Modeling the relationship of diverse genomic signatures to gene expression levels with the regulation of long-range enhancer-promoter interactions

Zhen-Xing Feng^{1,2}, Qian-Zhong Li^{2,3}✉, Jian-Jun Meng²

¹ Department of Sciences, Inner Mongolia University of Technology, Hohhot 010051, China

² Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China

³ The State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, Inner Mongolia University, Hohhot 010070, China

Received: 15 January 2017 / Accepted: 7 December 2018 / Published online: 24 June 2019

Abstract Enhancer–promoter (E–P) interaction is an essential component of cis-regulatory regulation for gene expression. However, to comprehensively study the gene expression with the regulation of long-range E–P interactions is a major challenge in the regulatory networks. As these types of gene expression are regulated by diverse genomic signatures, we presented a computational method to study the relationships between gene expression levels and diverse genomic signatures. In this paper, based on the datasets of long-range E–P interactions, we extracted feature parameters from multiple signatures (*e.g.*, epigenetic marks, transcription factors) and used regression models to predict the gene expression levels. In our results, we found that the predicted expression values correlated well with the measured expression values in both the interacting and non-interacting sets, and the correlation values of the interacting set were higher than that of the corresponding non-interacting set in each cell line, which indicated that the distal enhancers would cooperate with diverse genomic signatures to facilitate the expression level of target genes. By comparing the important signature features for the gene expression levels between the interacting and non-interacting sets in the same cell line, we found that the important specific signatures affect the gene expression regulated by distal enhancers. Our research provided additional insights about the roles of diverse signatures in gene expression with the regulation of distal enhancers.

Keywords Enhancer–promoter interaction, Random Forest, Epigenetics, Gene expression, DNA methylation

INTRODUCTION

In various processes of eukaryotic cells, numerous DNA reactions are regulated by transcriptional promoters, enhancers and insulators (Bejerano *et al.* 2004; The

ENCODE Project Consortium 2007). Among these cis-regulatory elements, the enhancers can be away from their target genes over tens of kilobases distances (Miele and Dekker 2008; de Laat and Duboule 2013). It was reported that the enhancers can act as a key role in the regulation of tissue-specific gene expression (Visel *et al.* 2009a, b), and many specific enhancers would take part in the regulation of several human diseases (Williamson *et al.* 2011). Thus, elucidating the mechanism that how an enhancer selects its target promoter(s) and regulates gene expression is a major challenge so far.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s41048-019-0089-z>) contains supplementary material, which is available to authorized users.

✉ Correspondence: qzli@imu.edu.cn (Q.-Z. Li),

Recently, several types of experimental techniques enable us to study the E–P interactions, including fluorescence *in situ* hybridization (FISH) (Williamson *et al.* 2014), chromosome conformation capture (3C) and other 3C-based techniques (*e.g.*, 4C, 5C, Hi-C) (Dekker *et al.* 2002; Sanyal *et al.* 2012; Rao *et al.* 2014; van de Werken *et al.* 2012). Generally, the 3C-based techniques determine cross-linked interactions in a cell where a pair of sequence fragments are close enough in three-dimensional (3D) space, and these experimental techniques are expensive and time consuming. In the past years, based on the availability of sequence, epigenetics and other genomic data, computational approaches for understanding the interactions between enhancers and promoters have made a significant advance, such as PreSTIGE (Corradin *et al.* 2013), IM-PET (He *et al.* 2014), HIPPIE (Hwang *et al.* 2015), RIPPLE (Roy *et al.* 2015), Target Finder (Whalen *et al.* 2016) and our group's works (Feng and Li 2017; Feng *et al.* 2018). Furthermore, many deep sequencing data profiles of transcription factors and epigenetic marks were generated from different sequencing techniques. Thus, the omics data enable us to study the relationships between related features and gene expression levels, which offered insights into the roles of the regulatory proteins, RNAs and other factors in the regulation of gene expression on gene expression (Budden *et al.* 2015; Cheng *et al.* 2011; Cheng and Gerstein 2012; Karlič *et al.* 2010; Mcleay *et al.* 2012; Ouyang *et al.* 2009). However, for the regulation of gene expression related to distal enhancers, there is still much work to do.

