Computational and Structural Biotechnology Journal 17 (2019) 1226–1233

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

The spatial binding model of the pioneer factor Oct4 with its target genes during cell reprogramming

Hanshuang Li [a,1], Na Ta [b,1], Chunshen Long [a], Qiutang Zhang [a], Siyu Li [a], Shuai liu [b,*], Lei Yang [c,*], Yongchun Zuo [a,*]

[a] State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, College of Life Sciences, Inner Mongolia University, Hohhot 010070, China
[b] Inner Mongolia Key Laboratory of Social Computing and Data Processing, College of Computer Science, Inner Mongolia University, Hohhot 010020, China
[c] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

ARTICLE INFO

Article history:
Received 19 February 2019
Received in revised form 5 September 2019
Accepted 7 September 2019
Available online 11 September 2019

Keywords:
Spatial binding pattern
Cell reprogramming
Pioneer factor
Multivariate linear regression

ABSTRACT

Understanding the target regulation between pioneer factor and its binding genes is crucial for improving the efficiency of TF-mediated reprogramming. Oct4 as the only one factor that cannot be substituted by other POU members, it is urgent need to develop a quantitative model for describing the spatial binding pattern with its target genes. The dynamic profiles of pioneer factor Oct4-binding showed that the major wave occurs at the intermediate stage of cell reprogramming (from day 7 to day 15), and the promoter is the preferred targeting regions. The Oct4-binding distributions perform significant chromosome bias. The overall enrichment on chromosome 1–11 is higher than that on the others. The dramatic event of TF-mediated reprogramming is mainly concentrated on autosomes. We also found that the spatial binding ability of Oct4 binding can be represented quantitatively by using three parameters of peaks (height, width and distance). The dynamic changes of Oct4-binding demonstrated that the width play more important roles in regulating expression of target genes. At last, a multivariate linear regression was introduced to establish the spatial binding model of the Oct4-binding. The evaluation results confirmed that the height and width is positively correlated with the gene expression. And the additive interaction terms of height and width can better optimize the model performance than the multiplicative terms. The best average coefficients of determination of improved model achieved to 81.38%. Our study will provide new insights into the cooperative regulation of spatial binding pattern of pioneer factors in cell reprogramming.

1. Introduction

Somatic cell reprogramming is the process of reprogramming the differentiated somatic cells into pluripotency or even totipotency under specific induction conditions [1]. Since 2006, Yamanaka successfully induced fibroblasts into induced pluripotent stem cells (iPSCs) by screening, combination and over-expression of the four transcription factors of Oct4, Sox2, c-Myc and Klf4, this discovery has undoubtedly initiated the pioneer of transcriptional factor-mediated reprogramming and overturned the previous understanding of "Differentiation process is irreversible" [2–5]. The generation of iPSCs makes it possible to obtain patient-specific stem cells without ethical disputes, which has great potential in tissue and organ transplantation, stem cell therapy, molecular mechanism of specific diseases and personalized medicine [2,6–10].

With the deepening of research, many studies have shown that Oct4 as one of the four Yamanaka factors required for reprogramming somatic cells into iPSCs and that it is the only factor that cannot be substituted in this process by other members [11]. For example, OCT4 and Sox2 can reprogram human umbilical cord blood stem cells directly to pluripotency [5], but for endogenous neural stem cells with high expression of Sox2 and c-Myc, only forced expression of OCT4 can generate iPSCs [12]. Moreover, OCT4 can recruit other factors required for regulating the expression of its target genes, such as chromatin remodeling complexes as an important partner of OCT4 interaction, which is contribute to the reorganization of nucleosome positioning and is necessary for pluripotency [13,14].

* Corresponding authors.
E-mail addresses: cs_liushuai@imu.edu.cn (S. liu), leiyang@hrbmu.edu.cn (L. Yang), yczuo@imu.edu.cn (Y. Zuo).
[1] Equal contributors.

https://doi.org/10.1016/j.csbj.2019.09.002

Additionally, Oct4 acting as a pioneer transcription factor not only functions by recognizing and binding to DNA regulatory regions alone or in cooperation with other TFs [15–18], but also by identifying DNA in closed chromatin state [19]. Similarly, some studies have also revealed that in the early stage of reprogramming, Oct4 plays a pioneering role in the acquisition and maintenance of cellular pluripotency by effectively binding to noncoding regulatory regions of pluripotent regulatory genes, actively opening up the local chromatin, turning on the genome of downstream pluripotent regulatory network, and activating the expression of multiple genes [14,20,21]. In the studies of TFs impacting on transcriptional regulation, there are some quantitative models have been constructed, in which the gene expression profile is regarded as the response variable and various features related to Oct4 and other TFs are taken as the explanatory variables. Examples of such features include characteristics or counts of motifs recognized by the TFs [22–24], sum of motif occurrences weighted by their distances from the target gene [25], the weighted sum of the corresponding ChIP-Seq signal strength of TFs [15,26–28] and the number of TF-binding events [29]. Oct4 as the only one factor that cannot be substituted by other POU members in cell reprogramming, but there are few quantitative studies on the spatial binding pattern of Oct4 to regulate gene expression.

In our study, based on the high-throughput sequencing data generated by ChIP-seq, a comprehensive genome DNA binding map of Oct4 with different reprogramming time points was discussed. The chromosome bias of Oct4-binding distributions was further explored. Next, we delineated the genome-wide spatial binding pattern of Oct4 during reprogramming and presented the multivariate regression model of Oct4 binding pattern impacting gene expression levels. At last, we identified some target genes to analysis the cooperative regulatory relationship. The results of this study will be helpful for improving the efficiency of TF-mediated reprogramming.

2. Materials and methods

2.1. Dataset collection

The ChIP-seq data of this study were downloaded from Gene Expression Omnibus (GEO) database under accession number GSE67520. The ChIP-seq data of Oct4 contains nine reprogramming time points from mouse fibroblasts into iPSCs, including fibroblast status, days 1, 3, 5, 7, 11, 15, 18, induced pluripotent stem cell status (D0, D1, D3, D5, D7, D11, D15, D18, DIPSC). The microarray transcriptome data was also downloaded from the GEO database, and GEO accession no. GSE67462, which includes two biologically repeated gene expression microarray data of 18 samples at different time points of reprogramming (D0, D1, D3, D5, D7, D11, D15, D18 and DIPSC) in mice.

