Computational Modeling of Gene-Specific Transcriptional Repression, Activation and Chromatin Interactions in Leukemogenesis by LASSO-Regularized Logistic Regression

Nickolas Steinauer, Kevin Zhang, Chun Guo, and Jinsong Zhang

Abstract—Many physiological and pathological pathways are dependent on gene-specific on/off regulation of transcription. Some genes are repressed, while others are activated. Although many previous studies have analyzed the mechanisms of gene-specific repression and activation, these studies are mainly based on the use of candidate genes, which are either repressed or activated, without simultaneously comparing and contrasting both groups of genes. There is also insufficient consideration of gene locations. Here we describe an integrated machine learning approach, using LASSO-regularized logistic regression, to model gene-specific repression and activation and the underlying contribution of chromatin interactions. LASSO-regularized logistic regression accurately predicted gene-specific transcriptional events and robustly detected the rate-limiting factors that underlie the differences of gene activation and repression. An example was provided by the leukemogenic transcription factor AML1-ETO, which is responsible for 10-15 percent of all acute myeloid leukemia cases. The analysis of AML1-ETO has also revealed novel networks of chromatin interactions and uncovered an unexpected role for E-proteins in AML1-ETO-p300 interactions and a role for the pre-existing gene state in governing the transcriptional response. Our results show that logistic regression-based probabilistic modeling is a promising tool to decipher mechanisms that integrate gene regulation and chromatin interactions in regulated transcription.

Index Terms—Logistic regression, LASSO, machine learning, regulated transcription, repression, activation, AML1-ETO

1 Introduction

1.1 Overview of Transcriptional Regulation

In eukaryotic cells, DNA wraps around histones to form nucleosomes, which are the fundamental building blocks of chromosome. The interactions between histones and DNA are inhibitory to transcription and must be weakened in order for transcription to take place [1]. The histone-DNA interactions are regulated by covalent modifications of histone residues, such as lysine. Acetylation of the lysine residues (e.g., histone H3 K9/K18/K27) facilitates transcriptional activation by facilitating the recruitment of the basal transcription machinery, which includes RNA Pol II and its accessory proteins. Deacetylation facilitates repression by inhibiting the assembly of the basal transcription machinery. These acetylation and deacetylation reactions are catalyzed by histone acetyltransferases (HATs) and histone deacetylases (HDACs), respectively [2]. Other post-translational modifications of histones, such as methylation, also play a role in regulating gene transcription.

The recruitment of HATs and HDACs is mediated by sequence-specific transcription factors (TFs), which directly bind to the regulatory regions of the target genes. TFs can bind to enhancers, which are regulatory elements located at relatively far distances from the TSS (transcription start site). TFs can also bind to promoters near the TSS. Enhancers and promoters carry distinct histone marks. For example, mono-methylated histone H3K4 (H3K4me) is specifically enriched at the enhancer regions [3]. In addition, active enhancers are marked by acetyl-H3K27 (H3K27ac) and its corresponding HATs, such as p300.

1.2 Regulated Transcription

Certain house-keeping genes are constitutively expressed, however, the expression of most genes is repressed or activated in a regulated fashion. There are two common scenarios in regulated transcription depending on the absence and presence of signals or the function of regulatory proteins such as TFs. In the former scenario, the same genes are repressed or activated depending on the signals. This is best exemplified by the action of nuclear receptors [4]. Thus, in the absence of a nuclear receptor ligand (e.g., thyroid hormone), genes are repressed due to the binding of the unliganded nuclear receptors. These unliganded receptors recruit nuclear receptor corepressors (NCoR/SMRT) and HDACs to repress transcription. With the arrival of the ligand, the ligand can directly bind to the receptors to induce a conformational change that leads to the release of the corepressors and the subsequent recruitment of coactivators such as HATs, resulting in gene activation. In the second scenario, different genes

• *The authors are with Pharmacology and Physiology, Saint Louis University, Saint Louis, MO 63104 USA. E-mail: {nick.steinauer, chun.guo, jinsong.zhang}@health.slu.edu, kevinzhang2021.2@u.northwestern.edu.*

Manuscript received 30 July 2020; revised 23 Mar. 2021; accepted 28 Apr. 2021. Date of publication 7 May 2021; date of current version 8 Dec. 2021.
(Corresponding author: Jinsong Zhang.)
Digital Object Identifier no. 10.1109/TCBB.2021.3078128

are repressed or activated in response to the binding by the same regulatory factor. This has been commonly observed for various TFs, such as RUNX1 [5], [6], p53 [7], NFkB [8], and c-Jun [9]. Interestingly, gene-specific repression and activation have also been observed for certain aberrant TFs, including leukemogenic fusion proteins generated from chromosome translocations. Since the mechanisms of transcriptional regulation are not fundamentally different between normal and aberrant TFs, in this study, we provided an example of this latter scenario of the gene-specific repression and activation by a leukemogenic transcription factor, AML1-ETO, to illustrate how machine learning can be used to model the binary modes of regulated transcription mediated by TFs.

1.3 AML1-ETO-Mediated Activation and Repression are Dependent on DNA and Protein Interactions

AML1-ETO is generated by the t(8;21) chromosomal translocation [10], which combines the DNA-binding domain of the RUNX1 transcription factor with a nearly complete eight twenty-one (ETO) gene. This fusion results in the expression of a 752-amino acid fusion protein, AML1-ETO, which is directly involved in the development of t(8;21) acute myeloid leukemia (AML). Due to the presence of the DNA-binding domain from the RUNX1 gene, AML1-ETO can directly bind to RUNX1 target genes. This confers AML1-ETO with the ability to both activate and repress transcription depending on the nature of the genes. Activation and repression both contribute to the biological activities of AML1-ETO. For example, AML1-ETO inhibits myeloid differentiation by repressing target genes such as CEBPalpha. In addition, AML1-ETO represses tumor suppressor genes such as RASSF2 [11]. AML1-ETO can also activate genes such as Id1 and JUP to increase the self-renewal of hematopoietic stem/progenitor cells (HSPCs). In addition, AML1-ETO represses CBFA2T3 gene transcription, which is thought to contribute to the relatively favorable prognostic outcome of t(8;21) AML.

AML1-ETO-mediated repression and activation are also dependent on the ETO portion of the fusion protein. ETO contains 4 nervy-homology regions to mediate protein-protein interactions, which allow AML1-ETO to bind to class I basic helix-loop-helix transcription factors (also known as E-proteins) [12], [13], the p300 HAT coactivator [14], and the NCoR/SMRT and HDAC corepressors [15], [16]. Various mechanisms have been proposed based on these interactions to explain AML1-ETO-mediated activation and repression. For example, AML1-ETO-p300 interaction is thought to allow AML1-ETO to activate transcription [14], and AML1-ETO-NCoR/SMRT/HDAC interaction is thought to allow AML1-ETO to repress transcription [15], [16]. Additionally, the binding of E-proteins (HEB, E2A) by AML1-ETO is thought to allow AML1-ETO to repress E-protein-dependent transcription by dismissing coactivators such as p300 and GCN5 pre-occupied to E-proteins [12], [13]. The basis of this latter effect is that p300/GCN5 and AML1-ETO/ETO bind to overlapping surfaces of E-proteins and thus their interactions with E-proteins are mutually exclusive [13].

1.4 Machine Learning-Based Classification can be Used to Study Gene-Specific Repression and Activation

A concern of the above activation or repression mechanisms is that they were mainly derived from studies of a small number of candidate genes. The contribution of AML1-ETO-binding locations, which is the hallmark of gene-specific activation and repression, appears to be insufficiently considered. As a result, it is unclear why the various mechanisms do not apply to all genes. Because different genes respond differently to AML1- ETO, their pre-existing states may play an important role in dictating the effects of AML1-ETO binding. This idea should be generally applicable to many other TFs. Machine learning has been widely used to model gene expression and clinical outcomes in various physiological and pathological settings [17], [18], [19], [20], [21]. More recently, logistic regression has shown successful uses in various classification tasks such as transcriptional initiation [22], enhancer-promoter interaction [23], [24], single-cell classification [25], [26], chromatin territories [27], gene expression patterning [28] and functional TF motifs [29]. However, to our knowledge, there are no reports of using machine learning to model gene-specific repression and activation. To allow such studies, we used AML1-ETO as an example and modeled its binding site-specific effects, as binary factors, on transcription and chromatin interactions mediated by TFs, coregulators, and components of the basal transcription machinery. Unlike many other machine learning studies, our study was carried out in a controlled fashion (i.e., with and without AML1-ETO), thus providing the opportunity to draw conclusions about the causal roles of AML1-ETO. By comparing different classification algorithms, we found that LASSO (least absolute shrinkage and selection operator)-regularized logistic regression [30] had the best performance. It robustly detected rate-limiting steps underlying the differential effects of AML1-ETO, performed exceptionally well in predicting repression, and revealed pre-existing gene state as an important determinant of the regulatory response. Next, by iteratively modeling the changes of chromatin regulators, LASSO correctly identified p300 as the rate-limiting factor of histone acetylation and revealed a role for E-proteins in p300 recruitment by AML1-ETO. These results support a unified mechanism of transcriptional regulation by AML1-ETO. We also provided an algorithm to show how AML1-ETO-regulated chromatin interactions are coordinately changed during repression and activation. Our work suggests that LASSO-regularized logistic regression is a promising new tool to decipher global mechanisms of gene-specific repression and activation mediated by TFs and other regulatory proteins.

2 Materials and Methods

2.1 RNA-Seq

RNA-Seq was performed in t(8;21) Kasumi-1 AML cells pretreated with a control shRNA (shControl) or an AML1-ETO-specific shRNA (shAML1-ETO) [31]. Total RNA was extracted from the transduced cells and sequenced at Genome Technology Access Center at the Washington University in St. Louis. Data analysis was performed as

previously described, which estimated counts using Kallisto [32] and tested differential expression using EdgeR [31]. To facilitate machine learning, we selected high-confidence AML1-ETO-regulated genes by using the following cutoff values. These cutoff values were consistent with previous studies showing that AML1-ETO typically had higher fold changes in repression than in activation [33]. Thus, fold changes >= 2.5 and FDR (false discovery rate) $<$ 0.00001 were used for repression, fold changes >=2 and FDR $<$ 0.00001 were used for activation. This produced similar numbers of AML1-ETO-repressed and AML1-ETO-activated genes to mitigate possible class imbalance issues. Since the main goal of this study is to understand the mechanism of gene-specific repression and activation, a highly stringent cutoff for FDR was used to ensure that genes to be analyzed were truly regulated by AML1-ETO. Nevertheless, using less stringent FDR (e.g., 0.001) produced similar results (data not shown).

