PATTER, Volume *1*

Supplemental Information

An Experiment on *Ab Initio* Discovery of Biological Knowledge from scRNA-Seq Data Using Machine Learning

Najeebullah Shah, Jiaqi Li, Fanhong Li, Wenchang Chen, Haoxiang Gao, Sijie Chen, Kui Hua, and Xuegong Zhang

Supplemental Figures

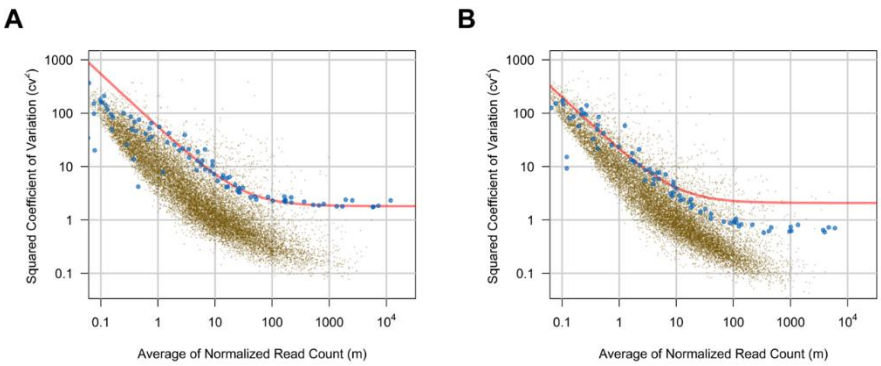

Figure S1. Selection of highly variable genes in human and mouse datasets.

(A) and **(B)** are illustrations for human and mouse embryonic datasets, respectively. 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.

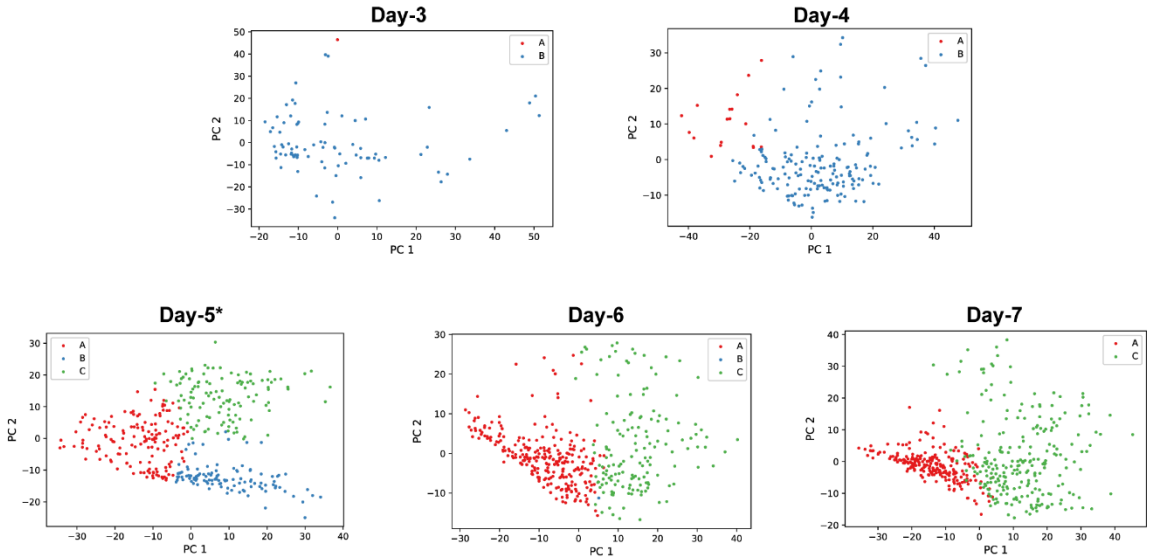

Figure S2. PCA plots of the top story using manually tuned Seurat parameters on the human embryonic data.

The k-means clustering method is replaced by Seurat while the classification method is still SVM. We manually tuned Seurat parameters for clustering to get the maximum ARI with k-means results for each day. Then we calculated the ARS using each clustering result as reference and found day-5 is the best reference day. The resulting developmental story is presented here. The results are similar to those from k-means clustering and SVM classification. "*" indicates that this time point is used as reference.

