IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON FEATURE REPRESENTATION AND LEARNING METHODS WITH APPLICATIONS IN LARGE-SCALE BIOLOGICAL SEQUENCE ANALYSIS

Received August 16, 2019, accepted August 26, 2019, date of publication September 5, 2019, date of current version September 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2939556

Identifying Essential Signature Genes and Expression Rules Associated With Distinctive Development Stages of Early Embryonic Cells

LEI CHEN[1,2,3], XIAOYONG PAN[4,5], TAO ZENG[6], YU-HANG ZHANG[7], TAO HUANG[7], AND YU-DONG CAI[1]

[1]School of Life Sciences, Shanghai University, Shanghai 200444, China
[2]College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
[3]Shanghai Key Laboratory of PMMP, East China Normal University, Shanghai 200241, China
[4]Key Laboratory of System Control and Information Processing, Institute of Image Processing and Pattern Recognition, Ministry of Education of China, Shanghai Jiao Tong University, Shanghai 200240, China
[5]IDLab, Department for Electronics and Information Systems, Ghent University, 9000 Ghent, Belgium
[6]Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China
[7]Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

Corresponding authors: Tao Huang (huangtao@sibs.ac.cn) and Yu-Dong Cai (caiyudong@staff.shu.edu.cn)

This work was supported in part by the Shanghai Municipal Science and Technology Major Project under Grant 2017SHZDZX01, in part by the National Key Research and Development Program of China under Grant 2018YFC0910403, in part by the National Natural Science Foundation of China under Grant 31701151, in part by the Natural Science Foundation of Shanghai under Grant 17ZR1412500, in part by the Shanghai Sailing Program under Grant 16YF1413800, in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) under Grant 2016245, in part by the Fund of the Key Laboratory of Stem Cell Biology of Chinese Academy of Sciences under Grant 201703, and in part by the Science and Technology Commission of Shanghai Municipality (STCSM) under Grant 18dz2271000.

ABSTRACT An embryo develops from a single-celled zygote, which produces a multi-cellular organism by mitosis. Due to the complication of processes and mechanisms, research on embryo cell clusters in different early embryo developmental stages with significant phenotypic differences is still lacking. In this work, we identified some gene characters and expression rules to classify these individual cells using several advanced computational methods. The single cell expression profiles of embryo cells were analyzed by the Monte Carlo feature selection (MCFS) method, resulting in a feature list. Then, the incremental feature selection (IFS) method, incorporating support vector machine (SVM), applied on such list to extract key gene characters. These gene characters include *KHDC1*, *HMGN1*, *DCP*, *GDF9*, *RNF11*, *DNMT3L*, and *CDX1*. Furthermore, a rule learning algorithm, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), was applied to the informative features yielded by MCFS method, producing a group of classification rules. These rules can clearly uncover different expression patterns on cells in different stages. This study provided a group of effective gene signatures and rules for embryo cell subtyping and presented an applicable computational tool to further dig into the regulatory mechanisms of embryo development.

INDEX TERMS Embryo development, single cell, expression pattern, rule, multi-class classification.

I. INTRODUCTION

An embryo develops from a zygote, which is produced by fertilization of female egg cell by the male sperm cell. A multi-cellular organism was produced by mitosis, which is the early developmental period of a multi-cellular diploid eukaryotic organism. The development of zygote into an embryo goes through multiple stages, i.e., blastula, gastrula, and organogenesis. The blastula stage is typically characterized by a fluid-filled cavity, the blastocoel, which is surrounded by blastomeres. In the gastrula stage, blastula cells form two or three tissue layers by coordinated processes of cell division, invasion, and migration. In the organogenesis stage, molecular and cellular interactions between germ layers prompt further differentiation of organ-specific cell types, based on the cells' developmental potentials or competence to respond.

There are many factors affecting embryo development. Tubulins and other cytoskeletal proteins play important roles in the gametogenesis, oocyte maturation, fertilization,

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

and preimplantation stages of embryo development [1]. Protein deubiquitination (DUBs) can influence sperm functions during embryo development and promotes the acquisition of developmental competence during oocyte maturation [2]. Actin nucleator Arp2/3 complex regulates cell division, thereby affecting preimplantation embryo development [3]. Adiponectin regulates metabolic processes involving several signaling molecules, which in turn regulate the female reproductive function that may influence preimplantation embryo development [4]. Moreover, some genes are related to embryo cell development. Kim Y et al. reported that the knockout of bromodomain-containing protein 7 (*BRD7*) induced a retardation in embryo development and mild changes in glucose metabolism [5]. *KPNA7* is an oocyte and cleavage stage embryo-specific karyopherin α subtype; it is required for procine embryo development [6]. As an important cellular regulator of mRNAs, *PTBP1* can influence the alternative splicing profile of a cell to change the stability, location, and translation of its regulated mRNA, which can regulate the embryonic and extra-embryonic structures required for embryonic development before gastrulation [7]. *ING2* regulates chromatin affecting the process of preimplantation embryo development [8]. In addition, *HUWE1* [9], *PS48* [10], and p38 MAPK signaling pathways [11] play important roles in embryo development. Although there have been many experiments that aimed to find these genes related to embryo development, the group of genes related to embryo development has not been completely identified.

The development of early embryonic cells can be divided into 3-, 4-, 5-, 6-, and 7-day stages, in which significant phenotypic differences exist. These cells should have different developmental models and regulatory mechanisms to distinguish developmental phases. Due to the complication of these processes and mechanisms, studies on the embryo cell clusters in different early developmental stages are still lacking. In this work, we found some gene characters and expression rules to classify these individual cells with several advanced computational methods. First, gene features were analyzed by Monte Carlo feature selection (MCFS) [12], generating a feature list. Based on such list, incremental feature selection (IFS) with a classic classification algorithm, support vector machine (SVM) [13], was executed to extract key gene features, with which a SVM classifier was built. Such classifier can efficiently classify embryonic mammalian cells into five types (3-, 4-, 5-, 6-, and 7-day stages). Key gene characters included *KHDC1*, *HMGN1*, *DCP*, *GDF9*, *RNF11*, *DNMT3L*, and *CDX1*. In addition, a rule learning algorithm, Incremental Pruning to Produce Error Reduction (RIPPER) [14], was adopted to produce classification rules based on informative features yielded by MCFS method. These rules were analyzed and they can clearly indicate the different expression patterns on cells in different types. Overall, our study may provide a group of effective gene signatures and rules for embryo cell subtyping and present an applicable computational tool for further digging on the regulatory mechanisms of embryo development.