2.2 ChIP-Seq

ChIP-Seq assays were performed in Kasumi-1 cells transduced with shControl, shAML1-ETO or shHEB shRNAs. Library construction and sequencing were performed as previously described [31]. Reads were aligned to the hg19 human genome using STAR. Peak calling and quantification of ChIP-Seq reads were performed using HOMER [34]. ChIP-Seq measured chromatin occupancies of the following proteins or histone modifications: (i) transcription factors (AML1-ETO, E2A, HEB, RUNX1), (ii) HATs/coactivators (p300, GCN5), (iii) HDACs/corepressors (HDAC1, HDAC3, SMRT), (iv) histone marks (H3K9/K18/K27ac, H3K4me1, H3K9/K27me3, H3K9me2 in untreated cells, H3K9/K18/K27ac/K9me2 in cells treated with trichostatin A (TSA) or DMSO for 1 hr), (v) basal transcription/elongation factor (CDK9), (vi) Pol II (total Pol II, and its elongating/active form, designated as S2P-Pol II, with Ser2 phosphorylated at the C-terminal domain) (see Table 2).

2.3 Machine Learning

2.3.1 Dataset

The final data matrix has 479 binding sites (rows) and 97 features (columns) (Table 2, Fig. 1). Each binding site (row) was associated with activation (1) or repression (-1) under the "activation" column. Each binding site was also associated with different cistromic features, totaling 95, derived from ChIP-Seq studies, plus a location variable "enhancer", which has the value of either 1 (for enhancer) or -1 (promoter).

In this work, enhancers were defined as binding sites that fall outside of the 500-bp upstream and 100-bp downstream range relative to TSSs ($d < -500$ or $d > +100$) similar to what was used in a previous study [35]. Using different upstream cutoffs (-500 bp, -2000 bp, -10000 bp), we found that -500 produced a slightly better performance compared to -2000 bp and -10000 bp (data not shown). Interestingly, revealing this difference required exclusion of histone variables. Different downstream cutoffs (100 bp, 500 bp) showed no significant differences. Based on these results, we chose -500 bp as the upstream cutoff for the definition of enhancers for this study. Nevertheless, this value may vary as a function of the TF. Our finding that histone variables can "mask" the distancing effect is consistent with the idea that enhancers function at a step upstream of histone modifications.

2.3.2 Dataset Pre-Processing

Following splitting the dataset into a training set and a test set (0.8/0.2 ratio), the training set was centered, scaled, and subsequently used for machine learning. The centering/scaling parameters derived from the training set were used to pre-process the test set (Fig. 1).

2.3.3 Performance Measurement

Modeling performance was evaluated using the test dataset based on AUC (area under curves), ACC (accuracy), TNR (true negative rate) and TPR (true positive rate) scores. The interpretations of TPR and TNR scores for this study are as follows: TPR measures the percentage of true positive samples (i.e., activation or increased chromatin occupancies associated with a binding site in the presence versus absence of AML1-ETO, Fig. 1) to be predicted as being positive, whereas TNR measures the percentage of true negative samples (i.e., repression or decreased chromatin occupancies associated with a binding site in the presence versus absence of AML1-ETO, Fig. 1) to be predicted as being negative. These scores were averaged from results of 500 independent runs and presented as boxplot and/or barplot (mean +/- standard error).

2.4 Training Algorithms Used in This Study

2.4.1 Overview

In the initial study, we compared the following algorithms implemented in python: logistic regression (LASSO) from the glmnet package and a set of classifiers from the sklearn package, including logistic regression (LASSO), random forest, support vector classifier (SVC), and k-nearest neighbor (KNN). All programs were run under their default settings except that a weight parameter for the class membership was manually set in glmnet (see below) and was set to be "balanced" in sklearn packages. In addition, 10-fold cross-validation was used in sklearn logistic regression, different from the default (5-fold). Unless otherwise noted, the input data is the complete 2D matrix (479X97) described in 2.3.1 with the "activation" column used as the dependent variable. Because glmnet showed the best overall performance, and has the ability to select the important features for interpretation purpose, it was used in subsequent studies with the R glmnet package. To evaluate different regularizations, we compared glmnet LASSO, Elastic Net, Ridge regularizations and another recently proposed feature reduction method using network-based regularization that has been implemented in the Regnet R package. All these studies showed that LASSO regularization had the best overall performance. This allowed us to exclusively use glmnet LASSO in subsequent mechanism-oriented studies. Below is a description of the algorithms used in this study, focusing on logistic regression.

2.4.2 Logistic Regression

In classification, logistic regression [30], [36] models the class membership by calculating the probability that a given

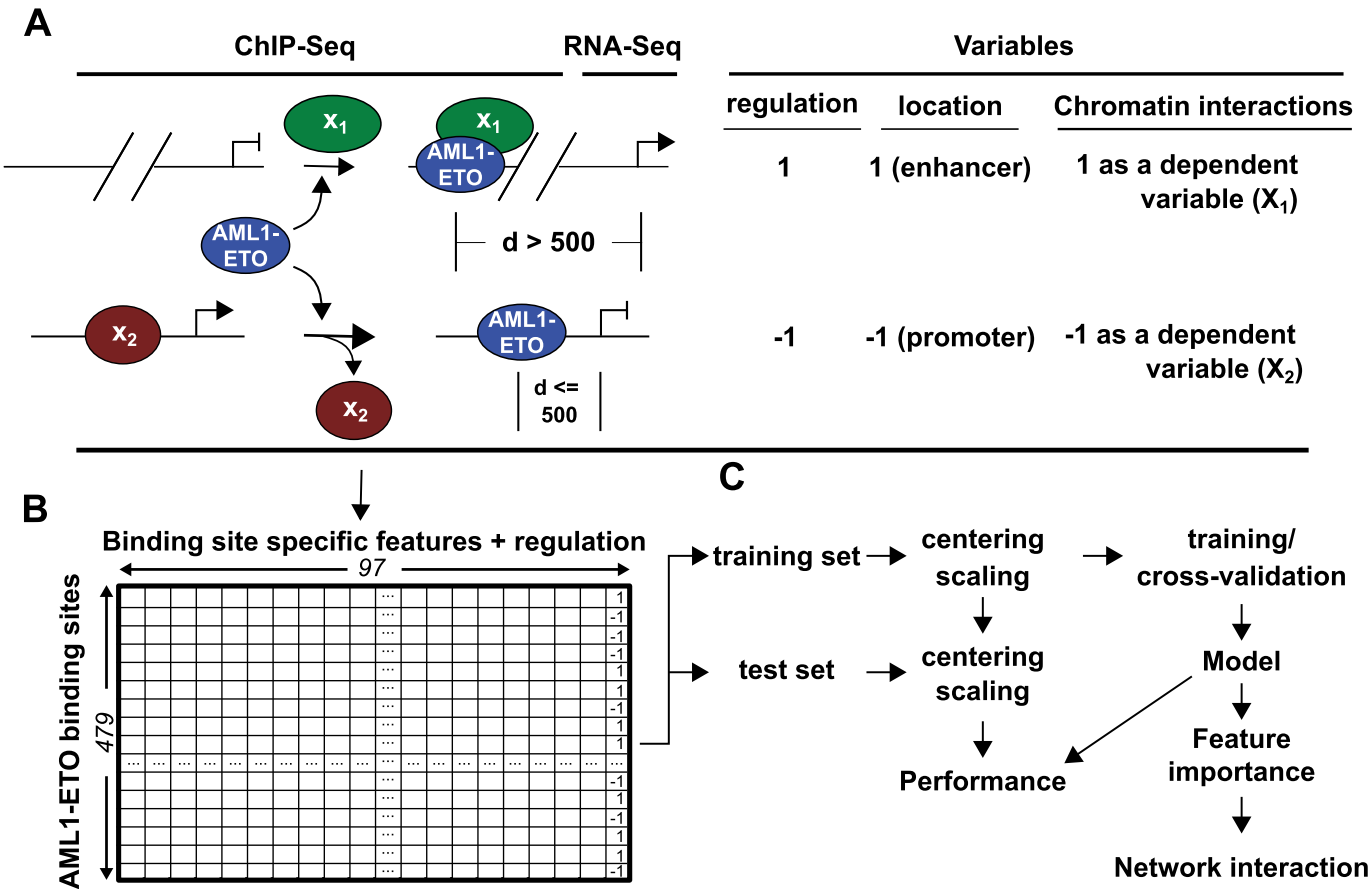

Fig. 1. Overall project design. (A) Schematic representations showing that binding of AML1-ETO to an enhancer site results in the recruitment of X_1 factor and concomitant activation of transcription, whereas binding of AML1-ETO to a promoter site results in dismissal of X_2 factor and concomitant repression. The right panel shows the corresponding values reflecting these changes. The values of X_1 and X_2 shown here represent the discrete numbers when both features were treated as dependent variables. (B) An illustration of the 2D matrix used for the machine learning study. (C) A diagram showing the workflow of the machine learning study.

sample belongs to a certain class, which takes the form of the sigmoid function of a linear combination of weighted predicting variables. To model gene regulation by AML1-ETO, the dependent variable takes the form of activation ($y = 1$) or repression ($y = -1$). Let p be the probability of a given gene to be activated by AML1-ETO, x_i be the ith feature of the binding site of AML1-ETO on that gene (e.g., chromatin occupancies determined by ChIP-Seq) ($1 \leq i \leq n$, n = number of features), w_i be the coefficient of the ith feature, b be the bias term. p is given by the following equation:

$$p = \frac{1}{1 + e^{-\left(\left(\sum_{i=1}^{n} w_i x_i\right) + b\right)}}.$$

From the above equation, one can deduce the odds ratio (the ratio of probabilities of activation versus repression) as

$$log\left(\frac{p}{1-p}\right) = \sum_{i=1}^{n} w_i x_i + b.$$

A cutoff (0.5) was applied such that if p>0.5, the gene is predicted to be activated by AML1-ETO, and if p<0.5, the gene is predicted to be repressed by AML1-ETO. The same rules were applied to classify the effects of AML1-ETO on chromatin occupancies of transcription factors, coactivators, corepressors, and histone marks.