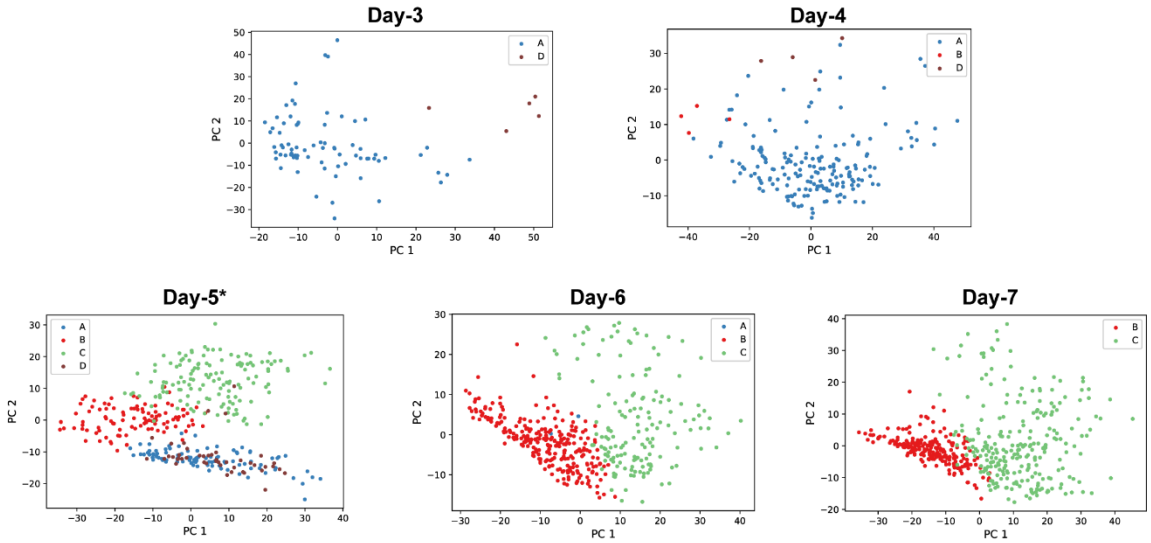

Figure S3. PCA plots of the top story using Seurat with exhaustive searching for parameters on the human embryonic data.

The clustering method is replaced by Seurat while the classification method is still SVM. The clustering on reference day-5 was achieved with dims=1:5, k.param=10 and resolution=0.28. "*" indicates that this time point is used as reference.

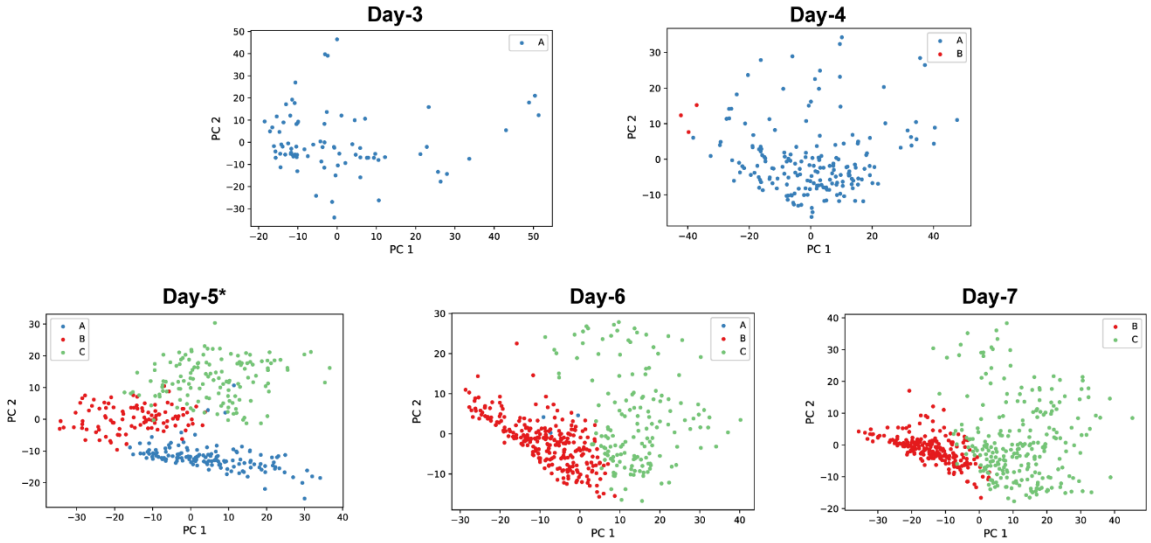

Figure S4. PCA plots of the second top story using Seurat with exhaustive searching for parameters on the human embryonic data.

The clustering method is replaced by Seurat while the classification method is still SVM. The clustering on reference day-5 was achieved with dims=1:5, k.param=10 and resolution=0.14. The results are similar to those from k-means clustering and SVM classification. "*" indicates that this time point is used as reference.

