Molecular BioSystems

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/c6mb00006a

Genome-wide targets identification of "core" pluripotency transcription factors with integrated features in human embryonic stem cells†

Leijie Li,‡[a] Zhaobin Chen,‡[a] Liangcai Zhang,[b] Guiyou Liu,[c] Jinlian Hua,*[d] Lianghui Jia*[a] and Mingzhi Liao*[a]

Received 3rd January 2016,
Accepted 16th February 2016

DOI: 10.1039/c6mb00006a

www.rsc.org/molecularbiosystems

Embryonic stem cells (ESCs) play an important role in developmental biology which is still lacking clear molecular mechanisms. The "core" transcription factors (TFs) including OCT4, SOX2 and NANOG are essential for maintaining the stemness of ESCs. But the downstream targets of these "core" TFs are still ambiguous. Based on support vector machine (SVM) technology, this study develops a label method algorithm (LMA) for genome-wide target identification of "core" TFs in humans, which eliminates the need for negative training samples. This method integrates histone modifications and TF binding motifs as identification features. Compared with a previous mapping-convergence (M-C) algorithm, the LMA can provide more stable and reliable predictions. 4796, 3166 and 4384 target genes of OCT4, SOX2 and NANOG, respectively, were identified with the LMA model. Then verifications of the predicted targets were carried out based on their functional consistency and their connection degree in networks from a computational system biology perspective. The results showed that the targets of "core" TFs present higher gene functional similarity and closer connection distance than background levels.

Introduction

Embryonic stem cells (ESCs) can be obtained from the inner cell mass (ICM) and cultured as immortalized cells *in vitro*.[1,2] Nowadays, research into the molecular mechanisms of early embryonic development has gained increasing attention, focusing on the genes that control the differentiation regulation of embryonic stem cells. Among these genes, as "core" pluripotency transcription factors, OCT4, SOX2 and NANOG play key roles in maintaining the pluripotency of ESCs.[3,4]

OCT4 is a member of the POU transcription factors family, which is very important for ESC multipotency and self-renewal.[5–7] It is mainly expressed in ESCs and germ stem cells (GSCs).[8] Through interactions with OCT3 and CDX2 (a primosome for trophectoderm differentiation) OCT4 prevents differentiation towards trophectoderm directly by forming a repressor complex.[9] As a transcriptional partner of OCT4, SOX2 is also involved in regulating downstream gene expression in ESC development.[10] During embryonic development, SOX2 has two expression peaks, including in the inner cell mass and neural stem cells (NSCs), which indicates the relationship between keeping pluripotency and inhibiting differentiation.[11] NANOG was also found to play significant roles in the self-renewal and differentiation potency of ESCs.[12] NANOG can maintain the ESCs self-renewal activity by regulating the expression of gata6 and gata4.[13] In a word, OCT4, SOX2, NANOG and their targets are very important in maintaining the pluripotency of ESCs.[14]

However, the exact targets of these human "core" TFs are ambiguous, which is an obstacle for downstream molecular experiments about maintaining the stemness of ESCs. According to a previous study, there are about 5000 targets of these "core" TFs in humans, which is nearly equal to the numbers found in mouse.[15] So there is a need to predict and supplement the targets of "core" TFs in human ESCs.

Small interfering RNA technology[16] and gene knockout technology[17] have been used in the identification of TF targets. But these methods are time-consuming and very expensive for the research of TFs targets on a genome-wide scale. What's more, chromosome immune precipitation (ChIP) combined with a DNA chip (ChIP–chip) is another popular method to identify protein binding motifs.[18,19] With the rapid development of next-generation sequencing technology, lower costs and higher

[a] College of Life Sciences, Northwest A&F University, Yangling, Shaanxi 712100, China. E-mail: liaomingzhi83@163.com, jialianghui@nwsuaf.edu.cn
[b] Department of Statistics, Rice University, 6100 Main St., Houston, TX 77005, USA
[c] Genome Analysis Laboratory, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, China
[d] College of Veterinary Medicine, Shaanxi Centre of Stem Cells Engineering & Technology, Northwest A&F University, Yangling, Shaanxi 712100, China. E-mail: jlhua2003@126.com

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c6mb00006a
‡ Contribute equal to this study.

efficiency have allowed ChIP-seq to be widely used in the study of "core" TFs binding motifs.[20] However there are large amounts of noise that can influence the results of ChIP-seq and ChIP–chip,[21] which is associated with sample quality and the experimental environment. Considering the "big data" property of genome-wide scale information, computational methods provide another way to identify the target genes of ESCs "core" pluripotency TFs.

This study presents a more accurate computational strategy to identify the targets of OCT4, SOX2 and NANOG from a genome-wide scale. Considering the basics of support vector machine (SVM) technology, the strategy was named the label method algorithm (LMA) method. The method integrates transcriptional regulation information and post-transcriptional modification information as identification features. The conservative motif sequences of TFs are important for binding site recognition. Histone modifications, as one of the major epigenetic modifications, may change the interaction between DNA and other nucleoproteins, which can have a substantial influence on gene expression and lead to a decrease in the efficiency of TFs.[22]

In summary, based on our prediction model LMA, this study tried to predict targets of OCT4, SOX2 and NANOG. Furthermore, we verified the predicted targets with their functional consistency and their connection degree in networks from a computational system biology perspective. Therefore, through this prediction model, researchers can obtain some reference biomarkers or pathways for downstream molecular studies about ESCs stemness.