To avoid overfitting, a regularization term was also added to the total cost, which, under LASSO, takes the form of the sum of the absolute value of all coefficients multiplied by a scaling factor (λ).

$$Cost = -\frac{1}{m}\sum_{i=1}^{m}(y_i\, log(p_i) + (1 - y_i)\ log\,(1 - p_i)) + \frac{\lambda}{m}\sum_{i=1}^{n}|w_i|,$$

($1 \leq i \leq m, n$, m = number of samples, n = number of features)

Similar to LASSO, Ridge Regression implements a penalization term to reduce overfitting. However, it uses the sum of squares of the coefficients, rather than the sum of absolute values. Thus, coefficients will not shrink to zero under ridge regression. The cost function is

$$-\frac{1}{m}\sum_{i=1}^{m}(y_i\, log\,(p_i) + (1 - y_i)\ log\,(1 - p_i)) + \frac{\lambda}{2m}\sum_{i=1}^{n} {w_i}^2.$$

The Elastic Net model combines both LASSO and Ridge Regression; the sum of the squares of coefficients, as well as the sum of the absolute value of coefficients, are both included in the penalization term, but depend on the parameter α. In the glmnet R package, the alpha parameter is the elastic net mixing parameter that balances L1 (LASSO) and L2 (ridge) regularization. To summarize, the LASSO

penalization induces sparsity, where coefficients may be reduced to zero. Ridge Regression's penalization induces shrinkage, which reduces coefficients, but does not eliminate them altogether. Elastic net compromises between these two. We utilize all three models.

2.4.3 Cross Validation

The default 10-fold cross-validation per the glmnet package was used to select the lambda parameter, which was chosen to have the lowest classification error. In addition, we assigned to each class (activation or repression) a weight parameter as the ratio of samples under a given class versus the total number of the samples

$$weight_{y=1} = 1 - (m_{y=1}/m), weight_{y=-1} = 1 - (m_{y=-1}/m).$$

2.4.4 Other Used Algorithms for Comparison

Other models which were implemented, as mentioned earlier, are random forest, support vector machine/classifier, and k-nearest-neighbors. Random forest models train via generating a large number of uncorrelated decision trees, then forming an ensemble with which to make predictions on new observations. The RandomForestClassifier function in sklearn was used with default parameters to generate a random forest model. Support vector machines generate a decision boundary which is used to predict new observations. Additionally, a kernel function is used to avoid computationally expensive calculations in high dimensional spaces. We used the default parameters for the SVC function in sklearn, which utilizes a RBF (radial basis function kernel). The k-nearest-neighbors model classifies new observations by examining k of the closest neighbors to the observation in terms of euclidean distance, then returning the mean of the values of those observations, i.e., the label which appears more frequently. We used the default k value of 5. Lastly, the regnet package in R was used for a network-based regularization approach, which is promising due to its ability to work with high-dimensional biological data by incorporating correlations among genomic features. Regnet was used to generate optimal lambda values via cross-validation which are then used to build network-based penalized logistic regression models. As explained in the regnet documentation, the first lambda value induces sparsity, while the second controls smoothness among coefficient profiles.

2.5 Statistical Analysis

Significance measurement comparing two groups of continuous variables was performed using Student's t-tests and adjusted for multiple hypothesis testing errors if necessary. P-values involving discrete variables such as the gene set enrichment analyses were from hypergeometric test. $*: P <= 0.05$; $**: P <= 0.01$; $***: P <= 0.001$, $****: P <= 0.0001$, "n.s." : $P > 0.05$.

3 Results

3.1 Overall Design

The overall design of the project was shown in Fig. 1. To collect the data, RNA-Seq was performed in Kasumi-1 t(8;21) AML cells with and without depletion of AML1- ETO. This allowed us to obtain the list of high-confidence AML1-ETO-activated and AML1-ETO-repressed genes (Table 1). Next, ChIP-Seq studies were performed in these cells under the same treatment conditions. The analyses of the AML1-ETO ChIP-Seq results led to the information regarding which of the RNA-Seq-determined genes were truly bound by AML1-ETO (and thus were justified as the direct AML1-ETO target genes), as well as the information regarding the binding sites of AML1-ETO on each of the target genes. ChIP-Seq also measured chromatin occupancies of the various proteins. Together, these allowed us to construct a 2D matrix (Fig. 1) with the rows corresponding to each of the binding sites and columns corresponding to the features of the binding sites, encompassing both cistromic and transcriptomic variables (Table 2). This dataset was then used for supervised machine learning which involves splitting of the dataset into a training set and a test set, and pre-processing the training set and subsequent modeling. This was followed by performance measurement using the test set following pre-processing using the pre-processing parameters of the training set. Our initial studies compared different classification algorithms. We found that LASSO-regularized logistic regression [30] outperformed other algorithms. It robustly detected the rate-limiting features for both activation and repression (Fig. 2) while making correct predictions about the underlying mechanisms (Figs. 3, 4, 5, 6, 7, and 8). Although LASSO is associated with sparse solutions, transcription is regulated by a small number of cofactors shared by all genes. We extended the modeling of transcription to modeling of chromatin interactions. This process was iterated to include all chromatin regulators whose changes in response to AML1-ETO were similarly treated as binary factors and modeled against other features. Clustering the coefficient matrix resulting from this study revealed networks of chromatin interactions that took place during both repression and activation, making it possible to draw conclusions about their unified regulatory mechanisms.

TABLE 1
Numbers of Target Genes and Their Binding Sites of AML1-ETO Used for the Machine Learning

	Activated Genes	Repressed Genes
Number of Regulated Genes	267	214
Number of the regulated genes bound by AML1-ETO	153	147
Binding sites of AML1-ETO	232	247
P-Value (probability of random binding)	4.6817e-07	4.0240e-15

3.2 Data Collection

3.2.1 AML1-ETO-Regulated Genes

Analysis of the RNA-Seq results showed that 267 genes were downregulated whereas 214 genes were upregulated by knocking down AML1-ETO. Combining these results with AML1-ETO ChIP-Seq showed that among the 267 AML1-ETO-downregulated genes, 153 were directly bound by AML1-ETO. Similarly, among the 214 AML1-ETO-upregulated genes, 147 genes were directly bound by AML1-ETO. These genes were enriched with known AML1-ETO-

TABLE 2
List of Features Used for the Machine Learning Study (Total 97)

Binding site specific features (cistromic)								
Occupancies with and without AML1-ETO or HEB					AML1-ETO/HEB-dependent changes			
wo_AE	with_AE	wo_HEB	with_HEB		delta_..._shae	delta_..._shheb	enhancer	activation
AE KD	AE CONTROL	HEB KD	HEB CONTROL	RATIO	DELTA_AE	DELTA_HEB	LOCATION	REGULATION
AE	AE	AE	AE	AE_E2A	AE	AE	1	1
CDK9	CDK9	HEB	HEB	AE_HEB	CDK9	HEB	-1	-1
DMSO_H3K18AC	DMSO_H3K18AC	P300	P300	AE_RUNX1	DMSO_H3K18AC	P300		
DMSO_H3K27AC	DMSO_H3K27AC	TOTAL_POLII	TOTAL_POLII		DMSO_H3K27AC	TOTAL_POLII		
DMSO_H3K9AC	DMSO_H3K9AC				DMSO_H3K9AC			
DMSO_H3K9ME2	DMSO_H3K9ME2				DMSO_H3K9ME2			
E2A	E2A				E2A			
GCN5	GCN5				GCN5			
H3K18AC	H3K18AC				H3K18AC			
H3K27AC	H3K27AC				H3K27AC			
H3K27ME3	H3K27ME3				H3K27ME3			
H3K4ME1	H3K4ME1				H3K4ME1			
H3K9AC	H3K9AC				H3K9AC			
H3K9ME2	H3K9ME2				H3K9ME2			
H3K9ME3	H3K9ME3				H3K9ME3			
HDAC1	HDAC1				HDAC1			
HDAC3	HDAC3				HDAC3			
HEB	HEB				HEB			
P300	P300				P300			
RUNX1	RUNX1				RUNX1			
S2P_POLII	S2P_POLII				S2P_POLII			
SMRT	SMRT				SMRT			
TOTAL_POLII	TOTAL_POLII				TOTAL_POLII			
TSA_H3K18AC	TSA_H3K18AC				TSA_H3K18AC			
TSA_H3K27AC	TSA_H3K27AC				TSA_H3K27AC			
TSA_H3K9AC	TSA_H3K9AC				TSA_H3K9AC			
TSA_H3K9ME2	TSA_H3K9ME2				TSA_H3K9ME2			

(The 3rd row depicts the notation of variable names for the study. "AE KD", "AE CONTROL", "HEB KD" and "HEB CONTROL" denote treatment conditions).

downregulated and AML1-ETO-upregulated genes [31], and thus, represent the bona fide AML1-ETO-regulated genes. This was also evidenced by the highly significant p-values under the null hypothesis that the binding is random (Table 1). To obtain the list of the binding sites, we extracted the top two sites most strongly bound by AML1-ETO based on their peak scores computed by HOMER [34]. We found that multiple binding sites of the same gene typically showed similar binding profiles. On the other hand, we also observed that AML1-ETO may bind to both enhancer and promoter regions of the same gene, and this kind of differential binding is often captured by the top two strongest binding sites.

3.2.2 Binding Site-Specific Features (Total 97, Table 2)

Information about the gene-regulatory effects of AML1-ETO was stored in the "activation" variable, which has a value of 1 (activation) if AML1-ETO increases gene expression or -1 (repression) if AML1-ETO decreases gene expression. This is different from modeling the basal-level expression as we found that highly-expressed genes tend to be repressed by AML1-ETO whereas lowly-expressed genes tend to be activated by AML1-ETO. Therefore, our study is different from many other machine learning studies that only modeled steady state gene expression based on variables such as histone marks [20].

An "enhancer" variable was constructed to model the binding site location, which has a value of 1 (enhancer) or -1 (promoter) depending on the distance (d) between the binding site and the TSS of the nearest gene. Enhancer-binding sites fall outside the 500-bp upstream and 100-bp downstream range relative to TSS and promoter-binding sites are within this range ($-500 <= d <= +100$, bps). As described in 2.3.1, the -500 was used as the upstream cut-off because it produced a slightly better modeling performance compared to -2000 and -10000 cutoffs when the modeling was conducted in the absence of histone variables (data not shown). All remaining variables were derived from ChIP-Seq, which measured chromatin occupancies of various regulatory proteins/events in Kasumi-1 cells treated with and without depletion ("KD") of AML1-ETO or HEB (Table 2). The following notations were used for the ChIP-Seq variables: (i) Variables beginning with "with_ae" or "with_heb" refer to absolute levels of chromatin binding in control shRNA-treated Kasumi-1 cells ("WT"); (ii) Variables beginning with "without_ae" or "without_heb" refer to absolute levels of chromatin binding in AML1-ETO- or HEB-depleted cells; (iii) Variables beginning with "delta" and ending with "shae" or "shheb" refer to the differences between WT and the AML1-ETO/HEB-depleted cells (WT-KD), (iv) Variables ending with "ratio" refer to the ratio between AML1-ETO and HEB, E2A or RUNX1, which were calculated from the ChIP-Seq results derived from control ("WT") Kasumi-1 cells.