2.2. Data processing

The downloaded ChIP-seq original fastq format data were controlled by FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to remove low-quality samples. Next Oct4 ChIP-seq reads were aligned to the mouse reference genome (Mm9 assembly) using Bowtie (version 0.12.7) [30] short read alignment software with parameter setting: -v2-m1 and other default values. In our study only tags uniquely mapped to the genome were retained. Then Oct4 binding peaks were called by MACS2 (version 2.1.0) [31] with the input data as the control. Finally using R package ChIPseeker [32] annotated with the position of the peaks in the genome, in which −5 kb to +0.5 kb of gene transcription start sites (TSS) were defined gene promoter, and gene enhancer were defined as upstream 5 kb to 50 kb of gene TSS. In addition, the expression value of replicates at different time points were averaged as the final gene expression value data.

2.3. Integration of Oct4 spatial binding parameters

Since each target gene is not usually corresponding to a unique peak (Fig. S2A), the many-to-single relationship between the peaks and the gene need to redefine. For each peak–gene pair, we integrated the binding parameters of the measured Oct4 peaks around each gene promoter into spatial binding parameters of a single peak, namely H_i, D_i, W_i (Height_i, Distance_i, Width_i) (Fig. 1A). We assumed that the height of peak binding on gene i was a weighted average of height values of all of the peaks on gene i:

$$H_i = \sum_{k=1}^{j} \mathrm{h}_{ki}/j(k \in j) \tag{1}$$

where h_{ki} is the height (the score of an individual peak binding site) of the kth peak of Oct4 binding on the gene i, j is a sum of the number of all of the peaks on gene i. And the distance (the distance from the peak center to the TSS) and width (the difference value between the endpoint coordinates and the starting coordinates of peak) of peak binding on gene i as follow:

$$W_i = \max\left[w_{\sum_{k=1}^{j}(i,k)}\right] \tag{2}$$

$$D_i = \min\left[d_{\sum_{k=1}^{j}(i,k)}\right] \tag{3}$$

The $d\,(i, k)$ is the distance of the kth peak to TSS of gene i, and D_i indicates the minimum distance between peak k and the gene i among all of the peaks on gene i. The $w\,(i, k)$ indicates the width of the kth peak binding on the gene i, and W_i is the maximum width of the peak binding on the gene i among all of the peak width values.

2.4. Quantitative correlation analysis and multivariate regression

Next, we calculated the Pearson correlation coefficient (PCC) between the three spatial parameters of Oct4 binding on its target genes and the gene expression to quantitatively describe the correlation between the two in the reprogramming process. The PCC value [33] was introduced as follow:

$$P_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} (-1 \le P_{X,Y} \le 1) \tag{4}$$

where X_i (H_i, D_i, W_i) is the spatial parameters of Oct4 binding on gene i and Y_i is the true expression of gene i. n is a total number of genes at the same reprogramming time point. The higher PCC indicates the stronger correlation.

Based on the obtained good quantitative correlations, we described the expression level of gene i by a linear regression model:

$$\widehat{Y}_i \sim \mu_i + \alpha_i H_i + \beta_i D_i + \gamma_i W_i \tag{5}$$

where $\widehat{Y}_i$ is the theoretical expression level of gene i, $\mu_i$is a constant, $H_i$is the height value of Oct4 peak for the ith gene and α_i is the regression coefficient for the peak height on ith gene. Similarly, D_i is the distance value of Oct4 peak for the ith gene and $\beta_i$is the regression coefficient for the peak distance, W_i is the width of peak for the ith gene and $\gamma_i$also is the regression coefficient.

The coefficients of regression model are estimated by using least squares:

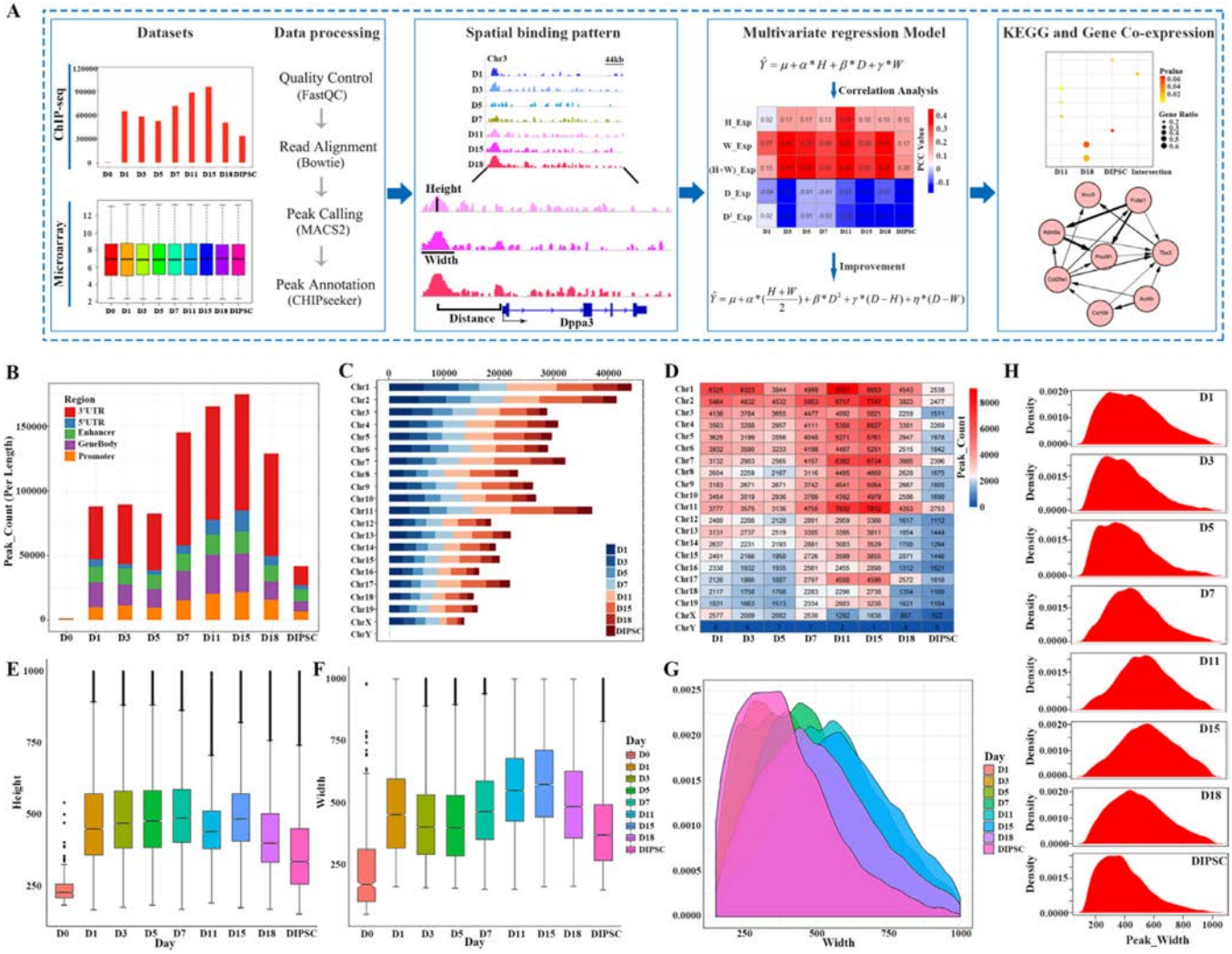


Fig. 1. Dynamic profiles of Oct4-binding and characteristic of chromosomal distribution. (A) Flowchart of this study. (B) Dynamic profiles of Oct4 binding in the whole genome during reprogramming. (C) Bar plot of the distribution of Oct4_Peak on chromosomes at different time points of programming. The thickness of connecting lines between chromosomes and time points is proportional of the number of peaks. (D) The number of Oct4 peaks distributed on different chromosomes during reprogramming. (E) The boxplot represents the dynamic change of height of peaks during reprogramming. (F) The boxplot represents the dynamic change of width of peaks during reprogramming. (G) Kernel Density plot of width of Oct4-binding on different target genes during reprogramming. (H) The dynamic change of width of Oct4 peaks during reprogramming.

$$(\mu_i, \alpha_i, \beta_i, \gamma_i) = \arg\min_{\mu_i,\alpha_i,\beta_i,\gamma_i} \left\{ \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 \right\} \tag{6}$$

So a simple multivariate linear regression model was given by

$$\hat{Y} = \mu + \alpha H + \beta D + \gamma W \tag{7}$$

where $\hat{Y}$ is the theoretical expression level, μ is a constant. And α, β, γ are the regression coefficients for H, D, W, respectively. To investigate the cooperative regulation of spatial binding pattern, we exhaustively searched interaction terms from our multivariate linear regression model, such as the additive and multiplicative interaction terms of height and width, the square interaction term of the distance. The interaction terms of parameters were added to the linear regression model to get five new models, which can be formulated as

$$\begin{cases} \hat{Y} = \mu + \alpha * \left(\frac{H*W}{2}\right) + \beta * (D+H) + \gamma * (D+W) & (8) \\ \hat{Y} = \mu + \alpha * \left(\frac{H+W}{2}\right) + \beta * (D-H) + \gamma * (D-W) & (9) \\ \hat{Y} = \mu + \alpha * \sqrt{(H*W)} + \beta * (D+H) + \gamma * (D+W) & (10) \\ \hat{Y} = \mu + \alpha * \sqrt{(H*W)} + \beta * D^2 + \gamma * (D+H) + \eta * (D+W) & (11) \\ \hat{Y} = \mu + \alpha * \left(\frac{H+W}{2}\right) + \beta * D^2 + \gamma * (D-H) + \eta * (D-W) & (12) \end{cases}$$

2.5. KEGG pathway enrichment analysis

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway-based analysis helps us to further understand genes different biological functions [34] at different time points of reprogramming. So we performed KEGG pathway enrichment analysis on the unique and shared genes at four time points of reprogramming (D11, D15, D18 and DIPSC) with the Database for Annotation, Visualization and Integrated Discovery (DAVID) Bioinformatics Resource [35] (Fig. 1A). The horizontal axis indicates time points of programming and vertical ordinates are the terms of the KEGG pathways. Gene ratio is the proportion of the number of genes vs. the total number of genes in the same KEGG pathway. The color represents p value.

2.6. Analysis of gene co-expression and network construction

The gene co-expression networks of Pou5f1, Xrcc5, Fubp1, Tbx3, Aurkb, Cd109, Col25a1 and Kdm5a the eight genes we screened. The WGCNA R package was used to establish co-expression networks of these genes [36] (Fig. 1A). Next the network was visualized by Cytoscape3.7.0 [37]. In the network, the nodes represent

the genes and the edge correlates with the regulation capacity for adjacent genes. The larger the edge, the stronger the regulatory relationship between the two genes and the more edges of a gene means the more central role it has within the network.

2.7. Data visualization

In this study, data visualization was carried out mainly by the R, including the R/Bioconductor (http://www.bioconductor.org). The heatmap and Upset plot were produced using Pheatmap and UpsetR [38], respectively. The genome browser view was obtained using the Integrative Genomics Viewer (IGV) [39] and the density graph, boxplot, bubble graph and so on were generated with the R packet ggplot2 (http://ggplot2.org/).

3. Results

3.1. Dynamic profiles and chromosomal distribution of Oct4-binding in the whole genome

Firstly, the counts of Oct4-binding peaks in five genomic regions, 5′UTR, 3′UTR, Enhancer, Promoter and GeneBody (Fig. 1B) were selected to describe dynamic profiles of Oct4-binding during reprogramming. And we proposed the peaks per kilobase mapped peaks (PPKM, $\mathrm{PPKM} = \sum \frac{\mathrm{Peak\ \ Count}}{\mathrm{The\ \ length\ \ of\ \ region}} * 1000$) to reduce the effects of region length on the statistical results.

We found that the major wave occurs at the intermediate stage of cell reprogramming (from day 7 to day 15). Especially in promoter regions, all of the PPKM reached more than 20000, showing that the promoter is the preferential targeting genomics regions. In addition, UTR and GeneBody regions also had high proportion of Oct4-binding (Fig. 1B). During the reprogramming, the quantity change of Oct4 targeted genes was consistent with the peaks and reached the most at day 15 (Fig. S1A and B). Interestingly, at the initial stage of reprogramming (D1-D3), the number of the genes and peaks also exhibited a high level, indicating that Oct4 not only has a targeted regulatory effect on pluripotent genes, but also plays a certain role in some housekeeping genes [40]. We further explored the number relationship between the peaks and genes (Fig. S2A). In the process of reprogramming (expect D1), each target gene corresponded to one or two peaks in all regions, and the one-to-one relationship was more obvious in promoter regions. Thus, it can be concluded that there is not a unique peak on the same regulatory element of the same target gene.

