ANNUAL REVIEWS

Annual Review of Genomics and Human Genetics

Enhancer Predictions and Genome-Wide Regulatory Circuits

Michael A. Beer,[1] Dustin Shigaki,[1] and Danwei Huangfu[2]

[1]Department of Biomedical Engineering and McKusick-Nathans Department of Genetic Medicine, Johns Hopkins University, Baltimore, Maryland 21205, USA; email: mbeer@jhu.edu

[2]Sloan Kettering Institute, New York, NY 10065, USA; email: huangfud@mskcc.org

Annu. Rev. Genom. Hum. Genet. 2020. 21:37–54

First published as a Review in Advance on May 22, 2020

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

https://doi.org/10.1146/annurev-genom-121719-010946

Keywords

enhancers, machine learning, gene regulatory networks, sequence-based prediction, cell fate switching

Abstract

Spatiotemporal control of gene expression during development requires orchestrated activities of numerous enhancers, which are *cis*-regulatory DNA sequences that, when bound by transcription factors, support selective activation or repression of associated genes. Proper activation of enhancers is critical during embryonic development, adult tissue homeostasis, and regeneration, and inappropriate enhancer activity is often associated with pathological conditions such as cancer. Multiple consortia [e.g., the Encyclopedia of DNA Elements (ENCODE) Consortium and National Institutes of Health Roadmap Epigenomics Mapping Consortium] and independent investigators have mapped putative regulatory regions in a large number of cell types and tissues, but the sequence determinants of cell-specific enhancers are not yet fully understood. Machine learning approaches trained on large sets of these regulatory regions can identify core transcription factor binding sites and generate quantitative predictions of enhancer activity and the impact of sequence variants on activity. Here, we review these computational methods in the context of enhancer prediction and gene regulatory network models specifying cell fate.

ENHANCERS IN DEVELOPMENT AND HUMAN DISEASE

Most of our understanding of the function of enhancers comes from developmental biology or studies of the genetics of human disease. Human traits typically have a hereditary component but demonstrate complex patterns of inheritance. Genome-wide association studies (GWASs) have been widely used to identify complex trait loci and have identified more than 25,000 single-nucleotide polymorphisms (SNPs) that are significantly associated with variation in more than 700 traits and diseases (74). Although GWASs can still explain only a small fraction of the phenotypic variance (50), the list of validated regulatory mutations responsible for heritable susceptibility to diseases is growing at a steady rate.

The significant role of regulatory variation in complex trait heritability is underscored by the finding that the vast majority of trait-associated SNPs are non-exonic (49) and occur within putative regulatory elements far more often than expected by chance (30, 52). This suggests that disruption of regulatory function is a common mechanism by which noncoding sequence variants contribute to human disease. When a regulatory variant is identified, it is often hypothesized that the variant disrupts a transcription factor (TF) binding site, creates a new binding site, or both. In a recently elucidated example, a GWAS showed that the common SNP rs339331 increases prostate cancer risk (68) (odds ratio $= 1.22$, $p = 1.6 \times 10^{-12}$). Huang et al. (32) dissected this locus and showed that the risk SNP allele TTTTATGAG is bound by HOXB13, while the protective allele TTTCATGAG is not bound by HOXB13. This particular TF, in combination with FOXA1 and AR, activates *RFX6* and promotes cell migration and metastatic disease (32). Since approximately 50 SNPs are typically in tight linkage disequilibrium with each GWAS-associated variant, similar detailed experimentation will be required to identify causal variants within disease-associated loci, but only a small number of these loci have been studied in detail.

A long-standing problem encountered when attempting to generalize known binding site disruptions is that the biological consequences of variation in a specific binding site are strongly dependent on both cell type and the neighboring local sequence context, which defines the combinatorial TF interactions with cell-specific cofactors. Because most TF binding sites are short and degenerate, there are usually thousands of what appear to be very good binding sites in the genome, yet only a fraction of these are occupied in a given cell type (7). Mutation of a binding site will have a functional consequence only in an occupied site. Therefore, the combinatorial code that determines cell-specific TF occupancy will determine which variants can alter regulatory element activity. Several computational methods have shown promise in detecting and quantitatively assessing the impact of variants in enhancers by training on uniformly processed genome-wide epigenomic data sets generated by the Encyclopedia of DNA Elements (ENCODE) Consortium and National Institutes of Health Roadmap Epigenomics Mapping Consortium (20, 62, 77).

In the context of embryonic development, multicellular organisms require cells to make fate decisions by integrating extracellular cues ranging from biochemical to mechanical signals. We now have a sophisticated understanding of how extracellular inputs, especially signaling molecules such as SHH or TGF-β, are transduced intracellularly and ultimately activate a relatively small set of TFs that play pivotal roles in cell fate determination. For instance, Nodal/TGF-β signaling is required for specifying definitive endoderm (DE) differentiation in gastrulating mouse and zebrafish embryos (2, 14, 21, 63), a process that has been recapitulated using human embryonic stem cells (hESCs) and mouse embryonic stem cells through directed differentiation (17, 38). Nodal/TGF-β signaling activates the SMAD2–4 TFs, which then cooperate with key lineage TFs, including FOXH1, EOMES, MIXL1, and GATA6, to activate the DE transcriptional program (46, 48, 58, 69, 72).

Indeed, Li et al. (46) uncovered all of these core TF genes (*FOXH1*, *EOMES*, *MIXL1*, *GATA6*, *SMAD2*, and *SMAD4*), along with additional new regulators, in pooled, genome-scale

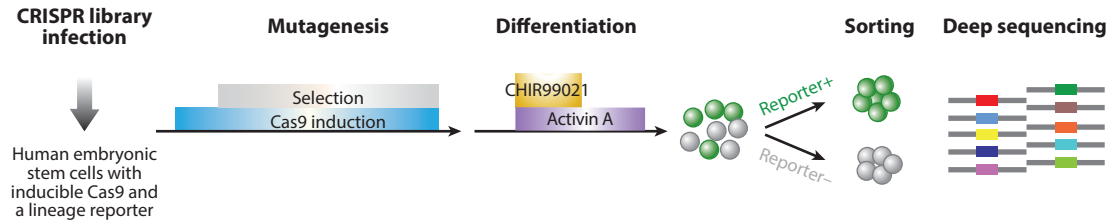


Figure 1

Overview of an in situ perturbation screening strategy to uncover core regulators based on lineage reporters. This general strategy is also applicable to other differentiation protocols.

CRISPR/Cas9 loss-of-function screens for genes that are required for DE differentiation from hESCs, as shown in **Figure 1**. This study used the ESC–DE differentiation system to interrogate DNA elements in the regulatory network that controls induction of hESCs into DE, induced in this system by TGF-β and Wnt signaling. DE differentiation was triggered when hESCs were treated with CHIR-99021 and activin A (**Figure 1**). A *SOX17*/GFP knock-in allele reported the DE fate (35), as assessed by flow cytometry. To detect regulators of this process, iCas9 *SOX17*/GFP cells were infected with the human Genome-Scale CRISPR Knock-Out (GeCKO) v2 guide RNA (gRNA) library. After selection for cells with viral integration and induction of Cas9 expression, DE differentiation was performed, and *SOX17*/GFP+ DE and *SOX17*/GFP− non-DE cells were isolated by fluorescence-activated cell sorting. The abundance of individual gRNAs in each population was determined by high-throughput sequencing: gRNAs that target positive or negative regulators of endoderm specification should be depleted or enriched, respectively, in *SOX17*/GFP+ compared with *SOX17*/GFP– cells. A *Z*-score was calculated for each gRNA based on the ratio of gRNA reads in the populations. The top 20 hits included almost all of the nonredundant, cell-autonomous required genes in the Nodal pathway (*ACVR1B*, *SMAD2*, and *FOXH1*) (63) as well as the established DE TF genes *EOMES* and *MIXL1* (80). TFs required to maintain the ESC state can also be screened by sequencing the pool of gRNAs enriched in self-renewing conditions.

However, there is a major gap in our knowledge of how TFs control the gene regulatory networks that dictate cell fate decisions. We often know which TFs are required for the acquisition or maintenance of a cell state during development, but the exact cascade of molecular events that either drives the cell state transition or stabilizes the cell state is unclear. Genomic data such as chromatin immunoprecipitation sequencing (ChIP-seq) data can provide rich information regarding the chromatin association of a TF, and knockout studies can identify genes with altered expression levels when the TF is deleted. However, these experiments may not indicate direct transcriptional consequences. There are two challenges to establish the causality of the TF binding and gene expression changes relevant to cell fate determination. First, a TF (e.g., TF *A*) usually has multiple binding sites near a gene of interest (e.g., gene *X*). Thus, even if the deletion of TF *A* causes a change of gene *X* expression, the impact (if any) of individual TF *A* binding sites on the control of gene *X* expression is typically unknown. Second, differentiation involves a cascade of molecular events, so it is conceivable that TF *A* regulates TF *B* expression, which then directly regulates the expression of gene *X*, even though TF *A* may also bind to genomic regions near gene *X*. In fact, multiple TFs often bind to the same region, but they may or may not directly contribute to transcriptional regulation, and some of the TFs may have overlapping, additive, synergistic, or buffering effects. Globally, it is challenging to determine from the genomic occupancy pattern alone which TFs are required for the cell to make a specific cell fate decision (51).

