0.591	0.619	0.563	0.690	0.611	0.770
RXRA	0.731	0.813	0.648	0.663	0.793	0.533
SP1	0.561	0.571	0.551	0.579	0.539	0.620
TCF12	0.564	0.638	0.491	0.684	0.597	0.770
USF1	0.737	0.753	0.721	0.723	0.71	0.735
YY1	0.611	0.456	0.765	0.601	0.396	0.807

[†] The transcriptome data generated by siRNA-based TF knockdowns were obtained from Cusanovich et al [13].

[§] The results on the training sets was obtained using 10-fold cross validation.

Table 4. Intersection of direct targets and 500 genes with the most similar expression profiles

TF	Cell line	Number of targets	Size of intersection	Targets among the most similar 10 genes[§]
EGR1	K562	174	11	None
ELF1		79	5	None
ELK1		116	4	GNL1(8^{th})
ETS1		275	14	None
GABPA		530	24	TAF1(1^{st})
IRF1		472	11	None
YY1		1797	125	MRPL9(2^{nd}), BAZ1B(6^{th}), ENY2(7^{th}), NUB1(8^{th}), USP1(9^{th}), HNRNPR(10^{th})
	GM19238	1066	61	MED4(1^{st}), SURF6(3^{rd}), BAZ1B(6^{th})
BATF		193	4	MB21D1(4^{th}), C16orf87(9^{th})
JUND		44	2	None
NFE2L1		60	3	None
RELA		252	22	HMG20B(9^{th})
RXRA		183	7	None
SP1		1630	96	ACLY(1^{st}), SEC22B(7^{th}), GPX1P1(10^{th})
TCF12		669	19	None
USF1		309	20	None
PAX5		938	76	IL21R(9^{th})
POU2F2		550	21	CD86(3^{rd})

[§] The rank of each target in the list of similar genes in the descending order of Bray-Curtis similarity values is shown in the brackets immediately following the target.

717 **Table 5. Mutation analyses on promoters of direct targets**

TF	Target	Normal cluster	Normal allele[§]	SNP ID[§]	Variant allele[§]	Variant cluster[‡]	Classifier output		
							Variant[†]		Wild-type
EGR1 ($R_{sequence}$ = 12.2899 bits)	EID2B	Cluster 1 of 2	***G***A***G***GGGGC***AT***C (chr19:39540286, -, 7.22 bits)	rs538610162 (chr19:39540296C>G)	***C***AGGGGGCATC (chr19:39540286, -, 4.84 bits)	Abolished	√	×	√
				rs759233998 (chr19:39540294C>T)	GA***A***GGGGCATC (chr19:39540286, -, 0.06 bit)	Abolished	√		
				rs974735901 (chr19:39540288T>A)	GAGGGGGC***T***TC (chr19:39540286, -, 6.90 bits)	Cluster 1 of 2	√		
				rs978230260 (chr19:39540287A>T)	GAGGGGGCA***A***C (chr19:39540286, -, 5.31 bits)	Abolished	√		
		Cluster 2 of 2	***G***CGTGCGT***G***GG (chr19:39540162, +, 1.59 bits)	rs764734511 (chr19:39540162G>A) (chr19:39540162G>C)	***A***CGTGCGTGGG (chr19:39540162, +, -0.72 bit)	Cluster 2 of 2	√		
					CCGTGCGTGGG (chr19:39540162, +, -0.79 bit)	Cluster 2 of 2	√		
				rs996639427 (chr19:39540170G>C)	GCGTGCGT***C***GG (chr19:39540162, +, -5.21 bits)	Abolished	√		
			GCGT***G***GGC***G***C***T*** (chr19:39540166, +, 9.72 bits)		GCGT***C***GGCGCT (chr19:39540165, +, -0.85 bit)				
				rs1027751538 (chr19:39540174G>A)	GCGTGGGC***A***CT (chr19:39540166, +, 5.16 bits)	Abolished	√		
				rs887888062 (chr19:39540176T>A)	GCGTGGGCGC***A*** (chr19:39540166, +, 10.94 bits)	Cluster 2 of 2	√		
ELF1 ($R_{sequence}$ = 11.2057 bits)	HIST1H4H	Cluster 1 of 2	GC***GGA***AG***CG***TG (chr6:26286540, +, 9.92 bits)	rs760968937 (chr6:26286547C>T) (chr6:26286547C>A)	GCGGAAG***T***GTG (chr6:26286540, +, 10.71 bits)	Cluster 1 of 2	√	√	√
					GCGGAAG***A***GTG (chr6:26286540, +, 8.84 bits)	Cluster 1 of 2	√	×	
				rs1000196206 (chr6:26286542G>C)	GC***C***GAAGCGTG (chr6:26286540, +, -6.26 bits)	Abolished	√		
				rs144759258 (chr6:26286543G>A)	GCG***A***AAGCGTG (chr6:26286540, +, -3.60 bits)	Abolished	√		
				rs966435996 (chr6:26286544A>G)	GCGG***G***AGCGTG (chr6:26286540, +, 5.28 bits)	Abolished	√		
				rs950986427 (chr6:26286548G>A)	GCGGAAGC***A***TG (chr6:26286540,	Cluster 1 of 2	√		