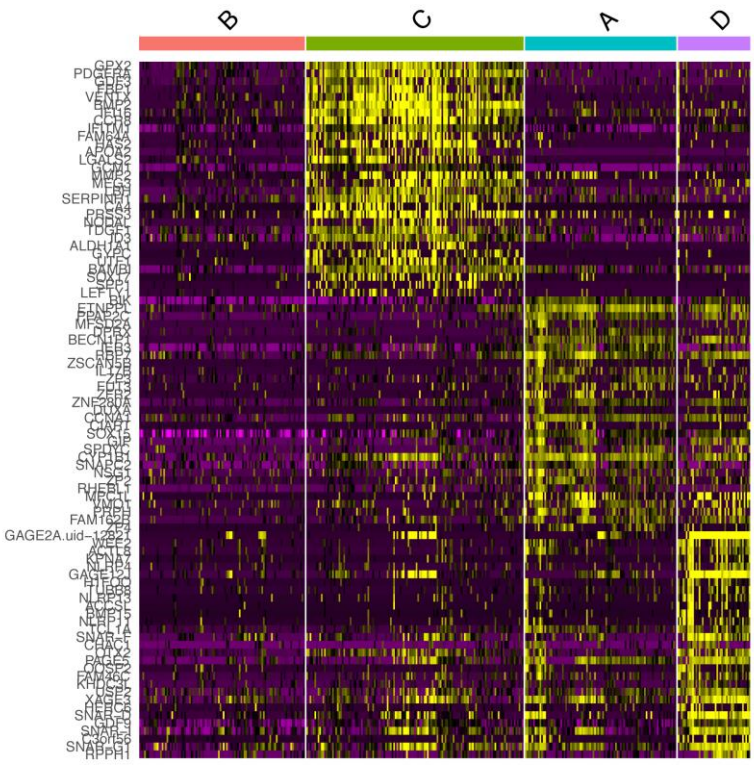

Figure S5. Expression heatmap of E5 cells in the story using day-5 with 4 clusters as reference on the human embryonic data.

We identified differentially expressed (DE) genes for each cluster using Seurat and visualized the expression patterns of E5 cells with heatmap. Each row represents one gene and each column represents one cell. The bar above shows the cluster labels of cells. Top 30 DE genes for each cluster are drawn in thie heatmap.

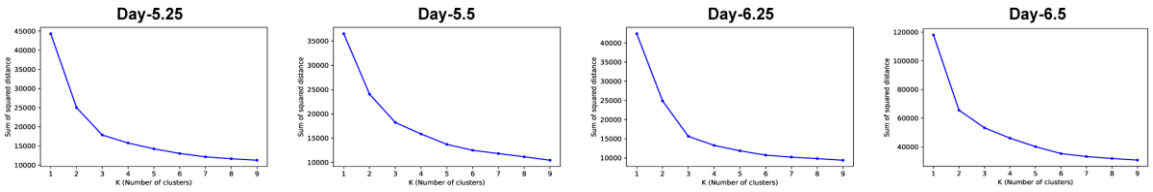

Figure S6. Scree plots of sum-of-errors of k-means clustering on each time point of the mouse embryonic data.

The horizontal axis is the cluster number k. The vertical axis is the sum of errors of samples to cluster centers. Weak elbow points can be identified for all the time points.

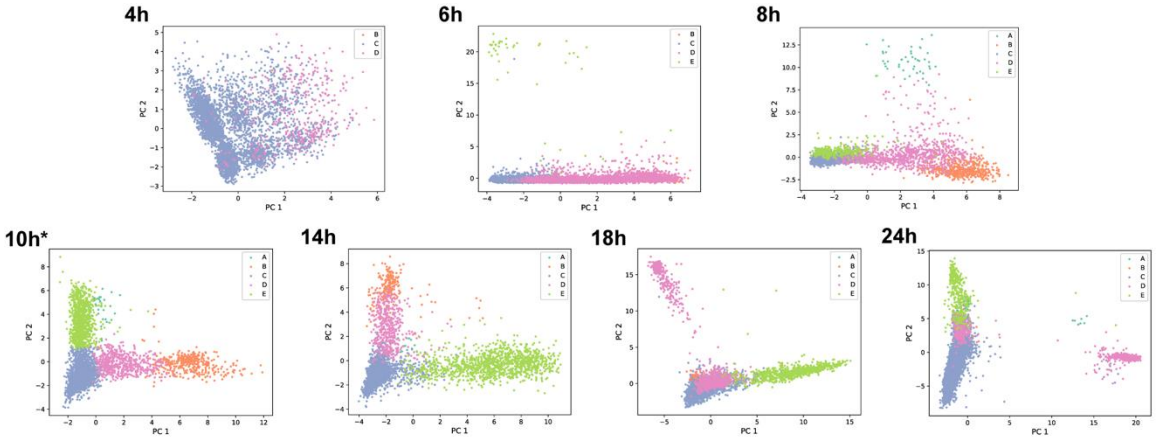

Figure S7. PCA plots of the h10_c5 story of zebrafish dataset.