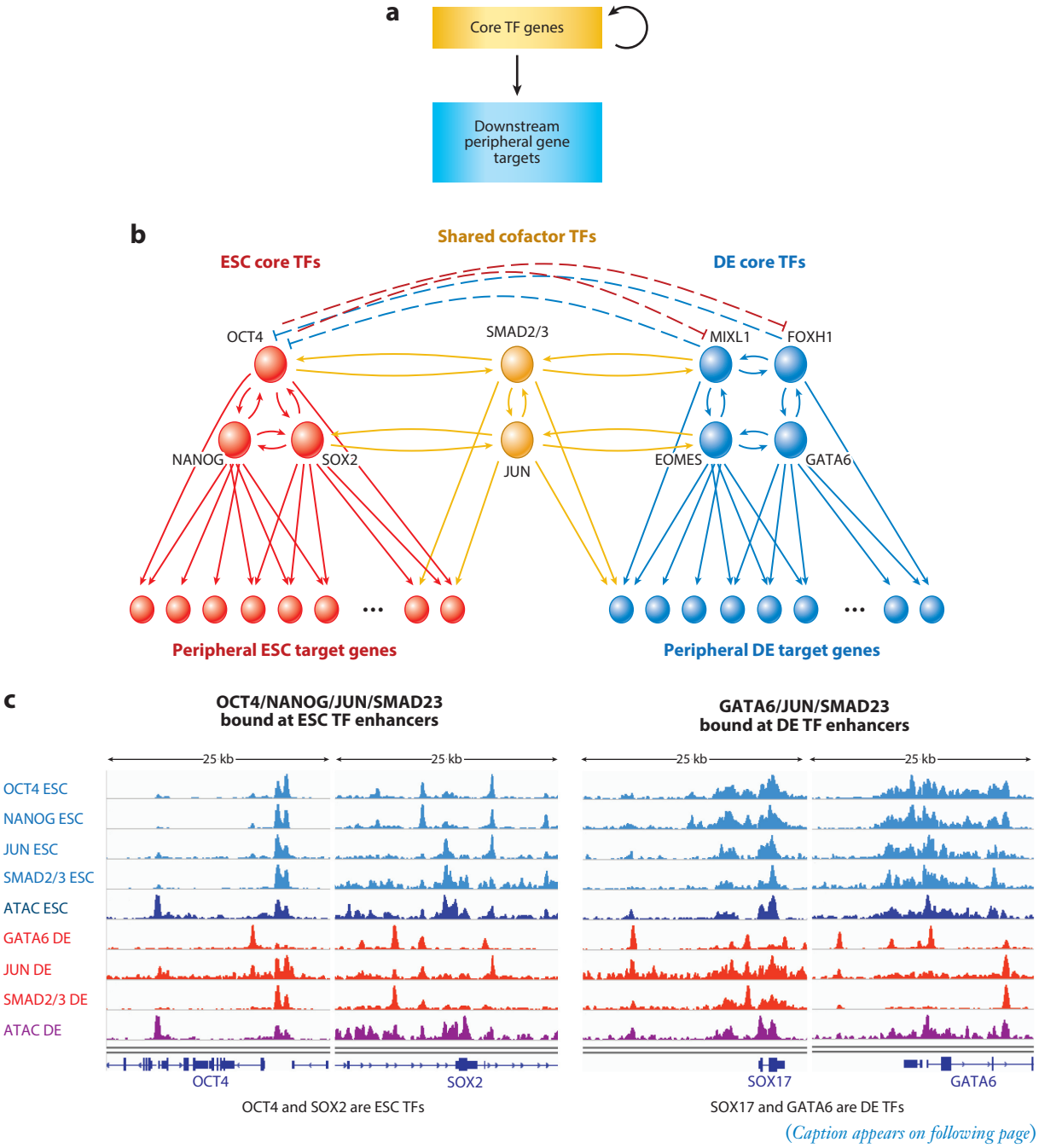
Therefore, in order to establish a predictive gene regulatory network for cell fate control, it is necessary to build on our knowledge of the TFs required for fate specification. This will allow us to identify cognate functional enhancer regions and measure the local and global consequences of perturbing these enhancers. For this purpose, we have been focusing on enhancers that mediate the ESC–DE transition because of their importance to the development of endoderm-derived organs, including the pancreas and liver. We expect that the identification of functional TF binding sites within noncoding regulatory elements will establish edges (causal regulatory interactions between genes) in the gene regulatory networks. Measurement of the dynamics of these regulatory networks will form the basis for building quantitative models of enhancer function. Ultimately, these models should describe how combinatorial TF genomic occupancy at multiple enhancers controls lineage-specific gene expression in embryonic development, tissue homeostasis, regeneration, and aging.

CORE REGULATORY GENOMIC CIRCUITS

We believe it is useful to develop a conceptual framework to address the issues raised above. We reason that instead of characterizing or identifying enhancers based solely on their ability to drive gene expression, it would be more productive to devise targeted strategies to interrogate enhancers that play distinct roles in the gene regulatory network. In particular, some enhancers, by virtue of their ability to regulate the expression of fate-determining genes, may play central roles in development. Previous studies of developmental regulatory control have identified general principles that govern the organization and structure of gene regulatory networks, which are consistent across a wide range of multicellular model organisms (18). Particularly useful for our purpose is the separation of genes and enhancers into two classes: those whose primary role is to specify cell fate, which we will call core genes, and downstream genes, whose role lies downstream of the fate specification genes and which perform the necessary functions of the cell once its fate is set (**Figure 2*a***). There are also therefore two classes of developmental enhancers based on their endogenous activity and position of the genes they regulate in the network: core enhancers, which have a global impact in terms of cell fate maintenance or transition, and peripheral enhancers, which regulate the expression of one or sometimes multiple adjacent genes in *cis* but have little or no global (developmental) impact. From the viewpoint of the gene regulatory networks that control developmental lineage decisions, the core enhancers are likely to regulate the expression of lineage-determining TFs that connect them into nonlinear networks that produce bifurcations between stable cell states. The peripheral enhancers are downstream targets that have an impact on specific gene expression levels (e.g., of differentiated cells) but do not feed back into the control circuit. This separation of core and peripheral regulators has been particularly useful in interpreting the heritability of complex traits (10, 13, 47).

The identification of separate core and peripheral genes and enhancers in the regulatory network has implications for how to model regulatory networks using computational genomics and machine learning and how to interrogate these classes of enhancers using distinct experimental methods. Computational machine learning and statistical methods rely on the existence of many examples that contain patterns that represent likely predictive causal mechanisms. In the case of enhancer prediction, the biological structure of gene regulatory circuits provides the redundant examples from which these patterns can be learned. Each peripheral gene is driven by a small set of enhancers (at least one per cell type in which the gene is expressed, and sometimes only one), each containing binding sites for the core TFs (18). The fact that core genes are greatly outnumbered by peripheral genes has two key consequences: First, the set of core TFs in each cell type (the TF vocabulary) is small enough that the TF vocabulary of a given cell type is of limited complexity and is learnable, and second, the large number of peripheral genes (say, 5,000–15,000) active in

any cell type requires a large set of peripheral gene enhancers that can be used as examples to train the computational model. Thus, the success of machine learning methods in predicting peripheral enhancers by identification of core regulator binding sites within them (25, 42, 44) is consistent with the general principles derived from targeted studies of developmental regulatory networks (18).



(*Caption appears on following page*)

Figure 2 (*Figure appears on preceding page*)

Cell-specific gene regulatory network model. (*a*) Schematic of the relationship between core TF genes, which feed back on each other and define the stability and dynamics of the genetic network, and downstream peripheral genes, which perform necessary functions of the differentiated cell states but whose disruption does not dramatically affect the overall regulatory state of the cell. (*b*) Model of the ESC and DE regulatory network consistent with observations from sequence-based computational analysis, perturbative studies, and functional studies of the ESC–DE transition, where a small set of core regulators interact through local enhancers and target a large number of peripheral gene enhancers. (*c*) Differential core TF binding at core enhancers flanking key ESC (*OCT4* and *SOX2*) and DE (*SOX17* and *GATA6*) regulators. These functional studies show that these core TFs bind cooperatively at enhancers specific to ESC or DE states and that shared cofactors shuttle between binding cooperatively with different sets of the core factors active in each state across the transition. Abbreviations: DE, definitive endoderm; ESC, embryonic stem cell; TF, transcription factor.