In this study, based on the mass-sequencing data, we mainly aim to develop computational models to study the relationship between diverse genomic signatures and gene expression levels with the regulation of distal enhancers. Thus, we extracted feature parameters from large numbers of epigenetic marks and other signatures including histone modifications, transcription factors, DNase I, enhancer RNAs (eRNAs), DNA methylation and nucleosome occupancy. Then, by integrating the combined features and ensemble classification strategy, we used Random Forest (RF), gradient boosting machine (GBM) and support vector regression (SVR) algorithms to predict the gene expression levels with the regulation of distal enhancers. Meanwhile, we studied the different types of important features for the gene expression associated with long-range E–P interactions and their location-specificity differences between the positive and negative sets, which would be beneficial to understand the distal regulatory mechanisms deeply.

RESULTS AND DISCUSSION

Modeling gene expression levels with the regulation of distal enhancers by diverse genomic features

We aimed to explore the quantitative relationships between diverse genomic signatures and gene expression levels in the long-range E–P interactions datasets. In this paper, we constructed three computational models (*i.e.*, the SVR, RF and GBM algorithms) to predict the gene expression levels. In the prediction models, we incorporated the signature features of different regulatory regions (enhancers, promoters and loop regions) together from diverse genomic signatures (*i.e.*, histone modifications, transcription factors, DNase I, FAIRE-seq, DNA methylation, nucleosome position and eRNAs). Using the combined features for SVR, RF and GBM models, we showed the prediction results of 10-fold cross-validation test in Table 1. Because of the advantages of ensemble learning algorithm in big data mining, the RF and GBM models obtained better performance results than that of SVR model in different cell lines. It showed that the best R^2 values of the interacting sets measured up to 0.7636, 0.8293, 0.9085 and 0.9021 in Gm12878 (B-lymphocyte), H1hesc (embryonic stem cells), Hela (cervical carcinoma) and K562 (leukemia) cell lines, respectively. However, the best R^2 values of the non-interacting sets were 8.23%, 24.45%, 14.05%, 16.48% lower in the same corresponding cell lines, respectively. Moreover, the above tendency was similar in the three models.

Based on the above results, we further discussed the Pearson correlation coefficient (PCC) between the predicted expression values and experimental measured levels in the RF model. In Fig. 1, we showed the results of interacting and non-interacting sets in four cell lines, respectively. In our observation, the PCC values of the interacting sets measured up to 0.8736, 0.9052, 0.9532 and 0.9498 in Gm12878, H1hesc, Hela and K562 cell lines, respectively. In contrast, the corresponding values of the non-interacting sets in above cell lines decreased to 0.8521, 0.7583, 0.8764 and 0.8587, respectively. Although the predicted expression values correlated well with the measured expression values in both the interacting and non-interacting sets, the PCC values of the interacting set were higher than that of the corresponding non-interacting set in each cell line, which indicated that the distal enhancers would cooperate with diverse genomic signatures to facilitate the expression level of target genes.

Table 1 The prediction results of the full feature sets using 10-fold cross-validation test

Model	Cell line	Interacting set		Non-interacting set	
		R^2	RMSE	R^2	RMSE
SVR	Gm12878	0.5606	0.7052	0.5031	0.8572
	H1hesc	0.5789	0.6646	0.3183	0.6993
	Hela	0.6693	0.6793	0.4887	0.7923
	K562	0.6357	0.7761	0.4291	1.0735
RF	Gm12878	0.7631 (mtry = 151)	0.5165	0.6808 (mtry = 300)	0.6844
	H1hesc	0.8195 (mtry = 222)	0.4353	0.5750 (mtry = 112)	0.5602
	Hela	0.9085 (mtry = 219)	0.3548	0.7680 (mtry = 110)	0.5516
	K562	0.9021 (mtry = 411)	0.4038	0.7373 (mtry = 206)	0.7305
GBM	Gm12878	0.7636	0.5016	0.6765	0.6850
	H1hesc	0.8293	0.4125	0.5738	0.5450
	Hela	0.9073	0.3559	0.7585	0.5431
	K562	0.8619	0.4677	0.7257	0.7457

