Ab initio Discovery of Biological Knowledge from scRNA-seq Data Using Machine Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Expectations of machine learning (ML) are high for discovering new patterns in high-throughput biological data, but most such practices are accustomed to relying on existing knowledge conditions to design experiments. There is a gap to investigate the power and limitation of ML in revealing complex patterns from data without the guide of existing knowledge. In this study, we conducted systematic experiments on such *ab initio* knowledge discovery with ML methods on single-cell RNA-sequencing data of early embryonic development. Results showed that a strategy combining unsupervised and supervised ML can reveal major cell lineages with minimum involvement of prior knowledge or manual intervention, and the *ab initio* mining enabled a new discovery of human early embryonic cell differentiation. The study illustrated the feasibility, significance, and limitation of *ab initio* ML knowledge discovery on complex biological problems.

1 Introduction

Machine learning (ML) is powerful in many pattern recognition tasks such as computer vision, natural language processing, as well as biological data analysis [1-4]. Machines can learn what scientist have already known with knowledge-derived labels and proper training settings. But it is still unclear whether ML methods can discover unknown patterns underlying the data that challenge human experts to analyze. A typical task is to identify unknown structures intrinsic in massive high-dimensional data and to infer underlying principles without the guide of existing knowledge. Instead of typical artificial intelligence scenarios, we expect ML methods to discover new knowledge that human experts cannot find.

Single-cell genomics is playing important roles in current biological studies. High-throughput single-cell RNA-sequencing (scRNA-seq) generates huge amount of high-dimensional data of cells, pushing to the boundary of existing biological knowledge. The reliance on existing knowledge may bury the value of the new technology in revealing new knowledge that could not be seen with previous technologies. Efforts are needed on systematically exploring the power and limitation of ML methods to discover biological knowledge from data in an *ab initio* manner with restricted or controlled involvement of existing knowledge and judgement by human experts.

In this study, we conducted experiments for *ab initio* knowledge discovery using basic ML methods with controlled involvement of human knowledge [5]. We selected a widely-used scRNA-seq dataset of early embryonic cell development [6]. We ignored all existing knowledge of embryonic development except the basic assumption that cells of a later day are developed from the earlier day in some unknown lineages, and experimented on the discovery of such lineages from the data using the combination of classical unsupervised and supervised ML methods. Results showed that a full ML-derived understanding on the developmental process can be well aligned to the latest knowledge, except for a new discovery that can be a renovation to existing knowledge. We also

Submitted to NeurIPS 2021 AI for Science Workshop.

observed the limitation of proposed method in discovering more complicated relations on zebrafish dataset [7]. These experiments highlighted the power and limitation of using current ML methods and scRNA-seq data to discover complicated biological knowledge *ab initio*, and showed the feasibility and significance of controlling the involvement of existing knowledge and subjective adjustment in mining new biological data.

2 Method

2.1 Data

The main dataset we worked on was the human early embryo development data [6] with single-cell gene expression data of 1,529 cells. Cells were captured during embryonic day-3 to day-7 (referred as E3, E4, ..., E7). We also worked on a zebrafish dataset [7] contains 36,749 embryonic cells collected at 7 time points during the development, i.e., 4, 6, 8, 10, 14, 18 and 24 hours post fertilization (referred as 4 hpf, 6 hpf, ..., 24 hpf). Detailed data pre-processing steps are described in *Appendix*. None of these steps is specific to any known biological knowledge or to the question to be studied.

2.2 Task and Strategy

The task of this *ab initio* knowledge discovery experiment is to identify the possible lineage relations among cells of each day (or hour) in the early embryonic development data.