For the distributions of Oct4-binding on different chromosomes, the results showed that Oct4 was more inclined to bind on autosomes, especially on chromosome 1, 2, 7 and 11 (Chr1, Chr2, Chr7 and Chr11) were the most significantly (Fig. 1C and S1C). The specific number of peak distributions on different chromosomes was marked in detail (Fig. 1D). The number of Oct4 peaks distributed on all autosomal chromosomes was more than 1000 throughout the reprogramming process. And at the days 7–18 Oct4 tended to bind on Chr1-Chr11 and Chr17. Surprisingly, the number of peaks enriched on chromosome 1, 2, and 11 (Chr1, Chr2, and Chr11) was more than 4000 (except DIPSC) and the ontologies for the genes represented by these peaks on these three chromosomes were been shown in Supplementary Table 1. However, there are a few peaks mapped to sex chromosomes, especially on the Y chromosome (ChrY) with almost no Oct4 binding. It indicated the dramatic event of TF-mediated reprogramming is concentrated on autosomes.

In addition, we selected the Oct4 peaks on day 3 of the early reprogramming, day 7 (the number of peaks significantly increased), and day 15 (the number of peaks reached the maximum), respectively to reveal the dynamic pattern of Oct4-binding. The results showed the median of peak height on all chromosomes (except ChrY) was around 400, and the change was not obvious (Fig. S1D). The significant change on ChrY was related to the low number of peaks distributed on this chromosome (Fig. 1D). The median of distance on D15 was lower than other two days, and the overall levels on D3 were the highest. However, the width of peaks was more than 400 in all chromosomes (except ChrY) on D15. Surprisingly, the height and width of peaks binding on Chr11, Chr13 and Chr17 showed evidently increased.

The dynamic changes of Oct4 spatial parameters in the whole genome as shown in Fig. 1E-1H, the height of peaks was generally 300–600 at different time points of reprogramming (except D0 and DIPSC), and the overall trend was not obvious (Fig. 1E). The width was 200–400 bp on days 1–5 (D1-D5) and days 18-IPSC (D18-DIPSC), but on days 7 and 11 the width gradually increased to 400–800 bp (Fig. 1F and 1G). Generally, the width of peaks tended to widen first and then narrow during the reprogramming (Fig. 1H). The above results demonstrated that the height of Oct4 combined with its target genes does not change significantly, but the width tended to widen, implying the width may play a vital role in promoting the reprogramming.

3.2. Dynamics of Oct4 binding in promoter regions during reprogramming

The dynamic profiles of Oct4-binding showed the promoter is the main targeting genomics regions. So we further illustrated the characteristic of Oct4 binding in promoter regions. Throughout the reprogramming process, the three spatial binding parameters of the peaks binding on each target gene promoter are processed, respectively (as shown in Methods).

We found that Oct4 extensively bound to the promoter from day 7 until day 18 (D7-D18) and the number of target genes was prominent increased at these days (Fig. 2A), implying that the targeted regulation of Oct4 is dynamic with time preference, as in the previous study [40]. As shown in Fig. 2B, the peaks of Oct4 were mainly located within 500 bp (−500, 500) of the TSS and relatively few located beyond upstream 2 kb of TSS (−3000, −2000) in the reprogramming process. Intriguingly, there was a tendency for Oct4 to bind to upstream 500 bp regions from TSS at the early stage of reprogramming (D3–D7) (blue box of Fig. 2B), but at the end stage of reprogramming (D11–D18) Oct4 mainly enriched in downstream 500 bp regions from the TSS (pink box of Fig. 2B). The results suggest that Oct4 first acts on the TSS upstream and then targeted binding on TSS downstream of target genes to facilitate the reprogramming process. Next, we questioned how Oct4 binding in promoter regions regulates gene expression. The quantitative analysis of Oct4 spatial binding parameters and its target gene expression levels during reprogramming were shown in Fig. 2C. The height of Oct4 peaks binding on its target genes remained relatively steady and there was a slight increase on day 15 (D15), but the distance of peaks tended to decrease from day 1 (D1) until day 18 (D18), and the overall changes were not obvious. Surprisingly, the width of peaks increased significantly from day 7 until day 18 (D7–D18) coincided with the increase of gene expression levels. Therefore, it can be inferred that during the reprogramming process, the spatial binding pattern of Oct4 plays a pioneering role in activating the expression of its target genes, which further proves that the height and width of Oct4 peaks may facilitate the gene expression, and distance may be a negative factor in this process.

To more intuitively profile the dynamic changes of Oct4 spatial binding pattern in the reprogramming process, we subsequently visualized the Oct4 peaks on several genes by IGV [39], and examined the expression levels of these genes (Fig. 2D). The results fur-

