Machine Learning of Single-Cell Transcriptome Highly Identifies mRNA Signature by Comparing F-Score Selection with DGE Analysis

Pengfei Liang,1,2 Wuritu Yang,1,2 Xing Chen,1 Chunshen Long,1 Lei Zheng,1 Hanshuang Li,1 and Yongchun Zuo1

1The State Key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, College of Life Sciences, Inner Mongolia University, Hohhot 010070, China

INTRODUCTION

Human preimplantation development is a complex process involving dramatic changes in transcriptional architecture. For a better understanding of their time-spatial development, it is indispensable to identify key genes. Although the single-cell RNA sequencing (RNA-seq) techniques could provide detailed clustering signatures, the identification of decisive factors remains difficult. Additionally, it requires high experimental cost and a long experimental period. Thus, it is highly desired to develop computational methods for identifying effective genes of development signature. In this study, we first developed a predictor called EmPredictor to identify developmental stages of human preimplantation embryogenesis. First, we compared the F-score of feature selection algorithms with differential gene expression (DGE) analysis to find specific signatures of the development stage. In addition, by training the support vector machine (SVM), four types of signature subsets were comprehensively discussed. The prediction results showed that a feature subset with 1,881 genes from the F-score algorithm obtained the best predictive performance, which achieved the highest accuracy of 93.3% on the cross-validation set. Further function enrichment demonstrated that the gene set selected by the feature selection method was involved in more development-related pathways and cell fate determination biomarkers. This indicates that the F-score algorithm should be preferentially proposed for detecting key genes of multi-period data in mammalian early development.

Recently, the single-cell RNA sequencing (RNA-seq) techniques are the main method for detecting developmental trajectories and cellular heterogeneity in early preimplantation embryos; however, such techniques could only provide detailed clustering signatures, and the identification of decisive factors remains difficult and requires high experimental cost and a long experimental period. As good complements to experimental techniques, computational methods play high potential roles for cancer diagnosis and sequence classification. For example, Capper et al. proposed random forest (RF) to classify approximately 100 known tumor types of the central nervous system based on DNA methylation data. Based on a single-cell transcriptome, single-cell variational inference (scVI) aggregates information across similar cells and genes by stochastic optimization and deep neural networks, and Scialdone et al. constructed a predictor for identifying cell-cycle stage. In addition, feature selection methods are independent of prior knowledge of biological dependencies, having been applied in bioinformatics, including protein prediction and biomarker discovery.

The QSPred-FL tool is based on...
the fact that quorum-sensing peptides in large-scale proteomic data can be detected by feature representation learning and machine learning algorithms, and the GF-ICF (gene frequency-inverse cell frequency) pipeline can provide an effective and simple workflow for feature selection and subsequent analyses. However, many studies have advanced new computational methods to interpret single-cell RNA-seq data, but most existing methods cannot build predictive models of development. To the best of our knowledge, so far there is no computational tool available for identifying signature genes of development.

Here, we develop the EmPredictor, a novel machine learning-based tool for predicting stages of human embryonic development. In this predictor, we compared three traditional differential gene expression (DGE) analyses with a feature selection method based on a single-cell RNA-seq dataset of preimplantation embryos. Figure 1 shows a schematic diagram of the model establishment workflow. The dataset was first integrated and removed genes with no expression in all cells. Then, we applied three DGE methods (edgeR, limma, and DESeq) and a feature selection method (F-score algorithm) to obtain signature genes. By comparing these method performances based on support vector machine (SVM) and functional enrichment analysis, the F-score algorithm had the highest performance and obtained an area under the receiver operating characteristic (ROC) curve (AUC) of 0.95. Our results also suggested that DGE analysis relied on pairwise comparison and overlap, inducing the loss of some key genes that were highly expressed at multiple stages, and the F-score algorithm considered gene expression at all stages and ignored low expression of transcripts.