Materials and methods

Targets of "core" pluripotency TFs and extraction of their features

In order to collect the targets of "core" TFs we manually collected the related articles in PubMed, and finally two articles were found.[23,24] 623 targets of OCT4, 2014 targets of SOX2, and 2676 targets of NANOG were obtained in this way. The detailed information about these datasets is shown in Fig. S1A (ESI†). In order to train our model, all the human genes were firstly divided into two types in the training data, including the MIX and POS sets. The MIX represented all the genome wide human genes with unknown labels, which may be a mixture of targets and non-targets of "core" embryonic stem cells transcription factors; while the POS set was the known targets of "core" embryonic stem cells transcription factors, which was extracted from previous research. In other words, if a gene had been verified as a target of "core" embryonic stem cells transcription factors in previous research, then it was labeled POS; otherwise, the gene was labeled MIX.

Considering both transcriptional regulation and post-transcriptional regulational effects during the development of embryonic stem cells, we integrated features from transcription factor binding sites and histone modifications.[25,26] We got the transcription factor binding sites from the UCSC database (http://genome.ucsc.edu/). At the table browser site of the UCSC database, we chose 'regulation' groups and 'TFBS conserved' track to get the output data. The histone modifications were downloaded from the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) database in NCBI with accession number GSE16256, which is originated from the Human Reference Epigenome Mapping Project and has 856 samples.[27–29] Then the data was normalized by local SISSRS software[30] which is an open source software implemented by Perl and can be downloaded at http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/sissrs/.

Table 1 2×2 contingency table used in the Fisher's exact test in feature extraction

Feature		
Gene set	0	1
MIX set	a	b
POS set	c	d

Where "a" represents the number of genes that are not combined with a given feature in the MIX set. "b" represents the number of genes that are combined with a given feature in the MIX set. "c" represents the number of genes that are not combined with a given feature in the POS set. "d" represents the number of genes that are combined with a given feature in the POS set.

In order to extract the features which can distinguish the targets of "core" embryonic stem cells transcription factors and other normal genes, we used Fisher's exact test and multiple testing correction in the following work. We calculated the number of each feature's targets and non-targets in the MIX set and POS set which were used to make a 2×2 contingency table (Table 1). Then we distinguished different features by Fisher's exact test (P-value < 0.05) using the table in R statistical software. Finally, in order to reduce the number of redundant features, we used multiple testing correction to deal with the P-value of Fisher's exact test and got a stringent Q-value for each feature. The Benjamini & Hochberg method ("BH" or its alias "FDR") which is commonly used for multiple testing correction was selected in our manuscript.[31] Multiple testing correction was performed by controlling the false discovery rate (FDR) in the BH algorithm. First we ranked the P-value of features with ascending order that were filtered by Fisher's exact test such as $P(1)$, $P(2)$,...,$P(m)$. Then if we found $P(i)$ that fulfilled the following formulas, the related features of $P(1)$, $P(2)$,...,$P(i)$ were retained.

$$P(i) \leq \frac{i \times q}{m} \quad (1)$$

$$P(i+1) \geq \frac{(i+1) \times q}{m} \quad (2)$$

Where q is the value of FDR (q-value = 0.05).

Finally we got 108, 81 and 36 features of OCT4, SOX2 and NANOG, respectively. To facilitate the use of these features, we provided them in the ESI.† In order to connect these two type sets and the feature, we constructed feature vectors for each gene. Then all the genes within the POS or MIX sets comprise a feature matrix. Thus the POS and MIX sets of OCT4 were a 469 column and 108 row (469-108) matrix and a (28 006-108) matrix, respectively. The size of the POS and MIX matrixes for SOX2

were (101-81) and (27 499-81), respectively. The POS and MIX matrixes of NANOG were (1303-36) and (26 943-36), respectively.

Datasets of protein–protein interactions (PPIN)

In order to avoid any preconceptions about the data sources, we downloaded protein–protein interaction data from two different databases: the Biological General Repository for Interaction Datasets (BioGRID) version 3.2.110 (http://theBioGRID.org/) and the Human Protein Reference Database (HPRD) (http://www.hprd.org/).[32,33] Both of the datasets were excluding pure high throughput experimental data. We then deleted multiple edges and self loops by using Cytoscape version 3.02.[34–36] Cytoscape was executed under Windows 2007, and the tool 'NetworkAnalyzer' was applied to deal with our data. The process of each PPIN was nearly ten minutes. The dataset from BioGRID contained 9698 nodes with 52 284 edges (excluding pure high throughput experimental data) in humans. The HPRD dataset contained 9453 nodes with 36 867 edges (excluding pure high throughput experimental data).

Building the prediction model

Based on SVM technology, the LMA prediction model is presented as a workflow in Fig. 1. The SVM was implemented using the freely downloadable software package LIBSVM (3.20, http://www.csie.ntu.edu.tw/~cjlin/libsvm/), and the radial basis function (RBF) was used.[37] There are two important parameters, the gamma and penalty parameters, which should be evaluated first in each SVM. Gamma is an important parameter of the RBF kernel, whose value has a big influence on the accuracy of the model. The penalty parameter represents the penalty of false samples in the model. So, in order to get an exact model, gamma and penalty parameters were confirmed by 5-fold cross validation in each cycle. In the model, the known target genes were treated as the POS set and all other genes as the MIX set (eqn (3)). NEG is extracted from the MIX set and the number of the NEG set is equal to the number of the POS set. 5-fold cross validation was performed in each SVM test in the LMA model. At the beginning of cross validation, the training set, including the positive set and negative set, was equally divided into 5 parts. Four parts among the training set were used for training and one part for testing, which was looped for 5 times in each cycle. And this process of 5-fold cross validation was done in every cycle of the LMA model, which was performed for a total of 10 000 cycles. In this model, all the genes have a label to indicate their status. We added an extra row in the MIX set to record the classification results, which is called label. If the label was set to plus one, it means the gene was predicted as the target of TFs. When the model is processing, the label of each may be increased from 0 to 10 000, which is the total number of cycles in the model. So the label can be used as a confidence score that indicates the probability of the gene being the target of TFs. In order to evaluate the model performance, a previous method was selected as a comparison model.[26]