II. MATERIALS AND METHODS

A. DATA SETS

We downloaded the single cell expression profiles of 26,178 genes in 1,529 embryo cells from ArrayExpress under accession number of E-MTAB-3929 [15]. A total of 81, 190, 377, 415, and 466 embryo cells were present at days 3, 4, 5, 6, and 7, respectively. The gene expression changes of these single cells may reveal mechanisms underlying dynamic embryo development.

B. FEATURE SELECTION

In this study, we wanted to identify the significant genes closely related to embryonic mammalian cells at five different stages. Here, the expression values of genes were used as input features for machine learning models, and a two-step feature selection method was used to select those important genes. MCFS was used to rank the input genes [16]–[21]. MCFS method is a powerful feature selection method and good at analyzing datasets with few samples and high dimensions. Considering that our dataset contained 1,529 samples and each sample was represented by 26,178 features, MCFS is very proper to tackle such dataset. After features were ranked by MCFS method, IFS with SVM was applied to select discriminative genes in order to classify different cells well [22]–[27].

MCFS is a decision tree-based feature selection method. It generates m bootstrap sample sets and n feature subsets, in which the number of features was smaller than the number of original features. The total m^*n decision trees were grown on each combination of the m bootstrap sample sets and n feature subsets. Based on the grown m^*n trees, a relative importance (RI) score was calculated. For each feature, RI was calculated based on how frequent this feature is involved in growing the m^*n trees and in the classification accuracy of individual trees. Accordingly, a feature list can be produced, in which all features are ranked by the decreasing order of their RI scores. MCFS implementation [12] was downloaded from http://www.ipipan.eu/staff/m.draminski/mcfs.html. Obtained MCFS program was performed with its default parameters, where two main parameters u and v were set to one.

After obtaining the ranked feature list, we further used IFS with a supervised classifier (i.e., SVM) to select the optimum features for distinguishing different cells. We first constructed a series of feature subsets with a step interval of 10 from the ranked feature list by MCFS. In each feature subset, there are 10 more features than the preceding feature subset. For example, the first feature subset has top 10 features. Thus, the second feature subset should have top 20 features, and so on. For each feature subset, we trained and evaluated SVM on the samples consisting of the features from this feature subset using 10-fold cross-validation. After running this process for all feature subsets, we selected that with the best performance, and the features in this subset are called optimum features.

C. SUPPORT VECTOR MACHINE

SVM is a widely used supervised classifier based on statistics theory [13], [28], [29]. It tries to find a hyperplane with maximum margin between two classes and can handle both linear and non-linear data. In non-linear data, it maps the data in low dimensional space into a high-dimensional space using a kernel function, and a linear model is fitted on the new data in the high-dimensional space.

In general, SVM is designed for binary classification. In this study, we needed to classify samples from five groups/stages of cells that can be formulated as a multi-class classification problem. One-vs-rest strategy was applied to adapt the SVM to handle multi-class classification. Multiple binary SVMs were trained, and each binary SVM was trained on positive samples from one class and negative samples from the other classes. Given a new sample, the multiple SVMs corresponding to individual classes will predict a probability score, and then, the predicted class with the highest probability score will be assigned for this new sample.

In this study, we adopted the tool 'SMO' in Weka, which implements one type of SVM. It was executed using its default parameters. In detail, the kernel was a polynomial function and parameter c was set to 1.0.

D. RULE LEARNING

To better understand how the predictors (optimum genes) make decision on the types of cells, we used RIPPER [14], [30] to learn the decision rules from the training data. RIPPER is based on the separate-and-conquer technique and reduced error pruning strategy. It obtained a good rule that fit some samples in the training data, and the samples covered by this rule were removed from the training data. Next, the abovementioned process of rule generation was repeated until all samples were removed from the training data or until other predefined conditions were met. Lastly, reduced error pruning was applied to reduce the redundancy of learned rules. Each rule consisted of IF-THEN statement, i.e., if the conditions are met, a decision is made. In this study, we produced the rules in the following form: IF gene1 $\geq$ 2.4 AND gene2 $\leq$ 10.8, THEN cell type stage $=$ 5 days. Because there were lots of gene features used to represent embryo cells, inducing difficulties to produce brief rules via RIPPER, informative features yielded by the MCFS method were fed into RIPPER to generate rules. The implementation of RIPPER was included in the MCFS package.

E. PERFORMANCE MEASUREMENT

In this study, we classified the samples into five cells at different stages, thereby resulting in multi-class classification. To objectively evaluate the performance of trained models, we measured the overall accuracy and Matthews correlation coefficient (MCC) [31]–[35]. Defining X as the binary matrix of the predicted class labels and Y as the binary matrix of the true class labels, MCC was calculated as follows:

$$MCC = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{cov}(X,X)\mathrm{cov}(Y,Y)}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{C}(x_{ij}-\bar{x}_j)(y_{ij}-\bar{y}_j)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{C}(x_{ij}-\bar{x}_j)^2\sum_{i=1}^{n}\sum_{j=1}^{C}(y_{ij}-\bar{y}_j)^2}} \quad (1)$$

where $\bar{x}_j$ and $\bar{y}_j$ are the mean values in the j^{th} column of X and j^{th} column of Y, respectively.

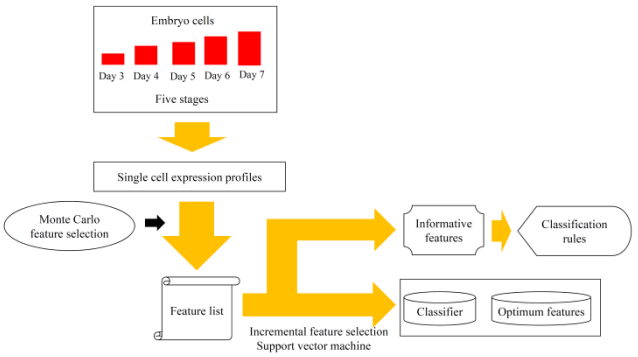


FIGURE 1. The entire procedures to analyze the single cell expression profiles of embryo cells in different developmental stages.