We used hour-10 cells of 5 clusters as reference for other hours. * means this time point is used as reference.

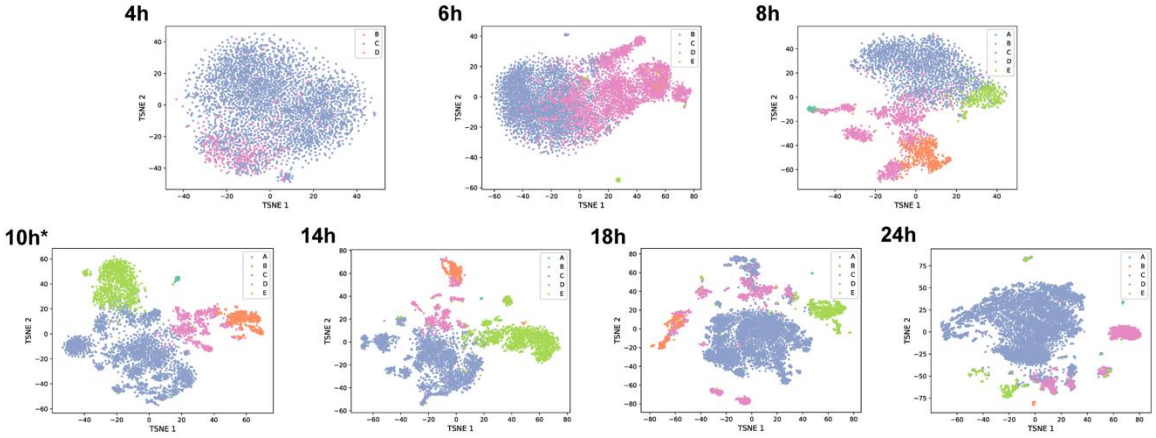

Figure S8. tSNE plots of the h10_c5 story of zebrafish dataset.

We used hour-10 cells of 5 clusters as reference for other hours. * means this time point is used as reference.

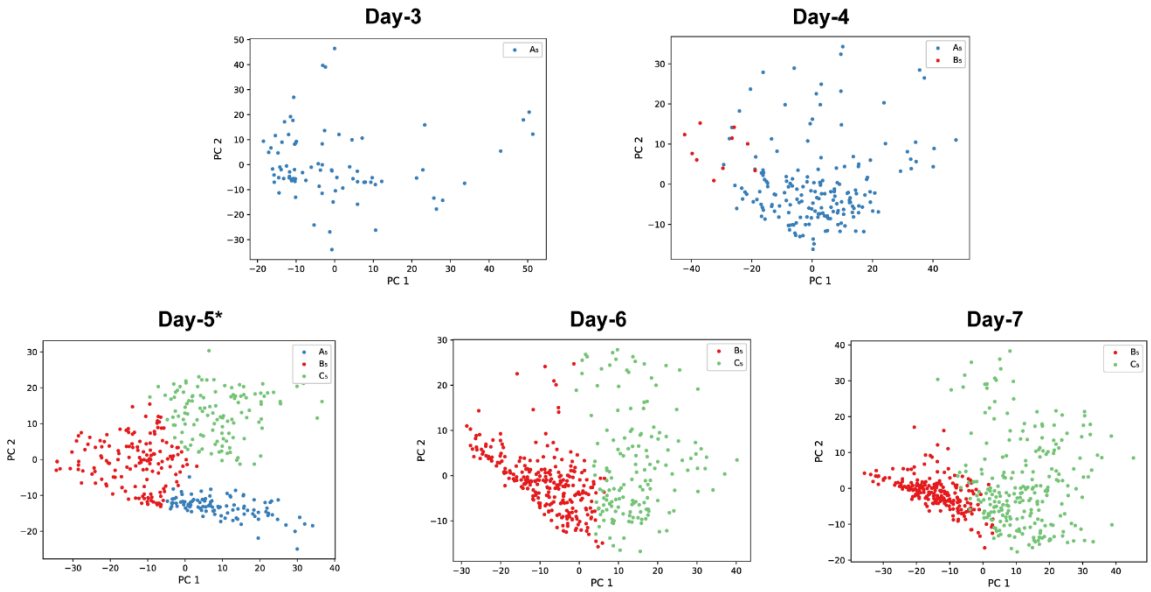

Figure S9. PCA plots for the story #5 using GMM clustering and SVM classification on the human embryonic data.