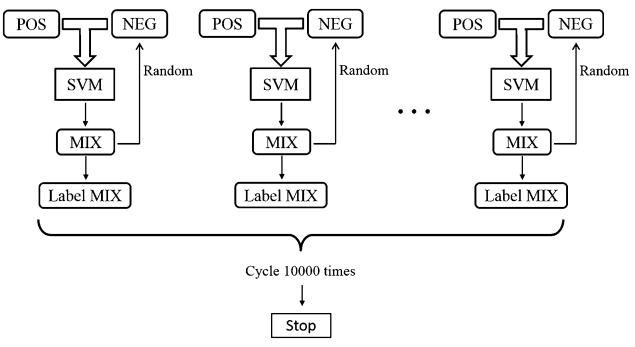

Fig. 1 Workflow of the LMA methods. In LMA methods we used POS and NEG that was selected from MIX set randomly to train SVM model. Then MIX set was detected by the SVM model and the label of MIX was changed according to the results.

$$\mathrm{MIX} = \mathrm{Universal} - \mathrm{POS} \tag{3}$$

Where universal means all of the human genes, the MIX set is unlabeled genes and the POS set is TF target genes. So MIX equals universal genes minus POS.

Both AUC and ACC were selected as measurements to determine the effectiveness of these models. AUC is computed as the area under the receiver operating characteristic curve (ROC curve).

$$\mathrm{AUC} = \int_0^1 f(\mathrm{FPR})\, d_{\mathrm{FPR}} \tag{4}$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{5}$$

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

Where FPR represents the false positive rate, TPR represents the true positive rate, FP represents the false positive set, FN represents the false negative set, TP represents the true positive set, and TN represents the true negative set. In fact, when our model was running, an equal number of genes from the POS set was randomly selected from the MIX set as the negative set (NEG). In each cycle, we randomly separated one-tenth of the POS set and NEG set as test data. Since we know the label of the test data, we can use the POS and NEG sets as judgment criteria. If one gene originated from POS was determined as POS, then the value of TP increases, otherwise FN increases; if one gene originated from NEG was determined as NEG, then the value of TN increases.

ACC represents the ability of correct classification. And ACC is computed as the following formula:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}} \tag{7}$$

where P represents the positive set and N represents the negative set.

The M-C model is also a SVM-based method which was trained by POS and NEG to classify the MIX set firstly. Then the predicted negative set was collected to the MIX. And the remaining predicted positive results were used in the next cycle. When the predicted negative was null, the cycle was finished.[26]

To evaluate the performance of the LMA method and the M-C algorithm, we used a Bayesian model to calculate the prediction accuracy of positive targets. This model is based on one biological assumption that more than 90 percent of genes aren't targets of "core" pluripotency transcription factors.[15] So we supposed event A_1 "gene A is target" and event A_2 "gene A is not target", the probabilities of event A_1 and event A_2 were 0.9 and 0.1, respectively. We supposed that event B was "Gene A is predicted as target in the LMA model" (we used the lowest 95% as an experimental threshold). And event C was "Gene A was predicted as target in the M-C model." The Bayesian formulas are presented as follows:

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{\sum_{i=1}^{2} P(A_i)P(B|A_i)} \quad (8)$$

$$P(A_1|C) = \frac{P(A_1)P(C|A_1)}{\sum_{i=1}^{2} P(A_i)P(C|A_i)} \quad (9)$$

where

$$P(A_1) = 0.9 \quad (10)$$

$$P(A_2) = 0.1 \quad (11)$$

$$P(B|A_1) = C_{10000}^{9500}\left(\overline{ACC}_1\right)^{9500}\left(1-\overline{ACC}_1\right)^{500} \quad (12)$$

$$P(B|A_2) = C_{10000}^{500}\left(\overline{ACC}_1\right)^{500}\left(1-\overline{ACC}_1\right)^{9500} \quad (13)$$

$$P(C|A_1) = \left(\overline{ACC}_2\right)^{n} \quad (14)$$

$$P(C|A_2) = \left(1-\overline{ACC}_2\right)^{n} \quad (15)$$

$\overline{ACC}_1$ represents the average accuracy of the 10 000 cycles in the LMA model, $\overline{ACC}_2$ represents the average accuracy of the M-C model, and n represents the number of M-C loops. The results are shown in Table 2.