III. RESULTS

In this study, some advanced computational methods were adopted to analyze the single cell expression profiles of embryo cells in different developmental stages. The entire procedures are illustrated in **Fig. 1**.

It is clear that not all genes were equally important in classifying samples from five embryonic cell stages. We first used MCFS to rank all the input features/genes, which were the expression values across 26,178 genes. The RI scores from MCFS for individual features and the feature list are given in **Table S1**.

To further select the discriminating features corresponding to different embryonic cell stages, we used IFS with SVM to classify the samples consisting of features from individually generated feature subsets based on the ranked feature list by MCFS. As shown in **Fig. 2** and **Table 1**, we achieved the best 10-fold cross-validation MCC value (0.996) when the top 3230 features were used. In addition, we can still yield an MCC value and overall accuracy of 0.908 and 0.931, respectively, if less number of features was used (i.e., the top 80 features). The corresponding performance of SVM using different number of features is given in **Table S2**. Our trained SVM was close to perfect in classifying the samples from five embryonic cell stages.

We justified the choice of SVM integrated with IFS by evaluating another widely used classifier, random forest (RF) [36], in the same way as SVM. This study employed the tool 'RandomForest' in Weka, which implements RF. As shown in

FIGURE 2. Optimal performances of IFS with SVM and RF.

TABLE 1. The performance and optimum number of features of IFS with the SVM and RF.

Classifier	Number of optimum features	MCC	Overall accuracy
SVM	3230	0.996	0.997
RF	960	0.948	0.961

Fig. 2 and **Table 1**, we achieved the best MCC value (0.948) if the top 960 features were used, with an overall accuracy of 0.961. The corresponding performance of RF using different number of features is given in **Table S3**. The performance of SVM is slightly better than that of RF. As such, we chose SVM over RF for IFS in this work.

SVM and RF are both black-box classifiers. To better understand how we can classify samples from five embryonic cell stages, we used RIPPER to learn the classification rules. The MCFS method produced 542 informative features, which were top 542 features in the ranked feature list. Based on these informative features, RIPPER produced 18 significant classification rules, as shown in **Table 2**, but we still needed to check their accuracy. Thus, we re-evaluated the prediction performance of these 18 classification rules using 10-fold cross-validation. We achieved an accuracy and overall accuracy of 0.906 and 0.918, respectively (**Fig. 3**). The learned 18 rules can classify samples from five embryonic cell stages with high accuracy.

IV. DISCUSSION

As mentioned above, 3230 features have been screened to describe the optimal classifier for early embryo cells subtyping. The expression levels of 3230 genes are screened for optimal features that contribute to early embryonic cell development. Considering the number of features, the top-ranked 80 genes were selected as the optimal genes

TABLE 2. The 18 classification rules for different cell stages learned by RIPPER algorithm.

Rules	Criteria	Embryonic cell stage
1	NME1 <= 82.158	3 day
2	ARID4A >= 37.482	3 day
3	(TPM4 <= 5.290) and (PRDX1 <= 3747.014)	4 day
4	(ETNPPL >= 99.665) and (ATP5A1 <= 509.291)	4 day
5	(KRT19 <= 171.765) and (SLC17A5 >= 22.892) and (FTL <= 816.657)	5 day
6	(CCKBR <= 15.962) and (CLDN10 >= 188.127) and (FABP3 <= 619.660)	5 day
7	(PRSS23 >= 61.762) and (MTRNR2L1 >= 470.139)	5 day
8	(KHDC1L >= 5254.623) and (S100A14 >= 724.067)	5 day
9	(CLDN4 <= 18.124) and (AK2 <= 34.609)	5 day
10	VCX >= 47.706	5 day
11	(CLDN10 >= 59.215) and (ATP5I >= 1559.751)	6 day
12	(FTL <= 1461.441) and (HMGCS1 >= 57.922) and (FADS2 <= 17.751)	6 day
13	(CGA <= 0) and (HINT1 <= 755.970)	6 day
14	(RGS13 >= 22.417) and (ERH >= 537.260)	6 day
15	HTR3B >= 0.112	6 day
16	(RPL9 >= 162.532) and (ACSL4 >= 5.347)	6 day
17	CTSV <= 13.020	6 day
18	others	7 day

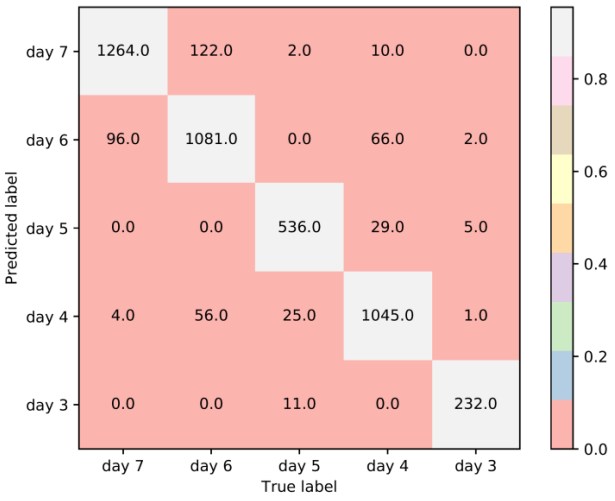


FIGURE 3. Confusion matrix of multi-classification model.

as supported by previous analysis, and we discussed the top-ranked 10 features as examples. In addition, we identified

a group of quantitative rules that contributed to the detailed classification of five different stages of embryonic cells, thereby providing a solid basic research.

A. GENES ASSOCIATED WITH EMBRYO DEVELOPMENT

The first identified gene in our optimal prediction list is *KHDC1L*. According to previous reports [37], *KHDC1* is expressed in oocyte and embryo cells and is related to atypical structure and phylogenomic evolution. Its family members are known to be KH-domain-containing RNA-binding proteins, which are highly expressed in oocytes and are unique to eutherian mammals [38]. In 2010, it is reported that *KHDC1B*, a novel CPEB-binding partner interacting with mCEPB1 and regulating oocyte maturation, is a member of a small family of KH-domain-containing proteins that have been identified in oocytes and early embryos [39]. Based on these evidences, we speculated that *KHDC1* may have a differential expression pattern in early embryonic cells, validating the efficacy and accuracy of our prediction.