The clustering method is replaced by Gaussian mixture model (GMM) while the classification method is still SVM. The results are similar to those from k-means clustering and SVM classification. "*" indicates that this time point is used as reference.

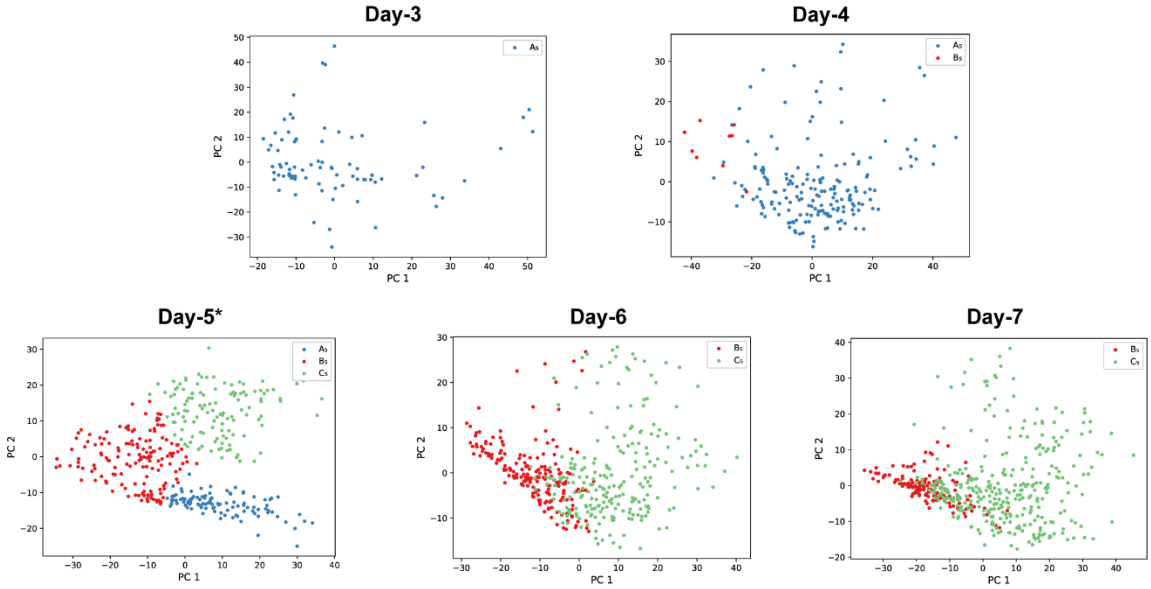

Figure S10. PCA plots for the story #5 using k-means clustering and logistic regression classification on the human embryonic data.

The clustering method is still k-means while the classification method is replaced by logistic regression. The results are similar to those from k-means clustering and SVM classification. "*" indicates this time point is used as reference.

Supplemental Tables

Table S1. Software Used in This Study

Algorithm or Calculation	Package	Version	Parameters
Feature Selection	statmod[5] (R)	1.4.32	default
Feature Selection	Seurat[4] (R)	3.1	nfeatures=500, other parameters as default.
Silhouette Score	scikit-learn[6] (Python)	0.21.2	metric='euclidean', other parameters as default.
K-means	scikit-learn (Python)	0.21.2	random_state=0, other parameters as default.
Seurat Clustering	Seurat (R)	3.1	dims, k.param and resolution parameters are searched for the highest ARS. Other parameters as default.
SVM	scikit-learn (Python)	0.21.2	kernel = 'rbf', gamma=0.0001, other parameters as default.
PCA	scikit-learn (Python)	0.21.2	n_components=2, other parameters as default.
t-SNE	scikit-learn (Python)	0.21.2	random_state=100, other parameters as default.
Plot Drawing	**Package**	**Version**	**Parameters**
Feature Selection	ggplot2[7] (R)	3.2.0	-
Other Plots	Matplotlib[8] (Python)	0.21.2	-

Table S2. Silhouette scores of different cluster numbers in each time point of the mouse embryonic data

k	Day-5.25	Day-5.5	Day-6.25	Day-6.5
2	0.3950	0.3187	0.4005	**0.4160**
3	**0.4159**	**0.3548**	**0.4343**	0.4075
4	0.4062	0.3160	0.4175	0.3623
5	0.3453	0.2760	0.3239	0.3759
6	0.2986	0.2694	0.2880	0.3282

* Note: we marked the highest Silhouette score in each time point in bold.