We designed a strategy integrating clustering and classification (Figure 1). Taking the human dataset as example, if cells of the same day are of different lineages, there must be distinct clusters in cells of that day; and if the clusters of different days belong to the same lineage, we can classify the lineage in one day using the model trained by the lineage in another day. We thus decomposed the task to the following sub-tasks: (1) Choosing one day as the candidate reference day for other days; (2) Building a candidate developmental process by finding relations among cells of different days based on the reference day; and (3) Assessing the plausibility of the candidate developmental process. The basic ML methods we used were k-means and SVM, other methods can also be applied (*Appendix*). As no prior knowledge available to decide a proper reference day, we took each day as the reference and got multiple candidate versions. We developed a method to infer the most plausible one by evaluating the self-consistency of each candidate process, illustrated as follows.

2.3 Evaluate Self-Consistency of Candidate Development Processes

We first calculate the level of concordance clustering and classification results on the same day using the adjusted random index (ARI):

$$concord(i|r) = ARI(S_i, C_{(i|r)}) \tag{1}$$

In the case of human dataset, $S_i, i = 3,, 7$ are the clustering results in each day, $C_(i|r), i = 3, \ldots, 7, i \neq r$ are the classification of day-*i* cells using day-*r* clusters as reference. The score is 1.0 when two results are identical, while it is 0 where classification result is similar to random assignment.

Then we measure the reliability of the clustering results of day-*i*, we define the *reliability score* of day-*i* as the average of *concord scores* of all other days using day-*i* as reference:

$$reliab(i) = \underset{\substack{j=3,\ldots,7 \\ j \neq i}}{average}\, concord(j|i) \tag{2}$$

A poor concordance of day-*i* based on day-*r* may be due to either the clustering result of day-*r* is not suitable as a reference for day-*i*, or the clustering result of day-*i* itself is bad. To take both factors into consideration, we further defined an adjusted reliability score (ARS) by weighting the *concord score* with the *reliab score* of each target day. In this way, we can judge the plausibility of candidate processes without using biological knowledge.

$$ARS(r) = \sum_{\substack{i=3,\ldots,7 \\ i \neq r}} (reliab(i) \cdot concord(i|r)) \tag{3}$$

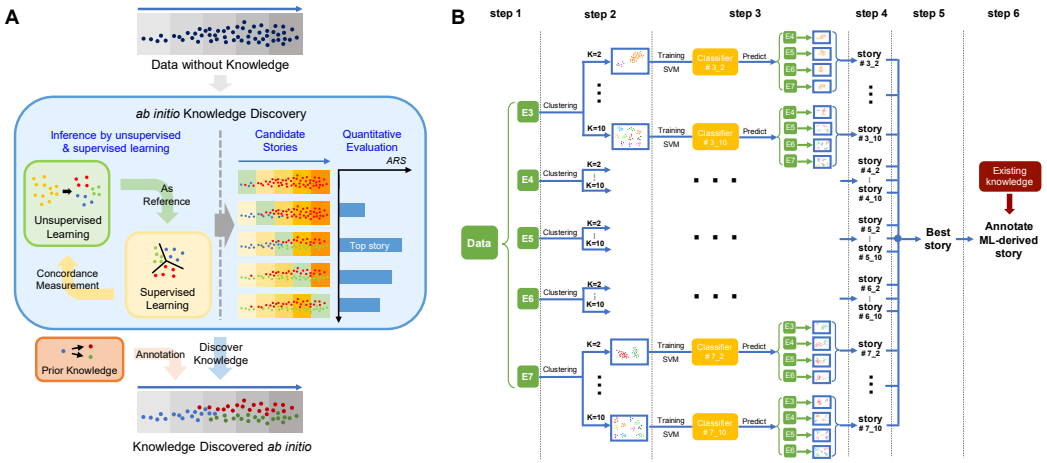

Figure 1: The strategy of *ab initio* knowledge discovery with ML. **(A)** Workflow. **(B)** Illustration of exhaustive search.

