Briefings in Bioinformatics, 21(1), 2020, 120–134

doi: 10.1093/bib/bby110
Advance Access Publication Date: 16 November 2018
Review Article

A comprehensive review of computational prediction of genome-wide features

Tianlei Xu, Xiaoqi Zheng, Ben Li, Peng Jin, Zhaohui Qin, Hao Wu

Corresponding author: Hao Wu Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA. Tel.: +1-404-727-8633; Fax: +1-404-727-1370; E-mail: hao.wu@emory.edu

Abstract

There are significant correlations among different types of genetic, genomic and epigenomic features within the genome. These correlations make the *in silico* feature prediction possible through statistical or machine learning models. With the accumulation of a vast amount of high-throughput data, feature prediction has gained significant interest lately, and a plethora of papers have been published in the past few years. Here we provide a comprehensive review on these published works, categorized by the prediction targets, including protein binding site, enhancer, DNA methylation, chromatin structure and gene expression. We also provide discussions on some important points and possible future directions.

Key words: machine learning; genomic features; prediction model

Introduction

Genome is a complicated yet coordinated system. A variety of genome-wide features (including genomic, epigenomic and transcriptomic) works in harmony to achieve complex functions. Advances in high-throughput technologies allow the profiling of different features on the whole genome scale. During the past decade, tremendous amount of different types of high-throughput omics data have been generated to measure these features, for example, by large consortia such as NIH Roadmap Epigenomic Consortium [1, 2] and ENCODE [3]. Analyses of these data greatly enhance our understanding of how genome works and provide opportunities for identifying diagnostic biomarkers and therapeutic targets.

Generally speaking, features in the genome can be categorized into two classes: the static features and the dynamic features. The static features are the ones defined by DNA sequence (without considering genetic variants) and genome annotation, thus do not change across cell types within one individual. These features only need to be measured once, then they can be applied for different biological conditions. Examples of static features are the following:

i. DNA sequence composition (GC content, CpG density, k-mer, etc.)
ii. DNA motifs (protein specific)
iii. gene annotation (coordinates of genes, exons, introns, untranslated regions (UTRs) and other regulatory regions)

Tianlei Xu is a PhD student in the Department of Biostatistics and Bioinformatics at Emory University. His research interest is to develop methods and tools for high-throughput epigenetics data.
Xiaoqi Zheng is a professor in the Department of Mathematics at Shanghai Normal University, China. His research is focused on developing methods for the analysis of high-throughput epigenomics data.
Ben Li is a PhD student in the Department of Biostatistics and Bioinformatics at Emory University. His research interest is to develop machine learning tools on biomedical and healthcare data.
Peng Jin is a professor in the Department of Human Genetics at Emory University School of Medicine. His research interests are the roles of epigenetics and non-coding RNAs in neurodevelopment and brain disorders.
Zhaohui Qin is an associate professor in the Department of Biostatistics and Bioinformatics at Emory University. His research interests are focused on developing and evaluating model-based methods to analyze genetics and genomics data.
Hao Wu is an associate professor in the Department of Biostatistics and Bioinformatics at Emory University, Rollins School of Public Health. His main research interest is to develop statistical methods and computational tools for analyzing large-scale biomedical data, in particular, different types of high-throughput omics data.

Submitted: 10 July 2018; **Received (in revised form):** 15 October 2018

iv. CpG island
v. sequence conservation.

Dynamic features are the ones that vary across individuals, cell types or under different biological conditions. They are closely related to the diverse functions of the cells, for example, different cell types from the same individual. Examples of dynamic features are the following:

i. DNA sequence variants such as the single nucleotide polymorphisms (SNPs)
ii. gene expression
iii. DNA methylation
iv. histone modification
v. protein–DNA interaction
vi. replication timing
vii. chromatin structure (spatial organization, open/close chromatin and nucleosome positioning).

Among these, the genetic features (ones defined by the DNA sequence and annotation) are mostly static. The only exception is the genetic variants, which are variable across individuals. Most of the genomic, epigenomic and transcriptomic features are dynamic. They are different not only across individuals but also across different cell types within the same individual. When studying a specific biological or clinical condition, it is ideal that all these features are measured. However, it's prohibitively expensive and laborious in practice. Analyses of existing high-throughput data reveal strong correlations among different features, for example, the binding of certain transcription factors (TFs) at the gene promoter regions are correlated with the gene expression. Even though the correlations do not necessarily imply any causal relationship, these intrinsic correlations, along with the availability of large amount of public high-throughput data, make it possible to predict the features. The prediction can fill gaps of experimental data and provide insight of how different genome features work as a coordinated system. Over the past several years, there have been great interests in computational genome-wide feature prediction, and a number of papers have been published. Here, we provide a rather comprehensive survey of existing works in this field, where the methods are grouped based on the targets of the prediction, including protein binding site, enhancer, DNA methylation, chromatin structure and gene expression. Within each group, the detailed nature and characteristics of the target, together with the rationale behind the selection of predictors and models, will be discussed in detail.

Prediction of protein binding sites

Predicting the binding sites of DNA-binding proteins such as TFs is useful due to the fact that experimental methods (such as chromatin immunoprecipitation sequencing (ChIP-seq)) can only determine the true binding sites of one type of protein, under one condition (tissue, cell, treatment/disease, etc.) at a time. It is impossible to profile the combinations of all TFs and cell conditions experimentally. Thus, computational prediction has become popular, where one can use existing data to learn the rules of TF binding and then impute the binding profile under a new biological condition without actually doing the experiment.

In general, there are a number of genomic features associated with protein binding, including DNA sequence motifs, chromatin accessibility, histone modification and DNA methylation status. Before the availability of high-throughput data, DNA sequence motif is the most widely used feature to predict transcription factor binding sites (TFBSs) *in silico*. The binding motifs can be found in databases such as JASPAR [4], TRANSFAC [5], ORegAnno [6], PAZAR [7] and Factorbook [8]. However, motif-based methods usually perform poorly for multicellular eukaryotic organisms such as human and mouse, because the static genome sequence features alone are unable to predict the cell type-specific binding events. Moreover, the sequence motif could vary in different cell types even for the same protein [9]. A recent study also shows that on individual level, the repository of TF binding activities may be affected by one's genetic variation [10]. In this section, we will not discuss existing prediction methods based on DNA sequence alone. Instead, we will focus on how sequence motifs can be combined with other genome features to predict *in vivo* binding site.

Prediction methods using histone modifications

Histone modifications are found to be closely correlated with regulatory activities in the human genome [11]. Different histone marks are associated with different regulatory elements, such as open chromatin, promoter, enhancers, etc. [11, 12]. Whitington *et al.* [13] first used DNA motifs to scan the genome and define potential binding sites. Then they used histone mark H3K4me3 as well as other features such as distance to known transcription start site (TSS) and sequence conservation to filter the candidates to reduce false positives. He *et al.* [14] used differential H3K4me2 ChIP-Seq signals to measure nucleosome positioning, followed by motif analysis to predict TF binding dynamics. Talebzadeh and Zare-Mirakabad used a composition of different histone marks within neighboring nucleosomes as predicting features. Then it is combined with position weighted matrix (PWM) to fit a logistic regression classifier to predict binding sites [15]. Ramsey *et al.* [16] used the local minima of histone acetylation ChIP-seq signals (referred to as 'valley score') combined with motif scanning score to predict TF binding sites. They defined a weighted sum of scores from different features as binding score where weighted parameters are trained by supervised learning. Won *et al.* [17] used a more comprehensive set of histone features to train a hidden Markov model (HMM) model to predict TFBS. Their method, named Chromia, fits a three-state Gaussian mixture model by taking both position-specific scoring matrix and binned histone modification signals as input. An interesting finding from Chromia is the heterogeneity of histone contribution; histone marks are predictive for TFBS in a TF-specific manner. They found that, among all TFs listed in the paper, H3K4me3 is the most frequent strong predictor.