RESULTS
Global Expression Profiles of Human Embryos
Global transcriptome profiles were first analyzed based on a dataset of early human embryos (Figure 2A, reads per kilobase transcript per million mapped reads [RPKM] > 0). The gene expression level of E3 cells was higher than that for other stages, indicating that the
zygote genome was activated and began to identify the genetic program that may control this process.\textsuperscript{15} Also, the gene expression level of the cells began to decrease during the E3–E6 stage but increased during the E7 stage, suggesting that E7 embryos may initiate new transcriptional activation to begin embryo implantation. To determine whether embryos at the same stage showed a high correlation, we analyzed RNA-seq data of the E3–E7 embryos using the SC3 package (Figure 2C).\textsuperscript{35} Most of the cells from the same stage were clustered into one cluster. E3 cells and E4 cells have a high correlation. However, E5 cells have two clusters, because after ZGA, differences of embryos begin to emerge and E5 cells appear to segregate ICM and TE.\textsuperscript{12} Interestingly, the embryos at the E6 and E7 stages were clustered together and divided into three clusters, suggesting that preimplantation of the early embryo resolved in the formation of three distinct cell lineages of blastocysts.\textsuperscript{36}

In order to investigate whether these gene expression profiles were related to developmental stages, we conducted t-distributed stochastic neighbor embedding (t-SNE) on all individual embryos (Figure 2B)\textsuperscript{37,38} and found that embryos at the same developmental stage were clustered together, and the primary segregating factor was developmental time. With the development of embryos, the heterogeneity of embryos increased gradually. In addition, we used differentially expressed genes (DEGs) to plot a heatmap by DESeq, which reflected that embryo cells segregated into five groups, that E6 cells were less different from cells in adjacent stages, and that E3 cells have more DEGs than do other stages (Figure 2D).

**Identification of the Developmental Signature by Comparing with F-Score and Differential Expression Analysis**

To identify the best signature genes related to embryonic development, we obtained 24,444 gene expression profiles of 1,529 individual cells (81 E3, 190 E4, 377 E5, 415 E6, and 466 E7 cells) from a public database. As DGE analysis usually applied sequence count data, count data were analyzed by comparing three DGE analyses and the F-score algorithm. By using the same parameter (fold change > 2, p < 0.05), limma, DESeq, and edgeR identified 3,754, 4,876 and 6,231 DEGs, respectively, and the number of overlapping genes based on these methods was 2,976 (Figure 3A; Figure S1A). E3 cells had the highest number of DEGs compared to other stages (Figure 3A), and edgeR had more differential genes than did other methods (Figure 3A). The F-score algorithm calculated and ranked each gene score, but we still did not know how many genes should finally be selected. To optimize signature gene selection, we tried the number of signature genes

![Figure 2. Global Transcriptome Profiles of Human Embryos](image-url)
(ranging from 10 to 24,444 genes) for training the support vector machine model and calculated their prediction performance, which applied incremental feature selection (IFS) (Figure 3B). More genes may bring the best performance but lead to more hardware and time loss, so we chose 1,881 of the top genes by considering gene number and their performance, the accuracy of which was 0.93, and this cost low memory consumption and took only 43 min to obtain the best model. Then, we plotted a heatmap of 1,881 genes based on F-score algorithms (FSGs) and using the SC3 package (Figure S1B). Interestingly, E5 cells were separated into two clusters, suggesting that embryo lineage separation showed the formation of TE and ICM.

We also compared DEGs and FSGs by gene function enrichment (Figure 3C; Table S1). Unique DEGs were enriched for regulation of the cellular process, positive regulation of the multicellular organismal process, positive regulation of the metabolic process, and multicellular organism development. Unique FSGs were enriched for embryo development, regulation of cell shape, cell-cell adhesion, and the cell cycle, which are most relevant to embryonic development. The genes overlapped by DEGs and FSGs were enriched for embryo implantation, regulation of transcription, DNA-templated, multicellular organism development, and positive regulation of cell proliferation, which related to transcription and development. Briefly, unique FSGs related more to embryonic development compared to unique DEGs.