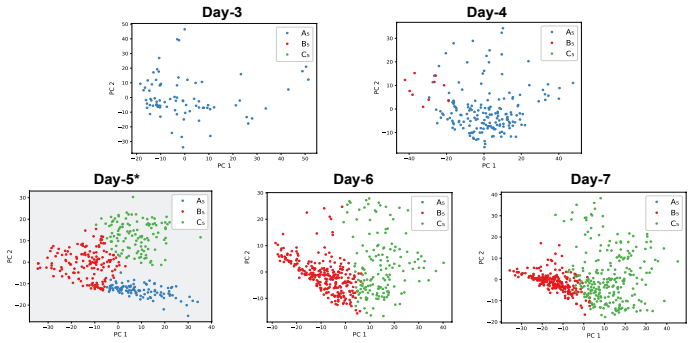

Figure 2: ML-derived developmental process using day-5 as reference (marked with *).

3 Experimental Results

3.1 ML-Derived Candidate Developmental Process

We first experimented the proposed strategy on the human embryonic development data. With no biological knowledge available, we conducted an exhaustive search on cluster numbers for each day as a potential reference (Figure 1B). For each day, we experimented with cluster numbers k being set from 2 to 10, respectively, and use the obtained clusters as reference to classify cells of other days. For each setting, we calculated the ARS for the particular reference day and cluster number. In this way, we enumerated the best possible candidate developmental processes using each day as a reference and each choice of cluster numbers within the given range (Table A1). Results showed that the developmental process derived using the 3 clusters of day-5 as reference gives the highest ARS (0.4674) among all enumerations. Figure 2 provides a PCA visualization of this ML-derived developmental process:

One lineage A_5 on day 3. A minor new lineage B_5 appears on day 4. It becomes larger on day 5, and another new lineage appears C_5 on day 5. The lineage A_5 disappears on day 6 and thereafter and B_5 and C_5 lineages continue thereafter.

3.2 Verification of *ab initio* Discovery with Known Biological Knowledge

We compared the ML-derived developmental process with existing biological knowledge and annotated the ML-derived lineages with biological lineages. According to the current understanding, from E3 to E7, human zygotes differentiate into 3 major embryonic cell types named pre-lineage,

trophectoderm (TE) and inner cell mass (ICM) lineages [6,8]. Cells of the pre-lineage are those that haven't started differentiation. TE lineage segregates first and then primitive endoderm (PE) and epiblast (EPI) cells come from the intermediate lineage of ICM [6,9]. Cells of different lineages play different roles in the embryogenesis. Cells in E3 and E4 belong to pre-lineage according to the current understanding. TE and ICM cells appear in E5 but there are still pre-lineage cells remaining in E5. ICM further segregates into EPI and PE in E5. In E6 and E7, all pre-lineage cells have differentiated into cells of either TE or ICM (EPI and PE) lineages.

Comparing this existing biological knowledge with the ML-derived developmental story in our discovery, it is straightforward to infer that cluster A_5 corresponds to the pre-lineage as it is the sole cell type in E3. To further identify TE and ICM, we took the clustering result of cells in day-5 with k=4 and compared it with the clusters of k=3 (Figure A1). We observed that cluster C_5 further split into two sub-clusters. Based on the existing knowledge that the ICM lineage is composed of two subtypes PE and EPI, we marked cluster B_5 as TE and C_5 as ICM. The ML-derived *ab initio* knowledge discovery on this particular dataset has been fully verified and annotated.

A minor disagreement between the ML-derived developmental process with the known biological lineages is that cells in E4 should be all of the pre-lineage according to the existing knowledge, but the ML-derived knowledge identified 10 "outlier" cells (out of the 190 cells) of E4 that are already differentiated. We drew the distribution of E5 cells in PCA plane and mapped all E4 cells to this plane. We observed that these 10 cells are more likely to be TE (Figure A2). Detailed experiment procedure is described in *Appendix*. Results indicate that a minor proportion of E4 cells grow faster and differentiate to cells with TE properties before E5, which updates existing knowledge.