In total, the final dataset matrix (Fig. 1B) has 479 rows (representing 232 activation sites and 247 repression sites) and 97 columns (95 derived from ChIP-Seq, plus "enhancer" and "activation" variables). This was used to model the functional effects of AML1-ETO with "activation" used as the dependent variable based on the set of cistromic features (Fig. 2).

3.3 Overall Performance

We first compared glmnet (python) and other classifiers from the sklearn package (version 0.22.2), including logistic regression (LogisticRegressionCV), random forest (RandomForestClassifier), support vector machine classifier

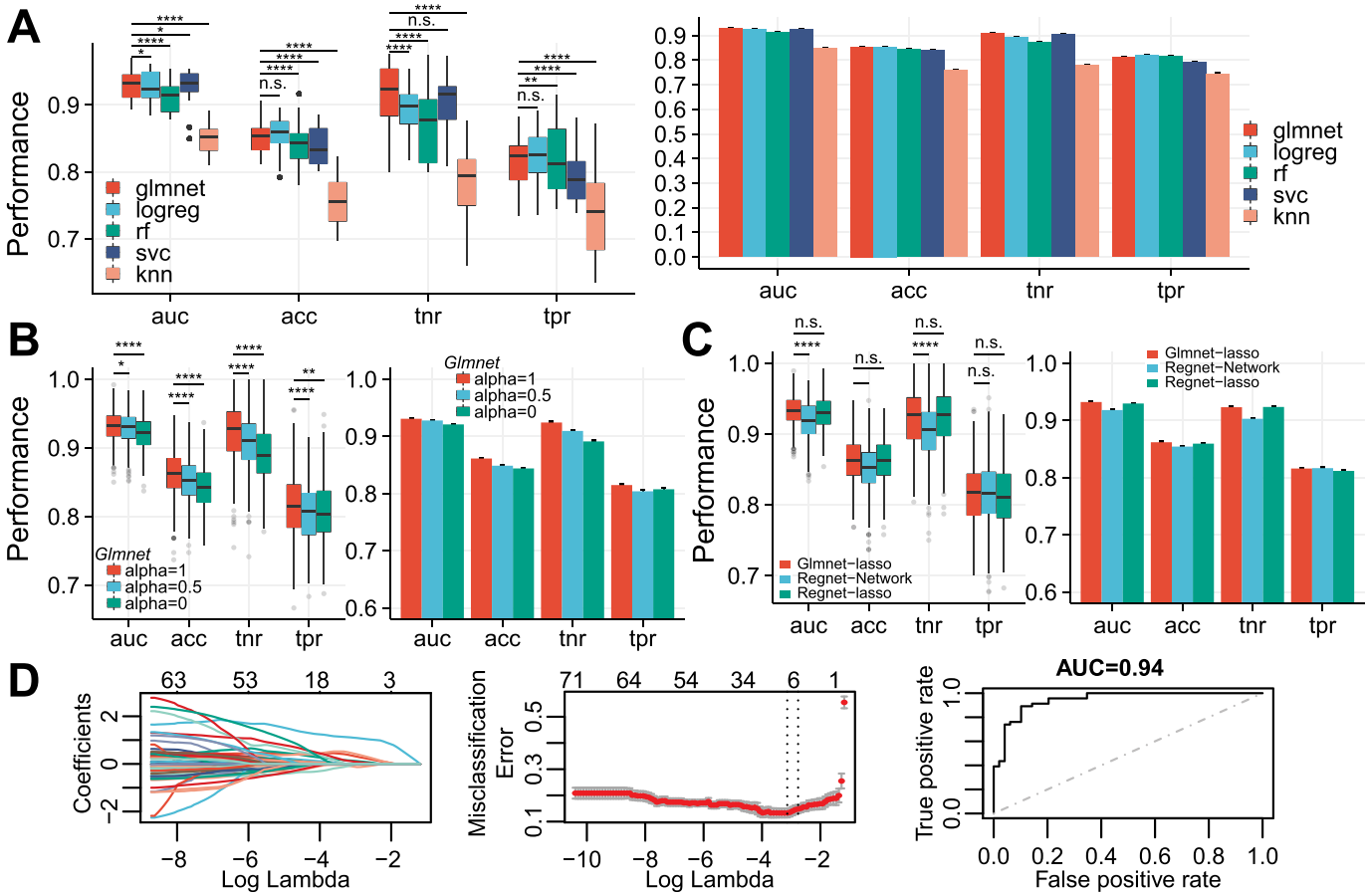

Fig. 2. Modeling the effects of AML1-ETO on gene-specific transcription. (A) Comparison of modeling performance using different algorithms. "glmnet", python port of the glmnet package (alpha=1); "logreg": logistic repression ("l1") from sklearn;"rf": Random Forrest from sklearn; "svc": Support Vector Machine Classifier from sklearn; "knn": K-Nearest Neighbors classifier from sklearn. (B) Comparison of LASSO, Elastic Net and Ridge regularizations using R glmnet packages. (C) Comparison of glmnet LASSO and regnet implemented algorithms using their R packages. (D) A typical run of logistic regression with LASSO using glmnet in R. In the middle panel, two dashed lines depict lambda.min and lambda.1se values. In (A-C), lambda.min was used for logistic regression with the glmnet package. In (A-D), the input is the full data matrix (479X97). The dependent variable is the "activation" column (Fig. 1B). Performance was measured using the test set.

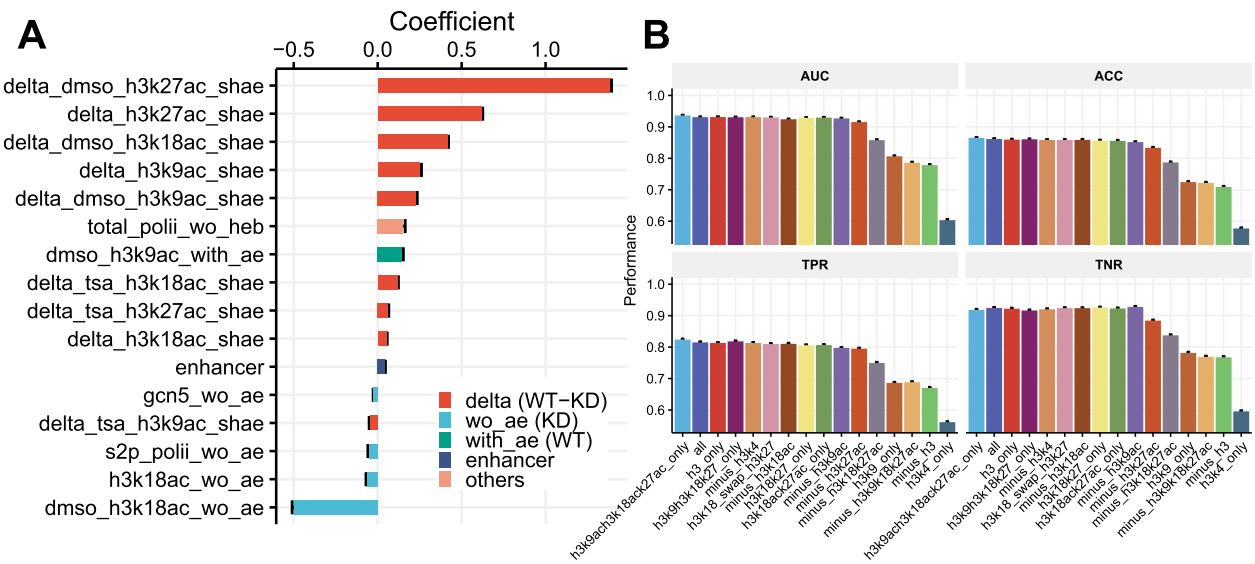

Fig. 3. Histone acetylation marks are among the most important variables that can predict the active and repressive effects of AML1-ETO. (A) The average values of coefficients from 500 runs. A positive value for a variable indicates its contribution to activation. A negative value indicates its contribution to repression. (B) Comparison of modeling performance for AML1-ETO's gene regulatory effects using features indicated at the bottom.

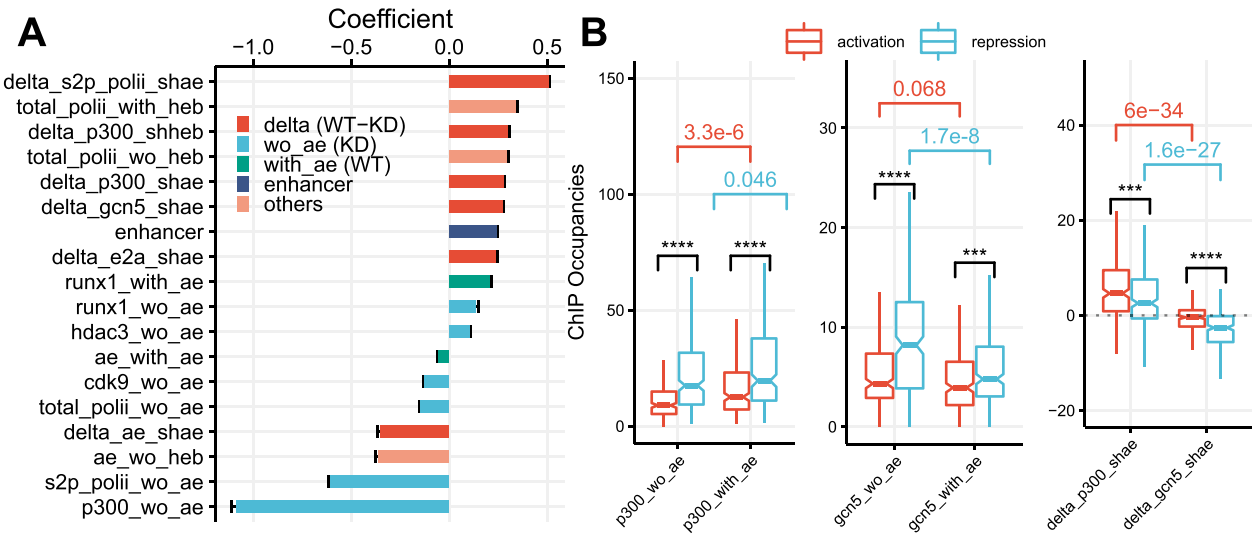

Fig. 4. AML1-ETO differentially affects p300 and GCN5 recruitment in activation and repression. (A) Average values of coefficients from 500 runs using non-histone variables. (B) ChIP occupancies of p300 and GCN5 in response to AML1-ETO expression at the activation and repression sites.