Analysis of targets connectivity and function similarity

The connectivity intensity and the functional consistency were analyzed for the predicted target genes to verify the effects of prediction. The analyses were all separated into different target sets, including targets regulated only by one TF, by two TFs and by all three "core" TFs. Then the relationships between different target sets were observed through both connectivity intensity and function similarity. Connectivity intensity reflects the network distance between two gene sets in molecular pathways. Function similarity reflects the phenotype distance between two gene sets based on gene function annotation systems. If two gene sets show strong connectivity intensity and high function similarity, there may be close relationship between them. The connectivity intensity was measured with the average shortest path length (ASPL), based on protein–protein interaction networks, including BioGRID and HPRD. Shortest path means the path that connects two nodes in a network whose length is smallest. ASPL computes the average length of all shortest paths between two sets of nodes in a network. The calculation of ASPL was processed in MATLAB version r2014a. The function similarity was analyzed with Lin's algorithm based on the Gene Ontology Consortium.[38,39] It was processed with R version 3.1.2, using the GOSim package.[40]

Table 2 Predicted target accuracy rate of the "core" pluripotency transcription factors in the M-C and LMA models

Model		OCT4	SOX2	NANOG
M-C	Positive accuracy	0.5738	0.7858	0.6684
	Average ACC	0.6554	0.6519	0.6727
	Average AUC	0.7423	0.7163	0.7103
LMA	Positive accuracy	>1-10*e-235	>1-10*e-235	>1-10*e-235
	Average ACC	0.6615	0.6622	0.6364
	Average AUC	0.6309	0.5898	0.5898

0 1

Results

Feature extraction in the LMA model

Considering both transcriptional regulation and post-transcriptional regulation influence, these two levels of factors were both included in this study. The primary positive (POS) sets and mixture (MIX) set used in our LMA method were the same as with the mapping-convergence (M-C) algorithm.[26] Primary POS sets of OCT4, SOX2 and NANOG were 469, 1015 and 1303 respectively, while the MIX sets were 28 006, 27 499 and 26 943 respectively. The numbers of candidate features of each TF were all 285, including the sequence motifs of both transcriptional factor binding sites and histone modification. Through examination with Fisher's exact test and multiple testing correction, the number of significant features of OCT4, SOX2 and NANOG remained 108, 81 and 36 respectively. These features are valuable for other research and are provided in the ESI.†

Comparison of the LMA model and the M-C model

In this portion, we used a Bayesian model to evaluate the accuracy of positive targets. As shown in Table 2, the LMA model accuracy of positive targets is near 1, compared with 0.7858 as the highest accuracy in the M-C model. As for robustness, it can be seen that the M-C model has different prediction results about targets in our experiments from Fig. S2 (ESI†), while the LMA model provides a threshold of 95% as a score to evaluate the stability in our 10 000 cycles, which means that our predictive positive targets were verified at least 9500 times during the 10 000 cycles from Fig. 2. Besides, both the LMA method and M-C method have an ACC average accuracy (ACC) of approximately 0.65 as shown in Table 2, while the M-C method has relatively higher average area under curve (AUC) compared with the LMA method. This is because of the following reason: the M-C model converged quickly to get the targets so the average ACC and AUC in the M-C method is based on several cycles in the inner model. Meanwhile, every cycle in the LMA model is