3.3 Limitation in Revealing More Complex Patterns

We conducted the same series of experiments on the zebrafish embryonic development dataset [7]. Exhaustive search identified the time point of 10 hpf of 5 clusters as the most plausible reference. Figure A3 and A4 in *Appendix* show the PCA and tSNE plots of cells in each time point colored with the predicted classes. Referring to the paper published this dataset [7], we found that the ML-derived developmental process only covers a draft outline of the true biological knowledge lineages with many details missed. Compared with E3 to E7 for human embryonic cells, the zebrafish cells sampled from 4hpf to 24hpf covers a longer period of embryonic development and is much more complex. There is no single time point in which the cells can represent all lineages that have appeared in this developmental period, which is beyond the scenario our method is designed for. Although the ML-derived developmental process from this zebrafish dataset makes basic sense as a coarse outline, it reveals the limitation of the proposed method when the assumptions underlying the method cannot be met. Detailed analysis on zebrafish dataset is available in *Appendix*.

4 Discussion

There are many different scenarios with the need for mining underlying patterns from massive complex data. Successful applications of ML in many fields may give people an illusion that ML has already been shown powerful for knowledge discovery, but actually most of the successes are the joint products of ML and human knowledge. Involvement of knowledge can be in many forms like known marker genes, models or labeled training data [10]. Efforts for using only ML methods to discover knowledge from data are still rare not only in biology but also in many other fields. Bridging the gap towards *ab initio* knowledge discovery with ML is crucial to build better understanding of AI for science.

In a recent work in physics, scientists explored a neural network method for the *ab initio* discovery of the basic physical understanding that Earth orbits the Sun based on observations on movements of the Sun and Mars appearing from Earth [11] commented as an "AI Copernicus" [12]. Our experiment shows an example of the *ab initio* discovery of knowledge on early embryonic development from data using basic ML methods. The method is still in its infancy if expected to work on more complicated biological processes, but its success sheds lights on the future possibilities of developing more advanced ML methods for *ab initio* scientific discovery from data in fields that lack existing knowledge and challenge manual interpretation. Such advancement will not only empower the discovery of new knowledge in biology and other fields of science, but also move machine intelligence to the higher level of automatic knowledge learning and discovery.

Appendix

References

[1] LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature* **521**(7553):436-444.

[2] Brynjolfsson, E. & Mitchell, T. (2017) What can machine learning do? Workforce implications. *Science* **358**(6370):1530-1534.

[3] Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X. & Yan, F. (2018) Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**(5):1122-1131.

[4] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**(7873): 583-589.

[5] Shah, N., Li, J., Li, F., Chen, W., Gao, H., Chen, S., ... & Zhang, X. (2020) An Experiment on Ab Initio Discovery of Biological Knowledge from scRNA-Seq Data Using Machine Learning. *Patterns*, **1**(5):100071.

[6] Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Reyes, A.P., Linnarsson, S., Sandberg, R. & Lanner, F. (2016) Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165**(4):1012-1026.

[7] Wagner, D.E., Weinreb, C., Collins, Z.M., Briggs, J.A., Megason, S.G. & Klein, A.M. (2018) Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**(6392):981-987.

[8] Cockburn, K. & Rossant, J. (2010) Making the blastocyst: lessons from the mouse. *The Journal of clinical investigation* **120**(4):995-1003.

[9] Niwa, H., Toyooka, Y., Shimosato, D., Strumpf, D., Takahashi, K., Yagi, R. & Rossant, J. (2005) Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. *Cell* **123**(5):917-929.

[10] Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M.J. & Mahfouz, A. (2019) A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome biology* **20**(1):194.

[11] Iten, R., Metger, T., Wilming, H., Del Rio, L. & Renner, R. (2020) Discovering physical concepts with neural networks. *Physical Review Letters* **124**(1):010508.

[12] Castelvecchi, D. (2019) AI Copernicus 'discovers' that Earth orbits the Sun. *Nature* **575**(7782):266-267.

1 Supplemental Experiment Procedures

1.1 Data Pre-processing Descriptions

As scRNA-seq data are sparse, noisy, and of very high dimensionality, original cell representation using all genes cannot highlight biological differences among cells. In this study, we selected highly variable genes that present significant differences in expression levels among cells, so that expression patterns get enhanced.