It is interesting to notice that among a variety of histone marks, only a few are proved to be relevant with TF binding regulation. These factors include open chromatin (Histone acetylation, HAc [16]), nucleosome distribution (H3K4me2 [14]), transcription (H3K36me3 [17]), promoter (H3K4me3) or enhancers (H3K4me1). Based on this idea, Ji *et al.* [18] used histone marks to define genomic categories and made TFBS prediction based on this information. The histone marks serve as a feature space in which the complete epigenome environment can be projected on to.

Prediction methods using chromatin accessibility

Since the open chromatin structure is often a necessary condition for a protein–DNA interaction, using chromatin accessibility data (for example from DNase-seq or Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq)) to predict protein binding is a natural idea. There are in general two major categories of methods using chromatin data: bin-based methods and candidate site-based methods. Bin-based methods include DNase2TF [19] and HINT [20]. This class starts with searching for footprints of TFBS shown on DNase-seq data, i.e. short valleys in DNase highly sensitive peak regions, under the assumption that binding sites of proteins are protected from enzymatic cleavage. The footprints are usually enclosed within signatures of histone modifications. HMM model was adapted to recognize this type of local-dependency relationship in HINT and Chromia. The latter class first scans the whole genome using known TF motifs and gets a list of candidate binding sites. Then the chromatin profile centered at these candidate sites are analyzed to detect potential TF binding sites. Methods in this class could use supervised or unsupervised learning algorithms. Unsupervised methods include CENTIPEDE [21], FootprintMixture [22], PIQ [23], Romulus [24] and Mocap [25]. These methods model the differences in data between binding and non-binding sites then phase out the labels (binding/no-binding) using prior knowledge. This type of method can learn the class structure directly from data, thus is useful when there is no training data available for supervised learning. While supervised methods use a variety of representations of DNase profile, some combine with additional epigenetic marks as prior [26] and train different types of models, including support vector machine (SVM) [9, 27] or random forest [28, 29].

One important problem associated with chromatin-based predicting methods is the sequence bias issue of the DNase footprinting experiments. The cutting efficiency of DNase I shows difference for certain DNA motifs, which might lead to false detection of enriched motif resulted from technique noise, rather than sequence signal of TF binding [30]. In addition, the residence binding time of different TFs shows great variance, which results in different reads coverage around the binding sites. Gusmao *et al.* [31] discussed these issues and computational solutions to address these challenges in detail in a recent review.

Other methods for protein binding prediction

Prediction methods using methylation profile [32] or DNA shape features [33] have been developed in addition to the popular methods using epigenetic marks discussed above. As an example, Xu *et al.* [32] showed that 5 mC DNA methylation profile including CG and CH methylation can accurately predict TF–DNA binding *in vivo*, in many cases, even better than using DNase data. In addition, recent popularity of deep learning technology brought applications in TFBS prediction as well. For example, DeepBind [34] used convolutional neural network (CNN) [35] to detect motif-like DNA sequence kernels and using them as input features in a feed-forward neural network to predict binding affinity of proteins. FactorNet [36] added an additional layer of recurrent neural network (RNN) [37] to model the spatial dependency of feature signals.

A list of protein binding site prediction methods reviewed above is listed in Table 1.

Prediction of enhancer

Definition of enhancer

Prediction of enhancer is difficult, partly because of the lack of a gold standard definition for an enhancer–gene pair. So far, the predicting methods mostly rely on experimentally validated enhancers from public data resources, such as FANTOM [38] or ENCODE [3], where the spatial chromatin organization that brings a regulatory enhancer to a distal gene promoter is considered as the validated evidence of an enhancer. However, this approach is not applicable to all the cells/tissues due to the cost. Alternative methods are proposed to use epigenetic features presented at enhancer regions as indirect gold standards. This approach, however, needs to be carefully designed since some epigenetic marks are used to define the outcomes, while some other epigenetic marks are used as predictors.

Challenges of enhancer prediction

With limited annotation resource, the task of predicting enhancers is challenging. Unlike well-annotated gene/TSS/promoter, this task only became possible with the accumulation of a large amount of TFBS/DNase I hypersensitive sites (DHSs)/histone data. Some histone modifications are considered associated with enhancer, acting as the markers to guide different TFs to form regulatory chromatin loops. These modifications include H3.3 and H2AZ [39], H3K4me1 and H4K27ac [11, 40, 41]. The latter is specifically studied as the marker for poised enhancer [42, 43]. Nucleosome positioning is also a key factor to define enhancers [14], which could potentially be the result of chromatin opening. Also, TFs are essential components of enhancer activity as well. Multiple studies showed the key role of TFs as the marker of enhancers, such as p300/CREB-binding protein (CBP) [44–47] or a combination of multiple TFs [48–51]. In prediction methods, common practice is to overlap predicted 'enhancer' with P300 + DHS + several TFBS known to be associated with enhancer.

It is equally important to choose proper negative set in the training data. For example, TSSs share a considerable number of common features with enhancers, thus the negative sites must include TSS regions should the user wish to differentiate them from distal regulatory regions. Tissue specificity is another issue that needs to be considered. Just like gene expression and other epigenetic profiles, enhancer is highly cell type-specific [52, 53]. For example, enhancers regulating developmental genes (e.g. SOX2/OCT4/NANOG) should not function in developed tissues in principle. Another challenge of enhancer prediction is the diversity of the enhancer itself. Studies have shown that the enhancers are highly heterogeneous [54, 55]. Thus, it is essential for the selection of training data to collect a comprehensive set of representing training samples, both for positive classes and negative classes.

Tools for enhancer prediction

Before epigenetic marks were used for predicting enhancer, DNA sequence was investigated as the potential predictor for enhancers [11], and this information has been used as an enhancer predictor for a long time. k-mer-based methods are often coupled with SVM classifier, and more sophisticated representation of k-mers allowing gaps was introduced as well [56, 57]. Other variants based on motifs [58] and DNA local structure [59] have also been proposed. As more epigenomic data become available, tools like REFCS [60] that based only on DNA sequence become less common.

Among all the epigenomic marks, histone modifications are the most relevant to enhancer. Firpi *et al.* [61] developed a tool in the early stage of enhancer prediction. They used neural network to model the histone modifications as features at enhancer regions. Their approach is considered as a standard, and many follow-up works adopt the same setting, including the size of negative sets, validation procedure, etc. Non-linear

