

## SOFTWARE ARTICLE

# EpiFIT: functional interpretation of transcription factors based on combination of sequence and epigenetic information

Shaoming Song, Hongfei Cui, Shengquan Chen, Qiao Liu, Rui Jiang\*

MOE Key Laboratory of Bioinformatics; Beijing National Research Center for Information Science and Technology; Department of Automation, Tsinghua University, Beijing 100084, China

\* Correspondence: ruijiang@tsinghua.edu.cn

Received April 3, 2019; Revised April 24, 2019; Accepted April 25, 2019

**Background:** Transcription factor is one of the most important regulators in the transcriptional process. Nevertheless, the functional interpretation of transcription factors is still a main challenge due to the poor performance of methods relating to regulatory regions to genes. Epigenetic information, such as chromatin accessibility, contains genome-wide knowledge about transcription regulation and thus may shed light on the functional interpretation of transcription factors.

**Methods:** We propose EpiFIT (Epigenetic based Functional Interpretation of Transcription factors), a tool to infer functions of transcription factors from ChIP-seq data. Briefly, we adopt a variable distance rule to establish associations between regulatory regions and nearby genes. The associations are then filtered to ensure that the remaining regions and associated genes are co-open. Finally, GO enrichment is applied to all related genes and a ranking list of GO terms is provided as functional interpretation.

**Results:** We first examined the chromatin openness correlation between regulatory regions and associated genes. The correlation can help EpiFIT purify regulatory region–gene associations. By evaluating EpiFIT on a set of real data, we demonstrated that EpiFIT outperforms other existing methods for precisely interpreting transcription factor functions. We further verify the efficiency of openness in interpretation and the ability of EpiFIT to build distal region–gene associations.

**Conclusion:** EpiFIT is a powerful tool for interpreting the transcription factor functions. We believe EpiFIT will facilitate the functional interpretation of other regulatory elements, and thus open a new door to understanding the regulatory mechanism.

**Availability:** The application is freely accessible at website: [bioinfo.au.tsinghua.edu.cn/openness/EpiFIT/](http://bioinfo.au.tsinghua.edu.cn/openness/EpiFIT/).

**Keywords:** transcription factor; functional interpretation; epigenetic information

**Author summary:** Transcription factors (TF) regulate the expression level of targeted genes and further effect biological functions. Hence, we developed EpiFIT to infer functions of TF using sequence and epigenetic data. Through a series of examination experiments, we verified that EpiFIT can precisely interpret TF functions and build distal TF binding sites – regulated genes associations with the help of epigenetic information. In a word, EpiFIT is a powerful tool for annotating the TF functions. We believe EpiFIT will facilitate the functional interpretation of other regulatory elements, and thus open a new door to understanding the regulatory mechanism.

## INTRODUCTION

Transcription factor plays an important role in gene

expression due to its involvement in eukaryotes' transcription process. By binding specific genomic regions and co-acting with other proteins (which are

also named as co-factors), transcription factors are able to regulate gene expression level. The invention of chromatin immunoprecipitation with massively parallel DNA sequencing, named ChIP-seq, opens a new era of transcription factors regulation mechanism analysis [1–3]. Thus far, many efforts have focused on the functional interpretation of the genomic locations of transcription factors binding events. However, transcription factor binding sites are mostly the genomic regions in non-coding areas, which results in a lack of quantitative measurements for analysis [4]. Therefore, many methods have been developed to collect genes picked by nearby genes for functional enrichment. For example, *sole-search* firstly connects transcription factor binding sites and nearest genes, and uses GO enrichment to generate functional annotations for target transcription factors [5]. DAVID (Database for Annotation, Visualization and Integrated Discovery) furtherly provides more annotation databases and gene sets cluster function [6,7]. Besides, GREAT (Genomic Regions Enrichment of Annotations Tool) introduces a variable distance rule to associate transcription factor binding sites and genes, with a binomial test coming to rank all annotations [8]. However, all methods mentioned above have a fundamental drawback: they use only distance criterion to build regulatory region – gene associations, which may result in numerous false positive functional interpretations.

Proposed studies have demonstrated that the frequency of transcription factors binding on specific transcription factor binding sites can effectively predict the expression level of nearby genes [9,10]. Besides, researchers also reveal high correlation between transcription factor binding events and many kinds of epigenetic characteristics. With this understanding, introducing epigenetic data into the functional interpretation may help reduce false positive regulatory region – gene associations, and thus achieve a more precise annotation performance. Among all epigenetic sequencing methods, DNase-seq (DNase I hypersensitive sites sequencing) can effectively capture genome-wide chromatin accessible regions. Studies have discovered the relationship between DNase-seq and transcription factor binding sites [12,13], which provides the theoretical basis of bringing epigenetic information into transcription factor related researches. In addition, the comprehensive web server, named *OPENANNO* [14,15], provides a service of generating a quantitative measure of chromatin openness based on ENCODE experiment data [16,17].

With the above understanding, we introduce EpiFIT (Epigenetic based Functional Interpretation of Transcription factors), a web-based tool integrating sequence information with epigenetic data to generate functional interpretation for transcription factors in specific cell types. In the following sections, we first introduce the

mainframe of EpiFIT, and then verify that the chromatin openness can effectively improve the interpretation performance of transcription factor functions. Furthermore, we apply EpiFIT to a real gold-standard dataset and compare its performance with GREAT [8], one of the most widely used tools, to demonstrate that EpiFIT can effectively and accurately interpret functions of transcription factors. We further test the performance of EpiFIT with randomly permuted chromatin openness and prove the efficiency of openness for the interpretation. At last, by comparing the statistical details and performance of EpiFIT with methods using different region-gene association building rules, we demonstrate the ability of EpiFIT to reveal real associations between regulatory regions and targeted genes.