					+, 8.28 bits)				
		Cluster 2 of 2	***C***AG***GAG***ATGC***G*** (chr6:26286473, -, 6.98 bits)	rs373649904 (chr6:26286483G>A)	***T***AGGAGATGCG (chr6:26286473, -, 0.61 bit)	Abolished	√		
				rs926919149 (chr6:26286480C>T)	CAG***A***AGATGCG (chr6:26286473, -, -6.53 bits)	Abolished	√		
				rs751263172 (chr6:26286479T>G)	CAGG***C***GATGCG (chr6:26286473, -, 1.24 bits)	Abolished	√		
				rs369076253 (chr6:26286473C>G)	CAGGAGATGC***C*** (chr6:26286473, -, 6.92 bits)	Cluster 2 of 2	√		
				<u>rs751263172 (chr6:1044474314C>T)</u>	<u>CAGGA***A***ATGCG (chr6:26286473, -, 11.43 bits)</u>	Cluster 2 of 2	√	√	
ELK1 ($R_{sequence}$ = 11.9041 bits)	G0S2	Cluster 1 of 2	C***AG***GGAAG***A***CC (chr1:209667959, -, 1.92 bits)	<u>rs146048477 (chr1:209667961T>A)</u>	<u>CAGGGAAG***T***CC (chr1:209667959, -, 2.24 bits)</u>	Cluster 1 of 2	√	√	√
				rs887606802 (chr1:209667968T>C)	C***G***GGGAAGACC (chr1:209667959, -, -3.35 bits)	Cluster 1 of 2	√	×	
				rs1021034916 (chr1:209667967C>T)	CA***A***GGAAGACC (chr1:209667959, -, -3.57 bits)	Cluster 1 of 2	√		
			GAGGA***A***ATGAG (chr1:209667969, +, 8.14 bits)	rs941962117 (chr1:209667974A>G)	GAGGA***G***ATGAG (chr1:209667969, +, 4.11 bits)	Abolished	√		
		Cluster 2 of 2	***C***TGGAAGA***GC***A (chr1:209673544, -, 5.91 bits)	rs896117033 (chr1:209673545G>A)	CTGGAAGAG***T***A (chr1:209673544, -, 3.95 bits)	Cluster 2 of 2	√		
				rs971962577 (chr1:209673546C>T)	CTGGAAGA***A***CA (chr1:209673544, -, 3.49 bits)	Cluster 2 of 2	√		
				rs1011969709 (chr1:209673554G>C)	***G***TGGAAGAGCA (chr1:209673544, -, 3.92 bits)	Abolished	√		
			CCA***G***AAGTCA***A*** (chr1:209673551, +, 7.44 bits)		CCA***C***AAGTCAA (chr1:209673551, +, -5.50 bits)				
				<u>rs1023312090 (chr1:209673561A>G)</u>	<u>CCAGAAGTCA***G*** (chr1:209673551, +, 8.40 bits)</u>	Cluster 2 of 2	√	√	
ETS1 ($R_{sequence}$ = 10.0788 bits)	TTC19	Cluster 1 of 1	GCA***G***GGAA***A***GG (chr17:16022293, +, 7.92 bits)	rs1022234223 (chr17:16022296G>C)	GCA***C***GGAAAGG (chr17:16022293, +, -4.98 bits)	Abolished	×	×	√
				<u>rs968299415 (chr17:16022301A>T)</u>	<u>GCAGGGAA***T***GG (chr17:16022293, +, 10.01 bits)</u>	Cluster 1 of 1	√	√	
GABPA ($R_{sequence}$ =	PLEKHB2	Cluster 1 of 1	A***C***A***G***GAAAGGG (chr2:131112770,	rs997328042 (chr2:131112771C>T)	A***T***AGGAAAGGG (chr2:131112770,	Abolished	×	×	√