The next gene in our prediction list, *HMGN1*, encodes the chromosomal protein HMG14, which helps maintain transcribable genes in a unique chromatin conformation. *HMGN1* can regulate embryogenesis by modulating *Sox9* expression and enhance the rate of DNA repair in chromatin [40], [41]. Chromatin structure plays a key role in regulating gene expression and embryonic differentiation. Thus, *HMGN1*, as a nucleosome-binding protein that is ubiquitously expressed in vertebrate cells; it regulates embryonic stem cells by modulating nucleosome occupancy [42]. Therefore, we speculated that *HMGN1* gene may be differentially expressed in embryonic cells, validating the efficacy and accuracy of our prediction.

The next gene in our prediction list is *DCP-1*. During meiotic maturation, *DCP-1* plays an important role in the regulation of mRNA stability, especially for proper mRNA degradation [43]. During oocyte maturation through the two-cell stage, the degradation of selected maternal transcripts is dramatic, which accounts for approximately 20% (total) RNA decrease in the oocyte [44]. The inhibition of increased *DCP1A* expression associated with maturation could prevent bulk degradation of maternal mRNAs and could affect embryonic genome activation during the two-cell stage. Therefore, *DCP-1* gene might be differentially expressed in embryonic cells.

The next gene, growth differentiation factor 9 (*GDF9*), is predicted to contribute to classify early embryonic cells. As a member of the transforming growth factor-beta (TGFβ) superfamily synthesized by ovarian somatic cells, it directly affects oocyte growth and function and is expressed in oocytes. Its expression level is closely associated with oocyte maturation, fertilization, embryo quality, and pregnancy outcome [45], [46], thereby suggesting its ability as a marker and in identifying embryonic cells. In a published report, *GDF-9* mRNA can be detected up to eight-cell stage in embryo but not in preantral follicles and early antral follicles [47]. Therefore, *GDF9*, a high-ranking predicted gene, has been confirmed to have differential expression pattern that contribute in embryonic cells.

RNF11, another gene identified in our predicted list, contributes to ubiquitination regulation. It is specifically present in early embryonic cells. Its transcripts are specifically present in presomatic mesoderm (PSM), and later in the brain and retina [48]. Another publication reported that it directly enhances TGFβ signaling by direct association with *Smad4*, which is a well-known signaling signal transducer and transcription factor that regulates TGFβ, BMP, and Activin pathways. Although it is associated with *Smad4* and other transcription factors, it may play a role in direct transcriptional regulation [49]. *Smad4* potentiates a subset of TGFβ-related signals during early embryonic development [50], thereby suggesting that it can regulate early embryonic development by affecting *Smad4*. These literature reports support that our predicted gene *RNF11* may have differential expression pattern in early embryonic cells, validating the efficacy and accuracy of our prediction.

DNMT3L, another gene identified in our predicted list, functions in CpG methylation that is an epigenetic modification relevant to embryonic development, imprinting, and X-chromosome inactivation. DNMT3 is the major DNA methyl-transferase expressed in gonocytes and is increased in spermatogonia at four and six days postpartum [51]. Another report showed that genomic DNA is methylated by de novo methyltransferases, Dnmt3a and Dnmt3b, during early embryonic development. The activity of both enzymes increases in the presence of Dnmt3L, a Dnmt3a/3b-like protein [52]. According to the results in these literatures, *DNMT3L* regulates embryonic development by increasing the expression of *Dnmt3a/3b* at days 4 and 6 to regulate CpG methylation, thereby strengthening our speculation that *DNMT3L* has a differential expression pattern in early embryonic cells.

The final gene to be discussed is *CDX1*, also known as homeobox protein CDX-1, which is expressed in the developing endoderm. *CDX1* is regulated by Wnt-3a, which is an important factor for somite specification along the anteroposterior axis of the embryo [53]. It has an early period of expression when the embryonic body axis is established, and its expression is maintained throughout adulthood in the proliferative cell [54]. Hox genes regulate axial extension and may be significant in mammalian embryo development. They regulate embryonic positional identities and are expressed in early embryonic cells [55]. Anther CDX gene, *CDX2*, is also expressed in the different stages and positions of embryonic cells [56], thereby suggesting that *CDX1* and its homologs are specifically expressed in embryonic cells. We speculated that the predicted gene *CDX1* has differential expression pattern in early embryonic cells.

B. RULES ASSOCIATED WITH EMBRYO DEVELOPMENT

Apart from the qualitative analyses of optimal genes and their expression patterns associated with embryo development, we identified 18 rules for detailed quantitative analysis on

the distinction of embryonic cells. According to recent publications, all expression tendencies have been confirmed. To validate the threshold parameters in these rules, we identified quantitative proofs from existing databases such as GEO. The detailed analysis of each rule can be seen below.

The first two rules are about distinguishing the three-day embryonic cells. According to our prediction, a relatively high expression level of *ARID4A* or a relatively low expression level of *NME1* may indicate the three-day stage of embryonic cells. *ARID4A* can regulate genomic imprinting, which is an important factor for embryo development [57]. *NME1* has not been found in embryonic development, suggesting that it may be a negative factor. Therefore, we regarded the non-detection of *NME1* expression and high detection of *ARID4A* expression as potential parameters for distinction of embryonic cells in the three-day stage.

The second two rules are related to four-day embryonic cells. According to our prediction, these rules are as follows: the relatively low expression levels of *TPM4* and *PRDX1* and the relatively high expression level of *ETNPPL* and a relatively low expression level of *ATP5A1*. It has been proven that *PRDX1* exists in embryos [58] and *ATP5A1* gene mutation leads to embryonic lethality [59], thereby suggesting that *PRDX1* and *ATP5A1* are important factors in embryo development. We regarded these genes as potential parameters for the distinction of embryonic cells.