Table S3. Concordance and reliability scores of each time point and candidate developmental process in the mouse embryonic data

Reference day (r)	$concord(i\|r)$				$reliab(r)$	$ARS(r)$
	$i = 5.25$	$i = 5.5$	$i = 6.25$	$i = 6.5$		
Day-5.25	**-**	**0.97**	**1**	**0.65**	**0.88**	**2.00**
Day-5.5	0.91	-	0.95	0.70	0.85	1.94
Day-6.25	0.95	0.83	-	0.71	0.83	1.91
Day-6.5	0.61	0.42	0.52	-	0.52	1.32

* Note: day-5.25 is selected as the reference day as it achieves the highest ARS value.

Table S4. Numbers of cells in the clusters of reference time point and in the classes of the other time points in the mouse embryonic data

Reference Day (# of clusters)	Number of cells in clusters/classes			
	Day-5.25	Day-5.5	Day-6.25	Day-6.5
Day-5.25 (3)	(137, 126,68)	(108,114,47)	(87,142,92)	(304,411,88)
Day-5.5 (3)	(139,133,59)	(109,116,44)	(93,143,85)	(335,388,80)
Day-6.25 (3)	(131,127,73)	(96,120,53)	(87,142,92)	(304,395,104)
Day-6.5 (2)	(132,199)	(87,182)	(90, 231)	(312, 491)

Table S5. Adjusted reliability scores (ARSs) of each enumerated candidate developmental process in the mouse embryonic data

Reference day & cluster number*	*ARS*	Reference day & cluster number*	*ARS*	Reference day & cluster number*	*ARS*
day5.25_clu2	1.9341	day5.5_clu5	1.5691	day6.25_clu8	1.5980
day5.25_clu3	**2.2598**	day5.5_clu6	-0.0307	day6.25_clu9	1.5980
day5.25_clu4	2.1879	day5.5_clu7	2.1482	day6.25_clu10	1.5560
day5.25_clu5	2.0604	day5.5_clu8	0.0	day6.5_clu2	1.9338
day5.25_clu6	2.0780	day5.5_clu9	1.1219	day6.5_clu3	2.2320
day5.25_clu7	1.5770	day5.5_clu10	1.7041	day6.5_clu4	2.0019
day5.25_clu8	1.5435	day6.25_clu2	1.9809	day6.5_clu5	1.4101
day5.25_clu9	1.8974	day6.25_clu3	2.2160	day6.5_clu6	1.4213
day5.25_clu10	1.8974	day6.25_clu4	2.1591	day6.5_clu7	1.3713
day5.5_clu2	1.9951	day6.25_clu5	1.8837	day6.5_clu8	1.4347
day5.5_clu3	2.0020	day6.25_clu6	1.6789	day6.5_clu9	1.5941
day5.5_clu4	1.8394	day6.25_clu7	2.0547	day6.5_clu10	1.3081

* Note: day5.25_clu2 means using Day-5.25 cells of 2 clusters as the reference for other days for building the candidate developmental process. The ARS measures the plausibility of each candidate story. The reference of day-5.25 with 3 clusters achieves the highest ARS value.

Table S6. Manual annotation on the ML-derived developmental process of zebrafish dataset

Cluster	Hour-4	Hour-6	Hour-8	Hour-10	Hour-14	Hour-18	Hour-24
A	-	-	Mesoderm	Mesoderm	Mesoderm	Mesoderm	Mesoderm
B	-	Mesoderm (Endoderm)	Mesoderm (Other)	Mesoderm	Mesoderm	Mesoderm	Mesoderm
C	Unknown	Epiblast, Mesoderm (Endoderm)	Mesoderm, Neural (Other)	Neural	Neural	Neural	Neural
D	Epiblast	Epiblast, Mesoderm, Endoderm	Mesoderm, Other (Neural)	Mesoderm	Mesoderm	Mesoderm	Mesoderm
E	-	Epidermal	Epidermal	Epidermal	Epidermal	Epidermal	Epidermal (Mesoderm, Endoderm)

Note: "-" means this cluster does not exist at certain time point (or has very few cells). "Unknown" means we cannot not map this cluster to any lineage (differentially expressed genes do not exist in the reference gene list). Lineages are colored similarly as reported in the Wagner's paper[9].