For the human embryonic development dataset, we followed the procedures and the model in original paper [1,2] to select highly variable genes. Assuming the expression of a gene follows negative binomial distribution, the relationship between square of variance (cv^2) and mean (m) is:

$$cv^2 = \frac{1}{m} + \frac{1}{r} \tag{1}$$

where r is the over-dispersion parameter following a negative binomial distribution. We filtered out reference data [1,3] with cv^2 less than 3 and fitted the model to the remaining reference data.

Then we used the reference model as the threshold to select genes with larger variances (Figure A5). We obtained 490 and 954 highly variable genes for human dataset, which were used as features to study the cells.

For the zebrafish embryonic development dataset, we selected highly variable genes with the widely-used pipeline Seurat v3.1 [4]. We used the "FindVariableFeatures" function with "vst" selection

method, which identifies genes with the highest standardized variance. We merged cells from all time points together and identify top 500 variable genes for the dataset.

1.2 Experiments with other clustering and classification methods

Besides k-means clustering and SVM classification methods as the basic unsupervised and supervised ML methods in the *ab initio* knowledge discovery strategy, we also used Seurat clustering, Gaussian mixture model (GMM) and logistic regression as the alternative clustering and classification methods, respectively. Using GMM to replace k-means and logistic regression to replace SVM produced the same results as we got with k-means and SVM. We drew the PCA plots of ML-derived developmental process on human embryonic data (Figure A6 and A7).

1.3 Experiment on E4 Cells of Human Dataset

The ML-derived understanding on the developmental process indicates that a minor proportion of cells in E4 already differentiated to TE cells (cluster B_5). We drew the distribution of E5 cells in the plane of the first 2 principle components of E5, and map all E4 cells to this plane. We can see that while most E4 cells map to the region of pre-lineage cells (cluster A_5), 10 E4 cells map to the area of TE cells (cluster B_5) in E5. This confirmed the existence of TE cells in E4. We also mapped all E3 cells to this plane, which all mapped to the pre-lineage region (cluster A_5) (Figure A2). It is interesting to see that most E3 cells tend to map to the far end of the pre-lineage cluster, while the E4 cells are scatted in an almost linear manner in the cluster with the 10 cells extending to the area of TE cells. Considering the observations from the scree plots that the distinction between clusters in the data are not sharp, we speculated that the gene expression patterns of pre-lineage cells with those of the TE cells are of a continuum rather than a clear switch. A minor proportion of E4 cells grow faster and differentiate to cells with TE properties before E5.

1.4 Analysis of Experiment Results on Zebrafish Dataset

The *ab initio* discovery of the developmental process from this dataset only covers a draft outline of the true biological knowledge lineages with many details missed (Figure A3 and A4). We compared the nature of the human, mouse and zebrafish datasets we experimented in this study to understand why the proposed method works well on the first two datasets but has limited success on the zebrafish data. Looking into the basic knowledge on vertebrate development [5-8], we realized that the sampling time points in the human data of 3-7 dpc (days post coitum) are approximately from Cargenie State 2 to 5, long before the development of the first somite. The mouse data of 5.25 to 6.5 dpc are approximately from Cargenie Stage 5 to 6, still before the first somite occurs. The zebrafish data from 4 to 24 hpf, however, actually span approximately from Cargenie Stage 7 to 12. During this period, the zebrafish goes through blastula (2.25 - 5.25 hpf), gastrula (5.25 - 10.33 hpf), segmentation stages (10.33 - 24 hpf) and entering pharyngula stage [9]. In the end of 24 hpf, the zebrafish embryo already has more than 26 somites. From these facts, we can conceive that the zebrafish development data are beyond the scenario the proposed method was designed for. The clustering of cells in the zebrafish data are decided not only by the developmental lineages, but also by many other developmental factors such as somites and locations.