The next six rules are about distinguishing the five-day embryonic cells. According to our prediction, these rules are as follows: 1) relatively low expression levels of *KRT19* and *FTL* as well as a relatively high expression level of *SLC17A5*; 2) relatively low expression levels of *CCKBR* and *FABP3* as well as a relatively high expression level of *CLDN10*; 3) relatively high expression levels of *PRSS23* and *MTRNR2L1*; 4) relatively high expression levels of *KHDC1L* and *S100A14*; 5) relatively low expression levels of *CLDN4* and *AK2*; and 6) a relatively high expression level of *VCX*. Although *FTL* and *KRT19* were not found in embryonic development, gene *SLC17A5* or sialin, is related to embryonic expression patterns [60]. According to published reports, many gene rules can be regarded as essential parameters in the distinction of embryonic cells. *CLDN10* is essential in blastocyst formation in preimplantation embryos [61]. *PRSS23* is essential in blastocyst development and hatching [62]. *MTRNR2L1* is important in human preimplantation epiblast [63]. *KHDC1L*, as an embryo-expressed gene, is a novel CPEB-binding partner that is specifically expressed in oocytes and early embryos [37], [39].

The following seven rules are associated with six-day embryonic cells: 1) relatively high expression levels of *CLDN10* and *ATP5I*; 2) relatively low expression levels of *FTL* and *FADS2* as well as a relatively high expression level of *HMGCS1*; 3) a relatively low expression level of *HINT1* and non-expression of *CGA*; 4) relatively high expression levels of *RGS13* and *ERH*; 5) a relatively high expression level of *HTR3B*; 6) relatively high expression levels of *RPL9* and *ACSL4*; and 7) a relatively low expression level of *CTSV*. *HMGCS1* is an essential factor during oocyte maturation and preimplantation embryo development [64]. *ERH*, a kind of embryonic RNA helicase gene, influences embryonic development by regulating RNA helicases [65]. *CTSV* gene is a predictor of human blastocyst hatching and is related to the launch of one of the direct hatching mechanisms [66]. *CGA* has not been found in recent reports, thereby suggesting that it is a negative factor for embryo development.

The last rule in our prediction list involves gene *CTSV*. As previously analyzed, a relatively lower expression level of *CTSV* turns to be the 6-day and 7-day embryonic cells.

We analyzed embryonic mammalian cells at two different levels. All the qualitatively analyzed genes have been confirmed to contribute to the distinction of embryonic cells. Moreover, most learned expression rules are supported by recent literature. Therefore, our newly presented computational approach identifies potential cell-signature genes and cell-cluster rules for different stages of embryogenesis and is significant for further research on the underlying mechanisms in early-stage embryonic cells.

ACKNOWLDEGMENT

(Lei Chen, Xiaoyong Pan, and Tao Zeng contributed equally to this work.) Supporting information: Table S1: Top features with their importance scores calculated by MCFS; Table S2: The 10-fold cross-validation performances of IFS with SVM; Table S3: The 10-fold cross-validation performances of IFS with RF.

REFERENCES

[1] H. Schatten and Q.-Y. Sun, *Posttranslationally Modified Tubulins and Other Cytoskeletal Proteins: Their Role in Gametogenesis, Oocyte Maturation, Fertilization and Pre-implantation Embryo Development*. New York, NY, USA: Springer, 2014.

[2] Y. J. Yi, M. Sutovsky, W.-H. Song, and P. Sutovsky, "Protein deubiquitination during oocyte maturation influences sperm function during fertilisation, antipolyspermy defense and embryo development," *Reprod. Fertility Develop.*, vol. 27, no. 8, pp. 1154–1167, 2015.

[3] S.-C. Sun, Q.-L. Wang, W.-W. Gao, Y.-N. Xu, H.-L. Liu, X.-S. Cui, and N.-H. Kim, "Actin nucleator Arp2/3 complex is essential for mouse preimplantation embryo development," *Reprod., Fertility Develop.*, vol. 25, no. 4, pp. 617–623, 2013.

[4] Š. Čikoš, "Chapter nine—Adiponectin and its receptors in preimplantation embryo development," in *Vitamins & Hormones*, vol. 90. New York, NY, USA: Academic, 2012, pp. 211–238.

[5] Y. Kim, M. A. S. Hernández, H. Herrema, T. Delibasi, and S. W. Park, "The role of BRD7 in embryo development and glucose metabolism," *J. Cellular Mol. Med.*, vol. 20, no. 8, pp. 1561–1570, Aug. 2016.

[6] X. Wang, K.-E. Park, S. Koser, S. Liu, L. Magnani, and R. A. Cabot, "*KPNA7*, an oocyte- and embryo-specific karyopherin α subtype, is required for porcine embryo development," *Reprod. Fertility Develop.*, vol. 24, no. 2, pp. 382–391, 2012.

[7] J. Suckale, O. Wendling, J. Masjkur, M. Jäger, C. Münster, K. Anastassiadis, A. F. Stewart, and M. Solimena, "PTBP1 is required for embryonic development before gastrulation," *PLoS ONE*, vol. 6, no. 2, 2011, Art. no. e16992.

[8] Z. Lin, W. Pei, J. Zhang, B. C. Heng, and Q. T. Guo, "ING2 (inhibitor of growth protein-2) plays a crucial role in preimplantation development," *Zygote*, vol. 24, pp. 89–97, Feb. 2016.

[9] L. J. Chen, W. M. Xu, M. Yang, K. Wang, Y. Chen, X. J. Huang, and Q. H. Ma, "HUWE1 plays important role in mouse preimplantation embryo development and the dysregulation is associated with poor embryo development in humans," *Sci. Rep.*, vol. 6, Nov. 2016, Art. no. 37928.

[10] L. D. Spate, A. Brown, B. K. Redel, K. M. Whitworth, and R. S. Prather, "PS48 can replace bovine serum albumin in pig embryo culture medium, and improve *in vitro* embryo development by phosphorylating AKT," *Mol. Reprod. Develop.*, vol. 82, no. 4, pp. 315–320, 2015.

[11] B. Sozen, S. Ozturk, A. Yaba, and N. Demir, "The p38 MAPK signalling pathway is required for glucose metabolism, lineage specification and embryo survival during mouse preimplantation development," *Mech. Develop.*, vol. 138, pp. 375–398, Nov. 2015.

[12] M. Draminski, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, and J. Komorowski, "Monte Carlo feature selection for supervised classification," *Bioinformatics*, vol. 24, pp. 110–117, Jan. 2008.

[13] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[14] W. W. Cohen, "Fast effective rule induction," in *Proc. 12th Int. Conf. Mach. Learn.*, 1995, pp. 115–123.