The bold values show the best results in each cell line

The contributions of diverse signatures to gene expression levels with the regulation of distal enhancers

To quantify the relative contributions of diverse signatures to gene expression levels, according to the previous works (Feng *et al.* 2018; Roy *et al.* 2015), we used the “%IncMSE” error in RF model to compute the relative importance of diverse variables in the prediction models of gene expression levels. One feature with the higher “%IncMSE” value contributes more to the prediction model. Here, we presented the importance scores of the top 50 variables of interacting and non-interacting sets in each cell line (Hela and K562 in Fig. 2; Gm12878 and H1hesc in Fig. S1).

In the prediction of gene expression levels regulated by distal enhancers, multiple types of signatures were identified as top-ranked variables, including histone modifications, DNase I, transcription factors, nucleosome occupancy, eRNAs and DNA methylation. Especially in promoter regions, several ones tended to appear robustly in the top-ranked positions in both the interacting and non-interacting sets of multiple cell lines, such as eRNA, H3k36me3, H3k79me2, H3k27me3, H4k20me1, H3k4me3, H3k9me3, H2az, Pol2, DNase I, DNA methylation and nucleosome occupancy. Among the top variables of interacting/non-interacting data sets, the histone modification features respectively accounted for 42%/36%, 52%/42%, 44%/48% and 48%/42% of the total 50 features in Gm12878, H1hesc, Hela and K562 cell lines. While the transcription factor features, respectively, accounted for 38%/50%, 36%/48%, 40%/38% and 34%/40% of the total 50 features in the four cell lines. Although the distal enhancers can interact with their target promoter regions in the complex cis-regulatory system, the promoter regions always occupy the more

important position in the gene expression regulation. In both of the interacting and non-interacting sets, the genomic signatures tended to be promoter-specific among the three regulatory regions. In the top 50 features, we observed that the features of promoter regions occupied the largest proportion in the interacting and non-interacting sets, which respectively accounted for 48%/60%, 44%/40%, 44%/36% and 58%/44% of the total features in Gm12878, H1hesc, Hela and K562 cell lines (Table 2). Among the above variables, there were many features with minor changes in score, and the features with great changes related to the specific gene expression regulation of distal enhancers. Thus, there would not be a higher percentage of important enhancer features in the interacting set than the non-interacting set. We observed that there was an obvious difference in the regulation of gene expression level between the histone modifications and transcription factors. For histone modifications, the features of promoter and loop regions took a large proportion of the top 50 variables of multiple cell lines. However, for transcription factors, the features of promoter and enhancer regions took a larger proportion (Blackwood and Kadonaga 1998; Pennacchio *et al.* 2013). Consistent with previous studies (Budden *et al.* 2015; Cheng *et al.* 2011; Cheng and Gerstein 2012; Karlič *et al.* 2010; Ouyang *et al.* 2009), the histone modifications and transcription factors in the promoter regions play an important role in the gene expression.

The changes of diverse signatures in the importance scores between interacting and non-interacting sets

Among the important features in one cell line, some signatures ranked well in both the interacting and non-interacting sets, however, their changes of importance

Fig. 1 Pearson correlation of experimental gene expression levels versus predicted expression levels in RF model. The scatter plots of interacting sets and non-interacting sets were plotted in red and blue, respectively. The horizontal axis is the measured expression value, and vertical axis is the predicted expression values

scores between interacting and non-interacting sets were not obvious, such as H3k36me3, H3k27me3, H3k4me1, H3K79me2 (Fig. 1). To find important signatures that regulate gene expression levels with the regulation of distal enhancers, we made a study on the score changes of top-ranked features between interacting and non-interacting sets in the same cell line. We observed that some features had significant changes in score, such as H3k27me3_p, Mef2e_e, Ikzf1_e, Nfya_e,