## METHODS AND IMPLEMENTATION

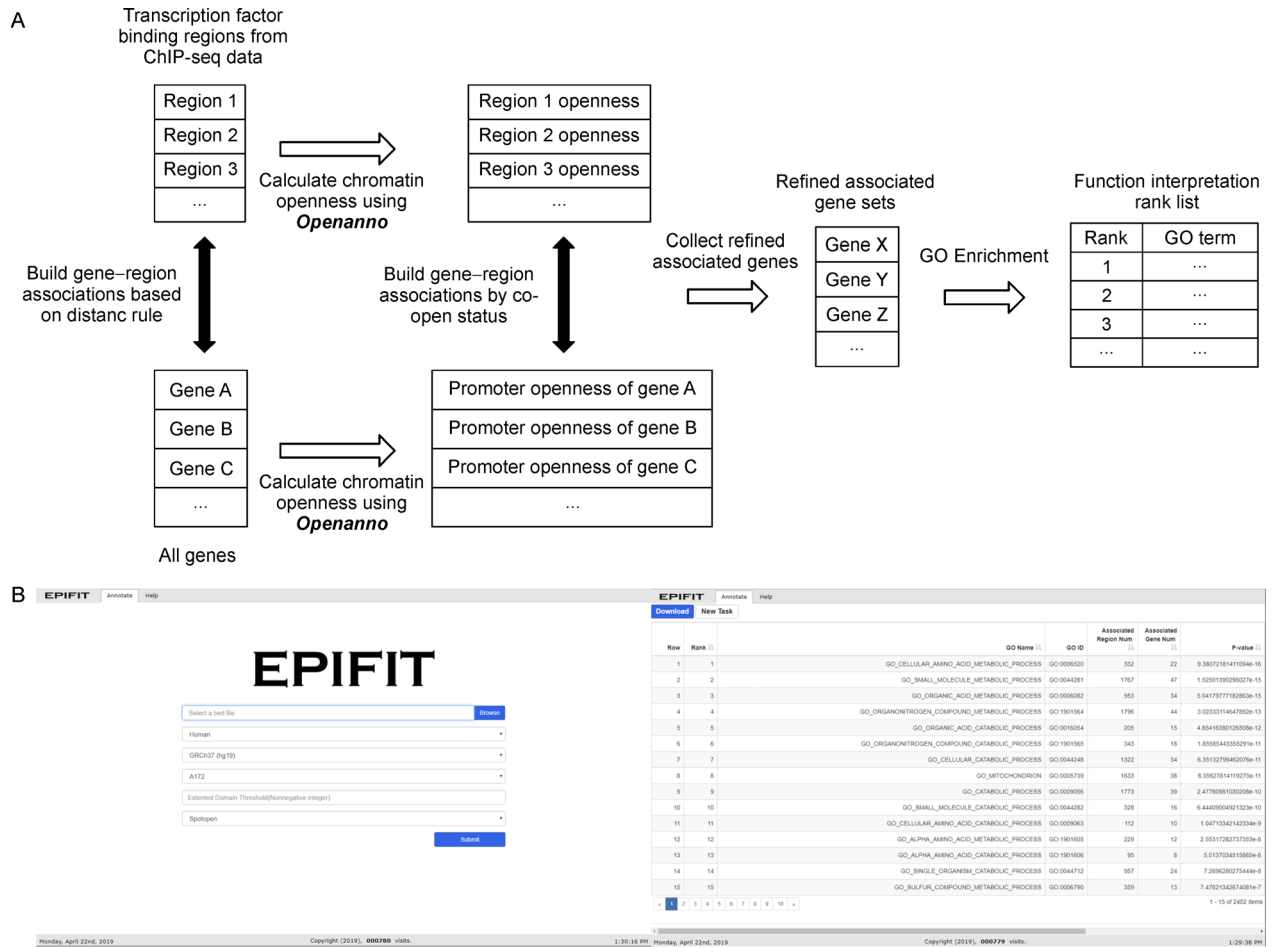
EpiFIT, as an integrated software, has a backend that assembles diverse algorithms developed using the programming language of python and R, and a shell profile for integral execution. The results are shown in web pages developed using jQuery, a JavaScript library for data visualization. The main framework of EpiFIT is illustrated in Figure 1A. In EpiFIT, we take the information of regulatory regions and other optional parameters as input. EpiFIT first associates regulatory regions with nearby genes using a variable distance rule. After this, EpiFIT calculates chromatin openness values for all regulatory regions and related genes using *OPENANNO* [14,15] with DNase-seq experiments of the same cell type (For regulatory regions, we calculate the value of whole area. For genes, we take areas from 2k base-pair upstream to 1k base-pair downstream of the transcription start site for calculating). For both sides of the regulatory region–gene association, if one of the two sides has a chromatin openness value smaller than 5, we denote this association as “not co-open”. EpiFIT then refines all associations by eliminating “not co-open” ones. Finally, based on gene-term association data from Gene Ontology Consortium [18,19], we implement GO enrichment using the hypergeometric test. The results are presented as a table on the web page and can be downloaded for later analysis.

The web interfaces of EpiFIT is shown in Figure 1B. We will describe more detailed information about EpiFIT in the following sections.

### Input of EpiFIT

In EpiFIT, one file and three optional parameters are required as input.

- (1) File of transcription factor binding sites (formatted



**Figure 1. Workflow and online illustration of EpiFIT.** (A) Main workflow schematic of EpiFIT. (B) Online illustration of EpiFIT. With inputting a ChIP-seq file and optional parameters set, EpiFIT generates results of functional interpretation in a table format and provides a service of downloading for later analysis.

as a bed file)

This file is generated from ChIP-seq experiments and contains all regulatory region information of targeted transcription factor in a specific cell type. Each row records a binding site satisfying the standard format of bed files.

(2) Other optional parameters

EpiFIT requires several optional parameters. A more detailed introduction is shown in Table 1.

## Chromatin openness calculation by *OPENANNO* and “co-open” status generation

*OPENANNO* is a web server developed by our group to annotate the epigenetic characteristics of genomic regions [14,15]. In this paper, we focus on chromatin openness, which can be quantified using *OPENANNO* with ENCODE DNase-seq data. Moreover, the effectiveness of chromatin openness has been proved in previous

**Table 1 The optional parameters of EpiFIT**

Parameter	Default	Effect
Openness calculation algorithm	Spotopen	<i>OPENANNO</i> offers 4 different openness quantification algorithms, and users can choose which one they want. More details are introduced below.
Cell type	K562	Cell type of transcription factors binding experiments. All supported cell types are listed in readme.txt
Region–gene association max distance	1000	Determine the max length (kilo base-pair) of each inferred gene regulatory domain.

researches [20,21].