Table 1. Prediction methods of protein binding sites

Publication	Features	Method	Software	Description
Whitington *et al.* [13]	Histone	PWM scan followed by filtering chromatin features	N/A	Scan the genome using motif PWM to detect candidate binding sites. Sites are then filtered by several genetic and epigenetic criteria using *ad hoc* thresholds.
He *et al.* [14]	Nucleosome-resolution histone	Dynamics of nucleosome occupancy and motif	N/A	Use differential H3K4me2 ChIP-seq signals to measure nucleosome positioning, followed by motif analysis to predict TF binding dynamics.
Won *et al.* [17]	Histone	HMM	Chromia (http://wanglab.ucsd.edu/star/job.php?s=chromia)	Use sequence counts of fixed-sized bins for a number of histone ChIP-seq data sets as inputs for a three-state HMM.
Ramsey *et al.* [16]	Histone acetylation ChIP-seq, nucleosome occupancy, DNA sequence	Weighted sum of scores	RamseyHAc2010 (http://magnet.systemsbiology.net/hac/)	Take 100 bp intervals of transcript-proximal regions as candidate regions. Sequence and epigenetic features are used as predictors. Weighted sum of thresholded predictors values are treated as prediction scores.
Pique-Regi *et al.* [21]	DNase-seq, DNA sequence	Two-component mixture model, expectation-maximization (EM) algorithm	CENTIPEDE (http://centipede.uchicago.edu/)	Compute prior for binding from sequence-based features. Binned read counts from DNase-seq are assumed to be from a mixture model, which is fitted by an EM algorithm. Posterior probabilities of binding are obtained.
Cuellar-Partida *et al.* [26]	DNase-seq, histone	Epigenetic data as prior, use motif to predict	FIMO, part of MEME (http://meme-suite.org/doc/fimo.html)	Prior probabilities of binding are derived from epigenetic data. Posterior probabilities are computed from prior and motif scores based on a Bayesian model.
Arvey *et al.* [9]	Histone, DNase-seq	SVM	N/A	Two SVMs are trained based on dinucleotide mismatch k-mer features of 100 bp regions and read counts from 100 bp bins.
Ji *et al.* [18]	Nine histones	Principal component analysis (PCA)-type unsupervised learning	dPCA (http://www.biostat.jhsph.edu/dpca/)	Binned read counts from several ChIP-seq data sets are obtained. Differences in counts between two conditions are decomposed using PCA. The motif sites with large PC1 scores are deemed differential binding sites.
Gusmao *et al.* [20]	DNase-seq, histone	HMM	HINT (http://www.regulatory-genomics.org/hint/introduction/)	Genome-wide read counts in 5000 bp bins are obtained and normalized. Eight-state HMM (with one of the states being the protein-binding footprint) with multivariate outcome is fitted.
Sung *et al.* [19]	DNase-seq, Motif (4-mer)	Tests based on counts	Dnase2TF (https://sourceforge.net/projects/dnase2tfr/)	Identify DHS from DNase-seq data. For each DHS, it looks for 'dip' within the peaks. The dips are then combined with motifs to define binding sites.
Yardimci *et al.* [22]	DNase-seq, bias adjustment	Two-component mixture model	FootprintMixture (https://ohlerlab.mdc-berlin.de/software/FootprintMixture_109/)	Two-component mixture model is used to predict binding sites. Binned read counts around candidate sites are modeled by factor-specific multinomial distribution.

Continued

Table 1. continued

Publication	Features	Method	Software	Description
Sherwood *et al.* [23]	DNase-seq (magnitude + shape around motif match sites)	Expectation propagation	PIQ (http://piq.csail.mit.edu/)	Scan the genome using motif PWM to detect candidate binding sites. Binned read counts from DNase-seq are analyzed to build TF-binding profiles (shapes and magnitude). Posterior probability of binding is computed using expectation propagation method.
Xu *et al.* [32]	Methylation, genomic features	Random forest + two-component mixture model for methylation	Methylphet (https://github.com/benliemory/Methylphet)	Compute methylation scores based on a mixture model, then use the scores with other features as predictor in a random forest to predict binding sites.
Alipanahi *et al.* [34]	DNA sequence (using binding array data as target)	CNN	DeepBind (http://tools.genes.toronto.edu/deepbind/)	Use a deep learning framework to model the relationships between DNA sequence patterns and TF binding sites. CNN is used to capture sequence patterns. Trained model can be applied to a new sequence with variations to estimate risks by predicting changes in TF binding affinity.
Quach and Furey [27]	DNase-seq (profile: mean and slope, centered at motif)	SVM	DeFCoM (https://bitbucket.org/bryancquach/defcom)	Find candidate regions based on motifs. Binned read counts from DNase-seq around the candidate regions are obtained for varying-sized bins. Read counts are used as predictors in an SVM.
Jankowski *et al.* [24]	DNase-seq (shape, on motif-matched sites)	Two-component mixture model, EM algorithm	Romulus (https://github.com/ajank/Romulus)	Similar to CENTIPEDE, except for the binning strategy. They use 20 bp bins outside the motif site and single-bp bins within the motif site. For non-binding sites, they put all the positions in a single bin.
Liu *et al.* [28]	DNase-seq (footprint score defined by counts) + genomic features	Random forest	BPAC (http://bioinfo.wilmer.jhu.edu/BPAC/)	Motif PWM scores and different types of read count features are used as predictors to train a random forest for prediction.
Ma *et al.* [33]	DNA sequence, shape kernel	Support vector regression	Sequence-shape (https://bitbucket.org/wenxiu/sequence-shape.git)	Use kernel functions to model both DNA sequence and shape features simultaneously. It then performs kernel-based regression and classification to predict TF–DNA interaction. It shows that incorporation of DNA shape information can improve prediction accuracy.
Kuang *et al.* [29]	Histone, DNase-seq, on known motif-matched sites	Random forest	DynaMO (https://github.com/spo111/DynaMO)	Use motif sites as candidate regions. Binned read counts around the motif sites are used as predictors in random forest for prediction.
Chen *et al.* [25]	DNase-seq, ATAC-seq and DNA sequence features	Three-component mixture model; sparse logistic regression; weighted least square	Mocap (https://github.com/xc406/Mocap)	Take motif sites as candidate regions. They used a three-component mixture model for the read counts and sparse logistic regression on a number of features. Cross-sample method uses weighted least square to minimize a loss function.
Quang *et al.* (unpublished)	DNase-seq	CNN + RNN	FactorNet (https://github.com/uci-cbcl/FactorNet)	Use a number of features to train a deep learning model (DanQ CNN-RNN hybrid architecture). Features include genome sequences and annotations, gene expression and DNase-seq data.

Table 2. Prediction methods of enhancers

Publication	Enhancer definition/control	Features used	Method	Software	Description
Heintzman *et al.* [11]	P300/random	DNA sequence	Correlation	N/A	Use correlation between sequence and P300 binding sites to predict enhancers. Some predicted enhancers lack p300 binding. Hold-out cross-validation is used to select the optimal combination of histone modifications.
Firpi *et al.* [61]	74 validated from Heitzman *et al.* [11]	Histone	ANN (Time delay neural network (TDNN))	CSIANN (http://tanlab4generegulation.org/CSIANNWebpage.html)	Use neural network model to predict enhancer from histone modification. It was one of the early works that defined the settings of enhancer prediction that later works adopted.
Lee *et al.* [56]	P300/random	DNA sequence (k-mer with k = 3∼10)	SVM	Kmer-svm (http://beerlab.org/kmersvm/)	Use SVM to model k-mer composition for enhancers. It is a non-cell-specific model. Performance varies in different cells. It explores the possibility to predict enhancer from DNA sequence.
Taher *et al.* [58]	Validated/random	TF motifs	LASSO regression	CLARE (http://clare.dcode.org/)	Use LASSO regression to model concurrence between TF motifs and enhancers. Consider the length and GC content of the target region.
Fernandez and Miranda-Saavedra [62]	P300 distal to TSS/random	Histone	SVM + genetic algorithm	ChromaGenSVM	Use SVM on histone data to predict enhancer. Use genetic algorithm for model selection.
Rajagopal *et al.* [63]	p300 overlapping with DHS, distal to TSS/validated	24 types of histone	Random forest	RFECS (https://github.com/kaizhang/RFECS)	Use an extended panel of histone data and evaluate the importance of different histone modifications. It can perform cross cell-type prediction.
Ghandi *et al.* [57]	p300 in mouse embryonic/random	Gapped-kmer	SVM	Gkm-svm (http://www.beerlab.org/gkmsvm/)	Use gapped-kmer as features for SVM model, which is an extension to the previous kmer-SVM method.
Erwin *et al.* [66]	VISTA enhancer/tissue-specific non-enhancer validated	Step 1: histone, TFBS, Dnase/FAIRE, conservation, motif; Step 2: histone, p300	Linear SVM as step 1; multiple kernel learning in step 2	EnhancerFinder (https://sourceforge.net/projects/enhancers/)	A two-step method for developmental enhancer. Step 1 is non-cell-specific; it detects developmental enhancer regions across the genome. Step 2 is trained in tissue-specific manner, thus unable to do cross-tissue prediction. It shows that functional genomics data are more informative about developmental enhancer tissue specificity than degree of conservation or sequence motifs.