Based on the study of the ESC–DE transition by Li et al. (46) and many previous works, we can summarize a schematic gene regulatory network model of the core regulators controlling the ESC–DE transition in **Figure 2*b***. The features of this model are built upon and consistent with the following observations from both perturbative functional studies and computational sequence analysis of cell-specific distal enhancers in many cell types (as described in more detail below):

- Predictive DNA sequence features map to a relatively small set of shared core TF binding sites in large sets of enhancers flanking both peripheral downstream target genes and core TFs.
- Cell-specific enhancers (both core and peripheral) are cooperatively bound by multiple core TFs.
- Cell-specific enhancers (both core and peripheral) contain multiple core TF binding sites.
- Shared cofactors bind cooperatively with distinct sets of core TFs in different cell states.
- Cell-specific core TF regulators are often controlled by autoregulatory self-bound enhancers.
- The stability of cell states is maintained by negative regulatory feedback between distinct core TF sets.

Examples of functional data across the ESC–DE transition supporting these global observations are shown in **Figure 2*c***, using ChIP-seq and chromatin accessibility [assay for transposase-accessible chromatin using sequencing (ATAC-seq)] data at two core ESC regulator genes (*OCT4* and *SOX2*) and two core DE regulator genes (*SOX17* and *GATA6*).

The fact that DNA sequence–based modeling can accurately predict a held-out test set, coupled with the observation that the features required to make this classification map to a relatively small set of TFs, implies that the set of core regulatory TFs is small (comprising 5–20 TFs) and that the much larger set of enhancers containing these binding sites map to peripheral target genes that do not typically affect the activity of the core regulator TFs directly. Additionally, each target enhancer typically contains TF binding sites for multiple core regulator genes.

Classifying enhancers into core and peripheral groups enables them to be interrogated separately using different methods. For the core enhancers, one only needs to focus on perturbing putative enhancers that regulate a relatively small number of core TF genes. Because core enhancers regulate core lineage-determining TFs, their perturbation would have a global impact on cell fate decisions. It is therefore possible to use a downstream cell fate–specific reporter (e.g., *OCT4*/GFP for hESC state or *SOX17*/GFP for the DE state) and large-scale, pooled CRISPR perturbation screens to determine the impact of perturbation on the cell fate. The regulation of hESC self-renewal and hESC–DE transition may involve both overlapping and distinct *cis*-regulatory sequences. It should be feasible to identify *cis*-regulatory elements that regulate the reporter gene expression (*OCT4* or *SOX17*). On the other hand, there are obvious risks for identifying enhancers that regulate an upstream lineage-determining TF, which in turn regulates the

cell fate decision. This identification would require the enhancer to have not only a relatively large effect on the transcription of the target gene but also secondary, tertiary, or even relatively indirect effects that would ultimately affect the cell fate.

GENERATING ENHANCER SETS FROM EPIGENOMIC DATA

Computational machine learning and statistical methods rely on the existence of many data points or training examples from which to extract patterns that represent likely predictive causal mechanisms. In the case of enhancer prediction, the biological structure of these circuits provides the redundant training examples from which these patterns can be learned. To train a DNA sequence model of enhancer activity, an appropriate enhancer training set must first be generated.

Enhancer activity is associated with both increased chromatin accessibility and histone modifications to chromatin state that contribute to the establishment and maintenance of activity (4, 8, 16, 31, 60, 61). Many epigenomic functional assays interrogate this active state and generate peaks of activity that can be used to define putative enhancer sets to train enhancer models, including ATAC-seq (11), DNase I hypersensitive site sequencing (DNase-seq) (9, 15, 64, 65, 70), histone ChIP-seq for acetylation of histone H3 on lysine 27 (H3K27ac) or monomethylation of histone H3 on lysine 4 (H3K4me1), and TF ChIP-seq (73) when core TFs are known. Trimethylation of histone H3 on lysine 4 (H3K4me3) marks are typically specific to promoters. ATAC-seq and DNase-seq reflect chromatin accessibility, which does not necessarily indicate enhancer activity but does have the advantage of higher spatial resolution relative to histone marks, which tend to flank the core TF binding sites. In addition, one can use ATAC-seq and DNase-seq without knowing the complete set of relevant TFs. For a complete set of TF ChIP-seq experiments in a given cell type, the full complement of TFs must be known, and good antibodies must exist—a tall order.

Once a set of appropriate marks are chosen, there are two common approaches to defining the training set. Many methods train on a limited positive set of 10,000–20,000 cell-specific peaks (1, 3, 5, 22, 25, 29, 41–44, 53, 66, 78) spanning 100–1,000 base pairs centered on the peak, along with a negative set of equal size or larger. In our work, we have found that a set of 20,000 300-base-pair sequences is usually close to optimal. This cell-specific training set approach has the advantage of focusing on identifying the core TF binding sites in that cell type. Other methods (36, 79) bin the genome in regularly spaced fixed-length (1,000-base-pair) bins that are not necessarily centered on an activity peak but have a multiclass label reflecting the epigenomic state of that bin for the full set of training data sets ($n = 919$ for the DeepSEA framework). This regular training bin approach has the advantage of generating a large set of sequences required for deep neural networks (DNNs) to obtain strong class-label accuracy but may miss the subtleties in the differences in TF vocabulary between specific biologically relevant cell states. This may be a particular concern for the less well-covered cell types in ENCODE, and such models should be retrained for these cases. In particular, the *RFX6* SNP example discussed above is missed by training on all ENCODE data sets at one time, even though the prostate cancer cell line LNCaP is included in the training set (6). When trained on focused ENCODE samples, sequence-based modeling can be used to refine the quality of the data sets (43).

DNA SEQUENCE–BASED MACHINE LEARNING ENHANCER MODELS

Support vector machines (SVMs) and DNNs are two of the main classes of machine learning methods that have been successful for enhancer prediction. These methods are trained to classify a set of positive and negative examples, and the result of training is a classifier score function that

can make predictions for the class of sequences outside the training set. SVM and DNN methods differ in the way the classifier score function is specified and how the parameters of this function are determined from the training data. They also differ in how the DNA sequence nucleotides are converted into an input vector of mathematical feature scores for each sequence element to be classified. In gkm-SVM (25), for example, each sequence is converted to a normalized vector of integer gapped *k*-mer counts. The parameters of this *k*-mer vocabulary are specified before training. In our studies, we use the full list of gapped *k*-mers of length L with k informative positions and $L - k$ free positions (gaps or wild cards) and typically use $(L,k) = (10,6)$ or (11,7); the latter is slightly more accurate and approximately half as fast. In the DNN, the input feature is later usually converted into a $4 \times L$ binary integer matrix, with each nucleotide represented by a permutation of (1,0,0,0).

The training sequence set determines the features detected and should be designed to most clearly reflect the specific biological processes one aims to model. For example, when building a sequence model to predict SNPs that affect chromatin accessibility [chromatin accessibility quantitative trait loci (caQTLs) or DNase I sensitivity quantitative trait loci (dsQTLs)] in a cell line or primary cells, it is important to include examples of all accessible regions that may be altered by genomic variants in the experiment. Thus, in order to predict dsQTLs in lymphoblasts (19, 42) or ATAC-QTLs in T cells (24), we trained gkm-SVM on a positive set of a large number (∼23,000) of peaks of length $L = 300$ base pairs centered on the peak signal versus a GC- and repeat-matched negative sequence set of the same size. For sets of this size, the gkm-SVM R package (27) is most convenient, but for larger training sets, LS-GKM (40) is recommended. Training the SVM yields a gkm-SVM score function $S(x_j) = \sum_i \alpha_i K(x_i, x_j)$ specified by the set of support vector coefficients that optimally separate the sequence elements in the positive and negative training sets (25). The gapped *k*-mer weight distribution is constructed from the gapped *k*-mer counts in the support vectors, $w_j = \sum_i \alpha_i K(x_i, x_j)$. We often map the gapped *k*-mer weights to full *k*-mer weights for ease of interpretability, which produces an equivalent scoring function after training (26). The tails of these weight distributions, shown in **Figure 3*a*,*b***, encode the features required to distinguish the cell-specific enhancer activity in the positive and negative training sets. In this case, these weights encode the TF binding sites required to predict chromatin accessibility in lymphoblasts, and the long positive tail of the weight distribution from 1 to 6 in **Figure 3*b*** maps to binding sites for the 10 TFs shown in **Figure 3*c***. In this case, there are three classes of features: CTCF, promoter-specific TF binding sites (NRF1, SP1, NFY, and ELK4), and lymphoblast distal enhancer-specific TF binding sites (IRF2, BATF, RUNX1, NF-κB, and PU.1). Lymphoblast dsQTL SNPs disrupt all of these TFs, and accurate prediction of lymphoblast dsQTLs requires resolution of all three classes of features. In DNNs, these important features are encoded in the first position weight matrix (PWM) layer of convolution filters (360 filters with a length of 8 base pairs are typically used).