Rfx5_e, Irf4_p and eRNA_p in Gm12878 cell line, and H3k36me3_p, H3k79me2_p, H3K4me3_p, H2az_p, nucleosome_p and eRNA_w in K562 cell line (Fig. 3). Among the features with great changes, the transcription factors and histone modifications took a large proportion, and other features (*e.g.*, DNA methylation, eRNAs) also had cell line-specific changes. In Fig. 3, we can see that in the top-ranked 30 features, the transcription factor features respectively accounted for 56.7%, 60%, 43.3% and 36.6%, and the histone modification features, respectively, accounted for 23.3%, 36.67%, 43.3% and 46.7% in Gm12878, H1hesc, HeLa and K562 cell lines. These features with great changes would play a cell line-specific role in the gene expression level by the regulation of distal enhancer. According to previous research (Heintzman *et al.* 2009), many specific histone modifications in enhancers regions would functionally regulate the gene expression in a tissue-specific manner. The eRNA transcription requires the presence of the promoter at the same time, and eRNA levels can correlate with mRNA levels at adjacent genes (Krivega and Dean 2012). In eukaryotes, by binding to enhancer regions of DNA adjacent to regulated genes, many transcription factors can precisely regulate the expression of various genes in the right cell and at the right time. For example, cohesin Rad21 is involved in chromatin looping and plays an important role in maintaining chromosomal loops (Merken-schlager and Odom 2013; Nitzsche *et al.* 2011). P300 was known to bind the enhancer regions and has been applied to predict enhancers (Blow *et al.* 2010; Visel *et al.* 2009a, b). MafK transcriptional activation can be mediated by an enhancer and regulated by different GATA factors in both hematopoietic and cardiac tissues (Katsuoka *et al.* 2000).

However, in the same cell line, there were also many top-ranked features with minor changes in the importance score between interacting and non-interacting sets, such as H3k36me3, H2az, H3k27me3, H3k4me1, H3K79me2 (Fig. 4). In our observation, the histone modifications in loop and promoter regions had a large proportion of the features. These features with minor changes would play a fundamental role in gene expression with the regulation of long-range E-P interactions in a specific cell line. For example, associated with elongating RNA polymerase II, the H3K36me3 is indicative of actively transcribed genes (Li *et al.* 2002). However, as clear markers of gene repression, the H3k27me3 is likely bound by other proteins to exert a repressive function (Cao *et al.* 2002) and H3K9me3 is a well-characterised marker for heterochromatin (Bannister *et al.* 2001; Lachner *et al.* 2001). As a multiprotein complex, Pol II catalyzes the transcription of DNA to synthesize precursors of mRNA and most

Table 2 The proportion of the features of enhancer, promoter and loop regions in the top 50 variables

Cell line	Interacting set			Non-interacting set		
	Enhancer (%)	Promoter (%)	Loop (%)	Enhancer (%)	Promoter (%)	Loop (%)
Gm12878	24	48	28	18	60	22
H1hesc	30	44	26	36	40	24
Hela	28	44	28	36	36	28
K562	12	58	30	28	44	28

Fig. 3 The 30 top-ranked features with large changes in importance scores between the interacting and non-interacting sets in the same cell line. The bars represent the changed D values between the interacting and non-interacting sets. The D values above the abscissa axis indicate the larger values of interacting set, while the values below the abscissa axis indicate the larger values of non-interacting set. The longer the bar, the greater the difference between the interacting and non-interacting sets

1995), SAP30 (Huang *et al.* 2003), FKBP3 (Yang *et al.* 2001) and RYBP (García *et al.* 1999).

Taken together, on the basis of the genomic signatures with basic biological functions, the distal enhancers would cooperate with multiple specific signatures and regulate the gene expression levels in complex regulatory networks.