(SVC), and k-nearest neighbor (KNeighborsClassifier), referred to, in Fig. 2A, as "logreg", "rf","svm" and "knn", respectively. LASSO was used in both of the logistic regression algorithms here. Comparison of the different penalty settings for logistic regression was conducted below (Fig. 2B). As shown in Fig. 2A, the glmnet LASSO-penalized logistic regression produced the best overall performance as judged by the AUC score. Its ACC score was similar to that of sklearn LASSO and better than all other algorithms. Some differences between glmnet LASSO and logreg LASSO could be related to the unique implementation of LASSO in sklearn, given that glmnet LASSO showed a similar performance as Regnet LASSO (Fig. 2C). Glmnet and sklearn LASSO algorithms used different solvers for the coordinate descent algorithm. In addition, a lambda (lambda.min) to minimize the classification error was used

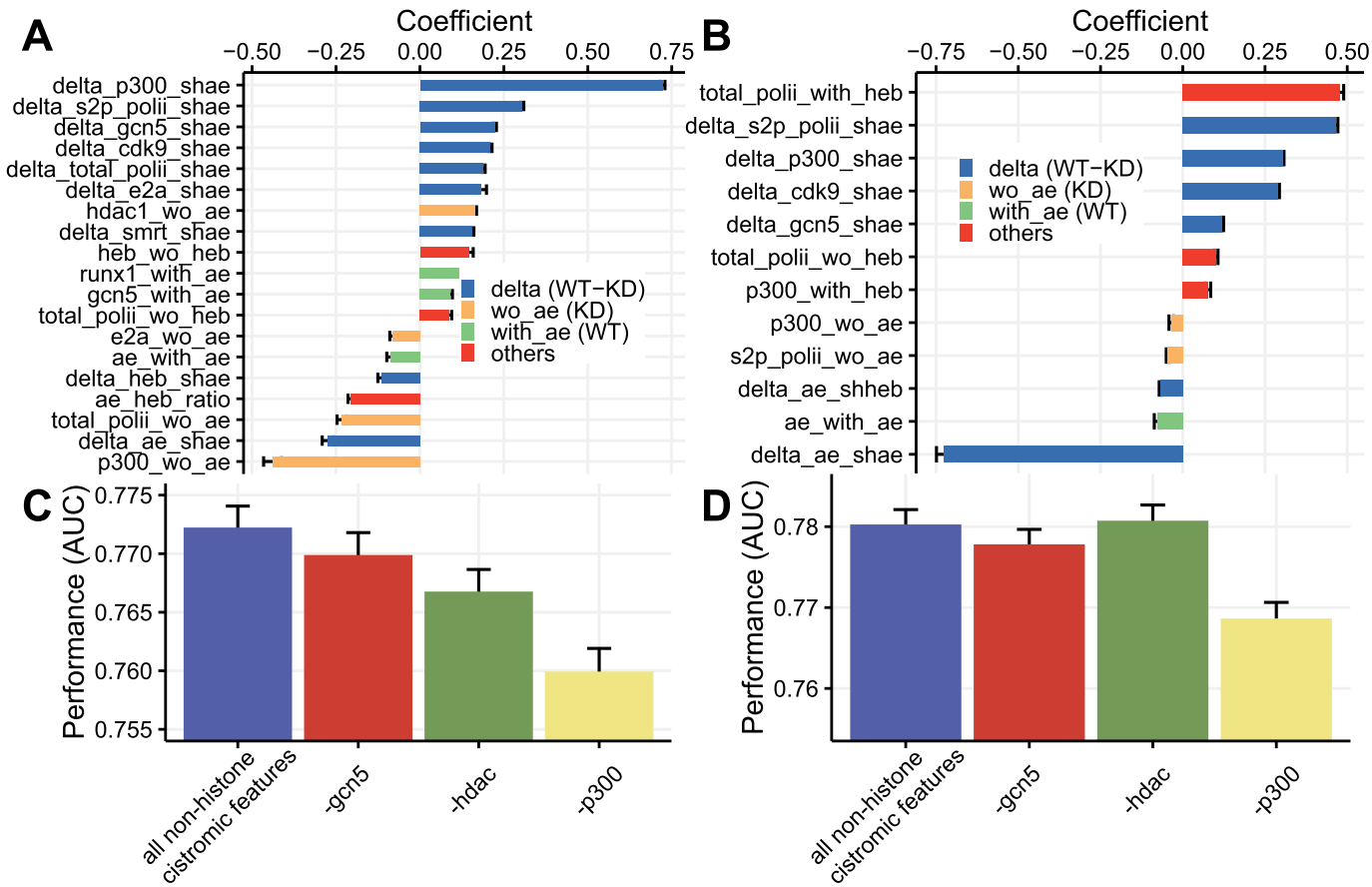

Fig. 5. Identification of p300 as the rate-limiting variable for AML1-ETO-regulated acetylation of histone H3K18 (A, C) and H3K27 (B,D). Shown are comparison of modeling using all non-histone cistromic features and modeling using one of the "gcn5", "hdac", or "p300" features removed.

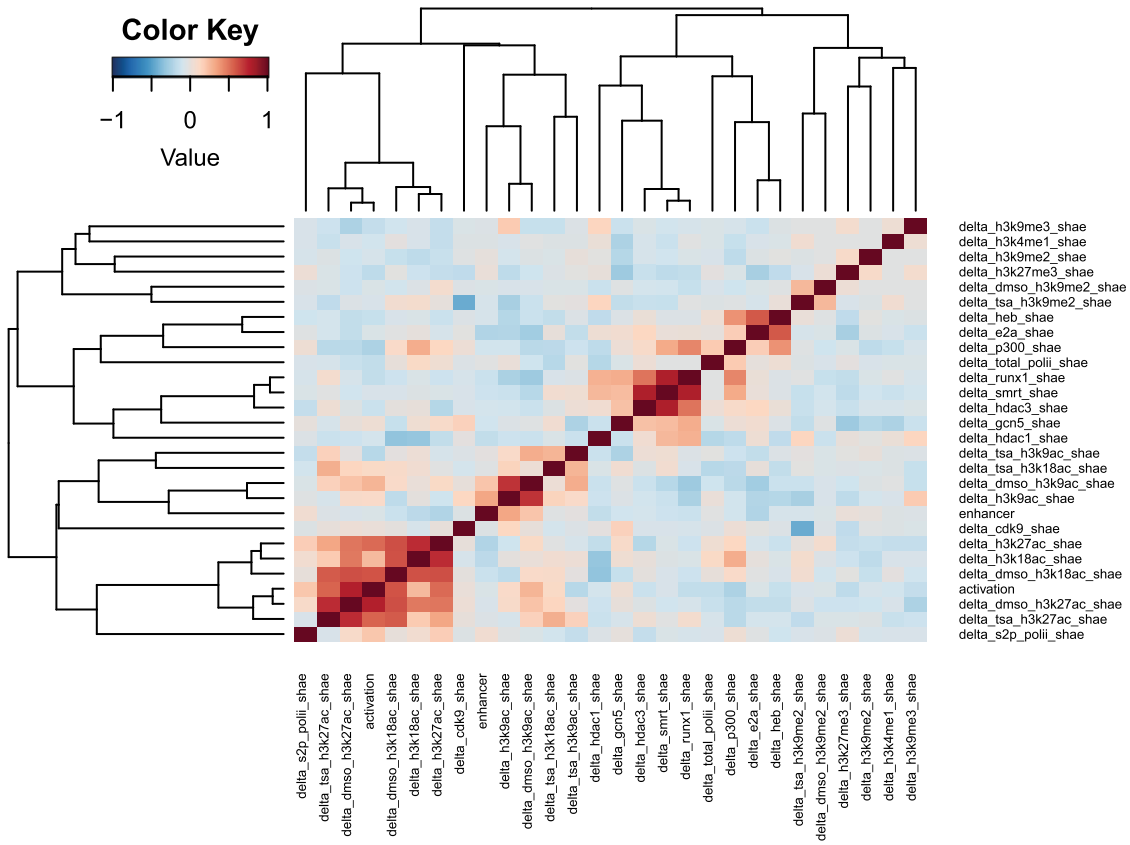

Fig. 6. Unsupervised clustering of cistromic variables using coefficients generated from the iterative modeling of their changes in chromatin occupancies upon AML1-ETO binding against other variables. All variables used for the modeling are shown in the figure. The results revealed coordinated changes of variable in response to AML1-ETO binding.

in glmnet as it gave a better performance than lambda.1st (data not shown). In sklearn, the equivalent parameter (C) was chosen by the program. Additionally, we manually set the class weights in glmnet, whereas the weights used the "balanced" setting in sklearn. Nevertheless, not setting this (i.e., the weight parameter) in glmnet LASSO did not noticeably compromise the performance in comparison with Regnet (Fig. 2C). We next compared LASSO, Elastic Net and Ridge regularization using the glmnet package in R, by setting the alpha parameter to 1 (LASSO), 0.5 (Elastic Net) or 0 (Ridge). The results showed that LASSO also performed the best compared to Elastic Net and Ridge (Fig. 2B).

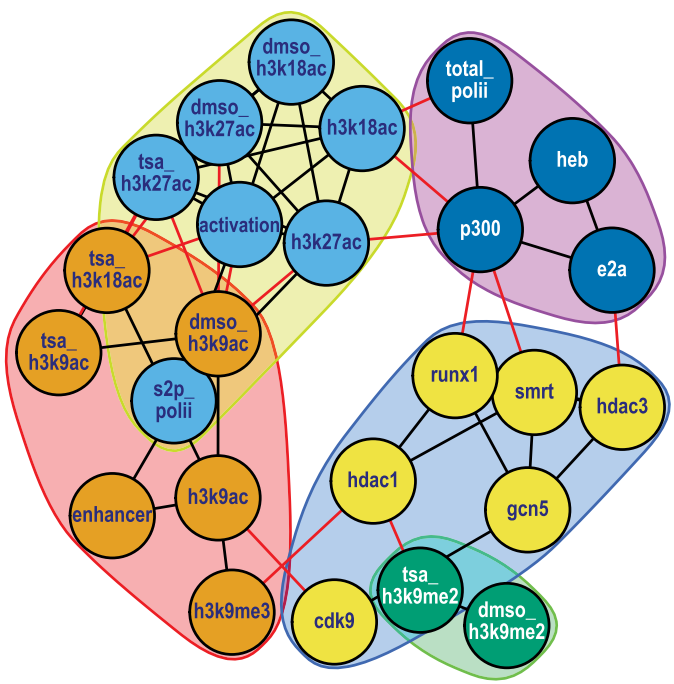

Fig. 7. Network graphs showing five detected clusters of variables with coordinated changes in response to AML1-ETO expression.