An alternative training set design can be used to detect more specific regulatory signals. For instance, to detect the TFs controlling the ESC–DE transition, instead of training versus inaccessible genomic negative sequence and comparing the weight vectors, one can isolate the differentially active TFs by choosing as a positive set the most differentially accessible peaks. **Figure 4*a*** shows the ESC and DE ATAC-seq signal from the study by Li et al. (46) at the union of all peaks from each state. Training a gkm-SVM model using the 5,000 most differentially accessible peaks in DE as a positive set (blue in **Figure 4*a***) and the 5,000 most differentially accessible peaks in an ESC as a negative set (red in **Figure 4*a***) yields a classifier with an area under the receiver operating characteristics (AUROC) value of 0.92. The tails of this gkm-weight vector contain binding sites for the core TF regulators of the DE (TCF, EOMES, SMAD2/3, GATA, AP1, and FOXH1)

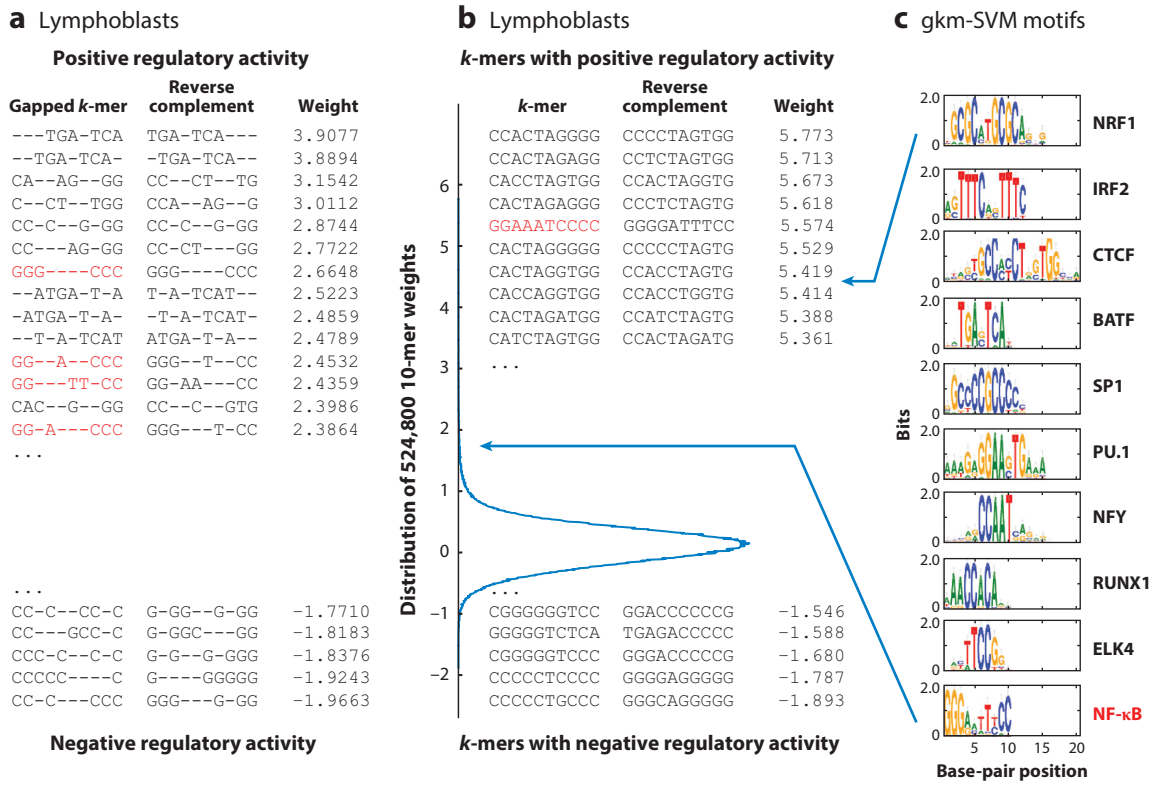


Figure 3

Quantification of the contribution of TF binding to cell-specific chromatin accessibility using a gkm-SVM gapped *k*-mer weight distribution. (*a*) Gapped *k*-mer weights for gkm-SVM trained on lymphoblast DNase I hypersensitive sites. (*b*) Mapping to full 10-mers, which produces an equivalent SVM scoring function (26). (*c*) The long positive tail of this weight distribution, which specifies the relative ranks of binding site strength for a set of active TFs in lymphoblasts. Highlighted in red are gapped *k*-mers (panel *a*), the top 10-mer GGAAATCCCC (panel *b*), and the PWM for NF-κB (panel *c*). Abbreviations: PWM, position weight matrix; SVM, support vector machine; TF, transcription factor.

and ESC states (OCT4, NANOG, SOX2, EBOX, and CTCF), whose PWMs and top weights are shown in **Figure 4*b*,*c***.

Training on ATAC-seq data from human tissue–derived cells or differentiated stem cells often detects very similar regulatory programs, as shown for human islet and pancreatic progenitor cells in **Figure 5**. These islet-specific enhancers form islet-specific DNA looping interactions in a type 2 diabetes–associated locus, as shown in **Figure 5*a***, as measured by promoter capture Hi-C (PCHi-C) (55). Although some computational models have been proposed to predict the gene targets of these enhancers (75), the predictive power of these methods is much lower than initially reported (12, 76), and additional higher-resolution enhancer–promoter interaction data are needed to develop improved models.

These sequence-based enhancer prediction models have been tested with luciferase and massively parallel reporter assays in a wide range of cell types (mouse liver, retina, neurons, and melanocytes as well as human T cells, lymphoblasts, and GM12878, K562, HepG2, and SK-N-SH cells) (6, 24, 34, 37, 39, 42, 53, 56, 59). Most recently, in a prediction assessment of massively parallel reporter assay data, Shigaki et al. (67) tested five enhancers and nine promoters using

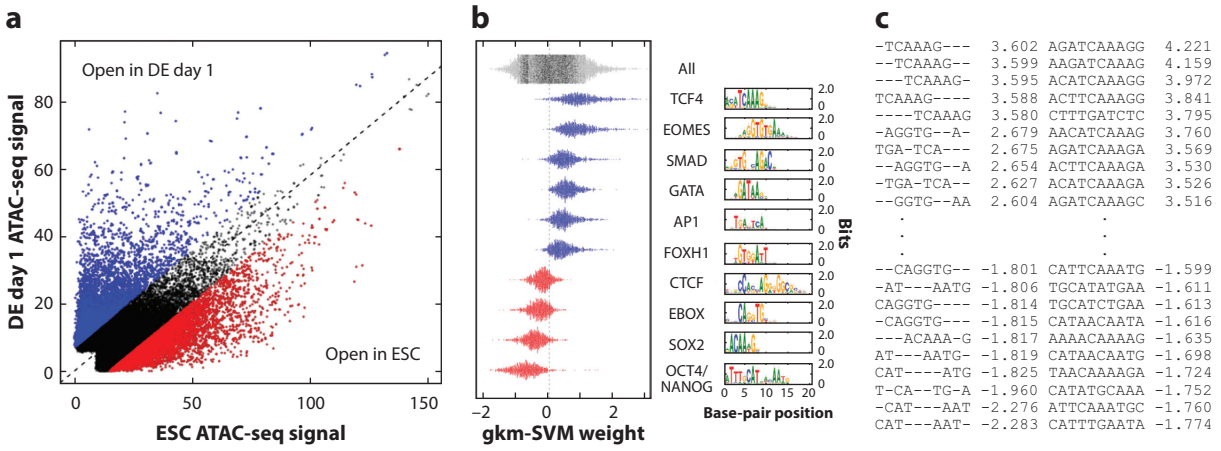


Figure 4

Detecting ESC and DE TF regulators via TF binding site mapping to the tail of the gkm-SVM weight distribution trained on differentially active ATAC-seq regions (AUROC = 0.92). (*a*) gkm-SVM is trained on DE day 1 open (*blue*) versus ESC open (*red*) ATAC-seq regions. (*b*) The core DE d1–specific TFs (*blue*) and ESC-specific TFs (*red*) are detected. Each dot is a distinct *k*-mer. From the two ATAC-seq experiments, a set of core regulators for the ESC and DE states can be found. (*c*) The top 10 positive and negative gapped *k*-mer and *k*-mer weights are shown, mapping to the TFs indicated in panel *b*. Abbreviations: ATAC-seq, assay for transposase-accessible chromatin using sequencing; AUROC, area under the receiver operating characteristics; DE, definitive endoderm; ESC, embryonic stem cell; SVM, support vector machine; TF, transcription factor.

saturation mutagenesis in disease-relevant cell types and found that the best models combined sequence features derived from enhancer prediction models that were trained on different data sets from ENCODE and the Roadmap Epigenomics Mapping Consortium. Using the same approach as previous studies (25, 42), the authors trained gkm-SVM on only the cell type–relevant DNase I hypersensitive site and ATAC-seq data set, which produced an average overall correlation with expression output of 0.39, as shown in **Figure 6*a***. Performance improved to a correlation of 0.58 when training on multiple data sets and combining the deltaSVM scores with random forest