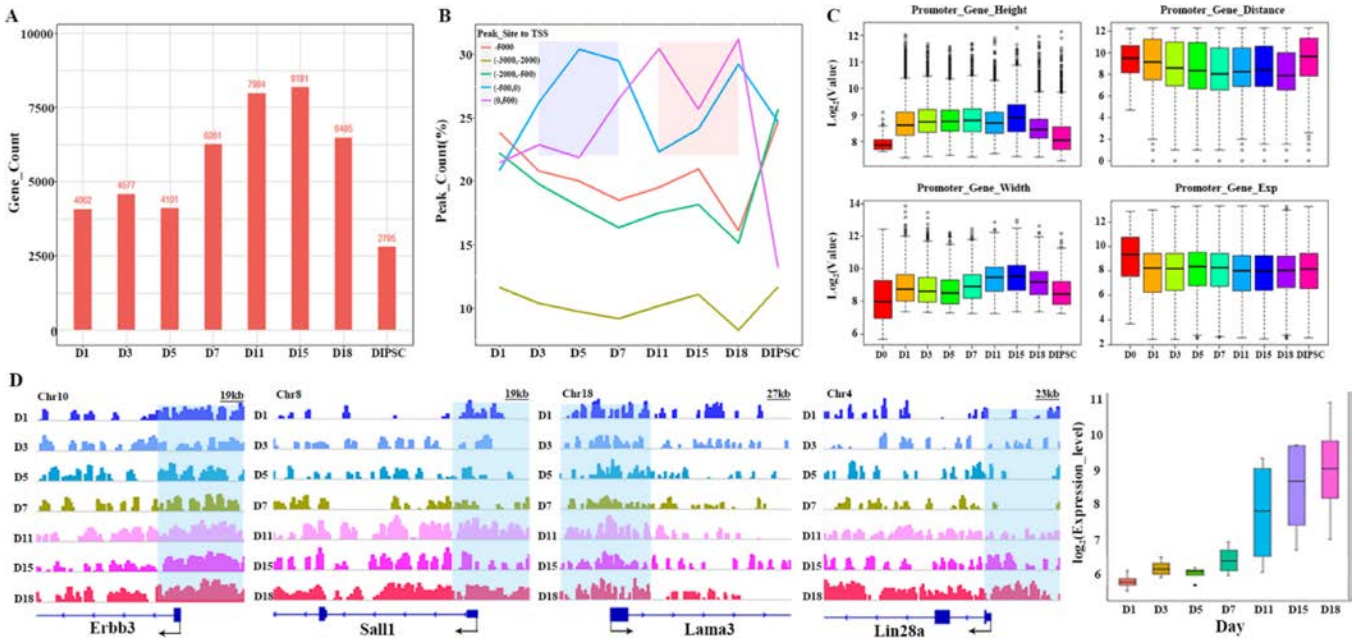


Fig. 2. The dynamics binding of Oct4 in promoter regions. (A) The number of Oct4 target genes at different reprogramming time points is represented as bar graph, and the specific number is marked at the top of the bar graph. (B) Distribution of Oct4-binding sites with respect to the TSS in promoter at different reprogramming time points is shown. (C) Boxplots of height (upper left), width (left bottom) and distance values of Oct4 peaks (upper right) located in target genes promoter and its target genes expression in promoter (right bottom). (D) Genome browser view at the Erbb3, Sall1, Lama3 and Lin28a locus of Oct4 binding at the different time points of reprogramming, respectively. And the blue boxes represent the promoters of these genes and their surrounding regions.

ther indicated that the increasing of height and width can activate gene expression, and the increasing of distance will inhibit gene expression, among which the width plays more important roles.

3.3. The spatial binding model of the pioneer factor Oct4 with its target genes

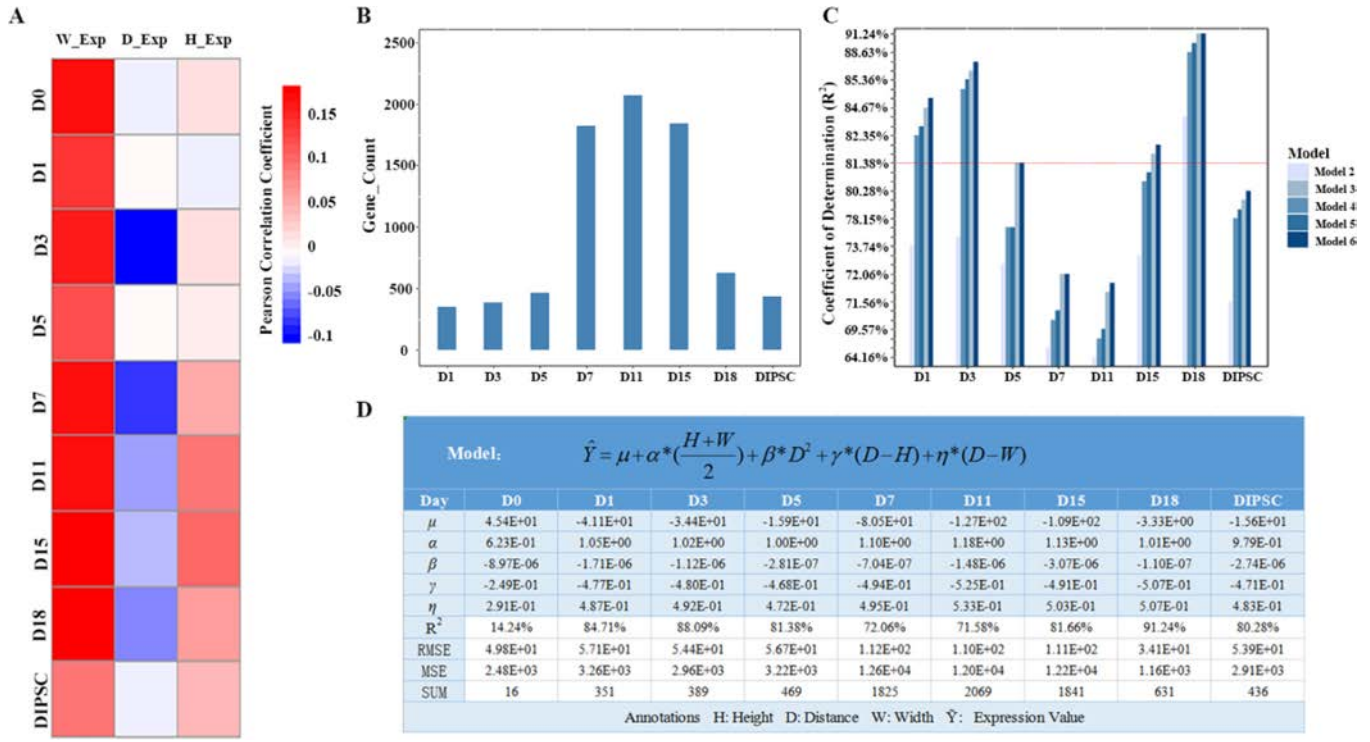
To assess the importance of spatial binding pattern of Oct4 in regulating the gene expression, we further quantitatively analyzed the correlation between spatial binding of Oct4 and gene expression levels (Fig. 3A). A careful examination of this correlation (Fig. 3A) revealed that there was prominent positive correlation between the width of peaks and gene expression levels where the correlation was particularly evident, and the height and gene expression levels were also positively correlated (except D1). By contrast, the distance from peaks to TSS was negatively correlated with the gene expression, suggesting that the peaks with higher, wider and binding closer to TSS contribute to promoting the expression of genes during reprogramming.

Given the good quantitative correlations above (Fig. 3A), the target genes of Oct4 at the promoter were appropriately screened by bioinformatics means under the premise of satisfying the correlation at the different time points of reprogramming (Fig. 3B). The minimum number of remained genes was also above 300 (Fig. 3B). Hence, a multivariate linear regression was introduced to establish the spatial binding model (model 1) of the Oct4, but the correlation of model 1 was inconsistent with our previous analysis. Then the interaction terms of parameters were added to the linear regression model to obtain other five new models (as shown in Methods). And the specific information of these six models and parameters were detailed in the Supplementary Table 2. Next, the coefficients of determination obtained were used to evaluate the other five models (except model 1) to determine the optimal model (Fig. 3C). The coefficients of determination can be used as a measure of the goodness of fit of sample observation values. The higher the coefficients of determination are, the better the model goodness of fit. On the contrary, the coefficient is small, indicating that the model fits the sample observations to a lesser extent. Compared to other five models (Supplementary Table 2) we constructed, model 6 showed the additive interaction terms of height and width can better optimize the model performance than the multiplicative terms. When adding the square interaction term of the distance, the average coefficients of determination of model 6 achieved to 81.38% (Fig. 3C), which was higher than other models. However, the coefficients of determination of model without width were less than 30% (Supplementary Table 3).