Table S7. ARIs between k-means clusters with different initial centroids on day-5 of the human data

Experiment ID	0	1	2	3	4	5
0	1	1	1	1	0.98	1
1		1	1	1	0.98	1
2			1	1	0.98	1
3				1	0.98	1
4					1	0.98
5						1

Note: Experiment 0 is the one reported in the main text.

Table S8. ARSs for each day with different initial centroids in k-means clustering on the human data

Experiment ID	Day-3	Day-4	Day-5	Day-6	Day-7
0	0.0	0.0034	0.4035	0.2633	0.1873
1	0.0	0.0034	0.4022	0.2675	0.1904
2	0.0	0.0034	0.4022	0.2675	0.1904
3	0.0	0.0034	0.4022	0.2675	0.1904
4	0.0	0.0033	0.4004	0.2638	0.1881
5	0.0	0.0034	0.4022	0.2675	0.1904

Table S9. Number of cells in each cluster with different initial centroids in k-means clustering on the human data

Experiment ID	Cluster 1	Cluster 2	Cluster 3
0	152	121	104
1	152	121	104
2	152	121	104
3	152	121	104
4	151	121	105
5	152	121	104

Supplemental Experimental Procedures

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 expressional patterns get enhanced.

For the human and mouse embryonic development datasets, 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:

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

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 $cv^2 \sim m$ model to the remaining reference data. Then we used the reference model as the threshold to select genes with larger variances (Figure S1). We obtained 490 and 954 highly variable genes for human and mouse datasets, respectively, 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.

Experimental procedure of exhaustive searching with Seurat clustering

Following the procedure of exhaustive searching on the reference day and cluster numbers using k-means, we conducted a new experiment and employed Seurat as the clustering method. There are 3 major parameters in Seurat that affect clustering results: "dims", "k.param" and "resolution". We used the exhaustive search strategy to look for the combination of parameters that results in the highest ARS after clustering and prediction. The search range is [5, 10], [10, 150], [0.1, 1.2] and the interval is 5, 10, 0.01 for "dims", "k.param" and "resolution" parameters, respectively. It is similar to the exhaustive search we used for k-means, but here the Seurat clustering results with each parameter setting in each day are used as individual reference. So it is possible there are multiple candidate references for each day with the same cluster number. For each setting, the predicted classes on the target days were compared with clustering results of those days. We chose the clustering result that has the highest *concord* score with predicted clusters, and calculated the corresponding *reliab* score. In this way, we enumerated the best possible candidate developmental processes using each parameter combination as a reference. Results showed that the developmental process derived using the 4 clusters of day-5 as reference gives the highest ARS (0.43) among all enumerations. The reference of day-5 with 3 clusters gives the second highest ARS (0.32). We visualized these two stories in PCA plots (Figure S3 and S4).

Consistency of k-means clustering in the experiments

In this study, we employed k-means clustering to group cells of reference day into clusters. The initial centroids of k-means algorithm are set randomly, which may cause instability of results. To check the consistency of using k-means in our experiments, we repeated k-means clustering experiments with other 5 initial centroids (by setting different "random_state" parameter in sklearn package) on human embryonic data. We calculated ARI between clustering results on day-5 (Table S7). Following the same procedure as previous work, we calculated the ARS for each day

(Table S8) and the number of cells in each cluster (Table S9) for each replicated experiment. The highest ARS scores in all experiments pointed to the same conclusion, and their ARS scores are also close. Results show that we achieved nearly the same results in the 5 new runs of the experiment as our previous one, which indicates k-means is a consistent clustering method in our experiments.

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. The experiments with Seurat clustering on human embryonic data is described in the main text and results are given in Figures S2 to S5. 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 story #5 on human embryonic data (Figure S9 and S10).

Pseudo-Code of Experiments

Pseudo-Code for Self-Consistency Evaluation Method

The self-consistency evaluation method calculates the *adjusted reliability scores* (ARS), which contains 3 algorithms. While running algorithm 3, we need to run algorithm 1 and 2 to obtain cluster labels, *concord* and *reliab* scores.