Recently, network-based regularization has been used in high-dimension genomic studies [37]. A version of it is implemented in the R Regnet package, which prioritizes features by taking consideration of feature correlations [38], [39], [40]. Since chromatin features are also correlated, we compared glmnet LASSO with the network regularization using the Regnet package. A LASSO implemented in Regnet was used as the control whose performance equaled that of glmnet LASSO (Fig. 2C). Glmnet (and Regnet) LASSO, however, outperformed the network regularization, which can be attributed to LASSO's better prediction power for the repressed genes. Comparing the set of features chosen by these algorithms showed that Ridge, Elastic Net and Network tended to shrink the important features outputted by LASSO, especially on the repression side, while including new features discarded by LASSO (data not shown), which explains why these algorithms performed less well than LASSO.

Fig. 2D showed a typical result of glmnet LASSO. The overall performance of glmnet LASSO from 500 independent runs was shown in Fig. 2B. The overall AUC score was 93 percent. The overall ACC score was 86 percent. The overall TNR score was 92 percent, indicating that 92 percent of truly

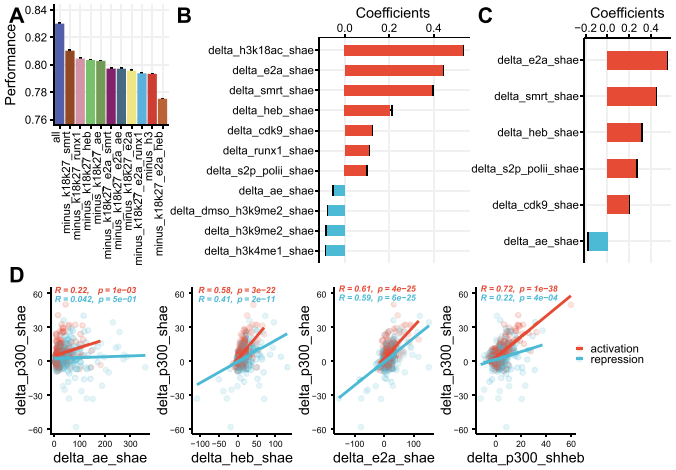

Fig. 8. Modeling AML1-ETO-dependent recruitment of p300. (A) AUC scores of modeling using features shown at the bottom. (B,C) Coefficient plots from modeling using total (B) and non-histone features (C). (D) Scatter plots showing correlation between p300 recruitment and other variables indicated at the bottom.

repressed genes were predicted to be repressed. The overall TNR score was 82 percent, indicating that 82 percent of the truly activated genes were predicted to be activated. One of the reasons why prediction of the activated genes had a lower accuracy than predicting the repressed genes may be related to the fact that AML1-ETO tended to bind enhancers to exert its effect in activation, indicating that the activation would be more influenced by the long-range communication between the enhancer and promoter/TSS regions. In this study, we modeled the location/enhancer effect by using a discrete factor (1 as enhancer and -1 as promoter) determined by the cutoff. Our results indicate that merely optimizing this cutoff may not be sufficient to fully capture the enhancer effect. It will be interesting to test in future studies whether including additional variables such as the transcriptomic/cistromic signals occurring at the TSS (both with and without AML1-ETO) may improve the prediction power for the activated genes. Additionally, this may also benefit from including features such as enhancer-promoter interactions measured by Hi-C, ChIA-PET and other assays [41]. Nevertheless, the overall performance of the model generated by glmnet LASSO is clearly in the excellent to outstanding range, setting the stage for subsequent studies to explore the rate-limiting factors underlying both activation and repression.

3.4 Acetyl-Histones are Among the Most Important Features to Differentiate Activation From Repression

LASSO regularization shrinks non-important features to an zero coefficient to select features that contribute to the difference of regulated transcription. In Fig. 3A, we ranked and plotted these features based on coefficients (w_i), which were determined from 500 independent runs, after filtering out those with relatively high p-values and small magnitudes. As shown in Fig. 3A, the most important features that contribute to activation (i.e., those with a positive coefficient) were the ones that measured the differences before and after AML1-ETO knockdown. These variables, starting with "delta", were mainly composed of acetyl-histone marks. In particular, AML1-ETO-regulated H3K27ac showed the highest positive coefficient. On the other hand, the important negative variables (i.e., variables with a negative coefficient) were enriched with features measuring the pre-existing levels of the chromatin occupancies in the absence of AML1-ETO. These variables, starting with "wo_ae" (Table 2), included H3K18ac, S2P-Pol II, and GCN5 (Fig. 2A). If histone marks are important for regulating transcription by AML1-ETO, removing them should impact the modeling performance. To test this, we compared the performance of modeling using only certain combinations of acetyl-histone marks as well as removing certain histone features, which, in total, contain h3k18ac, h3k27ac, h3k27me3, h3k9ac, h3k9ac, h3k9me2, h3k9me3, and h3k4me1. Remarkably, as shown in Fig. 3B, the only histone feature whose removal significantly impacted the modeling performance is h3k27ac, which is a mark of active enhancers and which is catalyzed by p300. The performance upon removing h3k27ac was further reduced by removing H3K18ac, and to a larger extent, by removing both H3K18ac and H3K9ac. In modeling using only acetyl-histone marks, the performance was even better than that using all 96 features. Together, these results suggest that AML1-ETO-regulated transcription is mainly impacted by acetyl-histone marks, especially H3K27ac, consistent with its ability to serve as a mark of active enhancers. When modeling was performed using all 8 histone variables, the "delta" acetyl-histone variables (especially H3K27ac) continued to have the highest positive coefficients whereas the "h3k18ac_wo_ae" had the highest negative coefficient (data not shown). The failure to detect "h3k27ac_wo_ae", i.e., the pre-existing level of H3K27ac, as an important negative variable was likely due to collinearity between H3K18ac and h3k27ac, both of which are catalyzed by p300. Supporting this idea, removing H3K18ac allowed detection of "H3K27ac_wo_ae" as the most important negative variable (data not shown).

3.5 Removing Histone Variables Allows Revealing HATs as the Important Variables

Since acetyl-histones are catalyzed by HATs, the failure to detect HATs as the rate-limiting proteins that contribute to activation may be caused by the "masking" effect of the histone variables. To test this, we performed modeling using only non-histone variables. Consistent with the idea, excluding histone variables allowed the detection of "delta_p300_shae" as one of the most important variables that contribute to activation, and detection of "p300_wo_ae", which measures the pre-existing level of p300, as the most important variable that contributes to repression (Fig. 4A). Other positive variables were AML1-ETO-regulated changes of GCN5 and E2A and the enhancer status. It has been previously shown that AML1-ETO dismisses GCN5 to mediate repression [12]. This idea seems contradictory to our ability to detect "delta_gcn5-shae" as a positive variable. However, a positive coefficient could apply to two scenarios. In one scenario, it means that AML1-ETO increases this feature to mediate activation. Alternatively, it could also mean that AML1-ETO decreases this feature to mediate repression. Consistent with these ideas, whereas AML1-ETO recruited p300 at the activation sites, it reduced GCN5 at the repression sites (Fig. 4B). Thus, our modeling results are consistent with the previously reported effects of

AML1-ETO on HAT recruitment [31]. In addition to the pre-existing level of p300, the pre-existing levels of total Pol II, S2P-Pol II and CDK9, as well as the binding level of AML1-ETO, were also among the LASSO-detected most important negative variables (Fig. 4A). Together with the negative variables detected in the modeling using all 96 binding site features, these results are consistent with the idea that, whereas the direction of gene expression changes (i.e., activation or repression) can be explained by the changes of histone marks, HATs or other features, the likelihood of these changes is pre-determined by the pre-existing state of the AML1-ETO binding sites. Thus, repression predominantly occurred at sites with high levels of HATs and transcriptional activities. These sites also experienced high levels of AML1-ETO binding in order to dismiss HATs. Activation was favored when AML1-ETO bound to enhancer sites carrying low levels of p300 because this is more likely to increase p300 (Fig. 4B).

3.6 p300, But Not HDACs, is the Rate-Limiting Enzyme Regulating H3K18/K27 Acetylation

Our finding that H3K27ac/K18ac are among the most important features dictating activation versus repression prompted us to further study the rate-limiting factors regulating these histone modifications. Although histone acetylation is catalyzed by HATs such as p300, it is also subjected to deacetylation by HDACs. Both p300 and HDACs have been shown to directly bind to AML1-ETO. To examine the regulatory factors for H3k27ac/K18ac, we similarly used logistic regression and modeled the effects of AML1-ETO on H3K18ac/K27ac levels as binary variables against all non-histone variables. The binary variable had a value of 1 if AML1-ETO increases H3K18ac or H3K27ac or -1 if AML1-ETO decreases them. The consensus from the modeling results of both H3K18ac and H3K27ac is that the p300-related variables were consistently detected both as the variable ("delta_p300_shae") that positively contributed to the regulation of H3K18ac/K27ac by AML1-ETO and as the variable ("p300_without_ae") that negatively contributed to this regulation. In contrast, HDACs were not identified as a consistent contributing factor for both H3K18ac/H3K27ac regulation by AML1-ETO. Further supporting the idea that p300, but not HDACs, is the rate-limiting factor for H3K18/K27 acetylation, removing p300 but not HDACs dramatically impacted the performance of modeling AML1-ETO regulation of H3K18ac/27ac (Figs. 5C and 5D).