The openness from four perspectives calculated using *OPENANNO* is offered in EpiFIT for chromatin openness calculation.

(1) Foreground: for each genomic region, this algorithm provides the count of raw reads which overlap with this region.

(2) Readopen: this algorithm provides a fold change of raw read count versus background read count (average count number from upstream 500 kbp to downstream 500 kbp).

(3) Peakopen: this algorithm uses narrowpeak calling methods to generate a fold change of foreground peak count versus background peak count.

(4) Spotopen: this algorithm uses broadpeak calling methods to generate a fold change of foreground peak count versus background peak count.

Because chromatin openness is highly cell type-specific, for specific ChIP-seq data, *OPENANNO* generates chromatin openness from DNase-seq experiments that have the same cell type. If there are multiple matched experiments in ENCODE DNase-seq database, multiple chromatin openness results will be generated for one region and we take the average of them as the final chromatin openness value.

Having acquired chromatin openness value calculated by *OPENANNO*, we are able to measure the “co-open” status for genomic regions-gene association using Equation (1).

$$S_{\text{co-open}} = \begin{cases} 0, & \text{open}_g < 5 \text{ or } \text{open}_r < 5, \\ 1, & \text{else} \end{cases} \quad (1)$$

where  $S_{\text{co-open}}$  denotes the co-open status,  $\text{open}_g$  the chromatin openness of gene, and  $\text{open}_r$  the chromatin openness of genomic regions.

### Genomic region-gene association

EpiFIT first assigns each gene a basic domain that extends 5k base-pair upstream and 1k base-pair downstream from its transcription site. Then, EpiFIT gives each basis domain an extension both upstream and downstream to the basal domain of the nearest gene. If the extended length is more than the pre-set max length, we only denote genomic regions with a set threshold as regulatory domains of the targeted gene. In this rule, most transcription factor binding sites can be assigned with genes to generate genomic region-gene associations.

### GO enrichment

EpiFIT uses a hypergeometric test to implement GO enrichment. This test identifies all genes who have at least one regulatory region associated and applies enrichment

with respect to the reference gene set using a hypergeometric distribution.

$$P\text{-value} = \sum_{i=k_t}^{\min(n, K_t)} \frac{\binom{K_t}{i} \binom{N-K_t}{n-i}}{\binom{N}{n}} \quad (2)$$

As shown in Equation (2), EpiFIT calculates the  $P$ -value of an observed GO term with the following four parameters:

1.  $N$ : the total number of genes in the reference genome.
2.  $K_t$ : the number of genes associated with term  $t$  based on GO database records.
3.  $n$ : the number of genes that are associated with regulatory regions.
4.  $k_t$ : the number of genes that are both associated with regulatory regions and associated with term  $t$ .

Notably, every gene is counted only once regardless of the number of regulatory regions it associates, which means that we take more concentration on the fraction of all selected genomic regions. GO terms with higher ranks tend to have high coverage of associated genes among all genes in the reference genome.

## VALIDATION AND EVALUATION

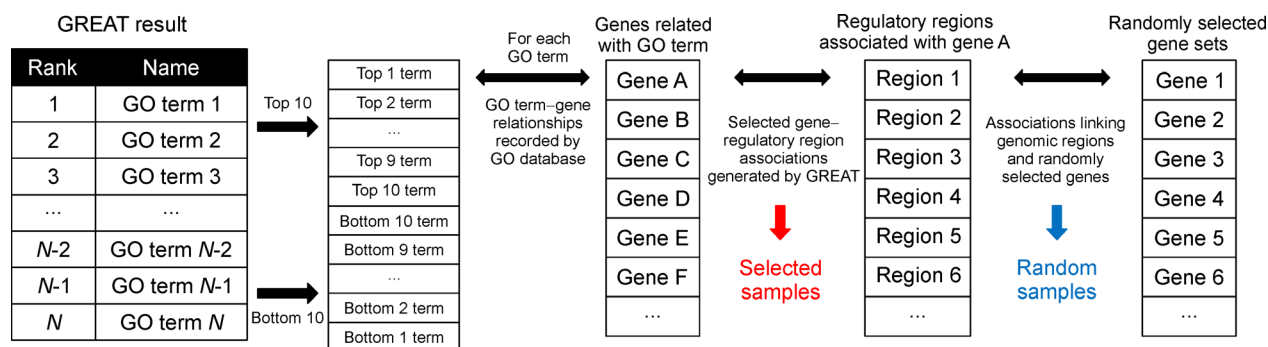
### Data and material

To evaluate the performance of transcription factor functional interpretation tools, we collected all *Homo sapiens* hg19 ChIP-seq broad peaks from the ENCODE Project [16,17]. Each experiment contains all binding sites of a targeted transcription factor in a specific cell type. Among all the 410 collected experiments, we filtered out the experiments that take non-specific transcription factor (e.g., CTCF) as a target to avoid data bias. Finally, 185 experiments are left for later use. This dataset is denoted as “ENCODE ChIP-seq data”.

### Examination of the chromatin openness correlation between regulatory regions and associated genes

To verify the hypothesis that a regulatory region tends to be co-open with its associated genes, we calculate the openness of all regulatory regions and their associated genes. We define “top” and “bottom” GO term sets according to the statistical significance of GO enrichment, which can be achieved using GREAT [8], a widely used method for the annotation of the existing regulatory region. The schematic is shown in Figure 2.

Briefly, for each GO term, we generate selected and random sample associations and calculate the median product of the chromatin openness of regulatory region and gene, and the proportion of “co-open” (which means



**Figure 2.** Flow chart of the evaluation sample generation. These samples are used for evaluation of chromatin openness correlation between transcription factor binding sites and associated genes.

both regulatory region and gene have chromatin openness higher than 5) as evaluation metrics for all samples. We compare these evaluation metrics between selected and random sample associations and examine whether they have significant differences to chromatin openness.