Continued

Table 2. continued

Publication	Enhancer definition/control	Features used	Method	Software	Description
Kleftogiannis *et al.* [65]	ENCODE validated enhancer/random	Histone, DNA sequence	Ensemble, with SVM as base classifier and ANN as final output classifier	DEEP (http://cbrc.kaust.edu.sa/deep/)	Take a comprehensive set of features from histone and DNA sequence and then use an ensemble learning framework with SVM as base classifiers. It performs feature selection based on an exhaustive search that identifies the optimal set of attributes that differentiates considerably between different cell lines. It shows that no model trained from a single cell-line data can effectively predict enhancers in all other cell lines.
Liu *et al.* [59]	Validated enhancer/non-enhancer	Sequence (k-mer), DNA local structure	SVM	iEnhancer-2 L (http://bioinformatics.hitsz.edu.cn/iEnhancer-2L/)	A two-step method that takes k-mer and DNA structure as input. Step 1 predicts enhancers; Step 2 distinguishes weak/strong enhancers. It downsamples negative class to solve imbalance problem.
Lu *et al.* [64]	P300, DHS; distal to TSS/random	Histone (shape of profile in addition to intensity)	Adaboost	DELTA (https://github.com/genereader/delta)	Take shape features from histone marks (instead of using intensity only) as input. An adaboost model is used for prediction.
Liu *et al.* [67]	H3K27ac peaks; multiple filters (distal to TSS, etc.)	Histone, 27 TFs and cofactors, 15 chromatin accessibility, transcription, RRBS, CpG islands, evolutionary conservation, sequence k-mers, motifs (TFBS)	Deep learning (DNN) + HMM; Iteratively train through cell types	PEDLA (https://github.com/wenjiegroup/PEDLA)	Take a comprehensive set of features, including histone, TF binding sites, chromatin accessibility, transcription and methylation. A deep neural network model is used for prediction. Model is trained across all available cell types iteratively then predict in one cell type. This is an early deep learning method for enhancer prediction.
Jia *et al.* [60]	Strong or weak enhancer/non-enhancer	400 bp: bi-profile Bayes (similar to 200 bp positive PWM and 200 bp negative PWM, Nucleotide frequency, pseudo-nucleotide frequency, 3-mer frequency)	SVM	EnhancerPred (http://server.malab.cn/EnhancerPRED/)	Build a set of sequence features based on nucleotide frequency and PWMs, then use an SVM model to predict enhancers. All features are derived from DNA sequence.
He *et al.* [68]	P300/random + promoters	Histone, methylation	Random forest	REPTILE (https://github.com/yupenghe/REPTILE)	Use histone and methylation data as features in a random forest model. It shows that enhancer overlaps with DMRs. Two random forest models are trained, one is based on pigenomic signature of the complete query region, and the second model is based on the epigenomic signature of DMRs within the query region. When predicting, the maximum score between the outputs form the two models are used as the final prediction.

Table 3. Prediction methods of methylation

Publication	Methylation type	Features used	Method	Software	Description
Bhasin *et al.* [76]	CpG sites from MethDB	DNA sequence	SVM	Methylator	MethDB sequences are fragmented into overlapping fragments of fixed length and used as predictors in an SVM.
Das *et al.* [78]	CpG sites, differentiate in/out CpG Island (CGI)	DNA sequence	SVM	HDFinder	∼100 sequence-based features are used as predictors. Feature selection was done by recursive feature elimination. A number of classifiers were applied, and SVM provides the best performance.
Fang *et al.* [77]	CGI	GC content, DNA motifs	SVM	MethCGI	Sequence-based features and TFBSs are used as predictors in an SVM model.
Whitaker *et al.* [79]	DNA methylation valleys	DNA sequence (motif)	LASSO feature selection + random forest prediction	Epigram (http://wanglab.ucsd.edu/star/epigram/)	Use a number of DNA sequence motifs as predictors. LASSO was used for feature selection, and random forest is the classifier.
Zhang *et al.* [86]	WGBS, 450 K	DNA sequence (composition, recombination rate, evolution rate), 450 K data, DHS site, TFBSs	Random forest	N/A	124 features are extracted from sequence, genome annotation and epigenetic modifications. Random forest is used for prediction.
Wang *et al.* [87]	RRBS	DNA sequence (composition, k-mer), both local and remote based on Hi-C data	DNN (stacked denoising autoencoders)	Deepmethyl (http://dna.cs.miami.edu/DeepMethyl/)	Features are obtained from Hi-C and a number of DNA sequence patterns. Stacked denoising autoencoder is used for prediction.
Fan *et al.* [85]	WGBS, 450 K	DNA sequence (k-mer), 450 K data	Random forest	N/A	Features are obtained from DNA sequence. The goal is to train the model using methylation array data and predict the methylation levels on CpG sites that are not covered by the array. It also proposes to integrate multiple types of epigenetic marks to improve prediction accuracy.
Zeng and Gifford [84]	RRBS, meQTL	DNA sequence variants	CNN of Keras	CpGenie (https://github.com/gifford-lab/CpGenie)	A deep learning framework to predict the impact of sequence variation on the methylation level of neighboring CpG sites. It adopts a CNN model to capture the DNA sequence patterns for prediction.
Angermueller *et al.* [83]	Single-cell RRBS	DNA sequence, neighboring CpG	CNN + RNN	DeepCpG (https://deepcpg.readthedocs.io/en/latest/)	A deep neural network model to predict methylation level on low-coverage CpG sites. It uses CNN to capture the DNA sequence patterns for prediction and an RNN framework to model the spatial dependency of methylation level among neighboring CpG sites.
Zou *et al.* [88]	WGBS, EPIC	ATAC-seq, Histone, TFBS, Genomic Features (CGI, GC content, recombination rate)	XGBoost	BoostMe (https://github.com/lulizou/boostme)	Imputation method to estimate methylation level for low-quality regions in WGBS data. It can integrate a diverse set of genomic and epigenetic features and leverage information from multiple samples.

effects among different histone modifications were modeled by machine learning approaches including SVM [62], random forest [63], adaboost [64] or combined with ensemble learning framework [65]. The current trend to predict enhancer is to utilize all the available epigenetic profiles and capture the inter-feature relationships by complex models. Such methods combine histones with chromatin accessibility [66, 67] or methylation [68]. A comprehensive list of the enhancer predicting tools is shown in Table 2.

Prediction of DNA methylation

DNA methylation is the most studied among all epigenetic phenomena. A variety of biological or clinical processes has been reported to be associated with methylation changes [69–72]. The majority of DNA methylation is on the Cytosine of CG dinucleotides (5 mC), or CpG sites, although other types of DNA methylation (non-CG methylation) exist. There are also variations of methylation types, such as hydroxymethylation (5 hmC) [73].

Before high-throughput sequencing became available, methylation level of CpG sites is investigated together with DNA sequence patterns using methylation-specific restriction enzymes [74], and prediction models aim to infer methylation levels based on the DNA sequence. MethDB [75] was the popular choice of the training resource then. Multiple works [76–78] were proposed to use sequence-derived features for methylation prediction on selected CpG sites. However, an intrinsic problem with these initial exploring works is that DNA methylation is highly cell/tissue type-specific, thus using only static genomic DNA sequence is not sufficient to capture its dynamic profile, or at least, this strategy can only be applied to some genomic regions containing specific functional elements [79].

High-throughput technologies enable researchers to profile methylation landscape in larger scale at low cost. Microarrays such as the Illumina Infinium Methylation 450 k and MethylationEPIC arrays are designed to cover CpG sites in important regulatory genomic regions. Sequencing-based technologies such as reduced representation bisulfite sequencing (RRBS) [80] and whole-genome bisulfite sequencing (WGBS) [81] provide wider genome coverage. Quantification of methylation from these diverse platforms is a non-trivial task, and detailed discussion of this topic can be found in this review [82]. Prediction of DNA methylation based on these high-throughput data has a distinct feature compared to other predictions tasks. The majority of tools are designed to perform imputation (instead of prediction) for the whole methylome [83–85], usually within the same cell/tissue. This is different from other prediction tasks, where the prediction for a new cell/tissue context is more important. This is largely the result of DNA methylation measurement techniques; array-based methods are cost effective but only cover a small portion of all CpG sites. On the other hand, WGBS provides genome-wide coverage but is prohibitively expensive to be applied to a large-scale study. Thus, it is desirable to train a model to predict genome-wide methylome (which is expensive to obtain) based on the data from a small portion of the genome (which is low cost).