As shown in Table 1, some key genes were selected from FSGs and DEGs. DPPA5, ZSCAN4, and SOX2 were selected by using these methods (rank 9, 22, and 1,520, respectively). DPPA5 stabilizes NANOG and supports human pluripotent stem cell (hPSC) self-renewal and cell reprogramming in feeder-free conditions. ZSCAN4 is a unique gene highly expressed at the zygotic genome activation stage. The POU5F1 gene is vital for PSC maintenance in the mammalian embryo. Interestingly, POU5F1 (rank 1,470) was not selected by the DGE method but was obtained by F-score. Therefore, we analyzed the expression of POU5F1 among E3–E7 stages, and
POU5F1 was highly expressed in E4 and E5 cells (Figure 3D). Therefore, DGE analysis mainly relies on pairwise comparison and overlap, so if a gene is highly expressed at two or more stages, differential expression analyses may lose the gene, suggesting that DGE analysis only considered DEGs highly expressed at a stage. The F-score algorithm showed the importance of a gene in all stages. If the expression level of a gene was too low, the F-score algorithm would give a low score for this gene, especially similar to ERVFRD-1 (Figure S1C; Table S2). Although transcripts of low expression may be important, most of these are outliers, suggesting that transcripts with low expression levels were preprocessed, in line with previous studies.48 In addition to known markers, several less described markers were identified, such as RNF168, CCKBR, PTN, CD24 and STC2 (rank 12, 151, 332, 349, and 1,851). CCKBR, a cholecystokinin B receptor, has been found in a diverse range of cancers.49 We found that in the late blastocyst, E6 and E7 cells expressed high levels of CCKBR, indicating that CCKBR may be involved in the ICM segregation of EPI and PE cells. PTN-encoded protein has significant roles in cell growth, migration, and tumorigenesis,50,51 and it was expressed in the late blastocyst, suggesting that PTN may be involved in embryonic cell migration. Then, we found that most of the 500 top-ranked genes were high relative expression genes of E3 stages (Table S2), similar to what has been previously reported.15

| Table 1. Comparison between Key Genes of FSGs and DEGs |
|---------------------------------|-----------------|--------|--------|--------|--------|
| Genes | Gene Ontology Terms | FSGs (1,881) | F-Score | limma | edgeR | DESeq |
|-------|---------------------|-------------|--------|--------|--------|--------|
| GDF9  | positive regulation of cell proliferation | ● | 1.97 | *** | *** | *** |
| DPPA5 | multicellular organism development | ● | 1.51 | • | • | • |
| RNF168| zinc ion binding | ● | 1.42 | *** | *** | *** |
| KLF17 | transcription factor activity, sequence-specific DNA binding | ● | 1.39 | *** | *** | *** |
| ZSCAN4| transcription, DNA templated | ● | 1.20 | • | • | • |
| TLE6  | repressing transcription factor binding | ● | 1.14 | *** | *** | *** |
| CCKBR |          | ● | 0.65 | *** | *** | *** |
| PTN   |          | ● | 0.49 | *** | *** | *** |
| CD24  |          | ● | 0.48 | | | |
| OSR2  | embryo development | ● | 0.34 | | | |
| DLX3  | transcription from RNA polymerase II promoter | ● | 0.31 | ** | ** | ** |
| SOX2  | transcription factor endodermal cell fate specification | ● | 0.22 | • | • | • |
| POU5F1| somatic stem cell population maintenance | ● | 0.22 | | | |
| STC2  | embryo implantation | ● | 0.19 | | | |
| ERVFRD-1 | | ○ | 0.09 | *** | *** | *** |

If a gene belongs to the dataset, replace it with ●; otherwise, replace it with ○. F-score shows the importance of features selected by the F-score algorithm. *p < 0.05, **p < 0.01, ***p < 0.001.

POU5F1 was highly expressed in E4 and E5 cells (Figure 3D). Therefore, DGE analysis mainly relies on pairwise comparison and overlap, so if a gene is highly expressed at two or more stages, differential expression analyses may lose the gene, suggesting that DGE analysis only considered DEGs highly expressed at a stage. The F-score algorithm showed the importance of a gene in all stages. If the expression level of a gene was too low, the F-score algorithm would give a low score for this gene, especially similar to ERVFRD-1 (Figure S1C; Table S2). Although transcripts of low expression may be important, most of these are outliers, suggesting that transcripts with low expression levels were preprocessed, in line with previous studies.48 In addition to known markers, several less described markers were identified, such as RNF168, CCKBR, PTN, CD24 and STC2 (rank 12, 151, 332, 349, and 1,851). CCKBR, a cholecystokinin B receptor, has been found in a diverse range of cancers.49 We found that in the late blastocyst, E6 and E7 cells expressed high levels of CCKBR, indicating that CCKBR may be involved in the ICM segregation of EPI and PE cells. PTN-encoded protein has significant roles in cell growth, migration, and tumorigenesis,50,51 and it was expressed in the late blastocyst, suggesting that PTN may be involved in embryonic cell migration. Then, we found that most of the 500 top-ranked genes were high relative expression genes of E3 stages (Table S2), similar to what has been previously reported.15