### Evaluation of functional interpretation results with an authorized dataset

To evaluate the performance of functional interpretation, we first build a benchmark test dataset. We downloaded the *Homo sapiens* GO term-protein association annotation dataset from EBI Gene Ontology Annotation Database [22]. The dataset consists of 423,655 association annotations, including 18,085 GO terms and 19,712 proteins. Transcription factor related entries were selected by “PubMed” or “Reactome” in “Source” property, indicating that these entries can be traced to published literature or Reactome database as authorization [23]. The transcription factor-function associations in the selected records can be considered as the “gold-standard” and formed as a “gold-standard dataset”. In total, 47,026 gold-standard GO term-protein associations are recorded in this dataset. We then filtered “gold-standard” associations that match input ChIP-seq target transcription factors. By applying methods for functional interpretation on this filtered gold-standard dataset, a function list sorted by the enrichment confidence from high to low can be achieved. By examining the rank of gold-standard associations in the predicted list, we can compare the performance of different methods.

### Examination of building associations between distal regulatory regions and targeted genes

EpiFIT builds the regulatory region-targeted gene association by not only chromatin openness but also

variable distance threshold criterion. Therefore, we are also interested in the ability of EpiFIT to reveal real associations among distal regulatory regions and targeted genes. Many published researches have focused on the targeted gene association of ChIP-seq. For example, TFAS [41] provides two criteria (binary and continuous) to integrate distal regions with regulated genes. However, it is still limited to reveal real distal regulatory regions. To evaluate the ability of EpiFIT to build associations between distal regulatory regions and targeted genes, we compared the performance of EpiFIT with two other methods, TFAS and DAVID. For TFAS, we selected binary criterion that associates regulatory regions with the nearest genes. For DAVID, we only connected genes with proximal regulatory regions within 2 kb from the transcription start site. Specifically, to demonstrate the contribution of chromatin openness, we also implemented EpiFIT-NR, which means EpiFIT without the refinement of openness. Therefore four results are generated for the final comparison.

In comparison, we first evaluated the numbers of regions associated of these methods to demonstrate that EpiFIT can associate distal regulatory regions with targeted genes. Using the gold-standard dataset described above, we then compared the performance of each method to show the regulatory region-gene associations built by EpiFIT are more accurate than that built by other methods.

### The method used for comparison

We used GREAT, one of the most commonly used tools, for performance comparison. Because GREAT does not have open-access codes and cannot provide any intermediate result, we implemented the method of GREAT, and the final results are agreed with that of the online version of GREAT[8].

## RESULTS

### Regulatory region–gene associations with higher interpretation ranks tend to be more co-open

To evaluate the chromatin openness of transcription factor binding sites (TFBS) and their associated genes (see section “Methods and Implementation”), we used 185 ENCODE ChIP-seq datasets, each of the datasets is a genomic region set recording transcription factor binding sites. For each dataset, we used GREAT to provide transcription factor functional interpretations, together with the predicted significance for each listed GO term. We selected the top 10 GO terms with the highest significance and gathered the genes that related to these GO terms, together with their corresponding regulatory regions. As a contrast, an analysis was also applied to the bottom 10 GO terms using a set of randomly selected genes that are generated.

The result shows that for most experiments, the top-ranked GO terms have higher median products of openness of both sides in the selected regulatory region–gene associations, while the low-ranked GO terms offer an opposite conclusion that random regulatory region–gene associations are more co-active in consideration of chromatin openness (Table 2). Take one of the experiments, ENCSR000BGE, as an example. This experiment was deployed on GM12878 cells and was targeted to Serum Response Factor (SRF), an important transcription factor during the development of the embryo. As illustrated in Figure 3, among the top-ranked GO terms, the median products and co-open proportions of selected samples are significantly larger than that of the random samples, which means the regulatory region–gene associations that are linked with higher rank GO terms tend to have more active chromatin status.

As a conclusion, chromatin openness has a strong correlation with transcription factor binding events, which indicates that we can apply chromatin openness value into the refinement of the regulatory region–gene associations.

### EpiFIT provides precise interpretations of the transcription factor functional interpretations

We evaluated the functional interpretation results using a

gold-standard dataset as the benchmark and 185 ENCODE ChIP-seq experiments as input on both EpiFIT and GREAT (see section “Methods and Implementation”).