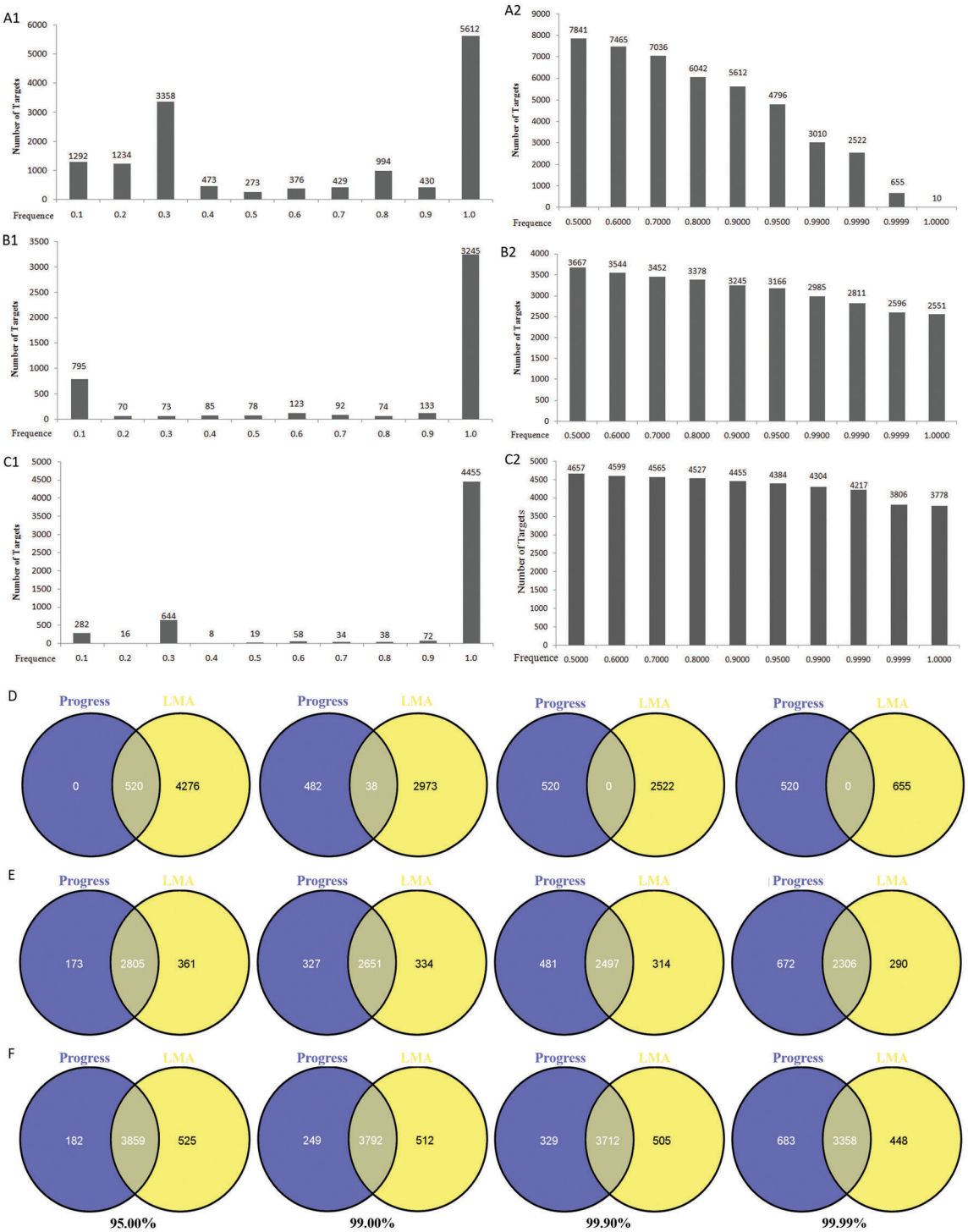

Fig. 2 (A–C) show frequency distributions of predicted target genes of OCT4, SOX2 and NANOG, respectively, in the 10 000 cycle experiments. (A1) is a full distribution of OCT4 prediction. For example 'Frequence 0.1' in (A1) means the number of genes whose frequency is greater than 0 and less than or equal to 0.1. And (A2) is the cumulative distribution from 0.5 to 1.0. For example 'Frequence 0.5' in (A2) means the number of genes whose frequency is greater than or equal to 0.5. (D–F) are Venn diagrams between the M-C algorithm results and LMA results with OCT4, SOX2 and NANOG, respectively. The LMA threshold was set to 95%, 99%, 99.9% and 99.99%.

independent, so the average ACC and AUC in the LMA model are just based on one cycle. After 10 000 cycles, the accuracy has been improved significantly with positive accuracy near to 1. So it is acceptable that the ACC of the LMA model is not too high. In our experiments, there are a large number of genes and functions in the genome. And our purpose is to extract the core targets instead of exactly each potential gene. Our biological analysis results also proved that the prediction results of the LMA model are excellent

from Fig. 3 and 4. So the LMA performance is suitable for follow-up verification through "wet" experiments, including molecular, cellular and tissue experiments.

Performance of the LMA model

As an improved method, the LMA model provided two advantages. First, the results of the LMA model are more robust. As a random stop method, the M-C algorithm will stop mapping convergence immediately when the negative set reaches null. So, the cycle number of the process may be small and vary tremendously (Fig. S3, ESI†). In order to obtain more robust results, the LMA model repeated the process up to 10 000 cycles. Second, the LMA model not only provided the final results, but also provided a confidence score, which is more readable for researchers. And as shown in Fig. 2(A1–C1), the targets are highly concentrated when frequency is higher than 90 percent, which means the LMA model is reliable. The frequency distribution showed an obvious threshold to determine whether a gene should be a target of TFs. So, in order to decide the precise threshold, this study zoomed in on the frequency distribution from 0.5 to 1 (Fig. 2A2–C2). Considering the variation tendency, 95% is a good critical value. In addition, we compared the results of LMA to the M-C algorithm with different thresholds from 95% to 1 (Fig. 2D–F). The results showed that if the threshold is too stringent, then the overlap between these two methods may decrease. So 95% was used as the prediction boundary, with the number of predicted targets for OCT4, SOX2 and NANOG being 4796, 3166 and 4384, respectively.

Connectivity analysis of the predicted targets in PPIN

In order to detect the extent of connection between the predicted targets of "core" TFs, this study used ASPL as a measurement to observe the relation between different target sets that are regulated by different TFs. This work is based on the hypothesis that if two gene sets have a similar function, they may be connected closely with each other for rapid response to complex environments.[15] Unsurprisingly, the results were in accordance with our expectations. For the purpose of elimination of data source bias, two protein–protein interaction databases were selected as our background biological networks, BioGRID and HPRD.[32,33] It was found that all the target sets in BioGRID had significantly smaller ASPL than random background (Fig. 3A). The results indicated that no matter how many "core" TFs they were regulated by, all the target sets connected closely. In the results of HPRD, the target sets that were regulated by OCT4 targets or regulated by both NANOG and SOX2 show smaller distance than the background value (Fig. 3B). No difference from the background value was found between target sets regulated by SOX2, regulated by both OCT4 and SOX2, or regulated by all three TFs. On the other hand, the ASPL of target sets that were regulated solely by NANOG or by both NANOG and OCT4 were higher than the background. The reason for this should be investigated in future research. This may be caused by variance between the HPRD and BioGRID databases or by the inexactness of the predicted targets.

Gene function similarity analysis of the predicted targets

Through the connectivity extent analysis with ASPL, a potential large scale molecular regulation pattern has been revealed. From the phenotype perspective, some patterns in function should also be detected. With this expectation, we analyzed the gene function similarity of the predicted targets based on the Gene Ontology (GO) database. The average gene function similarity