With the accumulation of epigenome data, using relevant genome features to predict DNA methylation has become possible. There are some works for DNA methylation prediction without data from a proportion of CpG sites (such as from microarray). These are the true 'prediction', instead of 'imputation' methods. For example, Zhang *et al.* [86] integrate sequence signatures together with other cell type-specific markers, including DHS and protein binding sites to predict methylation level. A different group introduced the other possibility to consume chromatin topological features from Hi-C experiments for the same task [87]. A recent method brings in histone modification to the already rich collection of features for the task of methylation prediction [88]. In addition to the prediction of methylation level for CpG sites in general, Zeng *et al.* [84] proposed to predict the effect of sequence variants on CpG methylation, which is similar to the application of DeepBind [34]. Also, imputation methods might find their new application in single-cell data, due to the large proportion of missing values in single-cell methylation data. Angermueller *et al.* [83] proposed a deep learning framework named DeepCpG, which used CNN to capture sequence features and RNN for spatial-dependent relationships among CpG sites to impute missing data for single-cell methylation sequencing results.

A list of the methods for methylation prediction is shown in Table 3.

Prediction of chromatin structure

The genome has a very complex three-dimensional structure. The particular spatial organization is closely linked to the biological functions. Recently developed chromosome conformation capture (3C) [89] and its extension Hi-C techniques can identify genome-wide long-range interactions of chromosome [81, 90, 91]. However, high-resolution Hi-C experiments are costly and difficult to implement, so it is desirable to predict the chromatin structure from other sources of information [92].

For this task, the prediction methods vary due to the complexity of characterizing chromatin structure. Different experimental techniques provide diverse views of the spatial organization of chromatin. A/B compartments are one way to divide the genome based on chromatin structure, where interactions between loci within one compartment are assumed to be independent to the other. The A compartment was found to be associated with open chromatin or euchromatin while the B compartment was with closed chromatin or heterochromatin. A common assumption behind the works to predict chromatin structure is that distal but interacting loci tend to harbor similar epigenomic features. Fortin and Hansen [93] used this strategy to reconstruct A/B compartment using methylation data and the other types of epigenomic marks across multiple cell lines. Genomic loci with high correlation of these marks are predicted to be within the same compartment. Similarly, Zhu *et al.* [94] developed a novel strategy to detect spatial chromatin interaction structure measured by capture-C experiments. They collected epigenomic marks including chromatin accessibility, histone modifications and gene expression levels from five different tissue types. Then correlations of distal loci are measured, and significantly associated loci are identified based on permutation. Huang *et al.* [95] adopted a different way to study the structural features by defining interaction hubs and train a machine learning classifier based on epigenomic marks. In addition to the efforts to utilize epigenomic marks, molecular thermal simulation has been applied to explore the structural feature of chromatin as well. Brackley *et al.* [96] proposed to use chromatin accessibility data to infer protein binding sites, which could serve as the interaction points to form protein bridge. Then polymer modeling was applied to simulate the thermal motion of the chromatin fiber to detect potential local interaction structures based on inferred protein bridging sites. A recent study extends the prediction target from the

Table 4. Prediction methods of chromatin structure

Publication	Targeted chromatin	Features used	Method	Software	Description
Fortin and Hansen [93]	A/B compartment	450 K Methylation microarray, DHS, scATAC-seq, scWGBS	Eigenvector of features + Correlation	N/A (R script provided)	Compute the methylation correlation matrix for fixed-length bin. The sign of the first principal components (PC) of the correlated matrix is used to estimate compartments.
Huang *et al.* [95]	Chromatin interaction hubs and topologically associated domain (TAD) boundaries	Hi-C, histones, DNA sequence	Bayesian additive regression tree	N/A	Summarized local pattern for each histone mark then signals are input to a BART model to predict hubs; call peaks of CTCF by MACS, TAD boundaries are predicted as CTCF peaks.
Zhu *et al.* [94]	Spatial association within TADs	Histone, DHS, RNA-seq	Tensor vector correlation + permutation	EpiTensor (http://wanglab.ucsd.edu/star/EpiTensor/)	Deconvolute 18 histone modification signals of five cell types into three tensor spaces, associations between two regions are calculated as their geometric mean of their scores in eigenlocus space.
Brackley *et al.* [96]	Local folding/interaction map on Alpha/beta globin loci in mouse erythoblasts	DHS, TF motifs	Polymer model	N/A	Use DHS as a proxy for binding of a generic type of protein bridge and then predict chromosome architecture f using a Polymer model.
Jung *et al.* [97]	Chromatin accessibility (ENCODE DNase-seq)	Transcriptomic data (ENCODE RNA-seq)	Hierarchical random forest	ChromAccPrediction (https://github.com/saschajung/ChromAccPrediction)	Use a hierarchical random forest model to predict chromatin accessibility from gene expression data.

looping structure of chromatin to accessibility prediction [97]. Transcriptome data are used to infer the chromatin landscape within gene-regulatory regions.

A list of the methods for chromatin structure prediction is shown in Table 4.

Prediction of gene expression

Gene expression is often considered as the target in the whole cascade of regulatory network. All the other elements, including chromatin accessibility, histone modification and protein binding, serve as intermediate steps to control the expressions of target genes. Gene expression can be easily profiled by a number of high-throughput experiments such as gene expression microarray [98], RNA-seq [99], Cap analysis gene expression (CAGE) [100, 101] and RNA-PET [102]. These methods are considerably cheaper than profiling several other epigenetic marks. Thus, the main purpose of gene expression prediction is to study the regulatory mechanisms of different epigenetic marks on gene expression.

Initially, there were efforts to study the patterns of DNA sequence at the regulatory regions of expressed genes in limited tissue context. Sequence-based prediction method was developed in early works by Yuan *et al.* [103], where they select regulatory motif to predict gene expression patterns in yeast. In this work, genes are clustered into groups, and then predictive models are trained to use their sequence signature to classify genes into the corresponding class. Later methods shift their focus onto a broader range of features such as histone modifications to predict gene expression rates, either as a discrete level from a classifier or as a continuous output from a regression framework. Different types of histone modifications have been found to be correlated with gene expression. For example, Karlic *et al.* [104] used a number of histone marks to predict gene expression. They showed that only a small subset of histone marks is needed to accurately predict gene expression levels, while the subset selection of histone marks is dependent on the GC content of the gene promoter regions. Yu *et al.* [105] used a Bayesian network to jointly model the causal relationship between histone modifications and gene expression. Their result reveals not only the regulatory mechanism from histone modification to gene expression but also inter-regulatory relationships among different types of histone marks. With the popularity of deep learning, a recent work applies CNN with histone modification profile as the input and achieved good performance [106].

In addition to histone modifications, protein bindings are shown to be related to gene expression as well, which is not surprising due to the role of many DNA-binding proteins such as the TFs. Ouyang *et al.* [107] predict gene expression based on the binding profile of a number of TFs. They propose to decompose groups of TFBS signals into sets of combinations (PCs) using PCA then perform regression to model their relationships with gene expression. There are also attempts to use other features to predict gene expression. Park *et al.* used methylation levels as input for the first time in [108]. This strategy has been extended to non-linear SVM model by Kapourani and Sanguinetti [109]. Natarajan *et al.* [110] introduced DHS as the predicting feature when this type of cell-specific data were released by ENCODE, which makes the use of chromatin accessibility in study of cell-specific epigenetic regulation on a large scale possible. In addition, DNA shape was also found to be predictive for TFBS and gene expression [111]. Since 2011, researchers started to combine multiple resources for the gene expression prediction. Costa *et al.* [112] combined TFBS and histone marks to predict the expression levels of genes with low-CG promoters in diverse immune cells. They point out that genes with low-CG promoters tend to express in tissue-specific manner. Cheng *et al.* [113, 114] followed the same strategy to predict gene expression in *Caenorhabditis elegans* and mouse embryonic stem cell (ESCs), using SVM to accommodate non-linear model. There are also some works to use genetic variants to predict gene expression [115]. Note that this is different from the expression quantitative trait loci (eQTL) analysis, which focuses on detecting SNPs correlated with gene expressions, though the eQTLs or GWAS SNPs are often used as predictors.