CONCLUSION

The enhancers act as an important role in cell line-specific gene expression (Visel *et al.* 2009a, b). In the regulation, enhancers can specifically select their long-range target promoter(s) and facilitate expression levels

of related genes. In the cis-regulatory mechanism of gene expression, diverse genomic signatures would affect the enhancers and cooperatively establish a complex regulatory network. Recently, previous researchers have made great effort to study the relationships between the general gene expression levels and associated genomic signatures in promoters (Budden *et al.* 2015; Cheng *et al.* 2011; Cheng and Gerstein 2012; Karlič *et al.* 2010; Mcleay *et al.* 2012; Ouyang *et al.* 2009). However, due to the cell line-specificity of long-range E-P interactions and the limited experimental data, the gene expression levels with the regulation of distal enhancers does not been further studied. In this paper, based on the available experimental data from multiple platforms, we developed computational

Fig. 4 The top-ranked features with minor changes in scores between interacting and non-interacting sets in the same cell line. The bars represent the importance scores of interacting and non-interacting sets. The D values of the above signatures between the interacting and non-interacting sets were less than 10% of the corresponding positive values

methods to study the relationships between diverse genomic signatures and gene expression levels regulated by long-range E-P interactions in four human cell lines. On the basis of the long-range E-P interaction datasets from the 5C experiments (Roy *et al.* 2015), we extracted features from diverse genomic signatures and built three regression models of RF, GBM and SVR. First, through the results of regression models, we found that the predicted expression levels correlated well with the measured expression levels in the interacting sets, in which the best PCC values measured up to 0.8736, 0.9052, 0.9532 and 0.9498 in Gm12878, H1hesc, HeLa and K562 cell lines, respectively. However, the values of the non-interacting set were lower in the same cell line, which indicated that the distal enhancers would cooperate with diverse genomic signatures to facilitate the expression levels of far target genes. Second, according to analysis of the variable importance by RF models in different cell lines, we discovered the key roles of diverse signature features in the gene expression level

with long-range E-P interactions, such as eRNA, H3k36me3, H3k79me2, H3k27me3, H3k4me3, H2az, Pol2, DNase I, DNA methylation and nucleosome occupancy. For histone modifications, features of promoter and loop regions took a large proportion of the top variables of multiple cell lines. However, for transcription factors, the features of promoter and enhancer regions took a large proportion of the top variables. Finally, through the comparison of the interacting and non-interacting sets in the same cell line, we found the important genomic signatures with great changes would play specific roles in the gene expression level of the regulation of distal enhancer, such as H3k27me3_p, Mef2e_e, Ikzf1_e, Nfya_e, Rfx5_e, Irf4_p and eRNA_p in Gm12878 cell lines. In addition, we also found the important features with minor changes would play a fundamental role in gene expression with long-range E-P interactions in a specific cell line, such as H3k36me3_p, H2az_p, H3k27me3_w, H3k4me1, H3K79me2. Taken together, this work studied the roles

of diverse genomic signatures on the gene expression levels with the regulation of long-range E-P interactions.

MATERIALS AND METHODS

Datasets of long-range E-P pairs

In our work, the long-range E-P pair datasets of four human cell lines came from Roy *et al.*'s work (2015), the four cell lines were K562, H1hesc, HeLa and Gm12878. The datasets of the above four cell lines respectively contained 1754, 674, 1530 and 952 E-P pairs. Each cell line respectively contains an interacting set and a non-interacting set, and the two sets have an equal number of samples. For an arbitrary E-P pair in the datasets, the distance between the enhancer and the transcription start site was larger than 2500 bp. Because the enhancers tend to interact with their target promoters depending on their chromatin distances, the non-interacting set and interacting set in each cell line were constructed with the same distribution of chromatin distance.

The extraction of feature parameters

Tissue-specific gene expression with the regulation of long-range E-P interactions would be affected by the diverse genomic signatures. For example, DNA methylation can regulate tissue-specific gene expression, and it is related to multiple human cancers (Ball *et al.* 2009; Song and He 2012). The nucleosome reorganization in regulatory regions of enhancers and promoters

accompany the gene activation by T cell receptor signaling (Schones *et al.* 2008). The eRNAs are short non-coding RNAs (Smith and Shilatifard 2014), and they would enhance or stabilize the long-range E-P interactions (Fan *et al.* 2013; Wang *et al.* 2011). In each cell line, we downloaded the available data of several genomic signatures, including DNase I, FAIRE-seq, transcription factors, histone modifications, DNA methylation, nucleosome position, eRNAs and RNA-seq. The sequencing techniques and the available data sources of corresponding signatures were shown in Table 3. For the sequencing data of above signatures, we processed data and extracted the corresponding features of enhancer, promoter and loop regions by using the same method in our previous works (Feng *et al.* 2018). For the two and more biological replicates, we averaged the feature values for the corresponding regions.