[15] S. Petropoulos, D. Edsgärd, B. Reinius, Q. Deng, S. P. Panula, S. Codeluppi, A. P. Reyes, S. Linnarsson, R. Sandberg, and F. Lanner, "Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos," *Cell*, vol. 165, pp. 1012–1026, May 2016.

[16] X. Pan, X. Hu, Y.-H. Zhang, L. Chen, L. Zhu, S. Wan, T. Huang, and Y.-D. Cai, "Identification of the copy number variant biomarkers for breast cancer subtypes," *Mol. Genet. Genomics*, vol. 294, no. 1, pp. 95–110, Sep. 2019.

[17] X. Pan, L. Chen, K.-Y. Feng, X.-H. Hu, Y.-H. Zhang, X.-Y. Kong, T. Huang, and Y.-D. Cai, "Analysis of expression pattern of snoRNAs in different cancer types with machine learning algorithms," *Int. J. Mol. Sci.*, vol. 20, p. 2185, May 2019.

[18] J. Li, L. Lu, Y.-H. Zhang, Y. Xu, M. Liu, K. Feng, L. Chen, and X. Kong, "Identification of leukemia stem cell expression signatures through Monte Carlo feature selection strategy and support vector machine," *Cancer Gene Therapy*, May 2019.

[19] L. Chen, X. Pan, Y.-H. Zhang, X. Kong, T. Huang, and Y.-D. Cai, "Tissue differences revealed by gene expression profiles of various cell lines," *J. Cellular Biochem.*, vol. 120, pp. 7068–7081, May 2019.

[20] X. Pan, X. Hu, Y. H. Zhang, K. Feng, S. P. Wang, L. Chen, T. Huang, and Y. D. Cai, "Identifying patients with atrioventricular septal defect in down syndrome populations by using self-normalizing neural networks and feature selection," *Genes*, vol. 9, no. 4, p. 208, 2018.

[21] L. Chen, J. Li, Y.-H. Zhang, K. Feng, S. Wang, Y. Zhang, T. Huang, X. Kong, and Y.-D. Cai, "Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method," *J. Cell Biochem.*, vol. 119, pp. 3394–3403, Apr. 2018.

[22] T.-M. Zhang, T. Huang, and R.-F. Wang, "Cross talk of chromosome instability, CpG island methylator phenotype and mismatch repair in colorectal cancer," *Oncol. Lett.*, vol. 16, pp. 1736–1746, Aug. 2018.

[23] J. Li and T. Huang, "Predicting and analyzing early wake-up associated gene expressions by integrating GWAS and eQTL studies," *Biochim. Biophys. Acta Mol. Basis Disease*, vol. 1864, pp. 2241–2246, Jun. 2018.

[24] L. Chen, S. Wang, Y.-H. Zhang, J. Li, Z.-H. Xing, J. Yang, T. Huang, and Y.-D. Cai, "Identify key sequence features to improve CRISPR sgRNA efficacy," *IEEE Access*, vol. 5, pp. 26582–26590, 2017.

[25] Y. Zhou, T. Huang, G. Huang, N. Zhang, X. Kong, and Y.-D. Cai, "Prediction of protein N-formylation and comparison with N-acetylation based on a feature selection method," *Neurocomputing*, vol. 217, pp. 53–62, Dec. 2016.

[26] N. Zhang, M. Wang, P. Zhang, and T. Huang, "Classification of cancers based on copy number variation landscapes," *Biochim. Biophys. Acta*, vol. 1860, no. 11, pp. 2750–2755, Nov. 2016.

[27] J. Li, C.-N. Lan, Y. Kong, S.-S. Feng, and T. Huang, "Identification and analysis of blood gene expression signature for osteoarthritis with advanced feature selection methods," *Frontiers Genet.*, vol. 9, p. 246, Aug. 2018.

[28] H. Cui and L. Chen, "A binary classifier for the prediction of EC numbers of enzymes," *Current Proteomics*, vol. 16, no. 5, pp. 381–389, 2019.

[29] L. Chen, X. Pan, X. Hu, Y.-H. Zhang, S. Wang, Tao Huang, and Y.-D. Cai, "Gene expression differences among different MSI statuses in colorectal cancer," *Int. J. Cancer*, vol. 143, no. 7, pp. 1731–1740, Oct. 2018.

[30] L. Chen, Y.-H. Zhang, X. Pan, M. Liu, S. Wang, T. Huang, and Y.-D. Cai, "Tissue expression difference between mRNAs and lncRNAs," *Int. J. Mol. Sci.*, vol. 19, no. 11, p. 3416, 2018.

[31] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta-Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.

[32] J. Gorodkin, "Comparing two K-category assignments by a K-category correlation coefficient," *Comput. Biol. Chem.*, vol. 28, nos. 5–6, pp. 367–374, Dec. 2004.

[33] X. Zhao, L. Chen, and J. Lu, "A similarity-based method for prediction of drug side effects with heterogeneous information," *Math. Biosci.*, vol. 306, pp. 136–144, Dec. 2018.

[34] X. Zhao, L. Chen, Z.-H. Guo, and T. Liu, "Predicting drug side effects with compact integration of heterogeneous networks," *Current Bioinform.*, Aug. 2019.

[35] L. Chen, C. Chu, Y.-H. Zhang, M. Zheng, L. Zhu, X. Kong, and T. Huang, "Identification of drug-drug interactions using chemical interactions," *Current Bioinform.*, vol. 12, no. 6, pp. 526–534, 2017.

[36] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[37] A. Pierre, M. Gautier, I. Callebaut, M. Bontoux, E. Jeanpierre, P. Pontarotti, and P. Monget, "Atypical structure and phylogenomic evolution of the new eutherian oocyte- and embryo-expressed *KHDC1/DPPA5/ECAT1/OOEP* gene family," *Genomics*, vol. 90, no. 5, pp. 583–594, Nov. 2007.

[38] C. Cai, J. Liu, C. Wang, and J. Shen, "KHDC1A, a novel translational repressor, induces endoplasmic reticulum-dependent apoptosis," *Dna Cell Biol.*, vol. 31, no. 9, pp. 1447–1457, 2012.

[39] C. Cai, K. Tamai, and K. Molyneaux, "KHDC1B is a novel CPEB binding partner specifically expressed in mouse oocytes and early embryos," *Mol. Biol. Cell*, vol. 21, no. 18, p. 3137, 2010.