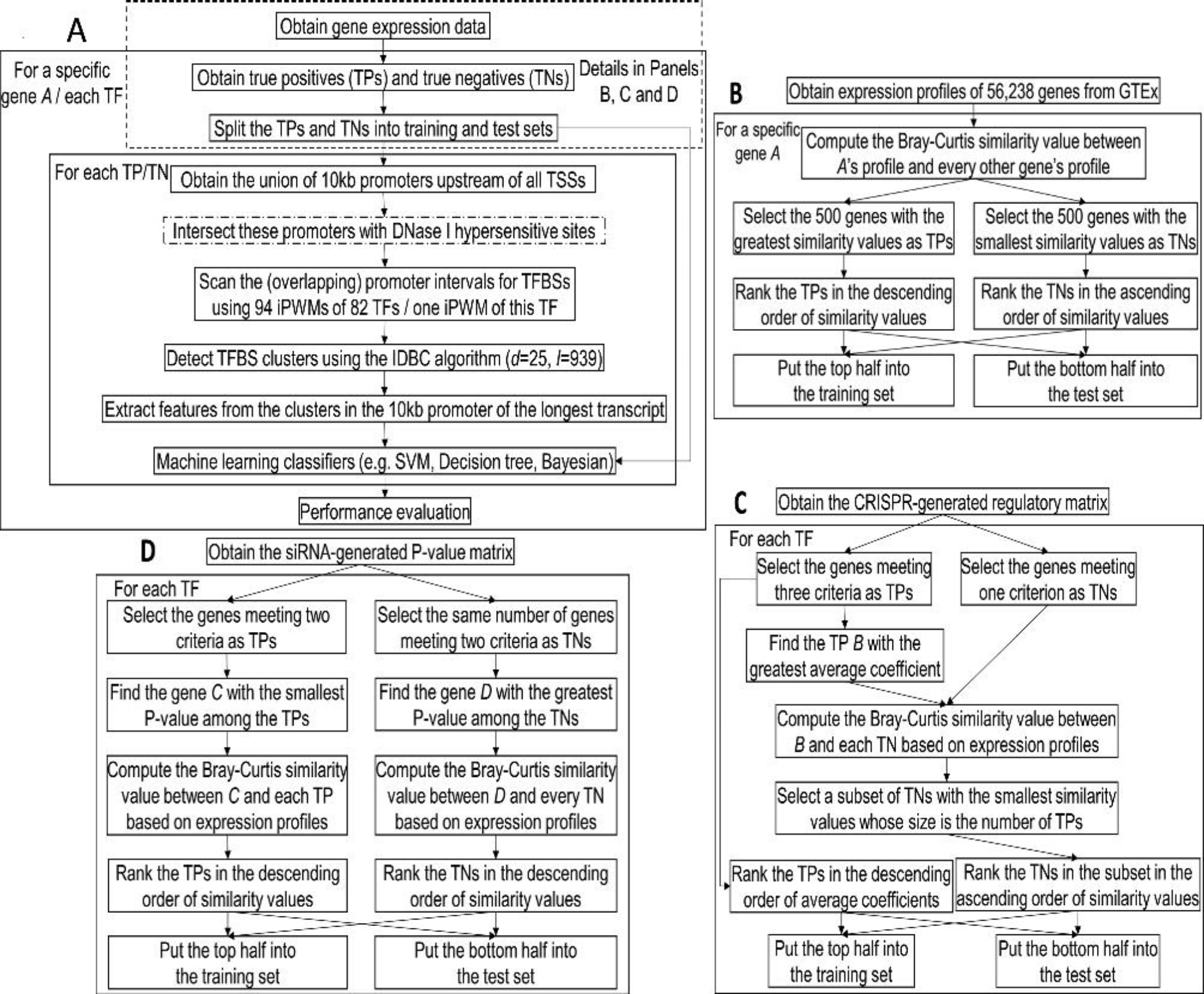
10.8567 bits)			+, 10.36 bits)		+, -3.68 bits)				
				rs1020720126 (chr2:131112773G>C)	ACA***C***GAAAGGG (chr2:131112770, +, -4.16 bits)	Abolished	×		
			T***CT***GGAAAC***T***A (chr2:131112760, +, 1.53 bits)	rs185306857 (chr2:131112761C>A)	T***A***TGGAAACTA (chr2:131112760, +, -2.86 bits)	Cluster 1 of 1	√		
				rs772728699 (chr2:131112762T>A)	TC***A***GGAAACTA (chr2:131112760, +, 5.23 bits)	Cluster 1 of 1	√		
				rs965753671 (chr2:131112769T>C)	TCTGGAAAC***C***A (chr2:131112760, +, 2.13 bits)	Cluster 1 of 1	√		
IRF1 ($R_{sequence}$ = 13.5544 bits)	SMIM13	Cluster 1 of 1	***GA***G***A***AT***G***AA***A***GC***A*** (chr6:11093663, +, 12.56 bits)	rs950528541 (chr6:11093663G>C)	***C***AGAATGAAAGCA (chr6:11093663, +, 8.97 bits)	Cluster 1 of 1	√	×	√
				rs886259573 (chr6:11093664A>G)	G***G***GAATGAAAGCA (chr6:11093663, +, 9.65 bits)	Cluster 1 of 1	√		
				rs982931728 (chr6:11093666A>G)	GAG***G***ATGAAAGCA (chr6:11093663, +, 8.09 bits)	Cluster 1 of 1	√		
				rs1020218811 (chr6:11093668T>G)	GAGAA***G***GAAAGCA (chr6:11093663, +, 9.36 bits)	Cluster 1 of 1	√		