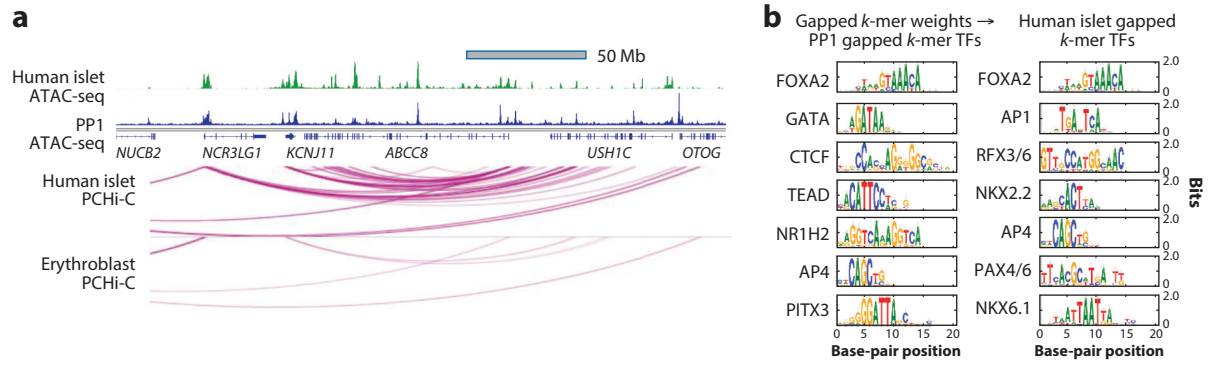


Figure 5

Similar TF vocabulary identified in human islets and stem cell–derived pancreatic progenitors. (*a*) ATAC-seq data from human islets (55, 71) and ATAC-seq data generated in PP1 cells (45) in the *KCNJ11-ABCC8* type 2 diabetes–associated locus detect peaks with islet-specific PCHi-C interactions (55). (*b*) gkm-SVM detects overlapping regulatory programs in ATAC-seq peaks from PP1 cells and islets and detects known islet regulators. Abbreviations: ATAC-seq, assay for transposase-accessible chromatin using sequencing; PCHi-C, promoter capture Hi-C; PP1, primary pancreatic progenitor; SVM, support vector machine; TF, transcription factor.

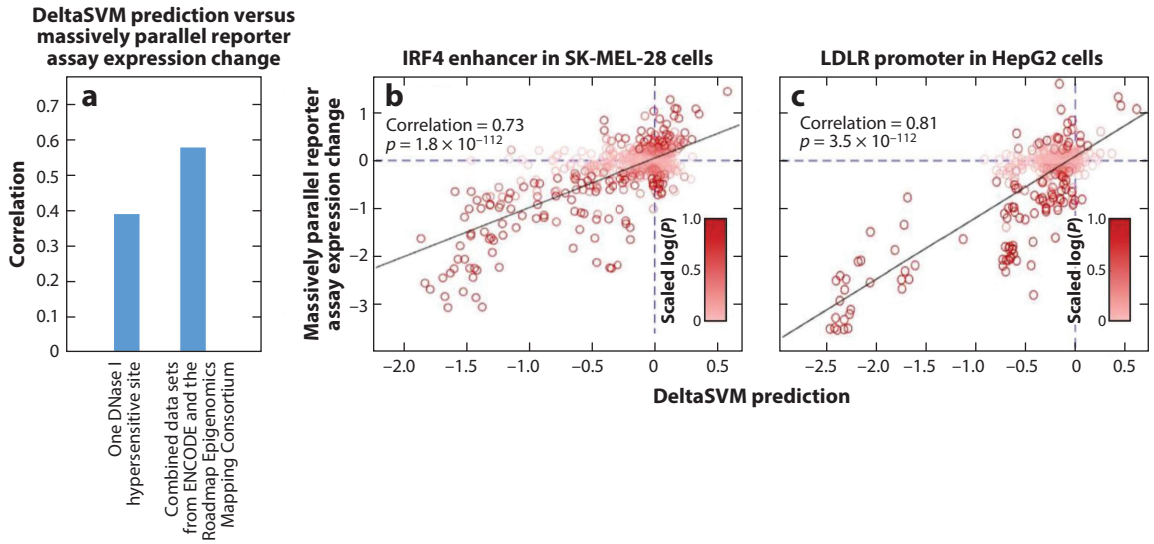


Figure 6

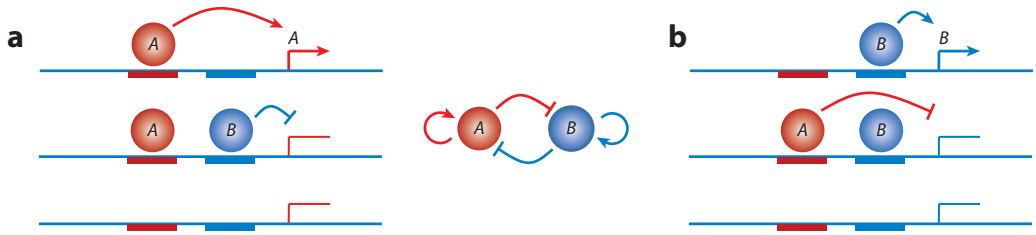
Comparisons of deltaSVM predictions and massively parallel reporter assay expression changes. (*a*) Overall correlation across 15 tested elements improves from 0.39 to 0.58 following training on multiple ENCODE and Roadmap Epigenomics Mapping Consortium data sets (67). (*b*) Correlation for the IRF4 enhancer improves to 0.73. (*c*) Correlation for the LDLR promoter improves to 0.81. Abbreviations: ENCODE, Encyclopedia of DNA Elements; SVM, support vector machine.

regression, as described in Reference 67. This method also improves correlations at individual enhancer and promoter loci, as shown in **Figure 6*b*,*c***.

REGULATORY NETWORK MODELS OF CELL FATE TRANSITIONS

Li et al. (46) introduced a simple continuum model of a gene regulatory network for a bistable genetic switch that described some features of the ESC–DE transition using reaction rate equation models similar to those previously used to model cell state transitions (23, 33, 57). It is reasonable to question the form of these reaction rate equation models on theoretical grounds, since the continuum limit upon which they are based may not be completely valid for some TFs expressed at low levels. A more technically rigorous and computationally much more challenging approach would utilize stochastic Langevin (28) or chemical master equations (54). Nevertheless, properly modeled reaction rate equations yield much clearer interpretations that lead to more facile biological insights. Also, the agreement between the experimental perturbations and the initial modeling in the study by Li et al. (46), in addition to the precision and robustness of embryonic developmental cell state transitions and the stability of cell states critical to multicellular life, suggests that a more theoretically rigorous model would lead to qualitatively similar conclusions.

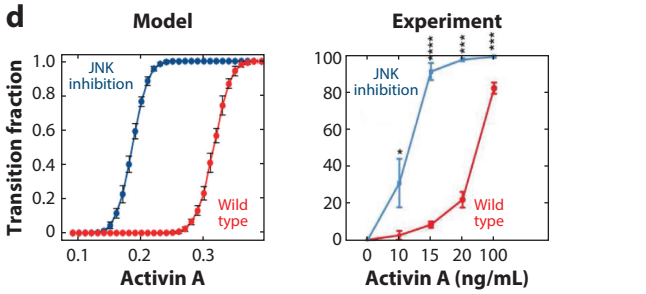
In the Li et al. (46) model, shown in **Figure 7*a*,*b***, TF genes *A* and *B* activate their own transcription by binding to nearby enhancers that negatively regulate the other TF. We will use lowercase letters (*a* and *b*) to describe the genes and uppercase letters (*A* and *B*) for the protein products. Gene *a* is transcribed (**Figure 7*a***) only when TF *A* is bound but TF *B* is not: When *B* is bound at the gene *a* locus, gene *a* is in a nonproductive transcriptional state, and thus the transcription rate of gene *a* is given by the bound concentration of *A* (in the absence of *B*) at its own gene, [*aA*]. We assume that the equilibrium occupancy of regulators *A* and *B* at genes *a* and *b* is established rapidly relative to rates of protein production, and therefore the equilibrium of *A* and *B* at their

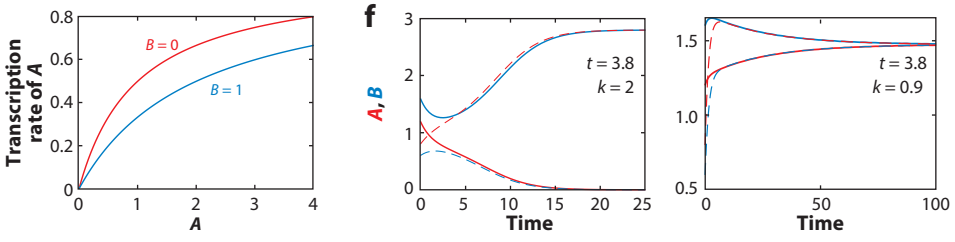


c

$$\frac{dA}{dt} = -rA + \frac{t_a A}{k_{aA} + A + \frac{k_{aA}}{k_{aB}} B}$$

$$\frac{dB}{dt} = -rB + \frac{t_b B}{k_{bB} + B + \frac{k_{bB}}{k_{bA}} A}$$





g Normalize time, concentrations, and symmetric case: $k_{aA}/k_{aB} = k$, $x,y \geq 0$:

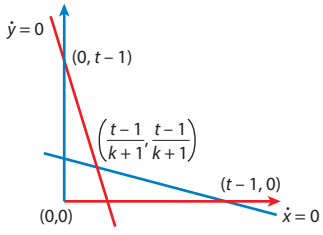
$$\frac{dx}{d\tau} = -x + \frac{tx}{1 + x + ky} \qquad \frac{dy}{d\tau} = -y + \frac{ty}{1 + y + kx}$$

h Equilibrium analysis:

$\dot{x} = 0$ Nullclines: $x = 0$ or $1 + x + ky = t$

$\dot{y} = 0$ Nullclines: $y = 0$ or $1 + y + kx = t$

Four fixed points at $(0,0)$, $(0, t-1)$, $(t-1, 0)$, $\left(\frac{t-1}{k-1}, \frac{t-1}{k-1}\right)$



i Stability analysis:

$$J = \begin{pmatrix} \left.\frac{\partial f}{\partial x}\right|_{x^*,y^*} & \left.\frac{\partial f}{\partial y}\right|_{x^*,y^*} \\ \left.\frac{\partial g}{\partial x}\right|_{x^*,y^*} & \left.\frac{\partial g}{\partial y}\right|_{x^*,y^*} \end{pmatrix} = \begin{pmatrix} -1 + \frac{t(1+ky)}{(1+x+ky)^2} & \frac{-tkx}{(1+x+ky)^2} \\ \frac{-tky}{(1+y+kx)^2} & -1 + \frac{t(1+kx)}{(1+y+kx)^2} \end{pmatrix}$$

System bistable if $k > 1$ or $k_{aA}/k_{aB} > 1$

Eigenvalues at $(x,y) = \left(\frac{t-1}{k+1}, \frac{t-1}{k+1}\right)$, $\lambda_1, \lambda_2 = \left(-\frac{t-1}{t}, \frac{(k-1)(t-1)}{t(k+1)}\right)$ Separatrix saddle point if $k > 1$

Eigenvalues at $(x,y) = (t-1, 0)$, $\lambda_1, \lambda_2 = \left(-\frac{t-1}{t}, -\frac{(t-1)(k-1)}{k(t-1)+1}\right)$ Stable bistable node if $k > 1$

(*Caption appears on following page*)

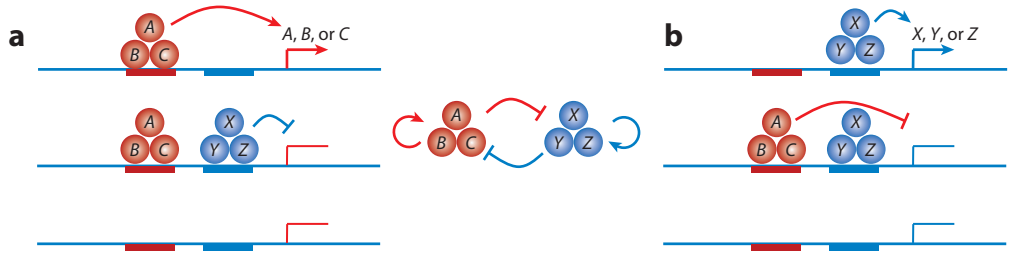
Figure 7 (*Figure appears on preceding page*)

Analysis of a simple noncooperative model of cell state bifurcation transitions driven by autoregulation and negative feedback. Lowercase letters are used for genes, and uppercase letters are used for protein products. (*a*,*b*) Bistable genetic circuit where TFs *A* and *B* autoactivate their own transcription by binding enhancers (*red* and *blue rectangles*) that drive their expression but interfere with or repress the transcription of the other TF. (*c*) Rate equations describing the evolution of the concentrations of TFs *A* and *B* under this model. (*d*) Stochastic simulations of this simple circuit, showing how transitions from the high *A* to high *B* state can be induced by external simulation and qualitatively agree with experimentally observed transition rates (46). (*e*) Concentration dependence of the transcription rate of TF *A* according to this model. (*f*) Bistable solutions and cell state transitions, which exist for some parameter choices ($t = 3.8$, $k = 2$) but not others ($t = 3.8$, $k = 0.9$). (*g*) Normalized system of equations for stability analysis. (*h*) Fixed points of this system. (*i*) Stability analysis showing that the system is bistable only for $k > 1$, which may require unrealistically strong negative feedback. Abbreviation: TF, transcription factor.

binding sites is given by Michaelis–Menten kinetics: $[aA] = [a][A] / k_{aA} = [a_0][A] / (k_{aA} + [A] + (k_{aA} / k_{aB})[B])$, where k_{aA} is the dissociation constant for TF *A* at its gene *a* binding site, k_{aB} is the dissociation constant for TF *B* at its gene *a* binding site, and $[a_0]$ is the total DNA concentration, which we will absorb into the transcription rate t_a. Similarly, $[bB]$ is equal to $[b_0][B] / (k_{bB} + [B] + (k_{bB} / k_{bA})[A])$. Both $[A]$ and $[B]$ are degraded at rate *r*, and the transcription of each is proportional to $t_a[aA]$ and $t_b[bB]$, yielding the model in **Figure 7*c***.

Stochastic simulations of induced cell state transitions can be modeled by adding a time-dependent impulse of activin that increases the transcription of one TF (here, *B* is a DE-specific TF), and weakening auto-activation models the effect of JNK inhibition and allows transition to DE at lower activin concentrations, in agreement with experiment (46) (**Figure 7*d***). The transcription rate for *A* is reduced by increased *B*, as shown in **Figure 7*e***. For parameters $t = 3.8$, $k = k_{aA} / k_{aB} = 2$, this system equilibrates at either a high *A*–low *B* or high *B*–low *A* state, depending on the initial conditions, as shown in **Figure 7*f***. The full stability of this model can be worked out simply when the parameters for *A* and *B* are symmetric, as in **Figure 7*g***. In this case, there are four fixed points, as shown in **Figure 7*h***, and both high *x*–low *y* and low *x*–high *y* states are stable if $k = k_{aA} / k_{aB} > 1$, in the usual case where $t > 1$. The fixed point at $x = y$ is an unstable saddle for $k > 1$, so for $k > 1$ this system exhibits bistability and can transition from one state to another with a significant perturbation, as shown in **Figure 7*d*** and Reference 46. However, this stability analysis shows that this system is somewhat sensitive to parameter choices. Since k_{aA} and k_{aB} are the dissociation constants for *A* and *B* at gene *a*, $k > 1$ requires that the repressive TF *B* binds at gene *a* with stronger affinity than the activating TF *A* and vice versa at the gene *b* locus. While understandable in the context of this mathematical model, this seems to be a rather difficult and unnatural requirement to satisfy for all mammalian cellular circuits.

A more realistic model is shown in **Figure 8**, which now incorporates the observation from **Figure 2** that there are multiple core TFs active in each cell state that bind cooperatively at activating and repressive regulatory DNA elements. Here, TFs *A*, *B*, and *C* bind cooperatively at their cognate enhancers with binding sites for TFs *A*, *B*, and *C* at genes *a*, *b*, and *c* and genes *x*, *y*, and *z*. However, at genes *a*, *b*, and *c*, the complex binding produces a transcriptionally productive DNA looping conformation, while at genes *x*, *y*, and *z* it produces a transcriptionally inactive conformation. Similarly, TFs *X*, *Y*, and *Z* bind cooperatively; activate genes *x*, *y*, and *z*; and produce a transcriptionally inactive DNA conformation when they bind at genes *a*, *b*, or *c*. Now the relevant kinetic parameters are the dissociation constants for the *ABC* and *XYZ* complexes at the relevant gene enhancers (e.g., k_{aABC} and k_{aXYZ}). The model equations for this situation are shown in **Figure 8*c***, and when simulated, they can produce a transition from the high *X*, *Y*, and *Z* state to the high *A*, *B*, and *C* state, as shown in **Figure 8*d***, with a sufficiently large perturbation. This system can also be studied with phase-plane analysis techniques if we assume $x = A = B = C$ and $y = X = Y = Z$, essentially modeling the complex concentrations, as in **Figure 8*f***, and the system

