Machine Learning of Single-cell Transcriptome Highly Identify Development Signature by Comparing F-score Selection with Differential Expression Analysis

*Pengfei Liang[1], **Xing Chen**[1], Chunshen Long[1], Lei Zheng[1], Lei Yang[2],* and Yongchun Zuo[1],**

[1] *State key Laboratory of Reproductive Regulation and Breeding of Grassland Livestock, College of Life Sciences, Inner Mongolia University, Hohhot, 010070, China.*

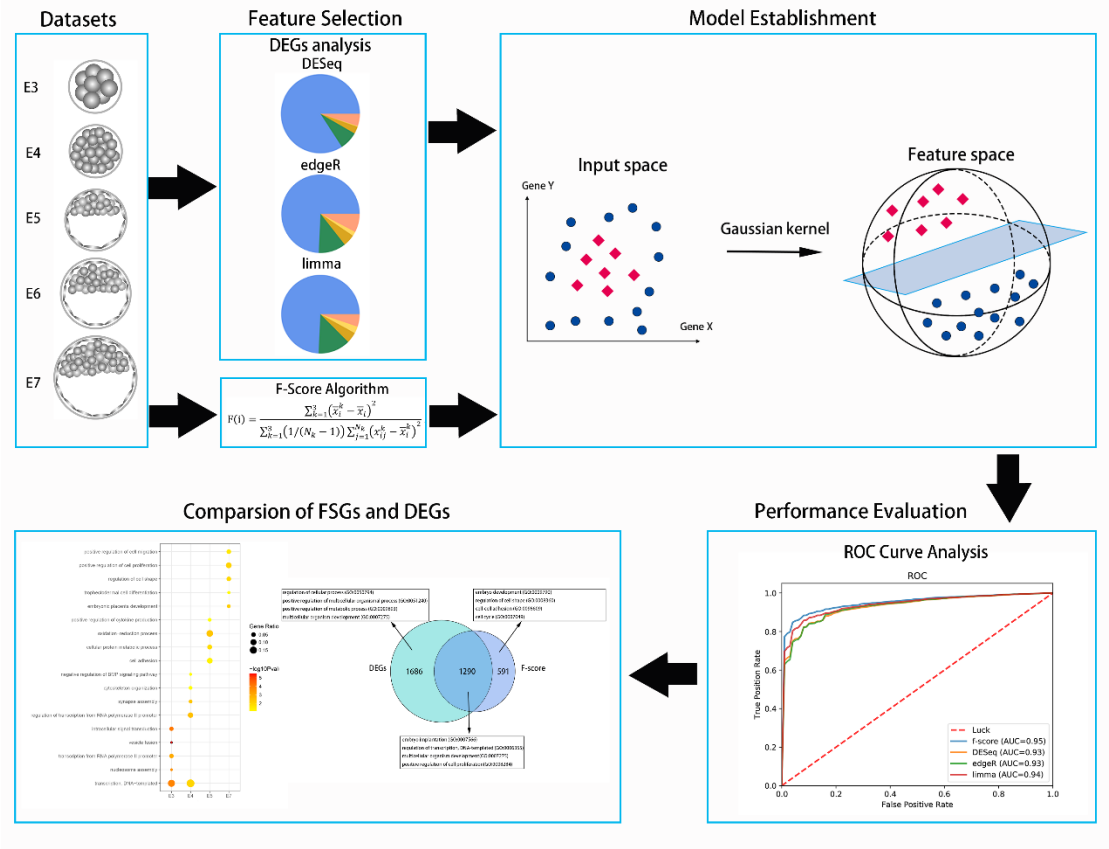
[2] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China.*

[*] *Corresponding author.*

Abstract:

Human preimplantation development is a complex process involving dramatic changes in the transcriptional architecture. Single-cell transcriptome analysis plays important roles in signature identification of transcription atlas in embryo development. However, the traditional analysis pipelines of enormous transcriptome datasets are not fully effective on single-cell RNA-Seq techniques. In this study, we firstly compared F-score of feature selection algorithms with differential gene expression (DGE) analysis to find specific signatures of development stage. By using machine learning evaluation, four types of signature subset were comprehensively discussed. The prediction results showed that a feature subset with 1881 genes based on F-score algorithm obtained that best predictive performance, which achieved the highest accuracy of 0.93 on the cross validation set. Further function enrichment demonstrated that the gene set selected by the feature selection method involved in more development-related pathways and cell fate determination biomarkers. It indicates that F-score algorithm should be preferential proposed for detecting key genes of multi-period data. At last, a free online server was established for the convenience of most experimental scientists at http://bioinfor.imu.edu.cn/empredictor/. A friendly guide is provided to describe how to use the web server.

Keywords: Embryo development, Expression signature, Single-cell transcriptome, Machine learning, Feature selection, Identification, Web server



The workflow consists of four steps**.** Firstly, using different feature selection methods found DEGs. Secondly, DEGs was trained by support vector machine. Thirdly, the model performances were compared to each other to evaluate different feature selection methods. Last, we use the best results to obtain biomarker of different stages.