				rs570723026 (chr6:11093672A>G)	GAGAATGAA***G***GCA (chr6:11093663, +, 8.01 bits)	Cluster 1 of 1	√		
				rs1004825794 (chr6:11093675A>C) (chr6:11093675A>T)	GAGAATGAAAGC***C*** (chr6:11093663, +, 10.47 bits)	Cluster 1 of 1	√		
					GAGAATGAAAGC***A*** (chr6:11093663, +, 10.42 bits)	Cluster 1 of 1	√		
			AA***G***ACCAA***AG***GCA (chr6:11093641, +, 2.43 bits)	rs1030185383 (chr6:11093649A>C)	AAGACCAA***C***GGCA (chr6:11093641, +, -3.39 bits)	Cluster 1 of 1	√		
				rs5874306 (chr6:11093650delG)	AAGACCAAAGCAG (chr6:11093641, +, 0.90 bit)	Cluster 1 of 1	√		
				rs558896490 (chr6:11093643G>A)	AA***A***ACCAAAGGCA (chr6:11093641, +, 7.06 bits)	Cluster 1 of 1	√	√	
YY1 ($R_{sequence}$ = 12.8554 bits)	CKLF	Cluster 1 of 1	***G***C***GG***C***C***ATCGGC (chr16:66549785, -, 10.06 bits)	rs865922947 (chr16:66549791G>A)	***C***CGGCCATCGGC (chr16:66549785, -, 6.80 bits)	Cluster 1	√	×	√
				rs946037930 (chr16:66549794C>A)	GC***T***GCCATCGGC (chr16:66549785, -, 8.02 bits)	Cluster 1	√		
				rs917218063 (chr16:66549793C>T)	GCG***A***CCATCGGC (chr16:66549785, -, 5.41 bits)	Abolished	×		