3.7 AML1-ETO Binding Results in Coordinated Changes of Network Modules

Notably, Fig. 5 also showed that H3K18/K27 levels were correlated with some other non-HAT proteins. This raised the possibility that the different chromatin regulators are inter-connected and, therefore, the changes of these levels may occur in a coordinated fashion in response to AML1-ETO expression. For example, components of the same complex are likely to show similar changes with and without AML1-ETO. Additionally, it has been proposed that AML1-ETO directly binds p300 to recruit it to chromatin. This would predict that modeling AML1-ETO recruitment of p300 should identify AML1-ETO binding as a significant positive variable. To test these ideas, we expanded the modeling of H3K18ac/H3K27ac in Fig. 5 to include all 26 cistromic "delta" variables that had paired values with and without AML-ETO (Table 2). Similar to what we did for H3K18ac/H3K27ac, all ChIP-Seq derived variables were converted into 1 (if AML1-ETO increased it) or -1 (if AML1-ETO decreased it). This was followed by LASSO-regularized logistic regression against all other variables shown in Fig. 6, which was performed for 500 independent runs. The means of coefficients were combined into a 2D 28 x 28 matrix (26 "delta" variables + activation + enhancer). This matrix was further converted into a Pearson correlation matrix in order to reveal the similarity/coordination among the changes of the features. Confirming the idea, unsupervised clustering (Fig. 6) and network graph (Fig. 7) analyses clearly revealed several modules that showed coordinated changes with and without AML1-ETO, in a manner consistent with previous studies and our current results. For example, activation was localized in the same module with H3K18ac/H3K27ac. Moreover, the active form of Pol II (S2P-Pol II) overlapped with both the activation module and the enhancer module. Consistent with the report that RUNX1 associates with the HDAC3/SMRT complex [31], RUNX1 was placed in the same module as HDAC3 and SMRT. Notably, this module also included GCN5, which may reflect the fact that AML1-ETO binding was competitive both with the binding of GCN5 (on E-proteins) and with RUNX1 (on DNA). Surprisingly, AML1-ETO binding was not detected as a positive contributor to p300 recruitment, but, paradoxically, as one of the most negative variables, when adjusted to other predicting variables (Fig. 8). These and other results support the view that AML1-ETO recruits p300 by recruiting E-proteins. First, the heatmap showed that p300 was clustered closely with E2A and HEB (Fig. 6). Second, results from the networked graph analysis showed that p300 was in the same module as E2A and HEB (Fig. 7). Third, AML1-ETO recruitment of E2A and HEB was among the most important positive variables that contributed to AML1-ETO recruitment of p300 (Fig. 8). Fourth, comparing the performance of modeling AML1-ETO recruitment of p300 revealed a significant reduction of modeling performance upon removing E2A and HEB (Fig. 8). Finally, providing the definitive involvement of E-proteins (HEB) in the recruitment of p300 by AML1-ETO, depleting HEB alone reduced p300 recruitment in a manner recapitulating the effect of depleting AML1-ETO. This occurred preferentially to the AML1-ETO activation sites (Fig. 8). We propose that AML1-ETO can play both a positive role and a negative role in the recruitment of p300. This corresponds to the ability of AML1-ETO to recruit p300 by recruiting E-proteins at the activation sites and dismissing p300 that has preoccupied E-proteins at the repression sites.

4 Discussion

4.1 LASSO-Regularized Logistic Regression Robustly Detects Rate-Limiting Factors for Gene-Specific Repression, Activation and Chromatin Interactions

Our results show that LASSO can robustly detect the rate-limiting factors that contribute to gene-specific repression, activation and chromatin interactions by AML1-ETO.

Clearly, if a given feature is determined by LASSO to have a zero coefficient, it would indicate that its levels are not causally important for the dependent variable, regardless of whether this variable is the difference of gene expression or the difference of chromatin interactions. Conversely, if a given feature is found to contribute to the differences of the dependent variable, it would indicate that the changes of its levels are important for the changes of gene expression and chromatin interactions. Subsequent studies comparing modeling with and without these features may definitively tell whether these variables are the rate-limiting factors.

Since transcription is a stepwise process, its inhibition or stimulation at a given step should only be determined by the immediate upstream regulators. We have shown that LASSO detected histone marks as the important features for both activation and repression. It is known that histone modifications play important roles both in activation and in repression. This explains why LASSO had a better performance compared to some other regularization methods. These methods tended to dampen the importance of histone variables, explaining why they had a relatively poorer performance than LASSO. There is a concern that since LASSO is associated with sparse solutions, the results may not generalize well to all genes. However, the activation and repression mechanisms of a given TF, such as AML1-ETO, may be shared by its genome-wide target genes. In support of this idea, our preliminary studies have shown that blocking the interaction between AML1-ETO and E-proteins using an E-protein-derived polypeptide impact genome-wide AML1-ETO-repressed and AML1-ETO-activated genes (C.G. and J.Z., unpublished data). Another concern is that, as a biased method, LASSO may produce different solutions in different runs, especially if the features are correlated. We found that averaging the results from multiple runs can help to overcome this issue. Alternatively, group lasso or other related algorithms [42], [43], [44], [45] may prove to be useful. Finally, while many TFs may behave similarly to AML1-ETO by sharing common mechanisms of repression and activation genome-wide, other TFs may behave differently. This will be the case if the repression and activation are controlled by different histone marks (e.g., acetyl- or methyl-lysine) at different locations. One potential solution to this may be to cluster these genes before performing logistic regression with LASSO or some other algorithms that can produce sparse and less biased solutions [37], [38], [39], [40], [42], [43], [44], [45].

4.2 Coefficients and the Direction of Changes

An interesting observation of our study is that both p300 and GCN5 have positive coefficients yet previous studies have shown that they are differentially involved in AML1-ETO-mediated activation and repression. Thus, whereas AML1-ETO recruits p300 to activate transcription, it dismisses GCN5 to repress transcription. This shows that correct integration of the machine learning results still requires consideration of the experimental data to determine the direction of the changes that actually occur to specific genes.

4.3 Rate-Limiting versus Regulatory Factors

The reason why HDACs were not detected as important features by LASSO may be explained by our finding that HDAC3 binds to both AML1-ETO and RUNX1. Because AML1-ETO and RUNX1 competitively bind to target genes, this would result in the situation in which HDAC3 constitutively occupies the binding site, making HATs the rate-limiting factors. Nevertheless, HDACs may still play a role in regulating gene transcription. As such, one of the functions of HATs may be to overcome the effects of HDACs.

4.4 Unified Model of Gene Regulation by AML1-ETO

Our findings that E-proteins mediate the AML1-ETO-dependent recruitment of p300 in gene activation, along with the previous report that AML1-ETO dismisses GCN5 from E-proteins in gene repression [11], support the idea that interactions between AML1-ETO and E-proteins are commonly involved in gene-specific activation and repression by AML1-ETO, and suggest a unified model of gene regulation by AML1-ETO. We propose that at the binding sites with low levels of pre-existing transcriptional activities and p300 (e.g., poised enhancers), binding by AML1-ETO to these sites will act to increase the level of p300 by recruiting more E-proteins, which leads to activation. In contrast, at the binding sites with high levels of pre-existing transcriptional activity and the related regulatory proteins, such as E-proteins and GCN5, a binding of AML1-ETO to these sites will preferentially dismiss GCN5, leading to repression.

It should be mentioned that, in this study, we have used RNA-Seq and ChIP-Seq to pre-select high-confidence AML1-ETO-regulated genes in order to meaningfully compare repression and activation mechanisms. The generated results will guide our future efforts by extending the modeling to genes less strongly regulated by AML1-ETO. Our ultimate goal is to predict the functional effect of AML1-ETO based only on its binding site location and the associated cistromic effects. Clearly, AML1-ETO binding to a given site may not always produce a strong effect on transcription. Conceivably, if AML1-ETO binding does not affect transcription, it would indicate that there is little communication between the binding site and the promoter/TSS of the gene. Therefore, including variables that can effectively model the communication should not only benefit the effort to predict the activated genes as indicated earlier, but also benefit the modeling of genes that are less influenced by AML1-ETO. Lastly, although this work has studied only a subset of AML1-ETO-regulated genes, these genes tend to play important roles in AML1-ETO's biological functions. Therefore, an interesting follow-up study is to predict the effects of AML1-ETO under new conditions. For example, we expect that manipulating the pre-existing transcriptional activity of the AML1-ETO binding sites may abolish the biological function of AML1-ETO by reversing both the activation and repression effects of AML1-ETO.

4.5 Implications in Other TFs

The basic transcriptional mechanisms of AML1-ETO are not different from those for other TFs, in that they share the same sets of histone marks, HATs, HDACs, and the components of the basal transcriptional machinery. Therefore, the logistic regression approach shown in this study should be generally applicable to studying the

mechanisms of diverse WT and aberrant TFs in their regulation of gene-specific repression and activation, which play important roles in their biological functions under physiological and pathological conditions. To do so, one will need to perform RNA-Seq and ChIP-Seq assays in relevant cells under paired conditions with and without the expression of the corresponding TF, followed by classification with logistic regression, which should benefit from LASSO-based regularization.

5 Conclusion

1. To our knowledge, this is the first study attempted to model regulated changes of gene expression and chromatin interactions influenced by a sequence-specific TF (AML1-ETO or HEB) under paired treatment conditions, e.g., with and without the expression of AML1-ETO. Although there are numerous examples of machine learning-assisted modeling of chromatin and transcriptional regulation [17], [18], [19], [20], [21], [27], [28], [29], most of these studies only analyzed features under one treatment condition (e.g., resting cells), limiting their ability to draw causal conclusions.
2. Our results show that logistic regression is a useful method to model binary changes of the regulated transcription and chromatin interactions and that LASSO is highly sensitive to detect the important rate-limiting factors underlying these changes.
3. An iterative approach can be used to reveal the coordinated changes and network interactions of the various features.

Acknowledgments

This work was supported in part by Grants R01HL093195 and R21CA178513 (NIH), in part by the President's Research fund (Saint Louis University), and in part by the Siteman Investment Program Award (Siteman Cancer Center at Barnes-Jewish and Washington University) to Jinsong Zhang. Nickolas Steinauer and Kevin Zhang contributed equally to this work.

References

[1] M. G. Rosenfeld, V. V. Lunyak, and C. K. Glass, "Sensors and signals: A coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response," *Genes Develop.*, vol. 20, no. 11, pp. 1405–28, 2006.

[2] F. Gong and K. M. Miller, "Mammalian DNA repair: Hats and HDACs make their mark through histone acetylation," *Mutat. Res./Fundam. Mol. Mech. Mutagenesis*, vol. 750, no. 1–2, pp. 23–30, 2013.

[3] Y. Dou *et al.*, "Physical association and coordinate function of the H3 K4 methyltransferase MLL1 and the H4 K16 acetyltransferase MOF," *Cell*, vol. 121, no. 6, pp. 873–875, 2005.