It is worth noting that there are very few works to predict gene expression from enhancers, likely due to the lack of gold standard definition for enhancers and enhancer–gene pairs.

Table 5. Prediction methods of gene expression. This table doesn't include a software column because most of the studies do not provide a software along with the method. If the method has a software, it is listed in the 'Description' column

Publication	Features used	Method	Target genes	Description
Yuan *et al.* [103]	TF motifs	Naïve Bayes	Microarray, 2587 yeast genes	Use TF binding motifs as features. A naïve Bayes model is trained to predict gene expression in yeast.
Yu *et al.* [105]	Histone	Bayesian Network	Microarray, ∼15 000 human genes from multiple tissues	Predict regulatory network between histone modifications and gene expressions.
Ouyang *et al.* [107]	TFBS	Linear regression	RNA-seq, mouse ESCs	Perform PCA on TF binding strength; use PCs as regressors. A linear regression model is fit for expression level prediction.
Karlic *et al.* [104]	Histone	Linear regression	Microarray, human T cell	Use different combinations of histone marks to predict gene expression. Linear regression models are used.
Costa *et al.* [112]	Histone + TFBS	Mixture of linear models	Microarray, human Th1, Th2, Th17 and iTreg cells	Use a combination of histone signals and TFBS as features. Use linear regression for each factor then use EM for estimating parameters.
Park *et al.* [108]	12 TFBS, Methylation, Histone, CpG island	Linear regression	RNA-seq, mouse ESCs	Combine TFBS, methylation, histone and CpG island annotation as features. Linear regression models are fit. Two classes of genes regulated by distinct combination of epigenetic marks were discovered.
Cheng *et al.* [113]	Histone, TFBS	SVM	RNA-seq, C. *elegans* and other species (modENCODE data)	Use SVM with histone and TFBS as features. It shows that histone features are redundant; positional contribution varies.
Natarajan *et al.* [110]	DHS	Logistic regression + L_1 norm	Microarray, 19 human cell lines (from ENCODE)	Use DHS and TF motifs on DNA sequence to infer binding sites. A logistic regression model with L-1 norm is used.
Cheng *et al.* [114]	Histone, TFBS	SVM	RNA-seq, mouse ESCs	Use Histone and TFBS in SVM model. TFBS and Histone show distinct spatial patterns.
Gamazon *et al.* [115]	DNA sequence variants	LASSO, elastic net and polygenic score	RNA-seq, Human (DGN, GEUVADIS, GTEx)	Use SNP genotypes to predict gene expressions with different penalized regression models.
Kapourani and Sanguinetti [109]	Methylation (RRBS)	SVM regression	RNA-seq, cell lines (K562, GM12878, H1-hESC)	Use methylation in SVM regression model for prediction. Methylation profiles are predictive of gene expression across cell lines.
Singh *et al.* [106]	Histone	CNN	RNA-seq (REMC)	Use a deep learning CNN model on histone data for prediction. Software (deepChrome) is available at deepchrome.org.
Peng and Sinha [111]	DNA shape features, TF motifs	Random forest	37 *Drosophila* genes	Combine DNA shape features with TF motifs as features for random forest model.

However, since the definition of enhancer is mostly based on epigenetic marks, predicting expression from epigenetic marks partially includes the effects from enhancer.

A list of the methods for gene expression prediction is provided in Table 5.

Discussion

In silico genome feature prediction has become a popular research field in recent years, mainly because of the availability of large-scale training data and the advances of machine learning methods. Since it is impossible to profile all genome features under all conditions, *in silico* prediction provides an economical solution to fill the gaps in the experimental data. In the meantime, the prediction models shed light on how the genome work as a coordinated system and greatly enhance our knowledge in gene regulation mechanism. In this work, we conduct a comprehensive review on the available *in silico* prediction methods for different genome features. All methods reviewed are summarized in a dynamic figure provided at https://stanleyxu.github.io/featureNet/, where each node is a genome feature, and an arrow represents the prediction direction (from predictor to outcome). It should be noted that for one method, there might be multiple arrows targeting the same node, which means that this method combines different types of marks as input in order to predict the target mark. It is of note that our aim here is to provide a summary of existing publications and methods instead of comparing different software, thus the performances of different software tools are not evaluated.

Despite the rapid developments and excitements in this field, there are some important points we want to make. First, the rationale for predicting features needs to be clarified. There could be economical reasons (e.g. to predict targets requiring higher cost to profile based on predictors that can be obtained from lower-cost experiments) or technical reasons (e.g. to predict target that is technically difficult to measure, for example, requires surgical procedure). Even though there are works that use multiple marks requiring higher cost to predict lower-cost targets, they are not meant to be used in practice. For example, predicting gene expression using a full collection of chromatin profiles would be economically unreasonable. The main goal in that practice is to provide insight to the biological mechanism. Secondly, in evaluating the prediction results, many methods train the model on half of the genome and predict on the other half. This evaluation is not practical and will provide inflated accuracies since the training and testing data share many technical characteristics such as the same experimental condition (e.g. no batch effect). It is advisable to use different data sets for training, testing and validation in order to provide evaluation in real-world scenario. Furthermore, it is more important and interesting to predict the features showing variability among distinct conditions. For example, in predicting DNA methylation, one can simply predict genome-wide hypermethylation and CpG islands hypomethylation and get reasonable results, but that would not be very interesting. The dynamic regions [such as the differentially methylated regions (DMRs) between different biological conditions] will be of more interests. In this regard, it is important to include dynamic features (such as DNase or protein binding data) as predictors, instead of using static features such as DNA sequence alone.

There are also some technical issues that require attention. The first one is the technical artifacts such as experimental protocol or batch effects. One needs to be careful in data normalization to make sure the model trained in one set of data can be applied to the other set. It is also highly likely that the trained model cannot be transferred between different platforms, for example, the model trained using sequencing data will not perform well when using microarray data as predictors. Secondly, the model trained using cell line data might not work well for data from clinical samples, which is often a mixture of different cell types. Even though there are methods for signal deconvolution for complex samples such as cancer [116, 117], how to incorporate these methods into the feature prediction framework is understudied and will be an interesting research direction worth exploring.

Key Points

- There are a number of features (genetic, genomic, epigenomic, etc.) in the genome, which can be measured on the genome-wide scale by high-throughput technologies.
- There are intrinsic correlations among different genome features. The correlation makes the *in silico* feature prediction possible.
- It is impossible to measure all features under all conditions. Thus, *in silico* feature prediction is useful to fill the gaps in experimental data.
- A plethora of methods and software are available for predicting different genome features based on statistical or machine learning techniques, using other genome features as predictors.

Funding

National Institutes of Health (R01GM122083 to H.W.; U54NS091859 to Z.Q. and P.J.; P01NS097206 to H.W. and P.J.); National Natural Science Foundation of China (61572327 to X.Z.).

References

1. Bernstein BE, Stamatoyannopoulos JA, Costello JF, *et al.* The NIH roadmap epigenomics mapping consortium. *Nat Biotechnol* 2010;**28**:1045–8.
2. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* 2015;**518**:317–30.
3. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**: 57–74.
4. Sandelin A, Alkema W, Engstrom P, *et al.* JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 2004;**32**:D91–4.
5. Matys V, Fricke E, Geffers R, *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* 2003;**31**:374–8.
6. Griffith OL, Montgomery SB, Bernier B, *et al.* ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res* 2008;**36**:D107–13.
7. Portales-Casamar E, Arenillas D, Lim J, *et al.* The PAZAR database of gene regulatory information coupled to the

ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Res* 2009;**37**:D54–60.