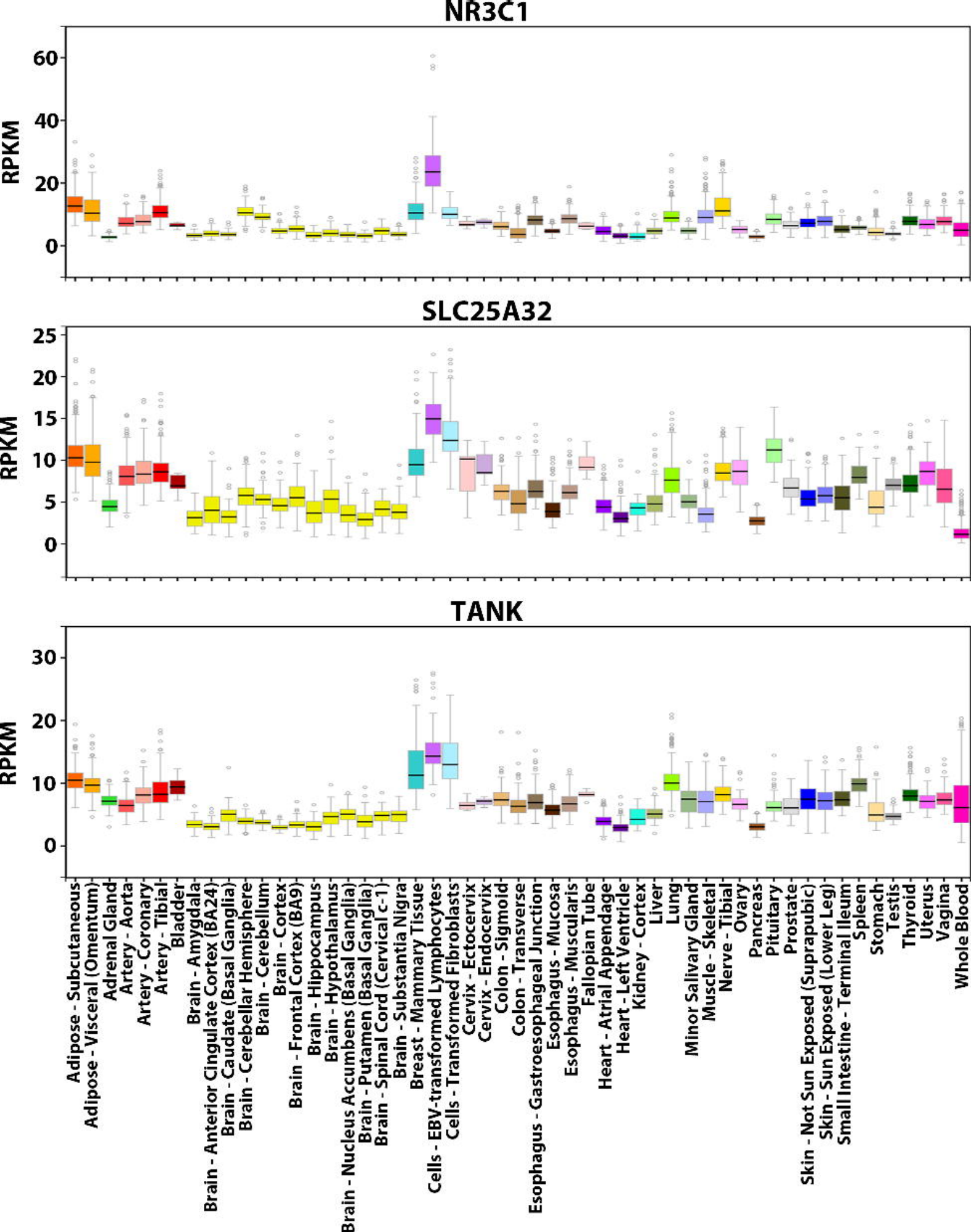
				rs928017336 (chr16:66549791G>A)	GCGGC***T***ATCGGC (chr16:66549785, -, -3.62 bits)	Abolished	×		
			GCCGCCCCCGTC (chr16:66549792, +, 1.34 bits)						