Predictive models for gene expression levels

In this work, we used three algorithms to predict the gene expression levels, including RF, GBM and SVR. The above prediction algorithms were implemented in R packages of 'randomForest', 'gbm' and 'e1071', respectively. For RF model, we set the parameter of n.trees = 500 (number of built trees), and selected the best mtry (number of variables randomly sampled as candidates at each split). For the SVR model, we selected the non-linear radial basis kernel. In the GBM model, we fit the parameters: interaction.depth = 5, n.trees = 5000, shrinkage = 0.01, n.minobsinnode = 10 and used R_{squared} to select the optimal model with the largest value.

Table 3 The sequencing techniques and the available data sources of corresponding signatures

Signatures	Sequencing techniques	Data sources
Histone modifications	Chromatin immunoprecipitation sequencing (ChIP-seq)	ENCODE ChIP-seq Experiment Matrix hg19 (http://genome.ucsc.edu/ENCODE/dataMatrix/encodeChipMatrixHuman.html)
Transcription factors	Chromatin immunoprecipitation sequencing (ChIP-seq)	
DNase I	DNase-seq	ENCODE Experiment Matrix hg19 (https://genome.ucsc.edu/ENCODE/dataMatrix/encodeDataMatrixHuman.html)
FAIRE-seq	FAIRE-seq	
DNA methylation	Reduced representation bisulfite sequencing (RRBS)	
Nucleosome position	Sequencing of micrococcal nuclease sensitive sites (MNase-seq)	ENCODE/Stanford/BYU (http://www.genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeSydhNsome)
Enhancer RNAs	Global run-on and sequencing (GRO-seq)	GEO(https://www.ncbi.nlm.nih.gov/geo/ ; GSM1480326, GSE62046, GSM1480325)
Expression level	RNA-seq	ENCODE Experiment Matrix hg19 (https://genome.ucsc.edu/ENCODE/dataMatrix/encodeDataMatrixHuman.html)

For the evaluation methods of statistical prediction, we used the 10-fold cross-validation test in the same cell line to evaluate the prediction performances. According to the previous works (Liu *et al.* 2016), we computed gene expression levels [$\log(\text{RPKM})$]. In the regression model, we calculated the standard deviation of regression model (RMSE) which was used to quantify the amount of dispersion of a set of data values. Then we calculated the R^2 (coefficient of determination) which indicates the proportion of variation of the gene expression levels that has been explained by the model (Cheng and Gerstein 2012; Everitt 2002). Finally, we computed the PCC (Stigler 1989) between the predicted expression values and experimental measured values in the testing data.

Abbreviations

5C	Carbon copy chromosome capture conformation
E-P	Enhancer-promoter
EPI	Enhancer-promoter interaction
GBM	Gradient-Boosting Machine
RF	Random Forest
RPKM	The reads per kilobase of exon model per million mapped reads

Acknowledgements The authors thank Roy *et al.* for sharing the datasets. This work was supported by the Grants from the National Natural Science Foundation of China (31870838, 31460234 and 11747315).

Compliance with Ethical Standards

Conflict of interest All they authors declares that they have no conflict of interest.

Human and animal rights and informed consent This article does not contain any studies with human or animal subjects performed by any of the authors.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Ball MP, Li JB, Gao Y, Lee JH, LeProust EM, Park IH, Xie B, Daley GQ, Church GM (2009) Targeted and genome-scale strategies