The model 6 as the optimal model quantitatively represented the spatial binding ability of Oct4 binding on its target genes by three parameters of peaks. And the detailed information about the model 6 can be found in Fig. 3D, in which the model has time and chromosome specificity and the performance on Day 18 (Chr1 and Chr2) is higher than on other chromosomes of different days (Fig. 3D and Supplementary Table 4). When reanalyzing the RNA-seq data, the correlation between Oct4-binding and target gene expression (Fig. S4A) was agreement with Fig. 3A and the coefficients of determination of model 6 were above 80.00% (Data are from Malik et al. 2019, Supplementary Table 5) [41]. Interestingly, the model also can apply into human data (Data were reanalyzed from Narayan et al. 2017) [42] and the comparative result was added in Supplementary Table 6. In a word, the width and height of Oct4 binding on the target genes does not alone but coordinately regulates gene expression.

3.4. The cooperative regulation of Oct4 target genes

Based on the above of the optimal model (model 6), we screened some genes to calculate the Pearson correlation between the Oct4-binding and these genes expression (Fig. 4A). Compared to the single spatial binding parameter, the interaction terms had stronger correlation with gene expression levels (Fig. 4A), which



Day	D0	D1	D3	D5	D7	D11	D15	D18	DIPSC
μ	4.54E+01	-4.11E+01	-3.44E+01	-1.59E+01	-8.05E+01	-1.27E+02	-1.09E+02	-3.33E+00	-1.56E+01
α	6.23E-01	1.05E+00	1.02E+00	1.00E+00	1.10E+00	1.18E+00	1.13E+00	1.01E+00	9.79E-01
β	-8.97E-06	-1.71E-06	-1.12E-06	-2.81E-07	-7.04E-07	-1.48E-06	-3.07E-06	-1.10E-07	-2.74E-06
γ	-2.49E-01	-4.77E-01	-4.80E-01	-4.68E-01	-4.94E-01	-5.25E-01	-4.91E-01	-5.07E-01	-4.71E-01
η	2.91E-01	4.87E-01	4.92E-01	4.72E-01	4.95E-01	5.33E-01	5.03E-01	5.07E-01	4.83E-01
R^2	14.24%	84.71%	88.09%	81.38%	72.06%	71.58%	81.66%	91.24%	80.28%
RMSE	4.98E+01	5.71E+01	5.44E+01	5.67E+01	1.12E+02	1.10E+02	1.11E+02	3.41E+01	5.39E+01
MSE	2.48E+03	3.26E+03	2.96E+03	3.22E+03	1.26E+04	1.20E+04	1.22E+04	1.16E+03	2.91E+03
SUM	16	351	389	469	1825	2069	1841	631	436

Annotations H: Height D: Distance W: Width Ŷ: Expression Value

Fig. 3. Modeling of Oct4 spatial binding and its target genes. (A) Pearson correlation analysis of spatial binding parameters of Oct4 and its target gene expression levels at different time points of reprogramming. (B) The number of genes screened based on Pearson correlation coefficient. (C) The evaluation of multivariate regression model. The red line indicated the average coefficients of determination (R^2) of model 6. The higher the coefficients are, the better the model goodness of fit. (D) The form and detailed parameters of the optimal model.

further demonstrated Oct4 acts on its target genes in a collaborative way.

Among in D11 – DIPSC, the correlation was more significant, we further explored the characteristic of the genes in these days. There were 52, 29, 29, 17 genes unique to these several days and 28 shared genes (Fig. 4B). The shared genes were associated with the signaling pathways regulating pluripotency of stem cells, but the unique genes in these days were mainly involved in Focal adhesion, Thyroid hormone signaling pathway and MAPK signaling pathway, indicating that they maintained basic biological functions as housekeeping genes (Fig. 4C). But the numbers of housekeeping genes [43,44] were steady around 30% during the reprogramming (Fig. S4B) and the change trend of the three binding parameters and expression of them (Fig. S4C-S4D) were consistent with our previous studies. We further selected three genes (Nanog, Dppa3 and Sall4) and visualized their peaks by IGV (Fig. 4D). The peak of Oct4 binding on these three genes tended to be wider and higher, but the change of distance was not very obvious, and the expression level of Sall4 also increased during reprogramming (Fig. 4D). It further proved that expression levels were positively correlated with the height and width of peaks but negatively correlated with distance.

Next, we identified eight genes to analysis the cooperative regulatory relationship (Fig. 4E). As shown in Fig. 4E, Fubp1, Oct4 (also known as Pou5F1) and Kdm5a genes had strong interactions, in which, Fubp1 not only can directly regulated Oct4 but also can directly regulated Kdm5a and in turn regulated Oct4 expression. Surprisingly, the expression of Fubp1 was decreased unlike other genes on day 11 (Fig. S3A), which illustrate Fubp1 may recruit Oct4 to promoter the pluripotent networks. Interestingly, chromatin regulators Kdm5a also had a strong regulatory effect on Oct4, which was in agreement with the previous finding [45]. Notably, the remaining six genes (except Xrcc5) were all related to Tbx3 and its expression level was significantly increased on day 15 (Fig. S3A), which was also different from other genes, these results highlighted the function of Tbx3 in the maintenance and induction of pluripotency during reprogramming [46]. We also deeply observed the dynamic binding pattern of Oct4 and expression levels of these target genes during reprogramming (Fig. S3B). As expected, the results were almost in agreement with our previous findings (Fig. 2C). In short, in the process of reprogramming, the cooperative regulatory relationship between genes can form a complex core pluripotent network.

4. Discussion

Understanding the quantitative binding pattern of TF to regulate gene expression is important. Here, we first systematically analyzed the dynamic binding profiles of Oct4. The result showed that the major wave of Oct4-binding occurs at the intermediate stage of cell reprogramming (from day 7 to day 15), and the promoter is the preferential target genomic regions. Interestingly, in the promoter regions, Oct4 first acts on the TSS upstream and then targeted binding on TSS downstream of target genes to active the expression of these genes. Next we addressed whether the Oct4-binding distributions perform significant chromosome bias. Compared with sex chromosomes, Oct4-binding tends to be on autosomes, and the overall enrichment on Chromosome 1–11 is higher than that on the others, which indicates the dramatic event of TF-mediated reprogramming is concentrated on autosomes.