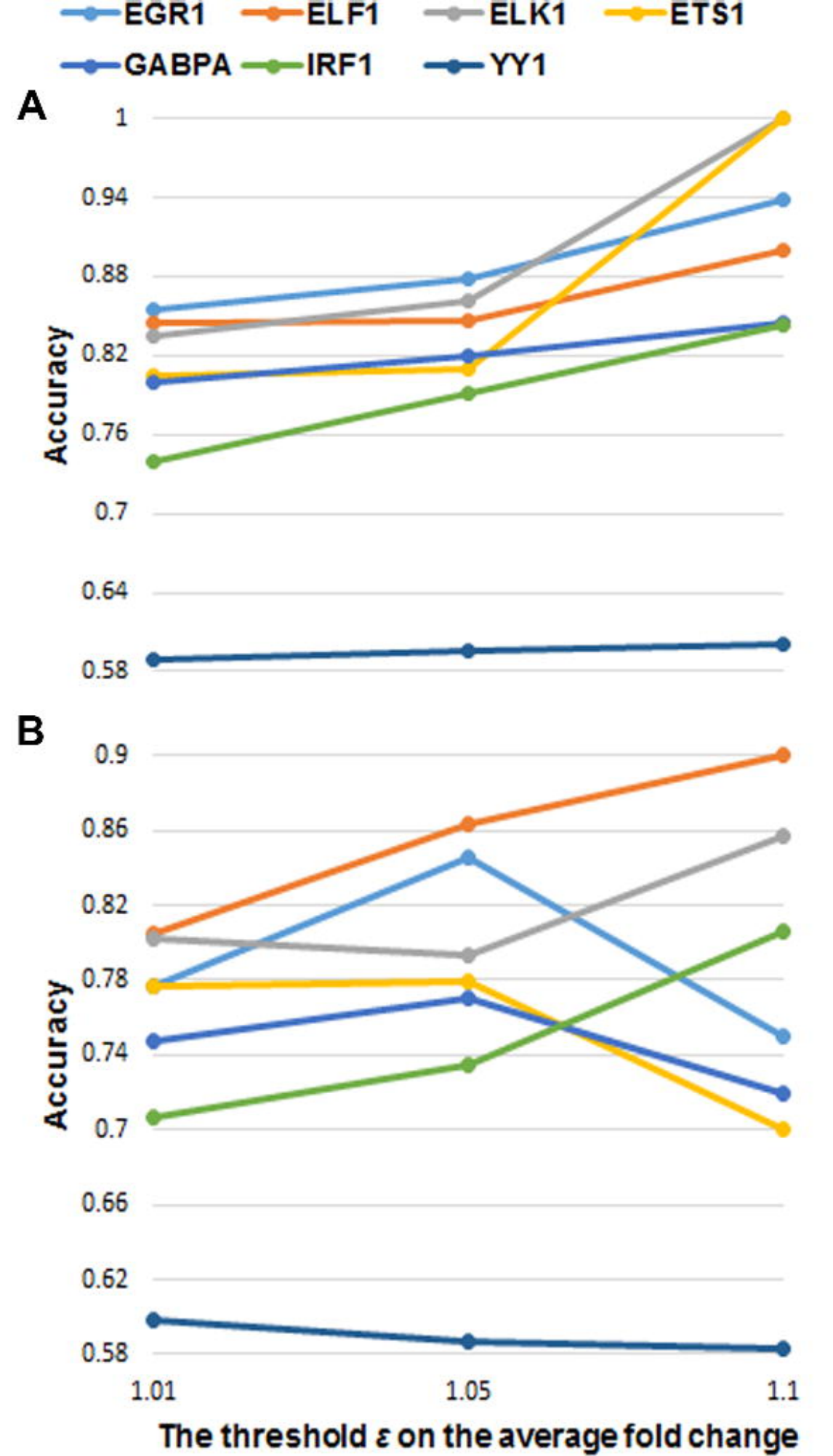
[§] All coordinates are based on the hg38 genome assembly. A bold italic letter in a binding site sequence indicates the base where a SNP occurs. The SNPs strengthening binding sites and corresponding variant binding site sequences are underlined.