8. Wang J, Zhuang J, Iyer S, *et al*. Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Res* 2013;**41**:D171–6.
9. Arvey A, Agius P, Noble WS, *et al*. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Res* 2012;**22**:1723–34.
10. Barrera LA, Vedenko A, Kurland JV, *et al*. Survey of variation in human transcription factors reveals prevalent DNA binding changes. *Science* 2016;**351**:1450–4.
11. Heintzman ND, Stuart RK, Hon G, *et al*. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* 2007;**39**: 311–8.
12. Schones DE, Cui K, Cuddapah S, *et al*. Dynamic regulation of nucleosome positioning in the human genome. *Cell* 2008;**132**:887–98.
13. Whitington T, Perkins AC, Bailey TL. High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Res* 2009;**37**:14–25.
14. He HH, Meyer CA, Shin H, *et al*. Nucleosome dynamics define transcriptional enhancers. *Nat Genet* 2010;**42**:343–7.
15. Talebzadeh M, Zare-Mirakabad F. Transcription factor binding sites prediction based on modified nucleosomes. *PLoS One* 2014;**9**:e89226.
16. Ramsey SA, Knijnenburg TA, Kennedy KA, *et al*. Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics* 2010;**26**:2071–5.
17. Won KJ, Ren B, Wang W. Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biol* 2010;**11**:R7.
18. Ji H, Li X, Wang QF, *et al*. Differential principal component analysis of ChIP-seq. *Proc Natl Acad Sci USA* 2013;**110**: 6789–94.
19. Sung MH, Guertin MJ, Baek S, *et al*. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol Cell* 2014;**56**:275–85.
20. Gusmao EG, Dieterich C, Zenke M, *et al*. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics* 2014;**30**:3143–51.
21. Pique-Regi R, Degner JF, Pai AA, *et al*. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Res* 2011;**21**:447–55.
22. Yardimci GG, Frank CL, Crawford GE, *et al*. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res* 2014;**42**:11865–78.
23. Sherwood RI, Hashimoto T, O'Donnell CW, *et al*. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat Biotechnol* 2014;**32**:171–8.
24. Jankowski A, Tiuryn J, Prabhakar S. Romulus: robust multi-state identification of transcription factor binding sites from DNase-seq data. *Bioinformatics* 2016;**32**:2419–26.
25. Chen X, Yu B, Carriero N, *et al*. Mocap: large-scale inference of transcription factor binding sites from chromatin accessibility. *Nucleic Acids Res* 2017;**45**:4315–29.
26. Cuellar-Partida G, Buske FA, McLeay RC, *et al*. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics* 2012;**28**:56–62.
27. Quach B, Furey TS. DeFCoM: analysis and modeling of transcription factor binding sites using a motif-centric genomic footprinter. *Bioinformatics* 2017;**33**:956–63.
28. Liu S, Zibetti C, Wan J, *et al*. Assessing the model transferability for prediction of transcription factor binding sites based on chromatin accessibility. *BMC Bioinformatics* 2017;**18**:355.
29. Kuang Z, Ji Z, Boeke JD, *et al*. Dynamic motif occupancy (DynaMO) analysis identifies transcription factors and their binding sites driving dynamic biological processes. *Nucleic Acids Res* 2018;**46**:e2.
30. He HH, Meyer CA, Hu SS, *et al*. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat Methods* 2014;**11**:73–8.
31. Gusmao EG, Allhoff M, Zenke M, *et al*. Analysis of computational footprinting methods for DNase sequencing experiments. *Nat Methods* 2016;**13**:303–9.
32. Xu T, Li B, Zhao M, *et al*. Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. *Nucleic Acids Res* 2015;**43**:2757–66.
33. Ma W, Yang L, Rohs R, *et al*. DNA sequence+shape kernel enables alignment-free modeling of transcription factor binding. *Bioinformatics* 2017;**33**:3003–10.
34. Alipanahi B, Delong A, Weirauch MT, *et al*. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;**33**:831–8.
35. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems 25 (NIPS 2012), 2012*. 1106–14. Curran Associates, Inc., Red Hook, NY.
36. Quang D, Xie W. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *BioRxiv* 2017;151274.
37. Mikolov T, Karafiat M, Burget L, *et al*. INTERSPEECH: recurrent neural network based language model. In: *11th Annual Conference of the International Speech Communication Association, 2010*. p. 1045–8. Curran Associates, Inc., Red Hook, NY.
38. Andersson R, Gebhard C, Miguel-Escalada I, *et al*. An atlas of active enhancers across human cell types and tissues. *Nature* 2014;**507**:455–61.
39. Jin C, Zang C, Wei G, *et al*. H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nat Genet* 2009;**41**:941–5.
40. Koch CM, Andrews RM, Flicek P, *et al*. The landscape of histone modifications across 1% of the human genome in five human cell lines. *Genome Res* 2007;**17**:691–707.
41. Cotney J, Leng J, Oh S, *et al*. Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. *Genome Res* 2012;**22**:1069–80.
42. Creyghton MP, Cheng AW, Welstead GG, *et al*. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci USA* 2010;**107**:21931–6.
43. Rada-Iglesias A, Bajpai R, Swigut T, *et al*. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* 2011;**470**:279–83.
44. Visel A, Blow MJ, Li Z, *et al*. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 2009;**457**: 854–8.
45. Blow MJ, McCulley DJ, Li Z, *et al*. ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* 2010;**42**: 806–10.