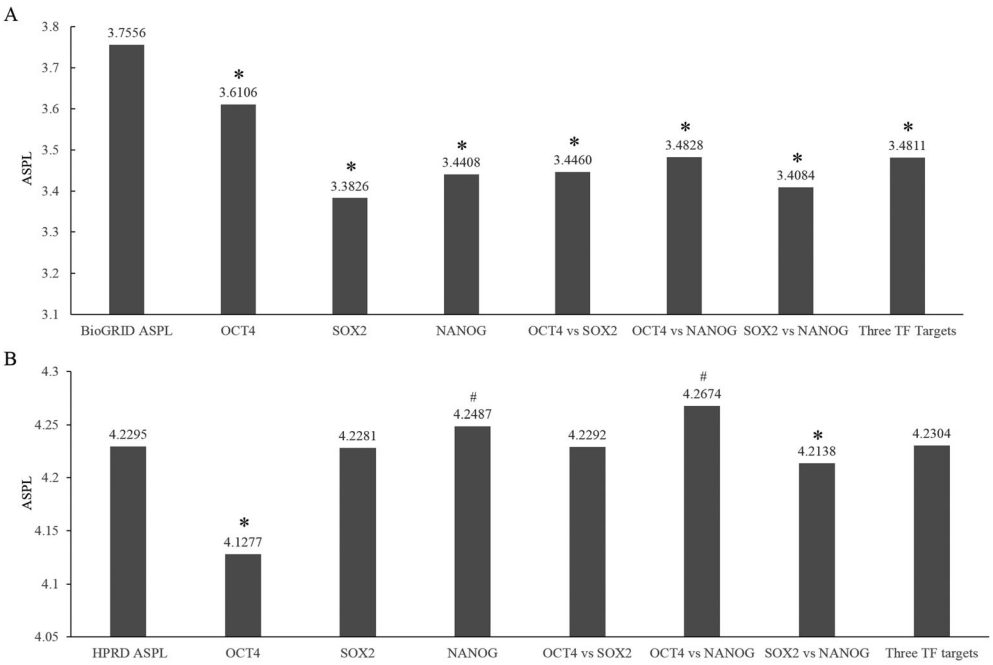

Fig. 3 The average shortest path length analysis of the predicted targets. The threshold was set to 0.01, with "*" meaning smaller than random background value, "#" meaning higher than random background value. (A) Results based on BioGRID database. (B) Results based on HPRD.

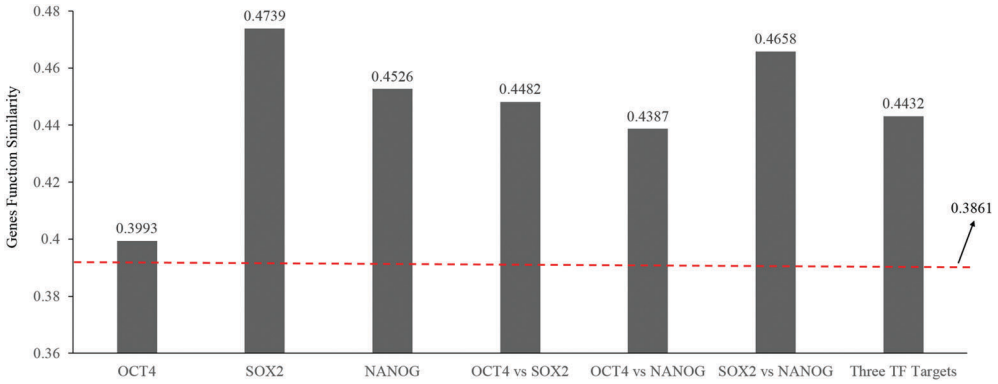

Fig. 4 Gene function similarity analysis results. 0.3861 is the average gene function similarity of all human genes. As shown in the figure, all the groups have higher than average gene function similarity.

among all the human genes was computed as a control value, which is 0.3861.[15] According to the analysis the functions of predicted target genes are more similar than the random background value with a significance level of 0.05, which means the predicted genes work together to process their common function (Fig. 4). The results showed that all the target sets, including those regulated by any one of the "core" TFs, regulated by two TFs or regulated by all three TFs, have higher function similarity than the random background. These results confirm the consistency of our predicted target genes at the function level.

Discussion

In this study we have built a reliable prediction model, with which we predicted the targets of the "core" TFs OCT4, SOX2 and NANOG. And the results are credible through comparison with the targets of "core" TFs in mouse. Through this prediction, the candidate targets of "core" TFs have increased remarkably (Fig. S1, ESI†), to near the known numbers in mouse.[15] In particular, the numbers of SOX2 and NANOG targets are most near the numbers of mouse targets after prediction. In addition, the ratio of their intersection was matched well. We also found that the number of predicted OCT4 targets in humans is still less than the number of targets of OCT4 in mouse.

From the point of view of eliminating the need for negative training samples, both the M-C method and the LMA method are excellent. But the LMA method can provide more readable prediction results. The LMA model contains 10 000 SVM experiments. The genes that were classified as positive more than 9500 times had a high probability to be targets. And we used a naive Bayes algorithm to calculate the positive accuracy of those genes. The result showed that the accuracy is very high and nearly 1. The accuracy of the M-C method which was created by Xue Xiao *et al.* largely depends on the first cycle.[26] If there are real targets in the negative set (NEG) of the first cycle, the process will need more cycles and create corruptions, which means that the model will develop in the wrong direction and result in the reduction of prediction accuracy. What's more, the LMA method can provide confidence scores for the predicted targets, which is more readable for researchers. The LMA method is based on the biological background that the targets of "core" TFs account for nearly 10 percent of all unclassified genes.[15] Based on this hypothesis, if the frequency of predicted targets reached 95 percent after ten thousand cycles, they have great probability to be a real target even though the accuracy of a single cycle isn't higher than the M-C algorithm. As a special case, there are some problems for the OCT4 predicted targets which resulted in the relatively poor prediction of OCT4 targets. OCT4 has the fewest features and targets compared with SOX2 and NANOG, which makes it harder to form a stable prediction model. And the OCT4 target set of the M-C algorithm has no elements the same as the LMA method whose frequency is above 99.9% (Fig. 2D).