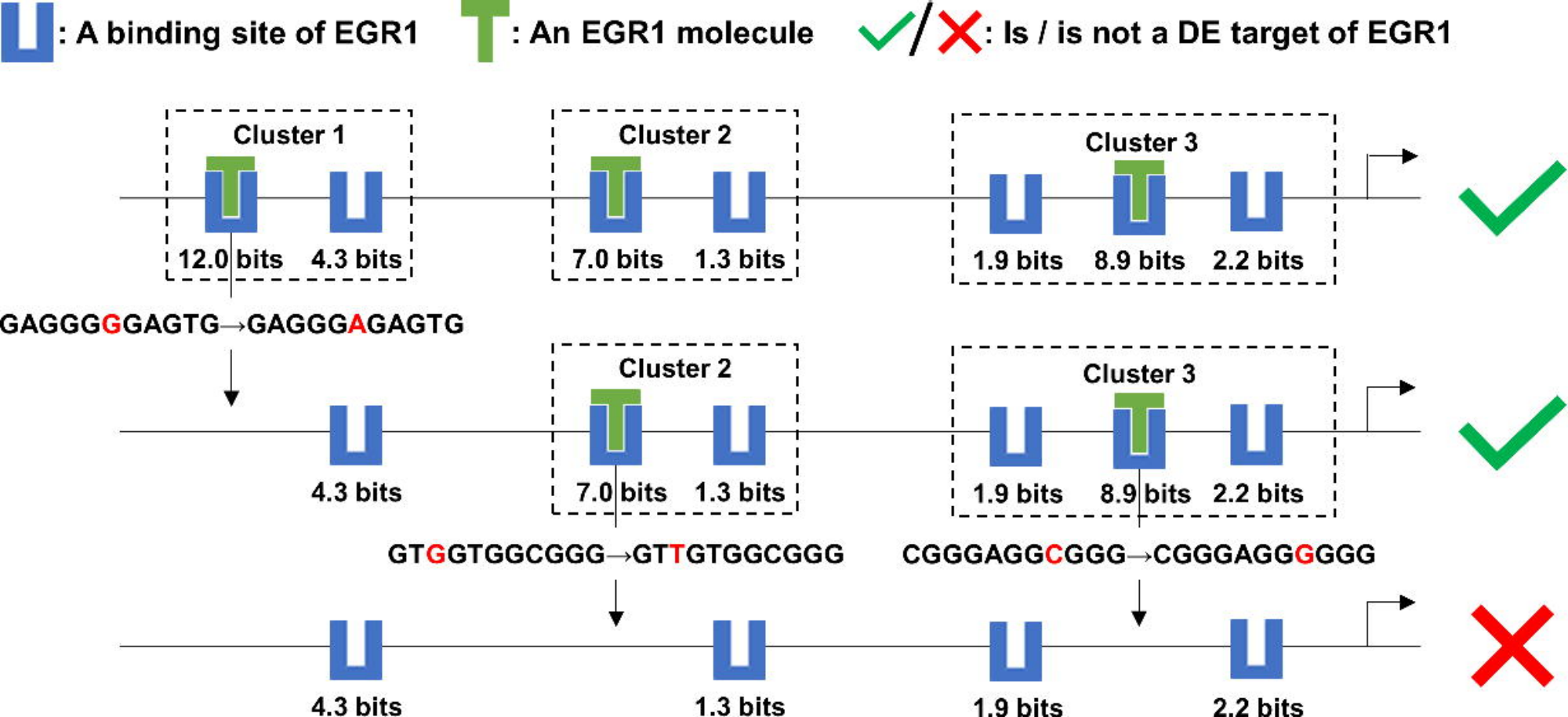
[‡] The impact on whether the occurrence of a single SNP resulted in the disappearance of the cluster containing it is shown.