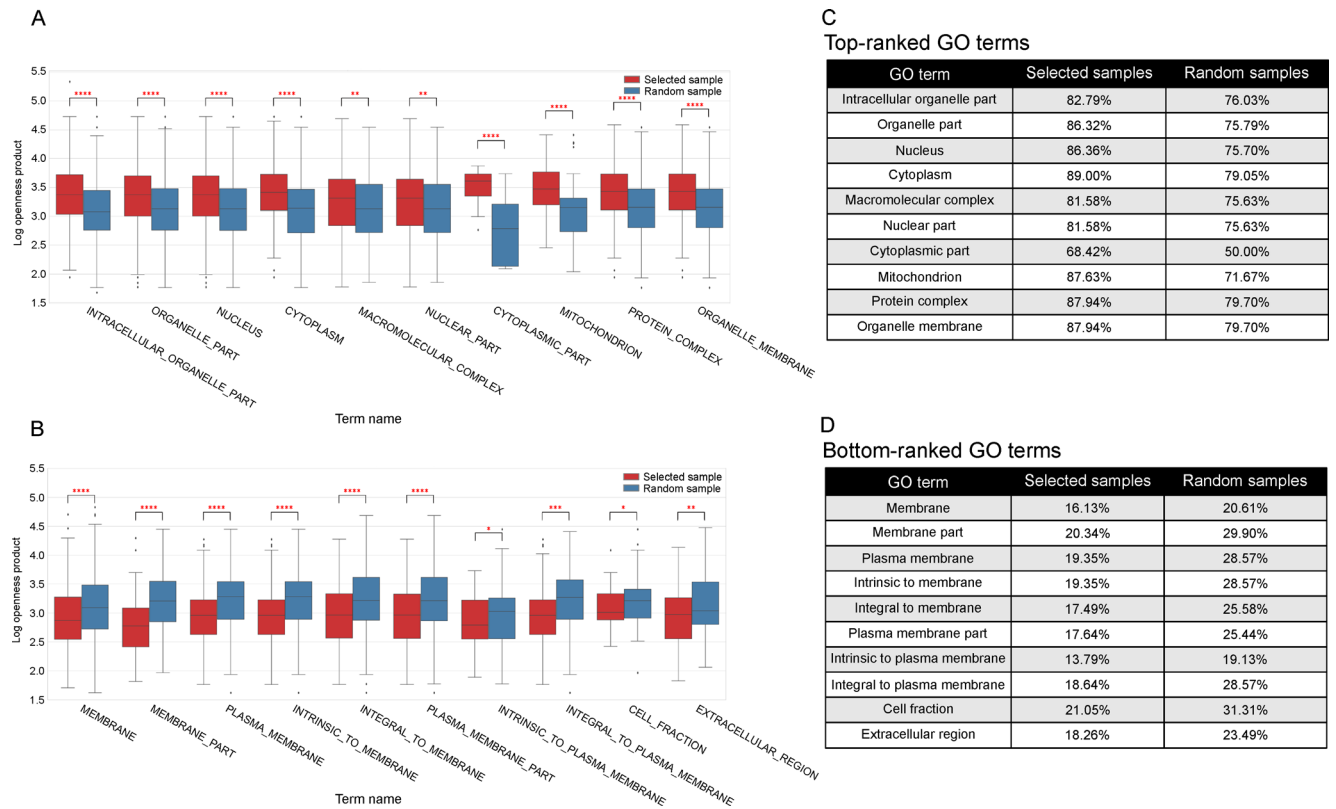
Among all the 185 ENCODE ChIP-seq experiments, 162 kinds of transcription factors are involved. In the gold-standard dataset, 409 out of 18,085 gold-standard GO terms are related to the 162 transcription factors and present in both EpiFIT and GREAT results.

We compared the ranks of all the 409 gold-standard GO terms in the ranking list using the two methods. The results are shown in Figure 4. Among all the gold-standard GO terms, 54% of them (221/409) have higher ranks in EpiFIT than those in GREAT [8], which indicates that EpiFIT is more likely to provide more precise functional annotations comparing with GREAT. For the rest 188 GO terms with their rank not improved in EpiFIT, most of them (123/409) are the same in the two methods. These GO terms are almost top 1 in both results, leaving them no space for improvement. Only 15.9% (65/409) of the GO terms are rank-down. The improved performance of EpiFIT is statistically significant (binomial test,  $P = 4.74e-21$ ).

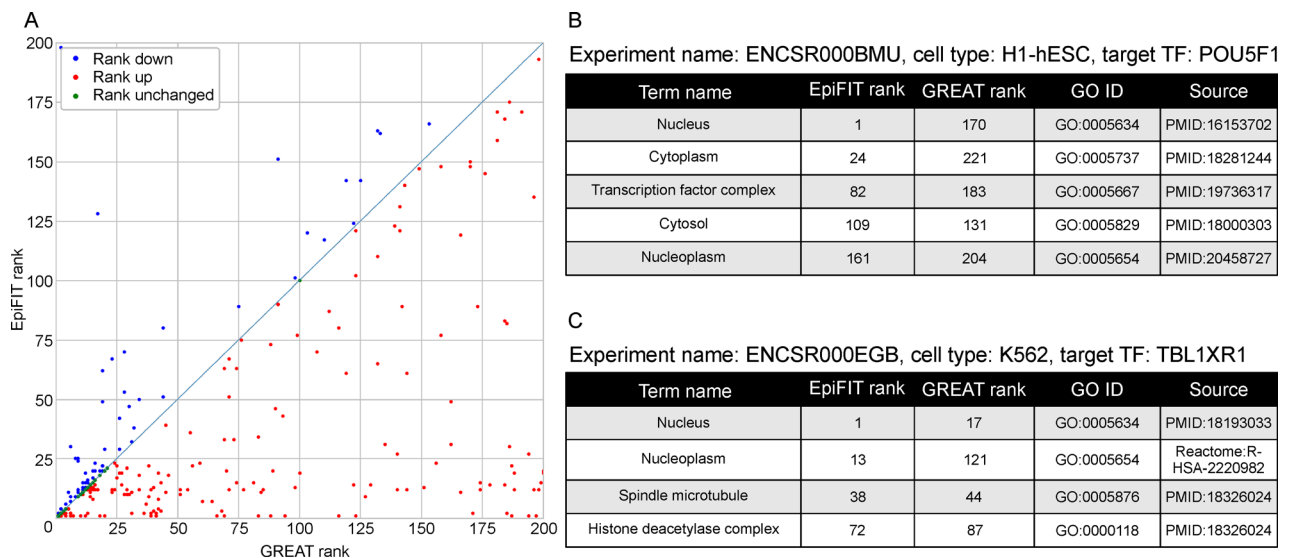
To explain the advantage of functional interpretation of EpiFIT for understanding biological mechanisms, we selected two ENCODE ChIP-seq experiments as examples. One is ENCSR000BMU (Figure 4B), an experiment deployed on *Homo sapiens* H1-hESC targeting to transcription factor POU5F1 (POU Domain, Class 5, Transcription Factor 1), a transcriptional repressor protein that is very active during stem cell development. There are five GO terms [24–34] from the gold-standard dataset that relate to this transcription factor and exist in the result of both EpiFIT and GREAT. All of them improved their ranks in our EpiFIT ranking list. For example, GO term “Nucleus”, which serves as a basic function of all transcription factors like POU5F1 [24–28], is ranked as the 170th by GREAT but ranked as the 1st by EpiFIT. GO term “Transcription Factor Complex” that reflects type characteristic of POU5F1 [31] has reached 82nd in the rank list of EpiFIT, moved up 101 places compared with the rank of 183<sup>rd</sup> in the results of GREAT. Another example is the experiment of ENCSR000EGB (Figure 4C), which studies the binding sites of TBL1XR1 (Transducin Beta Like 1 X-Linked Receptor 1, a