[40] T. Furusawa, J.-H. Lim, F. Catez, Y. Birger, S. Mackem, and M. Bustin, "Down-regulation of nucleosomal binding protein HMGN1 expression during embryogenesis modulates *Sox9* expression in chondrocytes," *Mol. Cell Biol.*, vol. 26, pp. 592–604, Jan. 2006.

[41] Y. Birger, K. L. West, Y. V. Postnikov, J.-H. Lim, T. Furusawa, J. P. Wagner, C. S. Laufer, K. H. Kraemer, and M. Bustin, "Chromosomal protein HMGN1 enhances the rate of DNA repair in chromatin," *EMBO J.*, vol. 22, no. 7, pp. 1665–1675, 2014.

[42] T. Deng, Z. I. Zhu, S. Zhang, F. Leng, S. Cherukuri, L. Hansen, L. Mariño-Ramírez, E. Meshorer, D. Landsman, and M. Bustin, "HMGN1 modulates nucleosome occupancy and DNase I hypersensitivity at the CpG island promoters of embryonic stem cells," *Mol. Cellular Biol.*, vol. 33, no. 16, pp. 3377–3389, 2013.

[43] J. Ma, M. Flemr, H. Strnad, P. Svoboda, and R. M. Schultz, "Maternally recruited DCP1A and DCP2 contribute to messenger RNA degradation during oocyte maturation and genome activation in mouse," *Biol. Reprod.*, vol. 88, no. 1, p. 11, 2013.

[44] L. M. Mehlmann, "Losing Mom's message: Requirement for DCP1A and DCP2 in the degradation of maternal transcripts during oocyte maturation," *Biol. Reprod.*, vol. 88, no. 1, p. 10, 2013.

[45] Y. Li, R.-Q. Li, S.-B. Ou, N.-F. Zhang, L. Ren, L.-N. Wei, Q.-X. Zhang, and D.-Z. Yang, "Increased GDF9 and BMP15 mRNA levels in cumulus granulosa cells correlate with oocyte maturation, fertilization, and embryo quality in humans," *Reprod. Biol. Endocrinol.*, vol. 12, pp. 1–9, Aug. 2014.

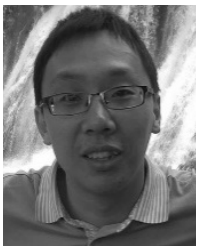
[46] L. M. Salvador, C. P. Silva, I. Kostetskii, G. L. Radice, and J. F. Strauss, III, "The promoter of the oocyte-specific gene, *Gdf9*, is active in population of cultured mouse embryonic stem cells with an oocyte-like phenotype," *Methods*, vol. 45, no. 2, pp. 172–181, 2008.

[47] Y. Sendai, T. Itoh, S. Yamashita, and H. Hoshi, "Molecular cloning of a cDNA encoding a bovine growth differentiation factor-9 (GDF-9) and expression of GDF-9 in bovine ovarian oocytes and *in vitro*-produced embryos," *Cloning*, vol. 3, no. 1, pp. 3–10, 2001.

[48] S. Maddirevula, M. Anuppalle, T.-L. Huh, S. H. Kim, and M. Rhee, "Rnf11-like is a novel component of NF-?B signaling, governing the posterior patterning in the zebrafish embryos," *Biochem. Biophys. Res. Commun.*, vol. 422, no. 4, pp. 602–606, 2012.

[49] A. Peter and S. Arun, "RNF11 is a multifunctional modulator of growth factor receptor signalling and transcriptional regulation," *Eur. J. Cancer*, vol. 41, no. 16, pp. 2549–2560, 2005.