[†] After a single SNP occurred or multiple SNPs simultaneously occurred, the classifier produced a new prediction on whether the TF is still capable of significantly affecting gene expression via the variant promoter.


A
Obtain gene expression data
For a specific gene A / each TF
Obtain true positives (TPs) and true negatives (TNs)
Details in Panels B, C and D
Split the TPs and TNs into training and test sets
For each TP/TN
Obtain the union of 10kb promoters upstream of all TSSs
Intersect these promoters with DNase I hypersensitive sites
Scan the (overlapping) promoter intervals for TFBSs using 94 iPWMs of 82 TFs / one iPWM of this TF
Detect TFBS clusters using the IDBC algorithm (d=25, l=939)
Extract features from the clusters in the 10kb promoter of the longest transcript
Machine learning classifiers (e.g. SVM, Decision tree, Bayesian)
Performance evaluation
B
Obtain expression profiles of 56,238 genes from GTEx
For a specific gene A
Compute the Bray-Curtis similarity value between A's profile and every other gene's profile
Select the 500 genes with the greatest similarity values as TPs
Select the 500 genes with the smallest similarity values as TNs
Rank the TPs in the descending order of similarity values
Rank the TNs in the ascending order of similarity values
Put the top half into the training set
Put the bottom half into the test set
C
Obtain the CRISPR-generated regulatory matrix
For each TF
Select the genes meeting three criteria as TPs
Select the genes meeting one criterion as TNs
Find the TP B with the greatest average coefficient
Compute the Bray-Curtis similarity value between B and each TN based on expression profiles
Select a subset of TNs with the smallest similarity values whose size is the number of TPs
Rank the TPs in the descending order of average coefficients
Rank the TNs in the subset in the ascending order of similarity values
Put the top half into the training set
Put the bottom half into the test set
D
Obtain the siRNA-generated P-value matrix
For each TF
Select the genes meeting two criteria as TPs
Select the same number of genes meeting two criteria as TNs
Find the gene C with the smallest P-value among the TPs
Find the gene D with the greatest P-value among the TNs
Compute the Bray-Curtis similarity value between C and each TP based on expression profiles
Compute the Bray-Curtis similarity value between D and every TN based on expression profiles
Rank the TPs in the descending order of similarity values
Rank the TNs in the descending order of similarity values
Put the top half into the training set
Put the bottom half into the test set


NR3C1
RPKM
60
40
20
0
SLC25A32
RPKM
25
20
15
10
5
0
TANK
RPKM
30
20
10
0
Adipose - Subcutaneous
Adipose - Visceral (Omentum)
Adrenal Gland
Artery - Aorta
Artery - Coronary
Artery - Tibial
Bladder
Brain - Amygdala
Brain - Anterior Cingulate Cortex (BA24)
Brain - Caudate (Basal Ganglia)
Brain - Cerebellar Hemisphere
Brain - Cerebellum
Brain - Cortex
Brain - Frontal Cortex (BA9)
Brain - Hippocampus
Brain - Hypothalamus
Brain - Nucleus Accumbens (Basal Ganglia)
Brain - Putamen (Basal Ganglia)
Brain - Spinal Cord (Cervical c-1)
Brain - Substantia Nigra
Breast - Mammary Tissue
Cells - EBV-transformed Lymphocytes
Cells - Transformed Fibroblasts
Cervix - Ectocervix
Cervix - Endocervix
Colon - Sigmoid
Colon - Transverse
Esophagus - Gastroesophageal Junction
Esophagus - Mucosa
Esophagus - Muscularis
Fallopian Tube
Heart - Atrial Appendage
Heart - Left Ventricle
Kidney - Cortex
Liver
Lung
Minor Salivary Gland
Muscle - Skeletal
Nerve - Tibial
Ovary
Pancreas
Pituitary
Prostate
Skin - Not Sun Exposed (Suprapubic)
Skin - Sun Exposed (Lower Leg)
Small Intestine - Terminal Ileum
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Whole Blood


EGR1
ELF1
ELK1
ETS1
GABPA
IRF1
YY1
A
1
0.94
0.88
0.82
0.76
0.7
0.64
0.58
Accuracy
B
0.9
0.86
0.82
0.78
0.74
0.7
0.66
0.62
0.58
Accuracy
1.01
1.05
1.1
The threshold ε on the average fold change


: A binding site of EGR1
: An EGR1 molecule
: Is / is not a DE target of EGR1
Cluster 1
12.0 bits
4.3 bits
Cluster 2
7.0 bits
1.3 bits
Cluster 3
1.9 bits
8.9 bits
2.2 bits
GAGGGGGAGTG→GAGGGAGAGTG
4.3 bits
Cluster 2
7.0 bits
1.3 bits
Cluster 3
1.9 bits
8.9 bits
2.2 bits
GTGGTGGCGGG→GTTGTGGCGGG
CGGGAGGCGGG→CGGGAGGGGGG
4.3 bits
1.3 bits
1.9 bits
2.2 bits