**Table 2** Statistical results of the chromatin openness of regulatory region–gene associations

	Median product				Co-open proportion			
	Top 10 GO terms		Bottom 10 GO terms		Top 10 GO terms		Bottom 10 GO terms	
	Amount	Proportion(%)	Amount	Proportion(%)	Amount	Proportion(%)	Amount	Proportion(%)
Selected samples	1397	75.5	114	6.2	1669	90.2	85	4.6
Random samples	453	24.5	1736	93.8	181	9.8	1765	95.4
Sum	1850	100	1850	100	1850	100	1850	100



**Figure 3.** Comparison of selected and random samples. (A) Box plots representing log openness product of selected and random association samples from 10 top-ranked GO terms. Statistical significance marks represent the rank-sum test  $P$ -value (\*:  $P$ -value < 0.05; \*\*:  $P$ -value < 0.01; \*\*\*:  $P$ -value < 0.001; \*\*\*\*:  $P$ -value < 0.0001). (B) Box plots representing log openness product of selected and random association samples from 10 bottom-ranked GO terms. (C) The proportions of the co-open status of both kinds of samples in the 10 top-ranked GO terms. (D) The proportions of the co-open status of both kinds of samples in the 10 bottom-ranked GO terms.



**Figure 4.** Performance of EpiFIT implemented on an authorized dataset. (A) The rank changes of gold-standard GO terms in different interpretation tools. Red dots are terms with higher rank in EpiFIT. Green dots are terms with a lower rank in EpiFIT. Blue dots are rank-unchanged terms. (B) Rank change details of experiment ENCSR000BMU. (C) Rank change details of experiment ENCSR000EGB.

transcription factor that is required for transcriptional activation by a variety of transcription factors) on K562 cell line. GO terms that proved either by PubMed literature such as “Histone Deacetylase Complex” [29,34] or by Reactome database such as “Nucleoplasm” [35–40] all received increased rank in the results of EpiFIT.

As a conclusion, EpiFIT can provide a more precise interpretation of the transcription factor functions compared with other methods.

### EpiFIT with randomly permuted chromatin openness results in unordered performance

To study the contribution of chromatin openness to the superior performance of EpiFIT, we evaluate the functional interpretation results on the same gold-standard dataset using random openness refinement. In detail, for every experiment among all 185 ENCODE ChIP-seq results, we first shuffle the chromatin openness of genes and regulatory regions to randomize “co-open” association refinement while assuring the distribution of all openness values. Then we compare the final ranks of all 409 gold-standard GO terms with that of GREAT.

We conducted a series of experiments, and the results are shown in Table 3. With the randomly permuted chromatin openness, the regulatory region–gene associations are purified based on stochastic “co-open” status, leading to total unordered and statistical insignificant performances. Notably, common sense says if the refinement is stochastic, the number of these two kinds should be close. However, among all the three experiments, the number of gold-standard GO terms with improved ranks is slightly less than that with decreased ranks. This is probably because most of the rank-unchanged GO terms are enriched as the top 1 in GREAT interpretation results, having no space for improving its rank. Meanwhile, these GO terms possess a tremendous advantage in statistical significance that even random refinement is not capable to reduce their ranks (for example, many ChIP-seq experiments obtain top 1 gold-standard GO term “Nucleus” in both GREAT and EpiFIT surpassing the 2nd GO term by more than five orders of magnitude in terms of statistical significance).

Overall, EpiFIT with randomly permuted chromatin

openness generates unordered transcription factor functional interpretation, demonstrating the importance of chromatin openness in refining the regulatory region–gene associations to discard false positive interpretation results.

### EpiFIT helps build associations between distal regulatory regions and targeted genes

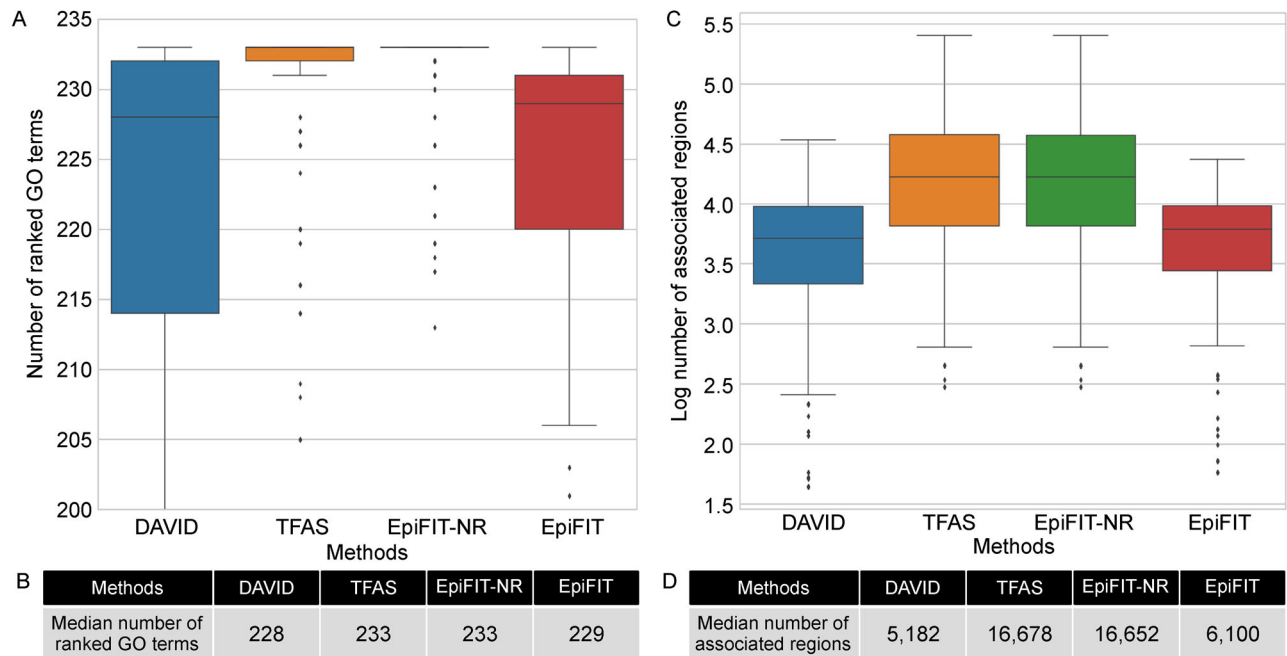
We examined the ability of EpiFIT to reveal real associations between distal regulatory regions and targeted genes by comparing the association details and performance with two other methods. We first compared the number of associated regions of all ENCODE ChIP-seq experiments among all four methods. The results are shown in Figure 5. As for numbers of ranked GO terms, these four methods show similar results, TFAS and EpiFIT-NR (EpiFIT with No chromatin openness Refinement) possess more ranked GO terms due to their loose criteria in building regulatory region–gene associations. However, when considering the number of associated regions of each method, significant diversity appears. For example, DAVID, which only takes account of proximal regions, associates the least number of regions. TFAS and EpiFIT-NR connect much more regulatory regions than other two methods. Using refinement based on chromatin openness, EpiFIT reduces more than 60% distal genomic regions which might not really regulate nearest genes. After refinement, EpiFIT retains about 18% more regions than DAVID, which might be considered as true distal regulatory regions.