2 Supplemental Figures

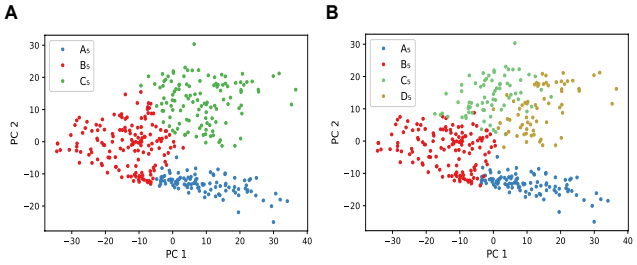

Figure A1: Comparison of clustering results on E5 cells with k=3 and k=4. (**A**) PCA plot of the 3 clusters. (**B**) PCA plot of the 4 clusters. The cluster C_5 when k=3 is further separated into two sub-clusters C_5 and D_5 when k=4.

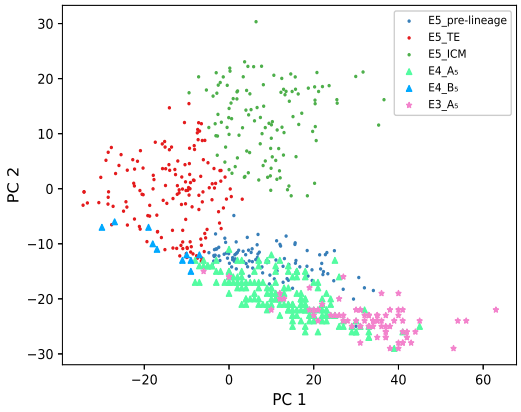

Figure A2: PCA plot of day-5 cells with day-3 and day-4 cells mapped onto it. We can see that all E3 cells map to the pre-lineage region of E5 cells, and most E3 cells are in the far end of this cluster. Most E4 cells map to the pre-lineage region along a linear shape, with 10 cells extended into the TE region.

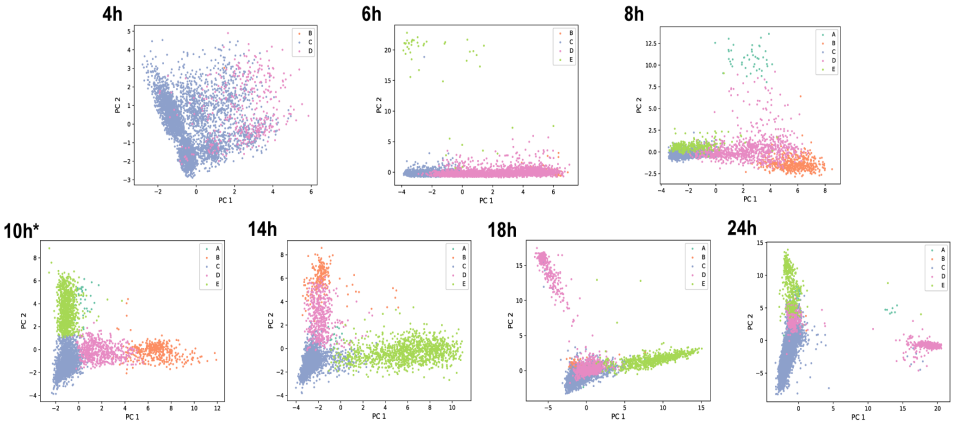

Figure A3: PCA plots of the ML-derived developmental process of zebrafish dataset. We used hour-10 cells of 5 clusters as reference for other hours. * means this time point is used as reference.