| Table 2. Performance of Stage Predicting Models with 5-Fold Cross-Validation |
|---------------------------------|--------|--------|--------|--------|--------|
| Method | Gene No. | Precision (%) | Recall (%) | Accuracy (%) | F1 Measure (%) |
|--------|---------|--------------|------------|--------------|---------------|
| DESeq  | 4,876   | 90.23        | 89.81      | 89.85        | 89.73         |
| limma  | 3,754   | 91.5         | 91.23      | 91.24        | 91.24         |
| edgeR  | 6,231   | 90.82        | 90.36      | 90.42        | 90.31         |
| F-score | 1,881  | 93.3         | 92.91      | 93.01        | 93.2          |

Underlined text represents the maximum value of every performance evaluation criterion.
DISCUSSION
Herein, we have proposed the first EmPredicitor, a novel machine learning-based tool for predicting stages of human embryonic development. Based on three DGE analyses (limma, edgeR, DESeq) and F-score algorithm, the single-cell transcriptomes data obtain 3,754, 4,876, 6,231, and 1,881 signature genes, respectively. Then, supervised machine learning is used to estimate the contribution of embryonic development to these signature genes. Toward the application of 5-fold cross-validation on a benchmark dataset, the F-score algorithm can achieve the highest accuracy of 0.93 and AUC of 0.95. Furthermore, functional enrichment analysis showed that the F-score algorithm can obtain key signaling pathways related to embryo development. Based on prior biological knowledge, some key genes were used to estimate the assessment of F-score and DGE analyses. DGE analyses rely on pairwise comparison and overlap to obtain differentially expressed genes. F-score detected key genes of multi-period data that contributed to identifying early embryo stages. In addition, we constructed a user-friendly and publicly accessible web server where users can upload or paste a dataset of the eight key genes to predict the stage of their samples.

There are still some disadvantages of this work. Here, we investigated only predicting embryonic days. However, embryonic development is a complex process involving lineage specification and X chromosome dosage compensation. Integrating genetic and epigenetic data with gene expression may provide a more comprehensive view of embryonic development. In addition, feature selection methods have irreplaceable advantages in processing single-cell transcriptome data and are independent of prior knowledge of biological dependencies, which extend the development analysis pipeline. In the future, we will use advanced feature selection methods to study embryonic development based on more accurate molecular events and multi-omics data.

MATERIALS AND METHODS
Data and Preprocessing
We downloaded a single-cell transcriptome dataset of human preimplantation embryos from ArrayExpress under accession E-MTAB-3929,12 including 1,529 samples. The dataset has five different cell stages, which are embryonic day (E)3, E4, E5, E6, and E7. The E3 stage has 81 cells, the E4 stage has 190 cells, the E5 stage has 377 cells, the E6 stage has 415 cells, and the E7 stage has 466 cells.

The data were processed using TrueSeq dual-index sequencing primers (Illumina) according to the manufacturer’s recommendations on an Illumina HiSeq 2000.12 The data quality was checked and reads were mapped to the human genome (hg19) using STAR with default settings.52 RPKM were calculated using rpkmforgenes53 by the uniquely mapped read counts. Genes were filtered, keeping 24,444 out of 26,178 genes that were expressed in at least 1 out of 1,529 cells (count > 0).

Feature Selection
Linear Model
During the past decade, the limma package54 has been a popular choice for gene discovery through differential expression analyses of microarrays. Recently, limma has also provided differential expression and differential splicing analyses of RNA-seq data. limma uses the voom function by converting mean variance to precision weights and using a linear model,

\[
E(\gamma_i) = \mu_i = X_i^T \beta_g; \quad \text{Equation 1}
\]

where \(X_i\) is a vector of covariates and \(\beta_g\) is a vector of unknown coefficients representing log2 fold changes between experimental conditions. In matrix terms

\[
E(\gamma_i) = X \beta_g; \quad \text{Equation 2}
\]

where \(\gamma_i\) is the vector of log cpm values for gene \(g\), and \(X\) is the design matrix with the \(X_i\) as rows. The limma package is available at https://bioconductor.org/packages/release/bioc/html/limma.html.