The spatial binding pattern of Oct4 is dynamic change. Specially, the height and width of peaks is increased with the reprogramming, but the distance is gradually decreased. Importantly, the significant increasing of width is coincided with the activation of gene expression. Similarly, the width and gene expression levels

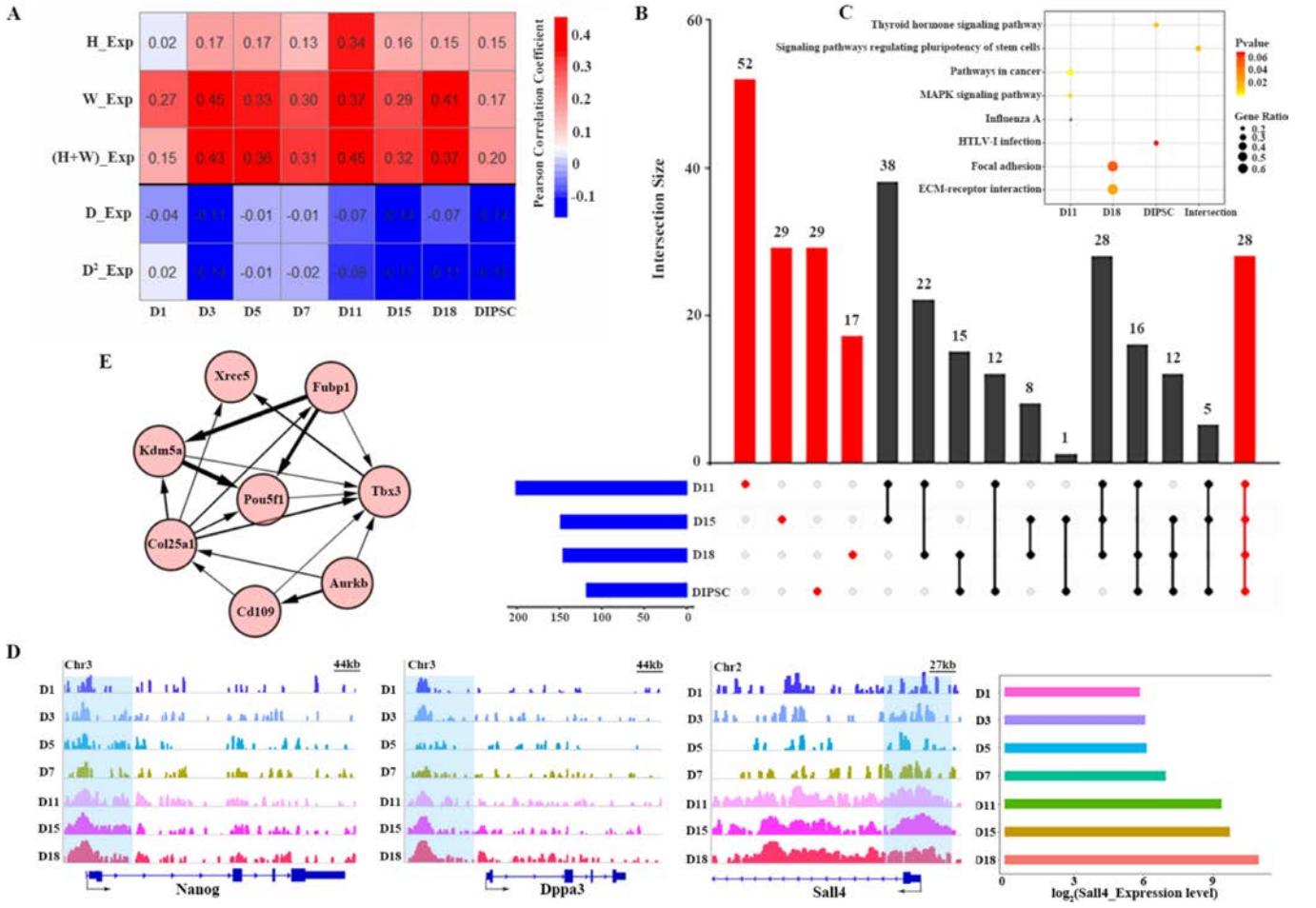


Fig. 4. Identification and characteristic analysis of strongly correlated target genes. (A) Pearson correlation analysis of spatial binding parameters and parameters combinations of Oct4 and its target genes expression values. (B) Upset chart showing the target genes screened by us at four time points of reprogramming (horizontal bar). The specific number of identified genes shared between different sets is indicated in the top bar chart corresponding to the solid points below the bar chart and each column represents shared genes between the different time points (linked dots). Figure generated using Upset R package. (C) The analysis KEGG pathway enrichment. (D) Genome browser view of the Oct4 density in the Nanog, Dppa3 and Sall4 region at different point of reprogramming, and the blue boxes represent the promoters of these genes and their surrounding regions. The bar chart presents the dynamic change of the expression level of Sall4, and the abscissa is the logarithmic conversion of the expression value. (E) Analysis of gene co-expression network, cycle nodes represent genes and the size of edges represents the power of the interrelation among the nodes.

have prominent positive correlation, and the height and gene expression levels were also positively correlated (except D1). But the distance from peaks to TSS was negatively correlated with the gene expression, among which the width is the most important factor for regulating expression of target genes.

Furthermore, we used three parameters of peaks to quantitatively represent the spatial binding ability of Oct4 binding on its target genes, include height, width (0.05–2 kb) and distance (0.5–5 kb). A multivariate linear regression was introduced to establish the spatial binding model of the Oct4-binding. After the interaction terms of parameters as features were added to improve the linear regression model, we presented the other five models. The evaluation results confirmed that the additive interaction terms of height and width can better optimize the model performance than the multiplicative terms. And then adding the square interaction term of the distance, the average coefficients of determination of improved model (model 6) can achieve to 81.38%. So model 6 as the optimal model showed that the spatial binding pattern of pioneer factors Oct4 acting on the target genes does not alone but coordinately regulates the gene expression. Especially, the cooperative relationship between width and height of Oct4 peaks plays the most important roles in regulating target genes. At last, we screened some genes with strongly correlation to analysis the cooperative regulatory relationship of these genes and speculated that Fubp1 may play vital roles in recruiting Oct4. Taken together, our study not only qualitatively analyzed the dynamic profiles of Oct4-binding but also represented the quantitative regulation model of Oct4 binding pattern impacting gene expression, hoping to provide a new insight into the cooperative regulation of spatial binding pattern of pioneer factors in cell reprogramming.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge the Chunshen Long for his valuable suggestions for improvement of the manuscript.

Funding

This work was supported by the National Nature Scientific Foundation of China (Nos: 61561036, 61702290), Program for Young Talents of Science and Technology in Universities of Inner

Mongolia Autonomous Region (NJYT-18-B01), the Fund for Excellent Young Scholars of Inner Mongolia (2017JQ04) and Student's Platform for Innovation and Entrepreneurship Training Program of Inner Mongolia University (No: 201714298). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2019.09.002.