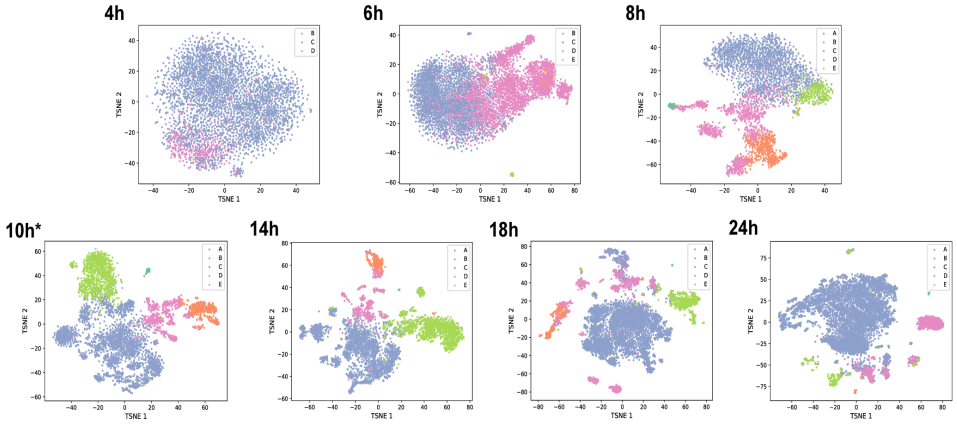

Figure A4: tSNE plots of the ML-derived developmental process of zebrafish dataset. We used hour-10 cells of 5 clusters as reference for other hours. * means this time point is used as reference.

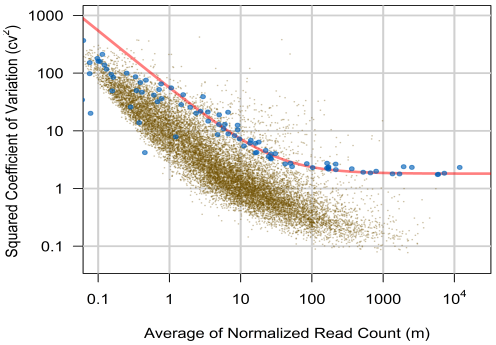

Figure A5: Selection of highly variable genes in human datasets. The horizontal axis is the average of normalized read count (m). The vertical axis is the squared coefficient of variation (cv^2). Each brown point represents one gene observed in the sequencing experiments. Blue points are the reference data. We chose the reference data with cv^2 larger than 3 and fitted negative binomial model, shown in red curve. We selected genes above the red curve as the highly variable genes.

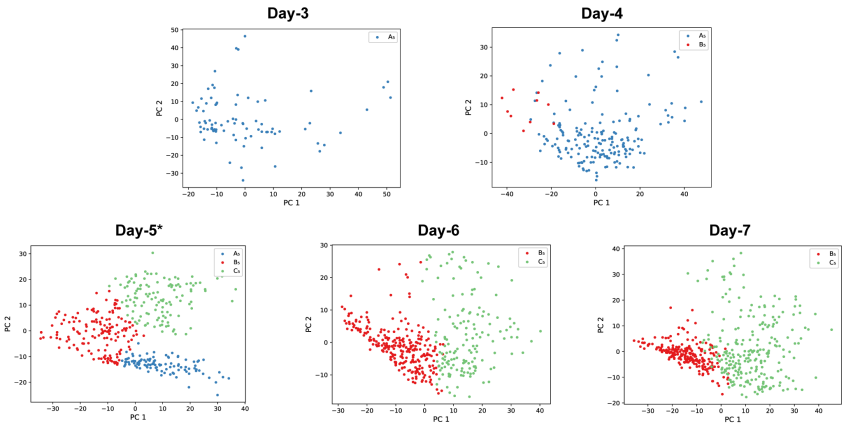

Figure A6: PCA plots for the ML-derived developmental process using **GMM** clustering and SVM classification on the human embryonic data.