This study is the first time the human "core" pluripotency TFs targets have been predicted genome wide without the gold standard negative sets. Considering the widespread lack of negative sets in biological research, the LMA prediction method will be helpful for researchers. With this method, 4796, 3166 and 4384 target genes were identified, which provide a large amount of candidate targets for follow-up "wet" experiments to reveal the pathways that maintain the stemness of cells. For example, CCND2 is known to be expressed in tumors, including liver cancer and diffuse large-B-cell lymphoma.[41,42] Through our methods, we found that CCND2 is a common target of OCT4, SOX2 and NANOG, which indicates that CCND2 is a potential factor in maintaining the pluripotency of ESCs. It is surprising that in the published literature we found that there are dozens of research papers about the relationship between CCND2 and stem cells.[43,44]

Then this study analyzed the consistency of predicted target genes in both their molecular mechanisms and from a phenotype function perspective. Protein–protein interaction networks were selected to reveal the molecular pattern among target gene sets of "core" pluripotency TFs with average shortest path length. The results showed that these target genes have a smaller distance between each other, which is helpful in quickly and synergistically maintaining the pluripotency and self-renewal of ESCs.[45] The results also showed that there are some differences between HPRD and BioGRID. In BioGRID, all the targets of the

"core" TFs are close to each other, while some target sets in HPRD didn't present the same results. One of the reasons may be the difference between these two databases. The phenotype function analysis was performed based on Gene Ontology with gene function similarity. Results showed that all the predicted target gene sets have higher function similarity with each other. This supports the accuracy of our predictions about target genes from a function perspective. These results indicate that there may be fast information flowing through those genes with complex pathways, which will help them to achieve the common function of maintaining pluripotency during the development of ESCs.[15]

Author contributions

M. Z. L. and L. H. J. conceived and initiated the project. L. J. L., Z. B. C. and J. L. H. performed the experiments. L. J. L., Z. B. C., L. C. Z., G. Y. L. and M. Z. L. analyzed the data. L. J. L., Z. B. C. and M. Z. L. wrote the manuscript. All authors reviewed the manuscript, and contributed to the final manuscript.

Competing financial interests

The authors declare no competing financial interests.

Acknowledgements

This work was supported by the Fund of Northwest A&F University; Natural Science Basic Research Plan in Shaanxi Province of China (Grant no. 2014JQ3110); the National Natural Science Foundation of China (Grant no. 31301938, 31572399 and 81300945); Fundamental Research Funds for the Central Universities (Grant no. 2452015077); National Major Project for Production of Transgenic Breeding (Grant no. 2014ZX08007-002); and National High Technology Research and Development Program of China (Grant no. SS2014AA021605).

References

1 M. J. Evans and M. H. Kaufman, *Nature*, 1981, **292**, 154–156.
2 G. R. Martin, *Proc. Natl. Acad. Sci. U. S. A.*, 1981, **78**, 7634–7638.
3 M. Murtha, F. Strino, Z. Tokcaer-Keskin, N. Sumru Bayin, D. Shalabi, X. Xi, Y. Kluger and L. Dailey, *Stem Cells*, 2015, **33**, 378–391.
4 S. H. Orkin, *Cell*, 2005, **122**, 828–830.
5 L. H. Looijenga, H. Stoop, H. P. de Leeuw, C. A. de Gouveia Brazao, A. J. Gillis, K. E. van Roozendaal, E. J. van Zoelen, R. F. Weber, K. P. Wolffenbuttel, H. van Dekken, F. Honecker, C. Bokemeyer, E. J. Perlman, D. T. Schneider, J. Kononen, G. Sauter and J. W. Oosterhuis, *Cancer Res.*, 2003, **63**, 2244–2250.
6 T. Burdon, A. Smith and P. Savatier, *Trends Cell Biol.*, 2002, **12**, 432–438.
7 G. M. Morrison and J. M. Brickman, *Development*, 2006, **133**, 2011–2022.
8 J. Bruix, L. Boix, M. Sala and J. M. Llovet, *Cancer Cell*, 2004, **5**, 215–219.
9 A. J. Harvey, D. R. Armant, B. D. Bavister, S. M. Nichols and C. A. Brenner, *Stem Cells Dev.*, 2009, **18**, 1451–1458.
10 A. A. Avilion, S. K. Nicolis, L. H. Pevny, L. Perez, N. Vivian and R. Lovell-Badge, *Genes Dev.*, 2003, **17**, 126–140.
11 Y. Tokuzawa, E. Kaiho, M. Maruyama, K. Takahashi, K. Mitsui, M. Maeda, H. Niwa and S. Yamanaka, *Mol. Cell. Biol.*, 2003, **23**, 2699–2708.
12 S. Y. Hatano, M. Tada, H. Kimura, S. Yamaguchi, T. Kono, T. Nakano, H. Suemori, N. Nakatsuji and T. Tada, *Mech. Dev.*, 2005, **122**, 67–79.
13 S. Masui, Y. Nakatake, Y. Toyooka, D. Shimosato, R. Yagi, K. Takahashi, H. Okochi, A. Okuda, R. Matoba and A. A. Sharov, *Nat. Cell Biol.*, 2007, **9**, 625–635.
14 Y. Tan, Y. Xue, C. Song and M. Grunstein, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 11493–11498.
15 L. Li, L. Zhang, G. Liu, R. Feng, Y. Jiang, L. Yang, S. Zhang, M. Liao and J. Hua, *PLoS One*, 2014, **9**, e105180.
16 M. E. Kleinman, K. Yamada, A. Takeda, V. Chandrasekaran, M. Nozaki, J. Z. Baffi, R. J. Albuquerque, S. Yamasaki, M. Itaya and Y. Pan, *Nature*, 2008, **452**, 591–597.
17 L. Galli-Taliadoros, J. Sedgwick, S. Wood and H. Körner, *J. Immunol. Methods*, 1995, **181**, 1–15.
18 V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder and P. O. Brown, *Nature*, 2001, **409**, 533–538.
19 B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett and E. Kanin, *Science*, 2000, **290**, 2306–2309.
20 C. Bock, E. Kiskinis, G. Verstappen, H. Gu, G. Boulting, Z. D. Smith, M. Ziller, G. F. Croft, M. W. Amoroso and D. H. Oakley, *Cell*, 2011, **144**, 439–452.
21 M. Ku, R. P. Koche, E. Rheinbay, E. M. Mendenhall, M. Endoh, T. S. Mikkelsen, A. Presser, C. Nusbaum, X. Xie and A. S. Chi, *PLoS Genet.*, 2008, **4**, e1000242.
22 M. Esteller, *Nat. Rev. Genet.*, 2007, **8**, 286–298.
23 L. A. Boyer, T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch and R. A. Young, *Cell*, 2005, **122**, 947–956.
24 I. Ben-Porath, M. W. Thomson, V. J. Carey, R. Ge, G. W. Bell, A. Regev and R. A. Weinberg, *Nat. Genet.*, 2008, **40**, 499–507.
25 J. W. Whitaker, Z. Chen and W. Wang, *Nat. Methods*, 2015, **12**, 265–272, 267 p following 272.
26 X. Xiao, Z. Li, H. Liu, J. Su, F. Wang, X. Wu, H. Liu, Q. Wu and Y. Zhang, *Gene*, 2013, **518**, 425–430.
27 R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren and J. R. Ecker, *Nature*, 2009, **462**, 315–322.
28 B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst,