Negative Binomial Distribution
edgeR55 is designed for the analysis of replicated count-based expression data. Data are modeled as negative binomial (NB) distributed

\[
Y_{gi} \sim NB(M_i p_{gi}, \phi_g) \quad \text{Equation 3}
\]

for gene \(g\) and sample \(i\). Here, \(M_i\) is the library size (total number of reads), \(\phi_g\) is the dispersion, and \(p_{gi}\) is the relative abundance of gene \(g\) in experimental group \(j\) to which sample \(i\) belongs. The edgeR package is available at https://bioconductor.org/packages/release/bioc/html/edgeR.html.
DESeq provides methods to test for differential expression by use of the negative binomial distribution and a shrinkage estimator for the distribution’s variance,

\[ K_i \sim NB(\mu_i, \sigma_i^2). \]  
(Equation 4)

which has two parameters, the mean \( \mu_i \) and the variance \( \sigma_i^2 \). The read counts \( K_i \) are non-negative integers. The DESeq package is available at https://bioconductor.org/packages/release/bioc/html/DESeq.html.

### F-Score Algorithm

F-score is a simple and basic but effective algorithm for evaluating the importance of each feature in the dataset. F-score is computed each feature values and

\[
F(i) = \frac{\left( \overline{x}_i^{(+)} - \overline{x}_i^{(-)} \right)^2 + \left( \overline{x}_i^{(-)} - \overline{x}_i^{(+)} \right)^2}{\frac{1}{n} \sum_{k=1}^{n} \left( x_{k,i}^{(+)} - x_{k,i}^{(+)} \right)^2 + \frac{1}{n} \sum_{k=1}^{n} \left( x_{k,i}^{(-)} - x_{k,i}^{(-)} \right)^2}.
\]  
(Equation 5)

where \( \overline{x}_i, \overline{x}_i^{(+)}, \overline{x}_i^{(-)} \) are the average of the \( i \)th feature of the whole, positive, and negative datasets respectively; \( x_{k,i}^{(+)} \) is the \( i \)th feature of the \( k \)th positive instance; and \( x_{k,i}^{(-)} \) is the \( i \)th feature of the \( k \)th negative instance. A Python program fselect.py can compute each feature value and rank the feature downloaded from https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/.

### Machine Learning Model Implementation

The support vector machine (SVM) was proposed by Vapnik et al. SVM shows many advantages in solving small sample, nonlinear, and high-dimensional pattern recognition. The idea of SVM is based on transforming the input vector into a high-dimensional Hilbert space and finding a separating hyperplane in this space. Gaussian radial basis function (RBF) kernel function is a widely used kernel function because of its high performance in non-line classification:

\[
K_{\text{Gaussian}}(x_i, x_j) = e^{-\frac{1}{2\sigma^2} \| x_i - x_j \|^2}.
\]  
(Equation 6)

We applied LIBSVM as an SVM model with a one-against-one strategy and RBF kernel. A grid search strategy with a cross-validation test is always utilized to obtain the best values of the regularization parameter \( C \) and kernel parameter \( \gamma \). We used the grid.py file (https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/) in LIBSVM to search for the best \( C \) value and \( \gamma \) value (the range of the \( C \) parameter is between \( 2^{-5} \) and \( 2^{10} \), and the range of the \( \gamma \) parameter is between \( 2^{-15} \) and \( 2^{5} \)).

Classifier performance was evaluated by 5-fold
cross-validation analysis, where each training dataset was randomly partitioned into four equal parts with one part being used for model training and the remaining part used for testing. We used the cross-validation method to limit overfitting of the classifier. To have a complete measurement of the prediction performance, four statistics, i.e., accuracy, recall, precision, and F1 measure, were calculated as follows:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FP} \quad (\text{Equation 7})
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (\text{Equation 8})
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (\text{Equation 9})
\]

\[
\text{F1 measure} = \frac{2 \times (\text{precision} \times \text{recall})}{\text{precision} + \text{recall}} = \frac{2 \times TP}{2 \times TP + FN + FP} \quad (\text{Equation 10})
\]

where \(TP\) is the true positive correct result, \(FP\) is the false unexpected result, \(FN\) is the false missing result, and \(TN\) is the true correct absence of result.

**Code Available**
The code for the implementation of the EmPredictor is available on GitHub: [https://github.com/lameihao/EmPredictor](https://github.com/lameihao/EmPredictor).