We also checked the functional interpretation performance of these four methods based on the gold-standard dataset described in the sections above. As shown in Table 4, compared with EpiFIT, for all the three other methods, most of the 409 authorized GO terms receive rank decrease, which means EpiFIT accurately associates real regulatory regions and targeted genes to implement transcription factor functional interpretation.

In a word, compared with existing methods, EpiFIT has superior ability to build real associations between distal regulatory regions and genes. This ability can be extended to other usages, such as the prediction of ChIP-seq targeted gene.

**Table 3 Gold-standard GO term rank change between EpiFIT and GREAT with randomly permuted chromatin openness**

Experiment No.	Rank improved	Rank decreased	Rank unchanged	Total term amount	Binomial test <i>P</i> -value
1	170	192	47	409	0.2697
2	173	188	48	409	0.4613
3	164	195	50	409	0.1132



**Figure 5.** Comparison of ranked GO term numbers and associated region numbers. (A) Boxplot of numbers of ranked GO terms derived by all four methods (DAVID, TFAS, EpiFIT-NR, EpiFIT). (B) The median number of ranked GO terms of four methods. (C) Boxplot of numbers of associated regulatory regions of all four methods. (D) The median number of associated regulatory regions of four methods.

**Table 4** Gold-standard GO term rank change compared with EpiFIT

Methods	Number of rank				Binomial test <i>P</i> -value
	Improved	Decreased	Unchanged	Total term amount	
EpiFIT-NR*	56	305	48	409	2.5163e−40
TFAS	69	282	58	409	1.0978e−31
DAVID	85	260	64	409	9.2544e−22

\*: EpiFIT-NR refers to EpiFIT with no chromatin openness refinement.

## CONCLUSION

We propose EpiFIT, a web-based tool that combines sequence information with epigenetic data to interpret the functions of transcription factors. EpiFIT highlights the idea that chromatin openness can reflect the correlation between transcription factor binding sites and related genes. The correlation can help EpiFIT to refine regulatory region–gene associations built by distance rules. Using a set of real data, we compare the performance of EpiFIT with that of other methods. Results demonstrate that EpiFIT can precisely interpret the functions of transcription factors. We furtherly examine the performance of EpiFIT with randomly permuted chromatin openness. Results demonstrate the effectiveness of openness in functional interpretation. At last, a series of experiments are conducted to demonstrate

the ability of EpiFIT to reveal real associations between regulatory regions and targeted genes, which can be used to other research aspects.

To sum up, EpiFIT can be used in the functional interpretation of not only transcription factors, but also all the *cis*-regulatory regions that have a correlation with epigenetic characteristics.

## ACKNOWLEDGMENTS

This work has been supported by the National Key Research and Development Program of China (No. 2018YFC0910404), the National Natural Science Foundation of China (Nos. 61873141, 61721003, 61573207, 71871019 and 71471016), and the Tsinghua-Fuzhou Institute for Data Technology.