[50] G. C. Chu, N. R. Dunn, D. C. Anderson, L. Oxburgh, and E. J. Robertson, "Differential requirements for *Smad4* in TGF β-dependent patterning of the early mouse embryo," *Development*, vol. 131, pp. 3501–3512, Apr. 2004.
[51] Y. Sakai, I. Suetake, F. Shinozaki, S. Yamashina, and S. Tajima, "Co-expression of de novo DNA methyltransferases Dnmt3a2 and Dnmt3L in gonocytes of mouse embryos," *Gene Expression Patterns*, vol. 5, no. 2, pp. 231–237, Dec. 2004.
[52] Y.-G. Hu, R. Hirasawa, J.-L. Hu, K. Hata, C.-L. Li, Y. Jin, T. Chen, E. Li, M. Rigolet, E. Viegas-Péquignot, and H. Sasaki, "Regulation of DNA methylation activity through *Dnmt3L* promoter methylation by Dnmt3 enzymes in embryonic development," *Hum. Mol. Genet.*, vol. 17, no. 17, pp. 2654–2664, 2008.
[53] M. Ikeya and S. Takada, "Wnt-3a is required for somite specification along the anteroposterior axis of the mouse embryo and for regulation of *cdx-1* expression," *Mech. Develop.*, vol. 103, nos. 1–2, pp. 27–33, 2001.
[54] H. Lickert, C. Domon, G. Huls, C. Wehrle, I. Duluc, H. Clevers, B. I. Meyer, J. N. Freund, and R. Kemler, "Wnt/(beta)-catenin signaling regulates the expression of the homeobox gene *Cdx1* in embryonic intestine," *Development*, vol. 127, no. 17, pp. 3805–3813, 2000.
[55] T. Young, J. E. Rowland, C. van de Ven, M. Bialecka, A. Novoa, M. Carapuco, J. van Nes, W. de Graaff, I. Duluc, J.-N. Freund, F. Beck, M. Mallo, and J. Deschamps, "*Cdx* and *Hox* genes differentially regulate posterior axial growth in mammalian embryos," *Develop. Cell*, vol. 17, no. 4, pp. 516–526, 2009.
[56] F. Beck, T. Erler, A. Russell, and R. James, "Expression of *Cdx-2* in the mouse embryo and placenta: Possible role in patterning of the extra-embryonic membranes," *Developmental Dyn.*, vol. 204, no. 3, pp. 219–227, 2010.
[57] S. Kacem and R. Feil, "Chromatin mechanisms in genomic imprinting," *Mammalian Genome*, vol. 20, nos. 9–10, pp. 544–556, 2009.
[58] A. R. Moawad, B. Xu, S. L. Tan, and T. Taketo, "L-carnitine supplementation during vitrification of mouse germinal vesicle stage–oocytes and their subsequent *in vitro* maturation improves meiotic spindle configuration and mitochondrial distribution in metaphase II oocytes," *Hum. Reprod.*, vol. 29, no. 10, pp. 2256–2268, 2014.
[59] A. A. Baran, K. A. Silverman, J. Zeskand, R. Koratkar, A. Palmer, K. McCullen, W. J. Curran, Jr., T. B. Edmonston, L. D. Siracusa, and A. M. Buchberg, "The modifier of *Min 2 (Mom2)* locus: Embryonic lethality of a mutation in the *Atp5a1* gene suggests a novel mechanism of polyp suppression," *Genome Res.*, vol. 17, pp. 566–576, Mar. 2007.
[60] B. Laridon, P. Callaerts, and K. Norga, "Embryonic expression patterns of Drosophila ACS family genes related to the human sialin gene," *Gene Expression Patterns*, vol. 8, no. 4, pp. 275–283, 2008.
[61] K. Moriwaki, S. Tsukita, and M. Furuse, "Tight junctions containing claudin 4 and 6 are essential for blastocyst formation in preimplantation mouse embryos," *Developmental Biol.*, vol. 312, no. 2, pp. 509–522, 2007.
[62] H. W. Chen and C. R. Tzeng, "The novel serine proteases, PRTN3 and PRSS23, in murine blastocyst development and hatching," *Fertility Sterility*, vol. 88, no. 1, p. S312, 2007.
[63] G. G. Stirparo, T. Boroviak, G. Guo, J. Nichols, A. Smith, and P. Bertone, "Integrated analysis of single-cell embryo data yields a unified transcriptome signature for the human preimplantation epiblast," *Development*, vol. 145, Feb. 2018, Art. no. 158501.
[64] W. S. A. El Naby, T. H. Hagos, M. M. Hossain, D. Salilew-Wondim, A. Y. Gad, F. Rings, M. U. Cinar, E. Tholen, C. Looft, K. Schellander, M. Hoelker, and D. Tesfaye, "Expression analysis of regulatory microRNAs in bovine cumulus oocyte complex and preimplantation embryos," *Zygote*, vol. 21, no. 1, pp. 31–51, 2013.
[65] J. Sowden, W. Putt, K. Morrison, R. Beddington, and Y. Edwards, "The embryonic RNA helicase gene (ERH): A new member of the DEAD box family of RNA helicases," *Biochem. J.*, vol. 308, no. 3, pp. 839–846, 1995.
[66] A. G. Syrkasheva, N. V. Dolgushina, A. Y. Romanov, O. V. Burmenskaya, N. P. Makarova, E. O. Ibragimova, E. A. Kalinina, and G. T. Sukhikh, "Cell and genetic predictors of human blastocyst hatching success in assisted reproduction," *Zygote*, vol. 25, no. 5, pp. 631–636, Oct. 2017.



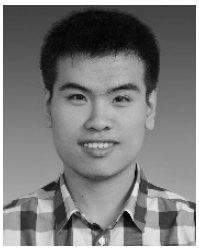
LEI CHEN received the B.S. degree in mathematics, the M.S. degree in operational researches, and the Ph.D. degree in system analysis and integration from East China Normal University, in 2004, 2007 and 2010, respectively, where he moved to the Software Engineering Institute, in 2007, to study computer science.

In 2010, he joined the College of Information Engineering, Shanghai Maritime University, where he is currently an Associate Professor. His interests include bioinformatics, computational biology, graph theory, and algorithm design.

Dr. Chen is a member of the China Computer Federation and the Chinese Association for Artificial Intelligence. He is the Editorial Board Member of *Current Bioinformatics* and *Current Proteomics*, and the Section Editor of *Combinatorial Chemistry & High Throughput Screening*.

XIAOYONG PAN received the master's degree from Shanghai Jiao Tong University, in 2011, and the Ph.D. degree in bioinformatics from Copenhagen University, Denmark, in 2017. He held a postdoctoral position with the Erasmus Medical Center, Rotterdam, The Netherlands, from 2016 to 2018. He is currently an Assistant Professor with Shanghai Jiao Tong University. His research interests include bioinformatics, deep learning, and electronic health record. He is a Lead Guest Editor of IEEE Access.

TAO ZENG received the B.S., M.S., and Ph.D. degrees from Wuhan University, Wuhan, China, in 2003, 2006, and 2010, respectively. Since 2013, he has been an Associate Professor with the Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. He is currently with the Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai, China. His research interests include bioinformatics, network biology, computational biology, machine learning, and graph theory.



YU-HANG ZHANG was born in Jinzhou, Liaoning, China, in 1992. He received the B.S. degree in medical laboratory from the Medical School, Shanghai Jiao Tong University, in 2014, and the Ph.D. degree in genetics from the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, University of Chinese Academy of Sciences, in 2019. He has authored more than 40 articles. His research interests include machine learning, liquid biopsy, and tumor immunotherapy. He was a recipient of the Merit Student from the University of Chinese Academy of Sciences, in 2017.

TAO HUANG received the B.S. degree in bioinformatics from the Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree in bioinformatics from the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, in 2012.

From 2012 to 2014, he was a Postdoctoral Fellow with the Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York, USA. Since 2014, he has been an Associate Professor and the Director of the Bioinformatics Core Facility, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai. He has published more than 100 articles. His works have been cited for more than 3000 times with an h-index of 26 and an i10-index of 64. His research interests include bioinformatics, computational biology, systems genetics, and big data research. He has been reviewer of more than 20 journals and editor/guest editor for seven journals and books.

YU-DONG CAI has been a Professor in bioinformatics with the School of Life Science, Shanghai University, since 2015. He has published more than 200 peer-reviewed scientific articles, including invited reviews. His works have been cited for more than 7500 times, with h-index of 51. His current research interests include systems biology and bioinformatics such as protein-protein interaction, disease biomarkers prediction, drug-target interaction, and protein functional sites prediction. He is an Editorial Board Member of *Biochimica et Biophisica Acta (BBA) - Proteins and Proteomics* and *Biochemistry Research International*. He has been the Guest Editor of *Computational Proteomics*, *Systems Biology & Clinical Implications*, and *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*.