**SUPPLEMENTAL INFORMATION**
Supplemental Information can be found online at [https://doi.org/10.1016/j.omtn.2020.02.004](https://doi.org/10.1016/j.omtn.2020.02.004).

**AUTHOR CONTRIBUTIONS**
Y.Z. designed this work. P.L. and W.Y. performed the whole bioinformatics analysis and wrote the manuscript. X.C. and C.L. performed the experiments and helped edit the manuscript. L.Z. and H.L. assisted the experiments.

**CONFLICTS OF INTEREST**
The authors declare no competing interests.

**ACKNOWLEDGMENTS**
We are grateful to our laboratory colleagues for their assistance with the bioinformatics analysis. We thank Prof. Fredrik Lanner (Karolinska Universitätssjukhuset) for sharing the single-cell RNA-seq datasets in ArrayExpress database. This work was supported by the National Nature Scientific Foundation of China (61702290, 61861036), the Program for Young Talents of Science and Technology in Universities of Inner Mongolia Autonomous Region (NJYT-18-B01), and by the Fund for Excellent Young Scholars of Inner Mongolia (2017/Q04). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**REFERENCES**
1. Cockburn, K., and Rossant, J. (2010). Making the blastocyst: lessons from the mouse. J. Clin. Invest. 120, 995–1003.
2. Zuo, Y., Gao, Y., Su, G., Bai, C., Wei, Z., Liu, K., Li, Q., Rou, S., and Li, G. (2014). Irregular transcriptome reprogramming probably causes the developmental failure of embryos produced by interspecies somatic cell nuclear transfer between the Przewalski’s gazelle and the bovine. BMC Genomics 15, 1113.
3. Graf, A., Krebs, S., Heininen-Brown, M., Zakhartchenko, V., Blum, H., and Wolf, E. (2014). Genome activation in bovine embryos: review of the literature and new insights from RNA sequencing experiments. Anim. Reprod. Sci. 149, 46–58.
4. Zuo, Y., Su, G., Cheng, L., Liu, K., Feng, Y., Wei, Z., Bai, C., Cao, G., and Li, G. (2017). Coexpression analysis identifies nuclear reprogramming barriers of somatic cell nuclear transfer embryos. Oncotarget 8, 65847–65859.
5. Ko, M. (2016). Zygotic genome activation revisited: looking through the expression and function of Zscan4. Curr. Top. Dev. Biol. 120, 103–124.
6. Zuo, Y., Su, G., Wang, S., Yang, L., Liao, M., Wei, Z., Bai, C., and Li, G. (2016). Exploring timing activation of functional pathway based on differential co-expression analysis in preimplantation embryogenesis. Oncotarget 7, 74120–74131.
7. Niakan, K.K., and Eggan, K. (2013). Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. Dev. Biol. 375, 54–64.
8. Kwon, G.S., Viotti, M., and Hadjantonakis, A.K. (2008). The endoderm of the mouse embryo arises by dynamic widespread intercalation of embryonic andextraembryonic lineages. Dev. Cell 15, 509–520.
9. Hendrickson, P.G., Doraís, J.A., Grow, E.J., Whiddon, J.L., Lim, J.W., Wike, C.L., Weaver, B.D., Pflueger, C., Emery, B.R., Wilcox, A.L., et al. (2017). Conserved roles of mouse DUX and human DUX4 in activating cleavage-stage genes and MERVL/HERVt retroposons. Nat. Genet. 49, 925–934.
10. De Iaco, A., Planet, E., Coluccio, A., Verp, S., Duc, J., and Trono, D. (2017). DUX family transcription factors regulate zygotic genome activation in placental mammals. Nat. Genet. 49, 941–945.
11. Niwa, H., Toyooka, Y., Shimozato, D., Strumpf, D., Takahashi, K., Yagi, R., and Rossant, J. (2005). Interaction between Oct4/4 and Cdx2 determines trophectoderm differentiation. Cell 123, 917–929.
12. Péterpalós, S., Edgard, D., Reinus, B., Deng, Q., Panula, S.P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R., and Lanner, F. (2016). Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. Cell 165, 1012–1026.
13. Eckersley-Maslin, M., Alda-Catalinas, C., Blotenburg, M., Kreibich, E., Krueger, C., and Reik, W. (2019). Dppa2 and Dppa4 directly regulate the Dux-driven zygotic transcriptional program. Genes Dev. 33, 194–208.
14. De Iaco, A., Coudray, A., Duc, J., and Trono, D. (2019). DPPA2 and DPPA4 are necessary to establish a 2C-like state in mouse embryonic stem cells. EMBO Rep. 20, 10.
15. Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell RNA-seq profiling of human preimplantation embryos and embryonic stem cells. Nat. Struct. Mol. Biol. 20, 1131–1139.
16. Farrell, J.A., Wang, Y., Riesensfeld, S.J., Shekhar, K., Regev, A., and Schier, A.F. (2018). Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. Science 360, eaar3131.
17. Cheng, S., Pei, Y., He, L., Peng, G., Reinus, B., Tam, P.P.L., Jing, N., and Deng, Q. (2019). Single-cell RNA-seq reveals cellular heterogeneity of pluripotency transition and X chromosome dynamics during early mouse development. Cell Rep. 26, 2593–2607.e3.
18. Hu, R., Zheng, L., Long, C., Song, M., Li, T., Yang, L., and Zuo, Y. (2019). EmExplorer: a database for exploring time activation of gene expression in mammalian embryos. Open Biol. 9, 190054.
19. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-seq whole-transcriptome analysis of a single cell. Nat. Methods 6, 377–382.
20. Wong, D., and Yip, S. (2018). Machine learning classifies cancer. Nature 555, 446–447.
21. Zuo, Y., Li, Y., Chen, Y., Li, G., Yan, Z., and Yang, L. (2017). PseKRAAC: A flexible web server for generating pseudo K-tuple reduced amino acids composition. Bioinformatics 33, 122–124.