References

[1] Mosteiro L, Pantoja C, Alcazar N, Marion RM, Chondronasiou D, et al. Tissue damage and senescence provide critical signals for cellular reprogramming in vivo. Science 2016;354.
[2] Takahashi K, Tanabe K, Ohnuki M, Narita M, Ichisaka T, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 2007;131:861–72.
[3] Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 2006;126:663–76.
[4] Ohnuki M, Takahashi K, Yamanaka S. Generation and characterization of human induced pluripotent stem cells. Curr Protoc Stem Cell Biol 2009. Chapter 4: Unit 4A 2.
[5] Nakagawa M, Koyanagi M, Tanabe K, Takahashi K, Ichisaka T, et al. Generation of induced pluripotent stem cells without Myc from mouse and human fibroblasts. Nat Biotechnol 2008;26:101–6.
[6] Daley GQ, Lensch MW, Jaenisch R, Meissner A, Plath K, et al. Broader implications of defining standards for the pluripotency of iPSCs. Cell Stem Cell 2009;4:200–1.
[7] Jung JH, Fu X, Yang PC. Exosomes generated from iPSC-derivatives: new direction for stem cell therapy in human heart diseases. Circ Res 2017;120:407–17.
[8] Takebe T, Sekine K, Enomura M, Koike H, Kimura M, et al. Vascularized and functional human liver from an iPSC-derived organ bud transplant. Nature 2013;499:481–4.
[9] Takebe T, Zhang RR, Koike H, Kimura M, Yoshizawa E, et al. Generation of a vascularized and functional human liver from an iPSC-derived organ bud transplant. Nat Protoc 2014;9:396–409.
[10] Sivapatham R, Zeng X. Generation and characterization of patient-specific induced pluripotent stem cell for disease modeling. Methods Mol Biol 2016;1353:25–44.
[11] Jerabek S, Merino F, Scholer HR, Cojocaru V. OCT4: dynamic DNA binding pioneers stem cell pluripotency. Biochim Biophys Acta 2014;1839:138–54.
[12] Yamanaka S. A fresh look at iPS cells. Cell 2009;137:13–7.
[13] Ho LN, Jothi R, Ronan JL, Cui KR, Zhao KJ, et al. An embryonic stem cell chromatin remodeling complex, esBAF, is an essential component of the core pluripotency transcriptional network. PNAS 2009;106:5187–91.
[14] van den Berg DL, Snoek T, Mullin NP, Yates A, Bezstarosti K, et al. An Oct4-centered protein interaction network in embryonic stem cells. Cell Stem Cell 2010;6:369–81.
[15] 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.
[16] Ambrosetti DC, Basilico C, Dailey L. Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. Mol Cell Biol 1997;17:6321–9.
[17] King HW, Klose RJ. The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells. Elife 2017;6.
[18] Chronis C, Fiziev P, Papp B, Butz S, Bonora G, et al. Cooperative binding of transcription factors orchestrates reprogramming. Cell 2017;168:442–459. e420.
[19] Soufi A, Donahue G, Zaret KS. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell 2012;151:994–1004.
[20] Pardo M, Lang B, Yu L, Prosser H, Bradley A, et al. An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. Cell Stem Cell 2010;6:382–95.
[21] Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. Genes Dev 2011;25:2227–41.
[22] Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. Nat Genet 2001;27:167–71.
[23] Zamanighomi M, Lin Z, Wang Y, Jiang R, Wong WH. Predicting transcription factor binding motifs from DNA-binding domains, chromatin accessibility and gene expression data. Nucleic Acids Res 2017;45:5666–77.
[24] Duren Z, Chen X, Jiang R, Wang Y, Wong WH. Modeling gene regulation from paired expression and chromatin accessibility data. Proc Natl Acad Sci USA 2017;114:E4914–23.
[25] Keles S, van der Laan M, Eisen MB. Identification of regulatory elements using a feature selection method. Bioinformatics 2002;18:1167–75.
[26] McLeay RC, Lesluyes T, Cuellar Partida G, Bailey TL. Genome-wide in silico prediction of gene expression. Bioinformatics 2012;28:2789–96.
[27] Budden DM, Hurley DG, Crampin EJ. Predictive modelling of gene expression from transcriptional regulatory elements. Brief Bioinform 2015;16:616–28.
[28] Zhang L, Xue G, Liu J, Li Q, Wang Y. Revealing transcription factor and histone modification co-localization and dynamics across cell lines by integrating ChIP-seq and RNA-seq data. BMC Genomics 2018;19:914.
[29] Shi W, Fornes O, Wasserman WW. Gene expression models based on transcription factor binding events confer insight into functional cis-regulatory variants. Bioinformatics 2018. bty992-bty992.
[30] Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 2009;10:R25.
[31] Feng J, Liu T, Zhang Y. Using MACS to identify peaks from ChIP-Seq data. Curr Protoc Bioinf 2011. Chapter 2: Unit 2.14.
[32] Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 2015;31:2382–3.
[33] Ahlgren P, Jarneving B, Rousseau R. Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. J Am Soc Info Sci Technol 2003:550–60.
[34] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 2016;44:D457–62.
[35] Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 2009;4:44–57.
[36] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9.
[37] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13:2498–504.
[38] Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics 2017;33:2938–40.
[39] Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013;14:178–92.
[40] Chen J, Chen X, Li M, Liu X, Gao Y, et al. Hierarchical Oct4 binding in concert with primed epigenetic rearrangements during somatic cell reprogramming. Cell Rep 2016;14:1540–54.
[41] Malik V, Glaser LV. Pluripotency reprogramming by competent and incompetent POU factors uncovers temporal dependency for Oct4 and Sox2. Nat Commun 2019;10:3477.
[42] Narayan S, Bryant G, Shah S, Berrozpe G, Ptashne M. OCT4 and SOX2 work as transcriptional activators in reprogramming human fibroblasts. Cell Rep 2017;20:1585–96.
[43] Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet 2013;29:569–74.
[44] Chang CW, Cheng WC, Chen CR, Shu WY, Tsai ML, et al. Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. PLoS ONE 2011;6:e22859.
[45] Wang WP, Tzeng TY, Wang JY, Lee DC, Lin YH, et al. The EP300, KDM5A, KDM6A and KDM6B chromatin regulators cooperate with KLF4 in the transcriptional activation of POU5F1. PLoS One 2012;7:e52556.
[46] Russell R, Ilg M, Lin Q, Wu G, Lechel A, et al. A dynamic role of TBX3 in the pluripotency circuitry. Stem Cell Rep 2015;5:1155–70.