46. Ghisletti S, Barozzi I, Mietton F, *et al*. Identification and characterization of enhancers controlling the inflammatory gene expression program in macrophages. *Immunity* 2010;**32**:317–28.
47. May D, Blow MJ, Kaplan T, *et al*. Large-scale discovery of enhancers from human heart tissue. *Nat Genet* 2011;**44**: 89–93.
48. Zinzen RP, Girardot C, Gagneur J, *et al*. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature* 2009;**462**:65–70.
49. He A, Kong SW, Ma Q, *et al*. Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. *Proc Natl Acad Sci USA* 2011;**108**:5632–7.
50. Yip KY, Cheng C, Bhardwaj N, *et al*. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* 2012;**13**:R48.
51. Cheng C, Alexander R, Min R, *et al*. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome Res* 2012;**22**:1658–67.
52. Wamstad JA, Alexander JM, Truty RM, *et al*. Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage. *Cell* 2012;**151**:206–20.
53. Paige SL, Thomas S, Stoick-Cooper CL, *et al*. A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. *Cell* 2012;**151**: 221–32.
54. Zentner GE, Tesar PJ, Scacheri PC. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res* 2011;**21**:1273–83.
55. Bonn S, Zinzen RP, Girardot C, *et al*. Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nat Genet* 2012;**44**:148–56.
56. Lee D, Karchin R, Beer MA. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res* 2011;**21**:2167–80.
57. Ghandi M, Lee D, Mohammad-Noori M, *et al*. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* 2014;**10**:e1003711.
58. Taher L, Narlikar L, Ovcharenko I. CLARE: Cracking the LAnguage of Regulatory Elements. *Bioinformatics* 2012;**28**:581–3.
59. Liu B, Fang L, Long R, *et al*. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics* 2016;**32**: 362–9.
60. Jia C, He W. EnhancerPred: a predictor for discovering enhancers based on the combination and selection of multiple features. *Sci Rep* 2016;**6**:38741.
61. Firpi HA, Ucar D, Tan K. Discover regulatory DNA elements using chromatin signatures and artificial neural network. *Bioinformatics* 2010;**26**:1579–86.
62. Fernandez M, Miranda-Saavedra D. Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines. *Nucleic Acids Res* 2012;**40**:e77.
63. Rajagopal N, Xie W, Li Y, *et al*. RFECS: a random-forest based algorithm for enhancer identification from chromatin state. *PLoS Comput Biol* 2013;**9**:e1002968.
64. Lu Y, Qu W, Shan G, *et al*. DELTA: a Distal Enhancer Locating Tool based on AdaBoost algorithm and shape features of chromatin modifications. *PLoS One* 2015;**10**:e0130622.
65. Kleftogiannis D, Kalnis P, Bajic VB. DEEP: a general computational framework for predicting enhancers. *Nucleic Acids Res* 2015;**43**:e6.
66. Erwin GD, Oksenberg N, Truty RM, *et al*. Integrating diverse datasets improves developmental enhancer prediction. *PLoS Comput Biol* 2014;**10**:e1003677.
67. Liu F, Li H, Ren C, *et al*. PEDLA: predicting enhancers with a deep learning-based algorithmic framework. *Sci Rep* 2016;**6**:28517.
68. He Y, Gorkin DU, Dickel DE, *et al*. Improved regulatory element prediction based on tissue-specific local epigenomic signatures. *Proc Natl Acad Sci USA* 2017;**114**:E1633–40.
69. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. *Nat Rev Genet* 2013;**14**:204–20.
70. Jones PA, Takai D. The role of DNA methylation in mammalian epigenetics. *Science* 2001;**293**:1068–70.
71. Baylin SB. DNA methylation and gene silencing in cancer. *Nat Clin Pract Oncol* 2005;**2**(Suppl 1):S4–11.
72. Jones PA. DNA methylation and cancer. *Cancer Res* 1986;**46**:461–6.
73. Yu M, Hon GC, Szulwach KE, *et al*. Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. *Cell* 2012;**149**:1368–80.
74. Rollins RA, Haghighi F, Edwards JR, *et al*. Large-scale structure of genomic methylation patterns. *Genome Res* 2006;**16**:157–63.
75. Grunau C, Renault E, Rosenthal A, *et al*. MethDB—a public database for DNA methylation data. *Nucleic Acids Res* 2001;**29**:270–4.
76. Bhasin M, Zhang H, Reinherz EL, *et al*. Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett* 2005;**579**:4302–8.
77. Fang F, Fan S, Zhang X, *et al*. Predicting methylation status of CpG islands in the human brain. *Bioinformatics* 2006;**22**: 2204–9.
78. Das R, Dimitrova N, Xuan Z, *et al*. Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci USA* 2006;**103**:10713–6.
79. Whitaker JW, Chen Z, Wang W. Predicting the human epigenome from DNA motifs. *Nat Methods* 2015;**12**:265–72, 267 p following 272.
80. Meissner A, Gnirke A, Bell GW, *et al*. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Res* 2005;**33**:5868–77.
81. Lister R, Pelizzola M, Dowen RH, *et al*. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 2009;**462**:315–22.
82. Qin Z, Li B, Conneely KN, *et al*. Statistical challenges in analyzing methylation and long-range chromosomal interaction data. *Stat Biosci* 2016;**8**:284–309.
83. Angermueller C, Lee HJ, Reik W, *et al*. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol* 2017;**18**:67.
84. Zeng H, Gifford DK. Predicting the impact of non-coding variants on DNA methylation. *Nucleic Acids Res* 2017;**45**:e99.
85. Fan S, Huang K, Ai R, *et al*. Predicting CpG methylation levels by integrating Infinium HumanMethylation450 BeadChip array data. *Genomics* 2016;**107**:132–7.
86. Zhang W, Spector TD, Deloukas P, *et al*. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol* 2015;**16**:14.
87. Wang Y, Liu T, Xu D, *et al*. Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks. *Sci Rep* 2016;**6**:19598.

88. Zou LS, Erdos MR, Taylor DL, *et al*. BoostMe accurately predicts DNA methylation values in whole-genome bisulfite sequencing of multiple human tissues. *BMC Genomics* 2018;**19**:390.
89. Dekker J, Rippe K, Dekker M, *et al*. Capturing chromosome conformation. *Science* 2002;**295**:1306–11.
90. Lieberman-Aiden E, van Berkum NL, Williams L, *et al*. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 2009;**326**:289–93.
91. Jin F, Li Y, Dixon JR, *et al*. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature* 2013;**503**:290–4.
92. Mifsud B, Tavares-Cadete F, Young AN, *et al*. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat Genet* 2015;**47**:598–606.
93. Fortin JP, Hansen KD. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol* 2015;**16**:180.
94. Zhu Y, Chen Z, Zhang K, *et al*. Constructing 3D interaction maps from 1D epigenomes. *Nat Commun* 2016;**7**: 10812.
95. Huang J, Marco E, Pinello L, *et al*. Predicting chromatin organization using histone marks. *Genome Biol* 2015; **16**:162.
96. Brackley CA, Brown JM, Waithe D, *et al*. Predicting the three-dimensional folding of cis-regulatory regions in mammalian genomes using bioinformatic data and polymer models. *Genome Biol* 2016;**17**:59.
97. Jung S, Angarica VE, Andrade-Navarro MA, *et al*. Prediction of chromatin accessibility in gene-regulatory regions from transcriptomics data. *Sci Rep* 2017;**7**:4660.
98. Schulze A, Downward J. Navigating gene expression using microarrays—a technology review. *Nat Cell Biol* 2001;**3**: E190–5.
99. Mortazavi A, Williams BA, McCue K, *et al*. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;**5**:621–8.
100. Shiraki T, Kondo S, Katayama S, *et al*. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci USA* 2003;**100**:15776–81.
101. Kodzius R, Kojima M, Nishiyori H, *et al*. CAGE: cap analysis of gene expression. *Nat Methods* 2006;**3**:211–22.
102. Ruan Y, Ooi HS, Choo SW, *et al*. Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). *Genome Res* 2007;**17**:828–38.
103. Yuan Y, Guo L, Shen L, *et al*. Predicting gene expression from sequence: a reexamination. *PLoS Comput Biol* 2007;**3**: e243.
104. Karlic R, Chung HR, Lasserre J, *et al*. Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci USA* 2010;**107**:2926–31.
105. Yu H, Zhu S, Zhou B, *et al*. Inferring causal relationships among different histone modifications and gene expression. *Genome Res* 2008;**18**:1314–24.
106. Singh R, Lanchantin J, Robins G, *et al*. DeepChrome: deep-learning for predicting gene expression from histone modifications. *Bioinformatics* 2016;**32**:i639–48.
107. Ouyang Z, Zhou Q, Wong WH. ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci USA* 2009;**106**: 21521–6.
108. Park SJ, Nakai K. A regression analysis of gene expression in ES cells reveals two gene classes that are significantly different in epigenetic patterns. *BMC Bioinformatics* 2011;**12**(Suppl 1):S50.
109. Kapourani CA, Sanguinetti G. Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics* 2016;**32**:i405–12.
110. Natarajan A, Yardimci GG, Sheffield NC, *et al*. Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res* 2012;**22**:1711–22.
111. Peng PC, Sinha S. Quantitative modeling of gene expression using DNA shape features of binding sites. *Nucleic Acids Res* 2016;**44**:e120.
112. Costa IG, Roider HG, do Rego TG, *et al*. Predicting gene expression in T cell differentiation from histone modifications and transcription factor binding affinities by linear mixture models. *BMC Bioinformatics* 2011;**12**(Suppl 1):S29.
113. Cheng C, Yan KK, Yip KY, *et al*. A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol* 2011;**12**:R15.
114. Cheng C, Gerstein M. Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res* 2012;**40**:553–68.
115. Gamazon ER, Wheeler HE, Shah KP, *et al*. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 2015;**47**:1091–8.
116. Carter SL, Cibulskis K, Helman E, *et al*. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 2012;**30**:413–21.
117. Zheng X, Zhang N, Wu HJ, *et al*. Estimating and accounting for tumor purity in the analysis of DNA methylation data from cancer studies. *Genome Biol* 2017;**18**:17.