- reveal gene-body methylation signatures in human cells. *Nat Biotechnol* 27:361–368
- Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, Allshire RC, Kouzarides T (2001) Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* 410:120–124
- Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D (2004) Ultraconserved elements in the human genome. *Science* 304:1321–1325
- Blackwood EM, Kadonaga JT (1998) Going the distance: a current view of enhancer action. *Science* 281:60–63
- Blow MJ, McCulley DJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Bristow J, Ren B, Black BL, Rubin EM, Visel A, Pennacchio LA (2010) ChIP-Seq identification of weakly conserved heart enhancers. *Nat Genet* 42:806–810
- Budden DM, Hurley DG, Crampin EJ (2015) Predictive modelling of gene expression from transcriptional regulatory elements. *Brief Bioinform* 16:616–628
- Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, Tempst P, Jones RS, Zhang Y (2002) Role of histone H3 lysine 27 methylation in Polycomb-group silencing. *Science* 298:1039–1043
- Cheng C, Gerstein M (2012) Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells. *Nucleic Acids Res* 40:553–568
- Cheng C, Yan KK, Yip KY, Rozowsky J, Alexander R, Shou C, Gerstein M (2011) A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets. *Genome Biol* 12:R15
- Corradin O, Saiakhova A, Akhtar-Zaidi B, Myeroff L, Willis J, Cowper-Sal R, Lupien M, Markowitz S, Scacheri PC (2013) Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res* 24:1–13
- de Laat W, Duboule D (2013) Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* 502:499–506
- Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. *Science* 295:1306–1311
- Everitt BS (2002) Cambridge dictionary of statistics. *J Appl Ecol* 36:842–843
- Fan L, Orom UA, Cesaroni M, Beringer M, Taatjes DJ, Blobel GA, Shiekhattar R (2013) Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* 494:497–501
- Feng ZX, Li QZ (2017) Recognition of long-range enhancer-promoter interactions by adding genomic signatures of segmented regulatory regions. *Genomics* 109:341–352
- Feng ZX, Li QZ, Meng JJ (2018) Recognition of the long range enhancer-promoter interactions by further adding DNA structure properties and transcription factor binding motifs in human cell lines. *J Theor Biol* 445:136–150
- García E, Marcos-Gutiérrez C, del Mar Lorente M, Moreno JC, Vidal M (1999) RYBP, a new repressor protein that interacts with components of the mammalian Polycomb complex, and with the transcription factor YY1. *EMBO J* 18:3404–3418
- He B, Chen C, Teng L, Tan K (2014) Global view of enhancer-promoter interactome in human cells. *Proc Natl Acad Sci USA* 111:2191–2199
- Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B (2009) Histone

- modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* 459:108–112
- Huang NE, Lin CH, Lin YS, Yu WC (2003) Modulation of YY1 activity by SAP30. *Biochem Biophys Res Commun* 306:267–275
- Hwang YC, Lin CF, Valladares O, Malamon J, Kuksa PP, Zheng Q, Gregory BD, Wang LS (2015) HIPPIEa highthroughput identification pipeline for promoter interacting enhancer elements. *Bioinformatics* 31:1290–1292
- Karlič R, Chung HR, Lasserre J, Vlahovicek K, Vingron M (2010) Histone modification levels are predictive for gene expression. *Proc Natl Acad Sci USA* 107:2926–2931
- Katsuoka F, Motohashi H, Onodera K, Suwabe N, Engel JD, Yamamoto M (2000) One enhancer mediates mafK transcriptional activation in both hematopoietic and cardiac muscle cells. *EMBO J* 19:2980–2991
- Kornberg RD (1999) Eukaryotic transcriptional control. *Trends Cell Biol* 9:M46
- Krivega I, Dean A (2012) Enhancer and promoter interactions—long distance calls. *Curr Opin Genet Dev* 22:79–85
- Lachner M, O'Carroll D, Rea S, Mechtler K, Jenuwein T (2001) Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* 410:116–120
- Lee JS, Galvin KM, See RH, Eckner R, Livingston D, Moran E, Shi Y (1995) Relief of YY1 transcriptional repression by adenovirus E1A is mediated by E1A-associated protein p300. *Gene Dev* 9:1188–1198
- Li M, Baumeister P, Roy B, Phan T, Foti D, Luo S, Lee AS (2000) ATF6 as a transcription activator of the endoplasmic reticulum stress element: thapsigargin stress-induced changes and synergistic interactions with NF-Y and YY1. *Mol Cell Biol* 20:5096–5106
- Li J, Moazed D, Gygi SP (2002) Association of the histone methyltransferase Set2 with RNA polymerase II plays a role in transcription elongation. *J Biol Chem* 277:49383–49388
- Liu L, Zhao WL, Zhou XB (2016) Modeling co-occupancy of transcription factors using chromatin features. *Nucleic Acids Res* 5:e49
- McLeay RC, Lesluyes T, Cuellar Partida G, Bailey TL (2012) Genome-wide in silico prediction of gene expression. *Bioinformatics* 28:2789–2796
- Merkenschlager M, Odom DT (2013) CTCF and cohesin linking gene regulatory elements with their targets. *Cell* 152:1285–1297
- Miele A, Dekker J (2008) Long-range chromosomal interactions and gene regulation. *Mol Biosyst* 4:1046–1057
- Nitzsche A, Paszkowski-Rogacz M, Matarese F, Janssen-Megens EM, Hubner NC, Schulz H (2011) RAD21 cooperates with pluripotency transcription factors in the maintenance of embryonic stem cell identity. *PLoS ONE* 6:e19470
- Ouyang Z, Zhou Q, Wong WH (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc Natl Acad Sci USA* 106:21521–21526
- Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G (2013) Enhancers: five essential questions. *Nat Rev Genet* 14:288–295
- Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159:1665–1680
- Roy S, Siahpirani AF, Chasman D, Knaack S, Ay F, Stewart R, Wilson M, Sridharan R (2015) A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res* 43:8694–8712
- Sanyal A, Lajoie BR, Jain G, Dekker J (2012) The long-range interaction landscape of gene promoters. *Nature* 489:109–113
- Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, Wang Z, Wei G, Zhao K (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132:887–898
- Shrivastava A, Saleque S, Kalpana GV, Artandi S, Goff SP, Calame K (1993) Inhibition of transcriptional regulator Yin-Yang-1 by association with c-Myc. *Science* 262:1889–1892
- Sims RJ, Mandal SS, Reinberg D (2004) Recent highlights of RNA-polymerase-II-mediated transcription. *Curr Opin Cell Biol* 16:263–271
- Smith E, Shilatfard A (2014) Enhancer biology and enhanceropathies. *Nat Struct Mol Biol* 21:210–219
- Song CX, He C (2012) Balance of DNA methylation and demethylation in cancer development. *Genome Biol* 13:173
- Stigler MS (1989) Francis Galton's account of the invention of correlation. *Stat Sci* 4:73–79
- The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 447:799–816
- van de Werken HJ, Landan G, Holwerda SJ, Hoichman M, Klous P, Chachik R, Splinter E, Valdes-Quezada C, Oz Y, Bouwman BA, Verstegen MJ, de Wit E, Tanay A, de Laat W (2012) Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat Methods* 9:969–972
- Visel A, Rubin EM, Pennacchio LA (2009a) Genomic views of distant-acting enhancers. *Nature* 461:199–205
- Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F, Afzal V, Ren B, Rubin EM, Pennacchio LA (2009b) ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* 457:854–858
- Wang KC, Yang YW, Liu B, Sanyal A, Corces-Zimmerman R, Chen Y, Lajoie BR, Protacio A, Flynn RA, Gupta RA, Wysocka J, Lei M, Dekker J, Helms JA, Chang HY (2011) A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 472:120–124
- Whalen S, Truty RM, Pollard KS (2016) Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nat Genet* 48:488–496
- Williamson I, Hill RE, Bickmore WA (2011) Enhancers from developmental genetics to the genetics of common human disease. *Dev Cell* 21:17–19
- Williamson I, Berlivet S, Eskeland R, Boyle S, Illingworth RS, Paquette D, Dostie J, Bickmore WA (2014) Spatial genome organization contrasting views from chromosome conformation capture and fluorescence in situ hybridization. *Genes Dev* 28:2778–2791
- Yang WM, Yao YL, Seto E (2001) The FK506-binding protein 25 functionally associates with histone deacetylases and with transcription factor YY1. *EMBO J* 20:4814–4825