Algorithm 1 Clustering of all day samples individually

```
gt = initialize clustering labels of each day samples
for k = 3, 4, 5, 6, 7 do
    X_k = Fetch day k samples
    n_k = Optimal clusters for day k samples using silhuoette coefficient
    gt{k} = KMEANS(X_k,n_k)
end for
```

Algorithm 2 Calculating concordance and reliability for each day

```
concord = initialize ARI scores of each day svm model
reliab = initialize reliability score of each day as reference
for i = 3, 4, 5, 6, 7 do
    X_i = Fetch day i samples
    lab_i = Get ground truth labels for day i from gt
    svm_model = TRAINSVM(X_i, lab_i)
    ari_scores = initialize ARI score of day i as reference
    ari_index = initialize with 0 for ari_scores array
    for j = 3, 4, 5, 6, 7 do
        if j != i then
            X_j = Fetch day j samples
            lab_j = svm_model-> PREDICT(X_j)
            gt_j = Get ground truth labels for day j
            ari_scores[ari_index] = ADJUSTEDRANDSCORE(gt_j, lab_j)
            ari_index++
        end if
    end for
    concord{i} = ari_scores
    reliab{i} = MEAN(ari_scores)
end for
```

Algorithm 3 Calculating Adjusted ARI Scores (ARS) for reference day selection

```
ars = initialize Adjusted ARI scores for each day as reference
for m = 3, 4, 5, 6, 7 do
    reliab_m = Fetch reliability score for all days except m
    concord_m = Fetch concordance scores for day m as reference
    ars{m} = SUM(DOTPRODUCT(reliab_m, concord_m))
end for
reference_day = output day with maximum Adjusted ARI Score
```

Pseudo-Code for the Exhaustive Searching Method

The exhaustive search method calculates the *adjusted reliability scores* (ARS) for multiple clustering results on each time point, which contains 2 algorithms. While running algorithm 2, we need to run algorithm 1 to obtain *concord* and *reliab* scores.

Algorithm 4 Calculating concordance, reliability and concordance map for each combination of day with clusters

```
concord = initialize ARI scores
concord_map = initialize to map concord
reliab = initialize reliability score
for k = 2, 3, 4, 5, 6, 7, 8, 9, 10 do
    for i = 3, 4, 5, 6, 7 do
        X_i = Fetch day i samples
        lab_i_k = KMEANS(X_i,k)
        svm_model = TRAINSVM(X_i, lab_i_k)
        ari_scores_i_k = initialize for day i with cluster k as reference
        concord_map_i_k = initialize for day i with cluster k as reference
        index = initialize with 0
        for j = 3, 4, 5, 6, 7 do
            if j != i then
                X_j = Fetch day j samples
                lab_j = svm_model-> PREDICT(X_j)
                clus_j = COUNTCLUSTERNUMBER(lab_j)
                if clus_j == 1 then
                    clus_j = 2
                end if
                gt_j = KMEANS(X_j,clus_j)
                ari_scores_i_k[index] = ADJUSTEDRANDSCORE(gt_j, lab_j)
                concord_map_i_k [index] = j, clus
                index++
            end if
        end for
        concord{i,k} = ari_scores_i_k
        concord_map{i,k} = concord_map_i_k
        reliab{i,k} = MEAN(ari_scores_i_k)
    end for
end for
```

Algorithm 5 Calculating Adjusted ARI Scores (ARS) for reference day with cluster selection

```
ars = initialize Adjusted ARI Scores for each day with cluster as reference
for concord_item, concord_map_item in concord, concord_map do
    index = initialize with 0
    reliab_m_k = initialize for day m with k clusters
    for m,k in concord_map_item do
        reliab_m_k[index] = Fetch reliab for day m with k clusters
        index++
    end for
    ars{m,k} = SUM(DOTPRODUCT(reliab_m_k, concord_item))
end for
reference_day_cluster = output day with cluster having maximum ARS
```

Supplemental References

1. Petropoulos, S., Edsgärd, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Reyes, A.P., 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.
2. Brennecke, P., Anders, S., Kim, J.K., Kołodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., and Marioni, J.C. (2013). Accounting for technical noise in single-cell RNA-seq experiments. Nature Methods 10, 1093.
3. Cheng, S., Pei, Y., He, L., Peng, G., Reinius, B., Tam, P.P., 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 reports 26, 2593-2607.
4. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck III, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell 177, 1888-1902. e1821.
5. Giner, G., and Smyth, G. K. (2016). Statmod: Probability calculations for the inverse Gaussian distribution. The R Journal 8, 339–351.
6. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., and Dubourg, V. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research 12, 2825-2830.
7. Wickham, H. (2016). ggplot2: elegant graphics for data analysis. (Springer).
8. Caswell, T., Droettboom, M., Hunter, J., Lee, A., Firing, E., Stansby, D., Klymak, J., de Andrade, E., Nielsen, J., and Varoquaux, N. (2019). Matplotlib: 3.1.1. https://zenodo.org/record/3264781.
9. Wagner, D.E., Weinreb, C., Collins, Z.M., Briggs, J.A., Megason, S.G., and Klein, A.M. (2018). Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science 360, 981-987.