E. S. Lander, T. S. Mikkelsen and J. A. Thomson, *Nat. Biotechnol.*, 2010, **28**, 1045–1048.

29 R. D. Hawkins, G. C. Hon, L. K. Lee, Q. Ngo, R. Lister, M. Pelizzola, L. E. Edsall, S. Kuan, Y. Luu, S. Klugman, J. Antosiewicz-Bourget, Z. Ye, C. Espinoza, S. Agarwahl, L. Shen, V. Ruotti, W. Wang, R. Stewart, J. A. Thomson, J. R. Ecker and B. Ren, *Cell Stem Cell*, 2010, **6**, 479–491.

30 R. Jothi, S. Cuddapah, A. Barski, K. Cui and K. Zhao, *Nucleic Acids Res.*, 2008, **36**, 5221–5231.

31 Y. Benjamini and Y. Hochberg, *Journal of the Royal Statistical Society.Series B (Methodological)*, 1995, 289–300.

32 A. Chatr-aryamontri, B.-J. Breitkreutz, S. Heinicke, L. Boucher, A. Winter, C. Stark, J. Nixon, L. Ramage, N. Kolas and L. O'Donnell, *Nucleic Acids Res.*, 2013, **41**, D816–D823.

33 T. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen and A. Venugopal, *Nucleic Acids Res.*, 2009, **37**, D767–D772.

34 M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang and T. Ideker, *Bioinformatics*, 2011, **27**, 431–432.

35 Y. Assenov, F. Ramírez, S.-E. Schelhorn, T. Lengauer and M. Albrecht, *Bioinformatics*, 2008, **24**, 282–284.

36 N. T. Doncheva, Y. Assenov, F. S. Domingues and M. Albrecht, *Nat. Protoc.*, 2012, **7**, 670–685.

37 D. L. Broomhead and D. Lowe, *Complex Systems*, 1988, **2**, 321–355.

38 G. O. Consortium, *Nucleic Acids Res.*, 2004, **32**, D258–D261.

39 D. Lin, *An information-theoretic definition of similarity*, Proceedings of the 15th International Conference on Machine Learning, San Francisco, CA, 1998.

40 H. Fröhlich, N. Speer, A. Poustka and T. Beißbarth, *BMC Bioinf.*, 2007, **8**, 166.

41 Q. Hu, J. Fu, B. Luo, M. Huang, W. Guo, Y. Lin, X. Xie and S. Xiao, *Oncol. Rep.*, 2015, **33**, 1965–1975.

42 I. S. Lossos, D. K. Czerwinski, A. A. Alizadeh, M. A. Wechser, R. Tibshirani, D. Botstein and R. Levy, *N. Engl. J. Med.*, 2004, **350**, 1828–1837.

43 L. Fu, J. Shi, K. Hu, J. Wang, W. Wang and X. Ke, *Oncotarget*, 2015, **6**, 8144–8154.

44 N. Whiffin, F. J. Hosking, S. M. Farrington, C. Palles, S. E. Dobbins, L. Zgaga, A. Lloyd, B. Kinnersley, M. Gorman, A. Tenesa, P. Broderick, Y. Wang, E. Barclay, C. Hayward, L. Martin, D. D. Buchanan, A. K. Win, J. Hopper, M. Jenkins, N. M. Lindor, P. A. Newcomb, S. Gallinger, D. Conti, F. Schumacher, G. Casey, T. Liu, H. Campbell, A. Lindblom, R. S. Houlston, I. P. Tomlinson and M. G. Dunlop, *Hum. Mol. Genet.*, 2014, **23**, 4729–4737.

45 D. J. Wong, H. Liu, T. W. Ridky, D. Cassarino, E. Segal and H. Y. Chang, *Cell Stem Cell*, 2008, **2**, 333–344.