## COMPLIANCE WITH ETHICS GUIDELINES

The authors Shaoming Song, Hongfei Cui, Shengquan Chen, Qiao Liu and

Rui Jiang declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

## REFERENCES

- Johnson, D. S., Mortazavi, A., Myers, R. M. and Wold, B. (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, 316, 1497–1502
- Mardis, E. R. (2007) ChIP-seq: welcome to the new frontier. *Nat. Methods*, 4, 613–614
- Tu, S. and Shao, Z. (2017) An introduction to computational tools for differential binding analysis with ChIP-seq data. *Quant. Biol.*, 5, 226–235
- Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., Birney, E., *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.*, 41, 827–841
- Blahnik, K. R., Dou, L., O’Geen, H., McPhillips, T., Xu, X., Cao, A. R., Iyengar, S., Nicolet, C. M., Ludäscher, B., Korf, I., *et al.* (2010) Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic Acids Res.*, 38, e13
- Huang, W., Sherman, B. T. and Lempicki, R. A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, 4, 44–57
- Huang, W., Sherman, B. T. and Lempicki, R. A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, 37, 1–13
- McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M. and Bejerano, G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, 28, 495–501
- Natarajan, A., Yardimci, G. G., Sheffield, N. C., Crawford, G. E. and Ohler, U. (2012) Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Res.*, 22, 1711–1722
- Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M. and Sidow, A. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, 5, 829–834
- Cao, S., Zhou, Y., Wu, Y., Song, T., Alsaihati, B. and Xu, Y. (2017) Transcription regulation by DNA methylation under stressful conditions in human cancer. *Quant. Biol.*, 5, 328–337
- Liu, Q., Xia, F., Yin, Q. and Jiang, R. (2018) Chromatin accessibility prediction via a hybrid deep convolutional neural network. *Bioinformatics*, 34, 732–738
- Sherwood, R. I., Hashimoto, T., O’Donnell, C. W., Lewis, S., Barkal, A. A., van Hoff, J. P., Karun, V., Jaakkola, T. and Gifford, D. K. (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. *Nat. Biotechnol.*, 32, 171–178
- Wang, Y., Jiang, R. and Wong, W. H. (2016) Modeling the causal regulatory network by integrating chromatin accessibility and transcriptome data. *Natl. Sci. Rev.*, 3, 240–251
- Chen, S., Wang, Y. and Jiang, R. (2019) OPENANNO: annotating genomic regions with chromatin accessibility. *BioRxiv*
- Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., *et al.* (2018) The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.*, 46, D794–D801
- ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489, 57–74
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, 25, 25–29
- The Gene Ontology Consortium. (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, 47, D330–D338
- Min, X., Zeng, W., Chen, N., Chen, T. and Jiang, R. (2017) Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding. *Bioinformatics*, 33, i92–i101
- Duren, Z., Chen, X., Jiang, R., Wang, Y. and Wong, W. H. (2017) Modeling gene regulation from paired expression and chromatin accessibility data. *Proc. Natl. Acad. Sci. USA.*, 114, E4914–E4923
- Huntley, R. P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M. J. and O’Donovan, C. (2015) The GOA database: gene Ontology annotation updates for 2015. *Nucleic Acids Res.*, 43, D1057–D1063
- Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, 42, D472–D477
- Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., Kumar, R. M., Murray, H. L., Jenner, R. G., *et al.* (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *cell*, 122, 947–956
- Zhao, M., Amiel, S. A., Christie, M. R., Muiresan, P., Srinivasan, P., Littlejohn, W., Relat, M., Arno, M., Heaton, N. and Huang, G. C. (2007) Evidence for the presence of stem cell-like progenitor cells in human adult pancreas. *J. Endocrinol.*, 195, 407–414
- Lee, J., Kim, H. K., Han, Y. M. and Kim, J. (2008) Pyruvate kinase isozyme type M2 (PKM2) interacts and cooperates with Oct-4 in regulating transcription. *Int. J. Biochem. Cell Biol.*, 40, 1043–1054
- Xu, H., Wang, W., Li, C., Yu, H., Yang, A., Wang, B. and Jin, Y. (2009) WWP2 promotes degradation of transcription factor OCT4 in human embryonic stem cells. *Cell Res.*, 19, 561–573
- Yoon, S. J., Wills, A. E., Chuong, E., Gupta, R. and Baker, J. C. (2011) HEB and E2A function as SMAD/FOXH1 cofactors. *Genes Dev.*, 25, 1654–1661
- Kristensen, D. M., Nielsen, J. E., Skakkebaek, N. E., Graem, N., Jacobsen, G. K., Rajpert-De Meyts, E. and Leffers, H. (2008) Presumed pluripotency markers UTF-1 and REX-1 are expressed in human adult testes and germ cell neoplasms. *Hum. Reprod.*, 23, 775–782
- Trubiani, O., Zalzal, S. F., Paganelli, R., Marchisio, M., Giancola,

- R., Pizzicannella, J., Bühring, H. J., Piattelli, M., Caputi, S. and Nanci, A. (2010) Expression profile of the embryonic markers nanog, OCT-4, SSEA-1, SSEA-4, and frizzled-9 receptor in human periodontal ligament mesenchymal stem cells. *J. Cell. Physiol.*, 225, 123–131
31. Stefanovic, S., Abboud, N., Désilets, S., Nury, D., Cowan, C. and Pucéat, M. (2009) Interplay of Oct4 with Sox2 and Sox17: a molecular switch from stem cell pluripotency to specifying a cardiac fate. *J. Cell Biol.*, 186, 665–673
32. Lei, X. X., Xu, J., Ma, W., Qiao, C., Newman, M. A., Hammond, S. M. and Huang, Y. (2012) Determinants of mRNA recognition and translation regulation by Lin28. *Nucleic Acids Res.*, 40, 3574–3584
33. Bard, J. D., Gelebart, P., Amin, H. M., Young, L. C., Ma, Y. and Lai, R. (2009) Signal transducer and activator of transcription 3 is a transcriptional factor regulating the gene expression of *SALL4*. *FASEB J.*, 23, 1405–1414
34. Kunarso, G., Chia, N. Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y. S., Ng, H. H. and Bourque, G. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.*, 42, 631–634
35. Li, J., & Wang, C. Y. (2008). TBL1–TBLR1 and  $\beta$ -catenin recruit each other to Wnt target-gene promoter for transcription activation and oncogenesis. *Nat. cell Biol.*, 10, 160–169.
36. Zhou, S., Fujimuro, M., Hsieh, J. J. D., Chen, L., Miyamoto, A., Weinmaster, G. and Hayward, S. D. (2000) SKIP, a CBF1-associated protein, interacts with the ankyrin repeat domain of NotchIC To facilitate NotchIC function. *Mol. Cell. Biol.*, 20, 2400–2410
37. Guenther, M. G., Barak, O. and Lazar, M. A. (2001) The SMRT and N-CoR corepressors are activating cofactors for histone deacetylase 3. *Mol. Cell. Biol.*, 21, 6091–6101
38. Yu, S. and Reddy, J. K. (2007) Transcription coactivators for peroxisome proliferator-activated receptors. *BBA-MOL Cell Biol. L.*, 1771, 936–951.
39. Feige, J. N., Gelman, L., Michalik, L., Desvergne, B. and Wahli, W. (2006) From molecular action to physiological outputs: peroxisome proliferator-activated receptors are nuclear receptors at the crossroads of key cellular functions. *Prog. Lipid Res.*, 45, 120–159
40. Ishii, S., Kurasawa, Y., Wong, J. and Yu-Lee, L. Y. (2008) Histone deacetylase 3 localizes to the mitotic spindle and is required for kinetochore-microtubule attachment. *Proc. Natl. Acad. Sci. USA*, 105, 4179–4184
41. Ouyang, Z., Zhou, Q. and Wong, W. H. (2009) ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proc. Natl. Acad. Sci. USA*, 106, 21521–21526