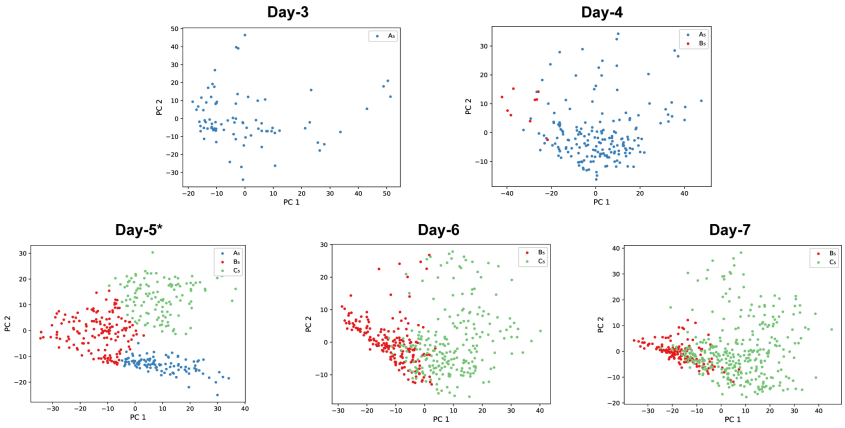

Figure A7: PCA plots for the ML-derived developmental process using k-means clustering and **logistic regression** classification on the human embryonic data.

3 Supplemental Tables

Table A1: Adjusted reliability scores (ARSs) of each enumerated candidate developmental process

Reference	ARS	Reference	ARS	Reference	ARS
day3-clu2	0	day4-clu8	0.1367	day6-clu5	0.1426
day3-clu3	-0.0002	day4-clu9	0.2079	day6-clu6	0.2565
day3-clu4	-0.0003	day4-clu10	0.2045	day6-clu7	0.1824
day3-clu5	-0.0003	day5-clu2	0.4220	day6-clu8	0.1848
day3-clu6	-0.0013	**day5-clu3**	**0.4674**	day6-clu9	0.1812
day3-clu7	-0.0002	day5-clu4	0.1936	day6-clu10	0.1655
day3-clu8	-0.0004	day5-clu5	0.2130	day7-clu2	0.2434
day3-clu9	-0.0004	day5-clu6	0.1703	day7-clu3	0.1706
day3-clu10	-0.0003	day5-clu7	0.2463	day7-clu4	0.2497
day4-clu2	0.0011	day5-clu8	0.2408	day7-clu5	0.2199
day4-clu3	0.0166	day5-clu9	0.2124	day7-clu6	0.2552
day4-clu4	0.0256	day5-clu10	0.2317	day7-clu7	0.1693
day4-clu5	0.1123	day6-clu2	0.4099	day7-clu8	0.2074
day4-clu6	0.0479	day6-clu3	0.2362	day7-clu9	0.2031
day4-clu7	0.0610	day6-clu4	0.1543	day7-clu10	0.1816

day3-clu2 means using Day-3 cells of 2 clusters as the reference for other days for building the candidate developmental process.

Supplemental References

[1] Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Reyes, A.P., Linnarsson, S., Sandberg, R. & Lanner, F. (2016) Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165**(4):1012-1026.

[2] Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A. & Marioni, J.C. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods* **10**(11):1093.

[3] Cheng, S., Pei, Y., He, L., Peng, G., Reinius, B., Tam, P.P., Jing, N. & Deng, Q. (2019) Single-cell RNA-seq reveals cellular heterogeneity of pluripotency transition and x chromosome dynamics during early mouse development. *Cell reports* **26**(10):2593-2607.

[4] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W.M., Hao, Y., Stoeckius, M., Smibert, P. & Satija, R. (2019) Comprehensive integration of single-cell data. *Cell* **177**(7):1888-1902.

[5] Otis, E.M. & Brent, R. (1954) Equivalent ages in mouse and human embryos. *The Anatomical Record* **120**(1):33-63.

[6] O'Rahilly, R. (1979) Early human development and the chief sources of information on staged human embryos. *European Journal of Obstetrics Gynecology and Reproductive Biology* **9**(4):273-280.

[7] Theiler, K. (1989) *The House Mouse.* Springer-Verlag.

[8] O'Rahilly, R. & Müller, F. (2010) Developmental Stages in Human Embryos: Revised and New Measurements. *Cells Tissues Organs* **192**(2):73-84.

[9] Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B. & Schilling, T.F. (1995) Stages of embryonic development of the zebrafish. *Developmental dynamics* **203**(3):253-310.