22. Liu, D., Li, G., and Zuo, Y. (2019). Function determinants of TET proteins: the arrangements of sequence motifs with specific codes.Brief. Bioinform. 20, 1826–1835.

23. Feng, C.Q., Zhang, Z.Y., Zhu, X.J., Lin, Y., Chen, W., Tang, H., and Lin, H. (2019). iTerm-PseKNC: a sequence-based tool for predicting bacterial transcriptional terminators. Bioinformatics 35, 1469–1477.

24. Chen, W., Feng, P., Song, X., Lv, H., and Lin, H. (2019). IRNA-m7G: identifying N7-methylguanosine sites by fusing multiple features. Mol. Ther. Nucleic Acids 18, 269–274.

25. Chen, W., Feng, P., Liu, T., and Jin, D. (2019). Recent advances in machine learning methods for predicting heat shock proteins. Curr. Drug Metab. 20, 224–228.

26. Zheng, L., Huang, S., Mu, N., Zhang, H., Zhang, J., Chang, Y., Yang, L., and Zuo, Y. (2019). RAAIClude: a web server of reduced amino acid alphabet for sequence-dependent inference by using Chou’s five-step rule. Database (Oxford) 2019, baz131.

27. Lai, H.Y., Zhang, Z.Y., Su, Z.D., Su, W., Ding, H., Chen, W., and Lin, H. (2019). iProEP: a computational predictor for predicting promoter. Mol. Ther. Nucleic Acids 17, 336–346.

28. Capper, D., Jones, D.W.T., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, K., Koelsche, C., Sahm, F., Chavez, L., Reuss, D.E., et al. (2018). DNA methylation-based classification of central nervous system tumours. Nature 555, 469–474.

29. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. Nat. Methods 15, 1053–1058.

30. Scialdone, A., Natrajan, K.N., Saraiwa, L.R., Prosperi, V., Teichmann, S.A., Stegle, O., Marioni, J.C., and Buettner, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods 85, 54–61.

31. Wei, L., Hu, J., Li, F., Song, J., Su, R., and Zou, Q. (2020). Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. Front. Bioinform. 21, 106–119.

32. Li, J., Lan, C.-N., Kong, Y., Feng, S.S., and Huang, T. (2018). Identification and analysis of blood gene expression signature for osteoarthritis with advanced feature selection methods. Front. Genet. 9, 246.

33. Wolf, P.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 19, 15.

34. Talwar, D., Mongia, A., Sengupta, D., and Majumdar, A. (2018). AutoImpute: Autocoder based imputation of single-cell RNA-seq data. Sci. Rep. 8, 16329.

35. Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Marioni, J.C., and Buettner, F. (2015). Computational assignment of cell-cycle stage from single-cell transcriptome data. Methods 85, 54–61.

36. Zuo, Y., Li, Y., Chen, Y., Li, G., Yan, Z., and Yang, L. (2017). Recent advances in conotoxin classification and analysis of early human development. Curr. Opin. Genet. Dev. 45, 1025–1030.

37. Ramsköld, D., Wang, E.T., Burge, C.B., and Sandberg, R. (2009). An abundance of non-coding RNA in human cells. Nature 455, 108–112.

38. McCarthy, D.J., Campbell, K.R., Lun, A.T., and Wills, Q.F. (2017). Scater: preprocessing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics 33, 1179–1186.

39. Huang, T., Zhang, J., Xu, Z.P., Hu, L.L., Chen, L., Shao, J.L., Zhang, L., Kong, X.Y., Cai, Y.D., and Chou, K.C. (2012). Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches. Biochimie 94, 1017–1025.

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

41. Rossant, J., and Tam, P.P.L. (2017). New insights into early human development: lessons for stem cell derivation and differentiation. Cell Stem Cell 20, 18–28.

42. Ortega, N.M., Winblad, N., Plaza Reyes, A., and Lanner, F. (2018). Functional genetics of early human development. Curr. Opin. Genet. Dev. 52, 1–6.

43. Qian, X., Kim, J.K., Tong, W., Villa-Diaz, L.G., and Krebsbach, P.H. (2016). DPPA5 supports pluripotency and reprogramming by regulating NANOG turnover. Stem Cells 34, 588–600.

44. Falco, G., Lee, S.L., Stanghellini, I., Bassey, U.C., Hamatani, T., and Ko, M.S. (2007). Zscan4: A novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. Dev. Biol. 307, 539–550.

45. Long, C., Li, W., Liang, P., Liu, S., and Zuo, Y. (2019). Transcriptome comparisons of multi-species identify differential genome activation of mammals embryogenesis. IEEE Access 7, 7794–7802.

46. Qian, X., Kim, J.K., Tong, W., Villa-Diaz, L.G., and Krebsbach, P.H. (2016). DPPA5 supports pluripotency and reprogramming by regulating NANOG turnover. Stem Cells 34, 588–600.

47. Falco, G., Lee, S.L., Stanghellini, I., Bassey, U.C., Hamatani, T., and Ko, M.S. (2007). Zscan4: A novel gene expressed exclusively in late 2-cell embryos and embryonic stem cells. Dev. Biol. 307, 539–550.

48. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550.

49. Roy, J., Potti, K.S., Coppola, D., Leon, M.E., Khalil, F.K., Centeno, B.A., Clark, N., Stark, V.E., Morse, D.L., and Low, P.S. (2016). Assessment of cholecystokinin 2 receptor (CCK2R) in neoplastic tissue. Oncotarget 7, 14605–14615.

50. Bai, P.S., Xia, N., Sun, H., and Kong, Y. (2017). Pleiotrophin, a target of miR-384, promotes proliferation, metastasis and lipogenesis in HBV-related hepatocellular carcinoma. J. Cell. Mol. Med. 21, 3023–3034.

51. Shen, D., Podolnikova, N.P., Yakubchenko, V.P., Ardell, C.L., Balabiyev, A., Ugurova, I., Shikhiev, S., and Zaman, S. (2012). Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches. Biochimie 94, 1017–1025.

52. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edger: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 134–140.
Supplemental Information

Machine Learning of Single-Cell Transcriptome
Highly Identifies mRNA Signature by Comparing
F-Score Selection with DGE Analysis

Pengfei Liang, Wuritu Yang, Xing Chen, Chunshen Long, Lei Zheng, Hanshuang Li, and Yongchun Zuo
Fig. S1 (A) The diagram shows DEGs number of each stage of embryos compared to three DEA methods. All DEGs represent overlapping genes from three DEA methods. DESeq-specific DEGs only appear in DESeq methods. edgeR-specific DEGs only appear in edgeR methods. limma-specific DEGs only appear in limma methods (B) Gene expression heatmap of 1881 genes of F-score algorithm selection across all cells (C) Boxplots of ERVFRD-1 relative expression level at E3-E7 stages.