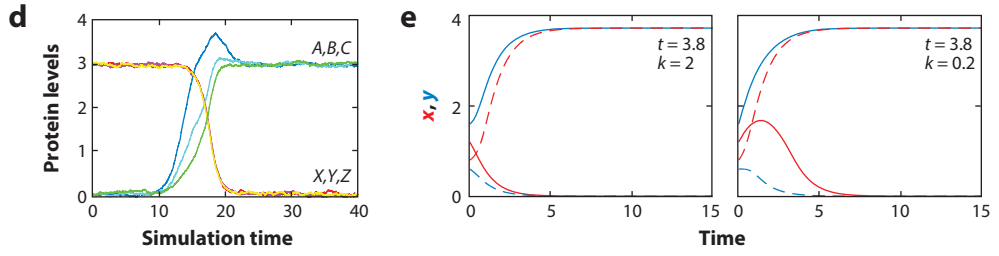


c

$$\frac{dA}{dt} = -rA + \frac{t_a ABC}{k_{aABC} + ABC + \frac{k_{aABC}}{k_{aXYZ}} XYZ} \qquad \frac{dX}{dt} = -rX + \frac{t_x XYZ}{k_{xXYZ} + XYZ + \frac{k_{xXYZ}}{k_{xABC}} ABC}$$

$$\frac{dB}{dt} = -rB + \frac{t_b ABC}{k_{bABC} + ABC + \frac{k_{bABC}}{k_{bXYZ}} XYZ} \qquad \frac{dY}{dt} = -rY + \frac{t_y XYZ}{k_{yXYZ} + XYZ + \frac{k_{yXYZ}}{k_{yABC}} ABC}$$

$$\frac{dC}{dt} = -rC + \frac{t_c ABC}{k_{cABC} + ABC + \frac{k_{cABC}}{k_{cXYZ}} XYZ} \qquad \frac{dZ}{dt} = -rZ + \frac{t_z XYZ}{k_{zXYZ} + XYZ + \frac{k_{zXYZ}}{k_{zABC}} ABC}$$



f Normalize time, concentrations, and symmetric case: $x = A = B = C$, $y = X = Y = Z$, $k_{aABC}/k_{aXYZ} = k$:

$$\frac{dx}{d\tau} = -x + \frac{tx^3}{1 + x^3 + ky^3} \qquad \frac{dy}{d\tau} = -y + \frac{ty^3}{1 + y^3 + kx^3} \qquad x, y \geq 0$$

g Stability analysis: at $x = 0$ and root of $1 + y^3 - ty^2 = 0$

Fixed points at $(x,y) \approx (0, t - 1/t^2)$ and $(x,y) \approx (t - 1/t^2, 0)$

Jacobian independent of k and stable bistable nodes for all k and $t > 1.9$

$$J(x = 0, y) = \begin{pmatrix} -1 & 0 \\ 0 & \frac{3ty^2}{1 + y^3} - \frac{5ty^5}{(1 + y^3)^2} - 1 \end{pmatrix}$$

Figure 8

Analysis of cooperative model of cell state bifurcation transitions. Lowercase letters are used for genes, and uppercase letters are used for protein products. (*a*,*b*) Bistable genetic circuit where TFs *A*, *B*, and *C* and TFs *X*, *Y*, and *Z* cooperatively autoactivate their own transcription by binding enhancers (*red* and *blue rectangles*) that drive their expression but interfere with or repress the transcription of the other three TFs. (*c*) Rate equations describing the evolution of the concentrations of TFs *A*, *B*, and *C* and TFs *X*, *Y*, and *Z* under this model. (*d*) Stochastic simulations of this simple circuit, showing how transitions from the high *ABC* to high *XYZ* state can be induced by external stimulation of *A*. (*e*) Bistable solutions and cell state transitions for the cooperative model, which exist for a much broader range of parameter choices; now both ($t = 3.8$, $k = 2$) and ($t = 3.8$, $k = 0.2$) support bistable behavior. (*f*) Normalized system of equations for stability analysis. (*g*) Stability analysis showing that the cooperative system is bistable for all choices of k as long as transcription is not weak ($t > 1.9$). Abbreviation: TF, transcription factor.

is now bistable for all k, as shown in **Figure 8*e***. Stability analysis shows that when $t > 1.9$, there are two fixed points at $(0, t - \frac{1}{t^2} + \mathcal{O}(\frac{1}{t^3}))$ and $(x, y) = (t - \frac{1}{t^2} + \mathcal{O}(\frac{1}{t^3}), 0)$, and since the Jacobian here is independent of k, these are stable nodes for all k, as shown in **Figure 8*g***. Thus, this cooperative model produces multiple stable states over a much wider range of kinetic parameter choices. It is possible that the ubiquity of cooperative binding by multiple TFs at mammalian developmental enhancers may have evolved because the state-switching dynamics of these cooperative circuits are more robust to binding site strength parameters than less cooperative gene regulatory networks involving single TFs.

SUMMARY AND FUTURE ISSUES

DNA sequence–based enhancer prediction methods and perturbative and functional studies are complementary methods that can be used to investigate the genetic regulatory networks controlling cell states and transitions between them. We have shown that these computational and experimental studies have detected features that are broadly consistent with each other. Predictive sequence features typically map to a relatively small set of core TF regulators, and cell-specific enhancers contain binding sites for multiple core TF regulators. When machine learning–based models are trained on these sets of enhancers as discriminative classifiers, they can predict the impact of mutations in reporter assays with reasonable quantitative accuracy. When these sequence-based models are used to design continuum dynamical models of genetic networks, these models can describe transitions between cellular states. Further analysis of these network models can yield insights into the features of genetic networks and the types of nonlinear interactions that support the stability and transitions between differentiated cellular states.

Yet many aspects of the modeling can be dramatically improved. Currently, linear and nonlinear classifiers often yield comparable overall accuracy, although there is ample evidence that nonlinear interactions between TFs contribute to enhancer activity and promoter interactions. Learning nonlinear interactions requires more training data, and it is likely that improved models in reduced feature spaces will be needed to allow models to learn the parameters of the nonlinear interactions between TFs and regulatory elements with existing amounts of training data. Methods of training on integrated large data sets can provide targeted training sequence data to detect more subtle regulatory events and should yield improved models. Higher-resolution three-dimensional chromatin looping measurements will likely be required to more fully understand and model the regulatory element interactions mediating promoter activation and repression.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants R01HG007348 and U01HG009380 to M.A.B. and R01DK096239 to D.H. We thank members of the Beer and Huangfu labs for useful discussions.

LITERATURE CITED

1. Agius P, Arvey A, Chang W, Noble WS, Leslie C. 2010. High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. *PLOS Comput. Biol.* 6:e1000916
2. Alexander J, Stainier DYR. 1999. A molecular pathway leading to endoderm formation in zebrafish. *Curr. Biol.* 9:1147–57

3. Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* 33:831–38
4. Allis CD, Jenuwein T. 2016. The molecular hallmarks of epigenetic control. *Nat. Rev. Genet.* 17:487–500
5. Arvey A, Agius P, Noble WS, Leslie C. 2012. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res.* 22:1723–34
6. Beer MA. 2017. Predicting enhancer activity and variant impact using gkm-SVM. *Hum. Mutat.* 38:1251–58
7. Beer MA, Tavazoie S. 2004. Predicting gene expression from sequence. *Cell* 117:185–98
8. Bernstein BE, Meissner A, Lander ES. 2007. The mammalian epigenome. *Cell* 128:669–81
9. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, et al. 2008. High-resolution mapping and characterization of open chromatin across the genome. *Cell* 132:311–22
10. Boyle EA, Li YI, Pritchard JK. 2017. An expanded view of complex traits: from polygenic to omnigenic. *Cell* 169:1177–86
11. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* 10:1213–18
12. Cao F, Fullwood MJ. 2019. Inflated performance measures in enhancer-promoter interaction-prediction methods. *Nat. Genet.* 51:1196–98
13. Chakravarti A, Turner TN. 2016. Revealing rate-limiting steps in complex disease biology: the crucial importance of studying rare, extreme-phenotype families. *BioEssays* 38:578–86
14. Conlon FL, Barth KS, Robertson EJ. 1991. A novel retrovirally induced embryonic lethal mutation in the mouse: assessment of the developmental fate of embryonic stem cells homozygous for the 413.d proviral integration. *Development* 111:969–81
15. Crawford GE, Holt IE, Mullikin JC, Tai D, Natl. Inst. Health Intramur. Seq. Cent., et al. 2004. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *PNAS* 101:992–97
16. Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *PNAS* 107:21931–36
17. D'Amour KA, Agulnick AD, Eliazer S, Kelly OG, Kroon E, Baetge EE. 2005. Efficient differentiation of human embryonic stem cells to definitive endoderm. *Nat. Biotechnol.* 23:1534–41
18. Davidson EH. 2010. *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*. Burlington, MA: Academic
19. Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, et al. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* 482:390–94
20. ENCODE Consort. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74
21. Feldman B, Gates MA, Egan ES, Dougan ST, Rennebeck G, et al. 1998. Zebrafish organizer development and germ-layer formation require nodal-related signals. *Nature* 395:181–85
22. Fletez-Brant C, Lee D, McCallion AS, Beer MA. 2013. kmer-SVM: a web server for identifying predictive regulatory sequence features in genomic data sets. *Nucl. Acids Res.* 41:W544–56
23. François P, Hakim V. 2004. Design of genetic networks with specified functions by evolution in silico. *PNAS* 101:580–85
24. Gate RE, Cheng CS, Aiden AP, Siba A, Tabaka M, et al. 2018. Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.* 50:1140–50
25. Ghandi M, Lee D, Mohammad-Noori M, Beer MA. 2014. Enhanced regulatory sequence prediction using gapped *k*-mer features. *PLOS Comput. Biol.* 10:e1003711
26. Ghandi M, Mohammad-Noori M, Beer MA. 2014. Robust *k*-mer frequency estimation using gapped *k*-mers. *J. Math. Biol.* 69:469–500
27. Ghandi M, Mohammad-Noori M, Ghareghani N, Lee D, Garraway L, Beer MA. 2016. gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics* 32:2205–7
28. Gillespie DT. 2000. The chemical Langevin equation. *J. Chem. Phys.* 113:297–306
29. Gorkin DU, Lee D, Reed X, Fletez-Brant C, Bessling SL, et al. 2012. Integration of ChIP-seq and machine learning reveals enhancers and a predictive regulatory sequence vocabulary in melanocytes. *Genome Res.* 22:2290–301

30. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, et al. 2014. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* 95:535–52
31. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* 39:311–18
32. Huang Q, Whitington T, Gao P, Lindberg JF, Yang Y, et al. 2014. A prostate cancer susceptibility allele at 6q22 increases *RFX6* expression by modulating HOXB13 chromatin binding. *Nat. Genet.* 46:126–35
33. Huang S, Guo Y-P, May G, Enver T. 2007. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Dev. Biol.* 305:695–713
34. Inoue F, Kircher M, Martin B, Cooper GM, Witten DM, et al. 2017. A systematic comparison reveals substantial differences in chromosomal versus episomal encoding of enhancer activity. *Genome Res.* 27:38–52
35. Kanai-Azuma M, Kanai Y, Gad JM, Tajima Y, Taya C, et al. 2002. Depletion of definitive gut endoderm in *Sox17*-null mutant mice. *Development* 129:2367–79
36. Kelley DR, Snoek J, Rinn JL. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* 26:990–99
37. Kreimer A, Zeng H, Edwards MD, Guo Y, Tian K, et al. 2017. Predicting gene expression in massively parallel reporter assays: a comparative study. *Hum. Mutat.* 38:1240–50
38. Kubo A, Shinozaki K, Shannon JM, Kouskoff V, Kennedy M, et al. 2004. Development of definitive endoderm from embryonic stem cells in culture. *Development* 131:1651–62
39. Kwasnieski JC, Fiore C, Chaudhari HG, Cohen BA. 2014. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* 24:1595–602
40. Lee D. 2016. LS-GKM: a new gkm-SVM for large-scale datasets. *Bioinformatics* 32:2196–98
41. Lee D, Beer MA. 2014. Mammalian enhancer prediction. In *Genome Analysis: Current Procedures and Applications*, ed. MS Poptsova, pp. 101–20. Norfolk, UK: Caister Acad.
42. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, et al. 2015. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* 47:955–61
43. Lee D, Kapoor A, Safi A, Song L, Halushka MK, et al. 2018. Human cardiac *cis*-regulatory elements, their cognate transcription factors, and regulatory DNA sequence variants. *Genome Res.* 28:1577–88
44. Lee D, Karchin R, Beer MA. 2011. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.* 21:2167–80
45. Lee K, Cho H, Rickert RW, Li QV, Pulecio J, et al. 2019. *FOXA2* is required for enhancer priming during pancreatic differentiation. *Cell Rep.* 28:382–93.e7
46. Li QV, Dixon G, Verma N, Rosen BP, Gordillo M, et al. 2019. Genome-scale screens identify JNK-JUN signaling as a barrier for pluripotency exit and endoderm differentiation. *Nat. Genet.* 51:999–1010
47. Liu X, Li YI, Pritchard JK. 2019. *Trans* effects on gene expression can drive omnigenic inheritance. *Cell* 177:1022–34.e6
48. Loh KM, Ang LT, Zhang J, Kumar V, Ang J, et al. 2014. Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations. *Cell Stem Cell* 14:237–52
49. Manolio TA. 2010. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med.* 363:166–76
50. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. 2009. Finding the missing heritability of complex diseases. *Nature* 461:747–53
51. Maston GA, Landt SG, Snyder M, Green MR. 2012. Characterization of enhancer function from genome-wide analyses. *Annu. Rev. Genom. Hum. Genet.* 13:29–57
52. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, et al. 2012. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337:1190–95
53. McClymont SA, Hook PW, Soto AI, Reed X, Law WD, et al. 2018. Parkinson-associated *SNCA* enhancer variants revealed by open chromatin in mouse dopamine neurons. *Am. J. Hum. Genet.* 103:874–92
54. McQuarrie DA. 1967. Stochastic approach to chemical kinetics. *J. Appl. Probab.* 4:413–78
55. Miguel-Escalada I, Bonàs-Guarch S, Cebola I, Ponsa-Cobas J, Mendieta-Esteban J, et al. 2019. Human pancreatic islet three-dimensional chromatin architecture provides insights into the genetics of type 2 diabetes. *Nat. Genet.* 51:1137–48

56. Mo A, Luo C, Davis FP, Mukamel EA, Henry GL, et al. 2016. Epigenomic landscapes of retinal rods and cones. *eLife* 5:e11613
57. Moris N, Pina C, Arias AM. 2016. Transition states and cell fate decisions in epigenetic landscapes. *Nat. Rev. Genet.* 17:693–703
58. Mullen AC, Orlando DA, Newman JJ, Lovén J, Kumar RM, et al. 2011. Master transcription factors determine cell-type-specific responses to TGF-β signaling. *Cell* 147:565–76
59. Patwardhan RP, Hiatt JB, Witten DM, Kim MJ, Smith RP, et al. 2012. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* 30:265–70
60. Rada-Iglesias A, Bajpai R, Swigut T, Brugmann SA, Flynn RA, Wysocka J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 470:279–83
61. Rivera CM, Ren B. 2013. Mapping human epigenomes. *Cell* 155:39–55
62. Roadmap Epigenom. Consort., Kundaje A, Meuleman W, Ernst J, Bilenky M, et al. 2015. Integrative analysis of 111 reference human epigenomes. *Nature* 518:317–30
63. Robertson EJ. 2014. Dose-dependent Nodal/Smad signals pattern the early mouse embryo. *Semin. Cell Dev. Biol.* 32:73–79
64. Sabo PJ, Hawrylycz M, Wallace JC, Humbert R, Yu M, et al. 2004. Discovery of functional noncoding elements by digital analysis of chromatin structure. *PNAS* 101:16837–42
65. Sabo PJ, Kuehn MS, Thurman R, Johnson BE, Johnson EM, et al. 2006. Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nat. Methods* 3:511–18
66. Setty M, Leslie CS. 2015. SeqGL identifies context-dependent binding signals in genome-wide regulatory element maps. *PLOS Comput. Biol.* 11:e1004271
67. Shigaki D, Adato O, Adhikar AN, Dong S, Hawkins-Hooker A, et al. 2019. Integration of multiple epigenomic marks improves prediction of variant impact in saturation mutagenesis reporter assay. *Hum. Mutat.* 40:1280–91
68. Takata R, Akamatsu S, Kubo M, Takahashi A, Hosono N, et al. 2010. Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. *Nat. Genet.* 42:751–54
69. Teo AKK, Arnold SJ, Trotter MWB, Brown S, Ang LT, et al. 2011. Pluripotency factors regulate definitive endoderm specification through eomesodermin. *Genes Dev.* 25:238–50
70. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, et al. 2012. The accessible chromatin landscape of the human genome. *Nature* 489:75–82
71. Thurner M, van de Bunt M, Torres JM, Mahajan A, Nylander V, et al. 2018. Integration of human pancreatic islet genomic data refines regulatory mechanisms at type 2 diabetes susceptibility loci. *eLife* 7:e31977
72. Tsankov AM, Gu H, Akopian V, Ziller MJ, Donaghey J, et al. 2015. Transcription factor binding dynamics during human ES cell differentiation. *Nature* 518:344–49
73. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, et al. 2012. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* 22:1798–812
74. Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucl. Acids Res.* 42:D1001–6
75. Whalen S, Truty RM, Pollard KS. 2016. Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat. Genet.* 48:488–96
76. Xi W, Beer MA. 2018. Local epigenomic state cannot discriminate interacting and non-interacting enhancer-promoter pairs with high accuracy. *PLOS Comput. Biol.* 14:e1006625
77. Yue F, Cheng Y, Breschi A, Vierstra J, Wu W, et al. 2014. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* 515:355–64
78. Zeng H, Edwards MD, Liu G, Gifford DK. 2016. Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics* 32:i121–27
79. Zhou J, Troyanskaya OG. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* 12:931–34
80. Zorn AM, Wells JM. 2009. Vertebrate endoderm development and organ formation. *Annu. Rev. Cell Dev. Biol.* 25:221–51

Annual Review of Genomics and Human Genetics

Volume 21, 2020

Contents