[4] V. Perissi and M. G. Rosenfeld, "Controlling nuclear receptors: The circular logic of cofactor cycles," *Nat. Rev. Mol. Cell Biol.*, vol. 6, no. 7, pp. 542–554, 2005.

[5] E. Reed-Inderbitzin plbibitalic- *et al.*, "RUNX1 associates with histone deacetylases and SUV39H1 to repress transcription," *Oncogene*, vol. 25, no. 42, pp. 5777–5786, 2006.

[6] K. Durst and S. Hiebert, "Mole of RUNX family members in transcriptional repression and gene silencing," *Oncogene*, vol. 23, no. 24, pp. 4220–4224, 2004.

[7] M. Murphy, "P53, transcriptional repression, and drug sensitivity: Fresh perspectives on an old activity," *Cell Cycle*, vol. 9, no. 22, 2010, Art. no. 4432.

[8] D. Datta De, A. Datta, S. Bhattacharjya, and S. Roychoudhury, "NF-kappaB mediated transcriptional repression of acid modifying hormone gastrin," *PLoS One*, vol. 8, no. 8, 2013, Art. no. e73409.

[9] C.-H. Wang, Y.-P. Tsao, H.-J. Chen, H.-L. Chen, H.-W. Wang, and S.-L. Chen, "Transcriptional repression of p21(Waf1/Cip1/Sdi1) gene by c-jun through Sp1 site," *Biochem. Biophys. Res. Commun.*, vol. 270, no. 1, pp. 303–310, 2000.

[10] H. Miyoshi *et al.*, "The T(8;21) translocation in acute myeloid leukemia results in production of an AML1-MTG8 fusion transcript," *EMBO J.*, vol. 12, no. 7, pp. 2715–21, 1993.

[11] C. H. Gow, C. Guo, D. Wang, Q. Hu, and J. Zhang, "Differential involvement of E2A-corepressor interactions in distinct leukemogenic pathways," *Nucleic Acids Res.*, vol. 42, no. 1, pp. 137–52, 2014.

[12] N. Steinauer *et al.*, "Myeloid translocation gene CBFA2T3 directs a relapse gene program and determines patient-specific outcomes in AML," *Blood Adv.*, vol. 3, no. 9, pp. 1379–1393, 2019.

[13] J. Zhang, M. Kalkum, S. Yamamura, B. T. Chait, and R. G. Roeder, "E protein silencing by the leukemogenic AML1-ETO fusion protein," *Science*, vol. 305, no. 5688, pp. 1286–1289, 2004.

[14] L. Wang *et al.*, "The leukemogenicity of AML1-ETO is dependent on site-specific lysine acetylation," *Science*, vol. 333, no. 6043, pp. 765–769, 2011.

[15] B. A. Hug and M. A. Lazar, "ETO interacting proteins," *Oncogene*, vol. 23, no. 24, pp. 4270–4274, 2004.

[16] B. Lutterbach *et al.*, "ETO, a target of t(8;21) in acute leukemia, interacts with the N-cor and mSin3 corepressors," *Mol. Cell Biol.*, vol. 18, no. 12, pp. 7176–7184, 1998.

[17] H. R. Frost and C. I. Amos, "Gene set selection via LASSO penalized regression (SLPR)," *Nucleic Acids Res*, vol. 45, no. 12, 2017, Art. no. e114.

[18] D. U. Gorkin *et al.*, "Integration of CHIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes," *Genome Res.*, vol. 22, no. 11, pp. 2290–2301, 2012.

[19] D. H. Oh, I. B. Kim, S. H. Kim, and D. H. Ahn, "Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning," *Clin. Psychopharmacol. Neurosci.*, vol. 15, no. 1, pp. 47–52, 2017.

[20] X. Xu, S. Hoang, M. W. Mayo, and S. Bekiranov, "Application of machine learning methods to histone methylation CHIP-seq data reveals H4R3me2 globally represses gene expression," *BMC Bioinf.*, vol. 11, 2010, Art. no. 396.

[21] M. Ram, A. Najafi, and M. T. Shakeri, "Classification and biomarker genes selection for cancer gene expression data using random forest," *Iranian J. Pathol.*, vol. 12, no. 4, pp. 339–347, 2017.

[22] M. Megraw, F. Pereira, S. T. Jensen, U. Ohler, and A. G. Hatzigeorgiou, "A transcription factor affinity-based code for mammalian transcription initiation," *Genome Res.*, vol. 19, pp. 644–656, 2009.

[23] J. M. Hariprakash and F. Ferrari, "Computational biology solutions to identify enhancers-target gene pairs," *Comput Struct Biotechnol. J.*, vol. 17, pp. 821–831, 2019.

[24] J. Moore, H. Pratt, M. Purcaro, and Z. Weng, "A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods," *Genome Biol.*, vol. 21, 2020, Art. no. 17

[25] A. Torang, P. Gupta, and D. Klinke, "An elastic-net logistic regression approach to generate classifiers and gene signatures for types of immune cells and T helper cell subsets," *BMC Bioinform.*, vol. 20, 2019, Art. no. 433.

[26] V. Ntranos, L. Yi, P. Melsted, and L. Pachter, "A discriminative learning approach to differential expression analysis for single-cell RNA-seq," *Nat. Methods*, vol. 16, no. 2, pp. 163–166, 2019.

[27] S. Ulianov *et al.*, "Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains," *Genome Res.*, vol. 26, no. 1, pp. 70–84, 2016.

[28] S. Liu, M. Lu, H. Li, and Y. Zuo, "Prediction of gene expression patterns with generalized linear regression model," *Front. Genet.*, vol. 10, 2019, Art. no. 120.

[29] H. Chaudhari and B. Cohen, "Local sequence features that influence AP-1 cis-regulatory activity," *Genome Res.*, vol. 28, no. 2, pp. 171–181, 2018.

[30] R. Tibshirani, "Regression shrinkage and selection via the lasso: A retrospective," *J. Royal Statist. Soc. Series B: Stat. Methodol.*, vol. 73, no. 3, pp. 273–282, 2011.

[31] C. Guo *et al.*, "Histone deacetylase 3 preferentially binds and collaborates with the transcription factor RUNX1 to repress AML1-ETO- dependent transcription in t(8;21) AML," *J. Biol. Chem.*, vol. 295, no. 13, pp. 4212–4223, 2020.
[32] N. Bray, H. Pimentel, P. Melsted, and L. Pachter, "Near-optimal probabilistic RNA-seq quantification," *Nat. Biotechnol.*, vol. 34, no. 5, pp. 525–527, 2016.
[33] C. Guo *et al.*, "Histone deacetylase 3 preferentially binds and collaborates with the transcription factor RUNX1 to repress AML1-ETO- dependent transcription in t(8;21) AML," *J. Biol. Chem.*, vol. 295, no. 13, pp. 4212–4223, 2020.
[34] S. Heinz *et al.*, "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities," *Mol. Cell.*, vol. 38, no. 4, pp. 576–589, 2010.
[35] T. Nguyen *et al.*, "High-throughput functional comparison of promoter and enhancer activities," *Genome Res.*, vol. 26, no. 8, pp. 1023–1033, 2016.
[36] J. Tolles and W. Meurer, "Logistic regression: Relating patient characteristics to outcomes," *JAMA - J. Amer. Med. Assoc.*, vol. 316, no. 5, pp. 533–534, 2016.
[37] X. Tian, X. Wang, and J. Chen, "Network-constrained group lasso for high-dimensional multinomial classification with application to cancer subtype prediction," *Cancer Inf.*, vol. 13, pp. 25–33, 2014.
[38] J. Ren *et al.*, "Network-based regularization for high dimensional SNP data in the case-control study of type 2 diabetes," *BMC Genet.*, vol. 18, no. 1, 2017, Art. no. 44.
[39] J. Ren, Y. Du, S. Li, S. Ma, Y. Jiang, and C. Wu, "Robust network-based regularization and variable selection for high-dimensional genomic data in cancer prognosis' *Genet. Epidemiol.*, vol. 43, pp. 276–291, 2019.
[40] L. Spirko-Burns and K. Devarajan, "Supervised dimension reduction for large-scale "omics" data with censored survival outcomes under possible non-proportional hazards," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jan. 10, 2020, doi: 10.1109/TCBB.2020.2965934.
[41] S. Schoenfelder and P. Fraser, "Long-range enhancer–promoter contacts in gene expression control," *Nat. Rev. Genet.*, vol. 20, no. 8, pp. 437–455, 2019.
[42] J. Ogutu and H.-P. Piepho, "Regularized group regression methods for genomic prediction: Bridge, MCP, SCAD, group bridge, group lasso, sparse group lasso, group MCP and group SCAD," *BMC Proc.*, vol. 8, 2014, Art. no. S7.
[43] J. Li, Y. Wang, H. Xiao, and C. Xu, "Gene selection of rat hepatocyte proliferation using adaptive sparse group lasso with weighted gene co-expression network analysis," *Comput. Biol. Chem.*, vol. 80, pp. 364–373, 2019.
[44] L.-Z. Liu, F.-X. Wu, and W.-J. Zhang, "A group lasso-based method for robustly inferring gene regulatory networks from multiple time-course datasets," *BMC Syst. Biol.*, vol. 8, 2014, Art. no. S1.
[45] G. Xie, C. Dong, Y. Kong, J. F. Zhong, M. Li, and K. Wang, "Group lasso regularized deep learning for cancer prognosis from multi-omics and clinical features," *Genes*, vol. 10, no. 3, 2019, Art. no. 240.

Nickolas Steinauer received the PhD degree in May 2020 from the Department of Pharmacology and Physiology, and is currently working toward the 3rd year of MD/PhD degree with Medical School, Saint Louis University. His research interests include bioinformatics analysis of patient data, ChIP-Seq data, and RNA-Seq data.

Kevin Zhang is currently working toward the undergraduate degree in his senior year with Northwestern University. He is currently working toward the major in statistics and a minor in computer science. His research interest focuses on big data analysis.

Chun Guo received the MS degree in biochemistry and molecular biology from the University of New Hampshire. She has expertise in shRNA-mediated gene knockdown assays and ChIP-Seq library construction. She is currently a research assistant with the Department of Pharmacology and Physiology, Saint Louis University.

Jinsong Zhang is currently an associate professor with the Department of Pharmacology and Physiology, Saint Louis University. His research interests include dissecting the roles of coactivators and corepressors in transcriptional regulation and the understanding of molecular and cellular mechanisms of gene regulation using mechanistic, high-throughput, and bioinformatics tools.