

Supplementary Information for

Time-series single-cell transcriptomic profiling of luteal-phase endometrium uncovers dynamic characteristics and its dysregulation in recurrent implantation failures

Dandan Cao¹, Yijun Liu^{2,3}, Yanfei Cheng¹, Jue Wang¹, Bolun Zhang^{1,4}, Yanhui Zhai¹, Kongfu Zhu¹, Ye Liu¹, Ye Shang¹, Xiao Xiao⁵, Yi Chang³, Yin Lau Lee^{1,6,7}, William Shu Biu Yeung^{1,6,7}, Yuanhua Huang^{2,7,8}, Yuanqing Yao^{1,9}

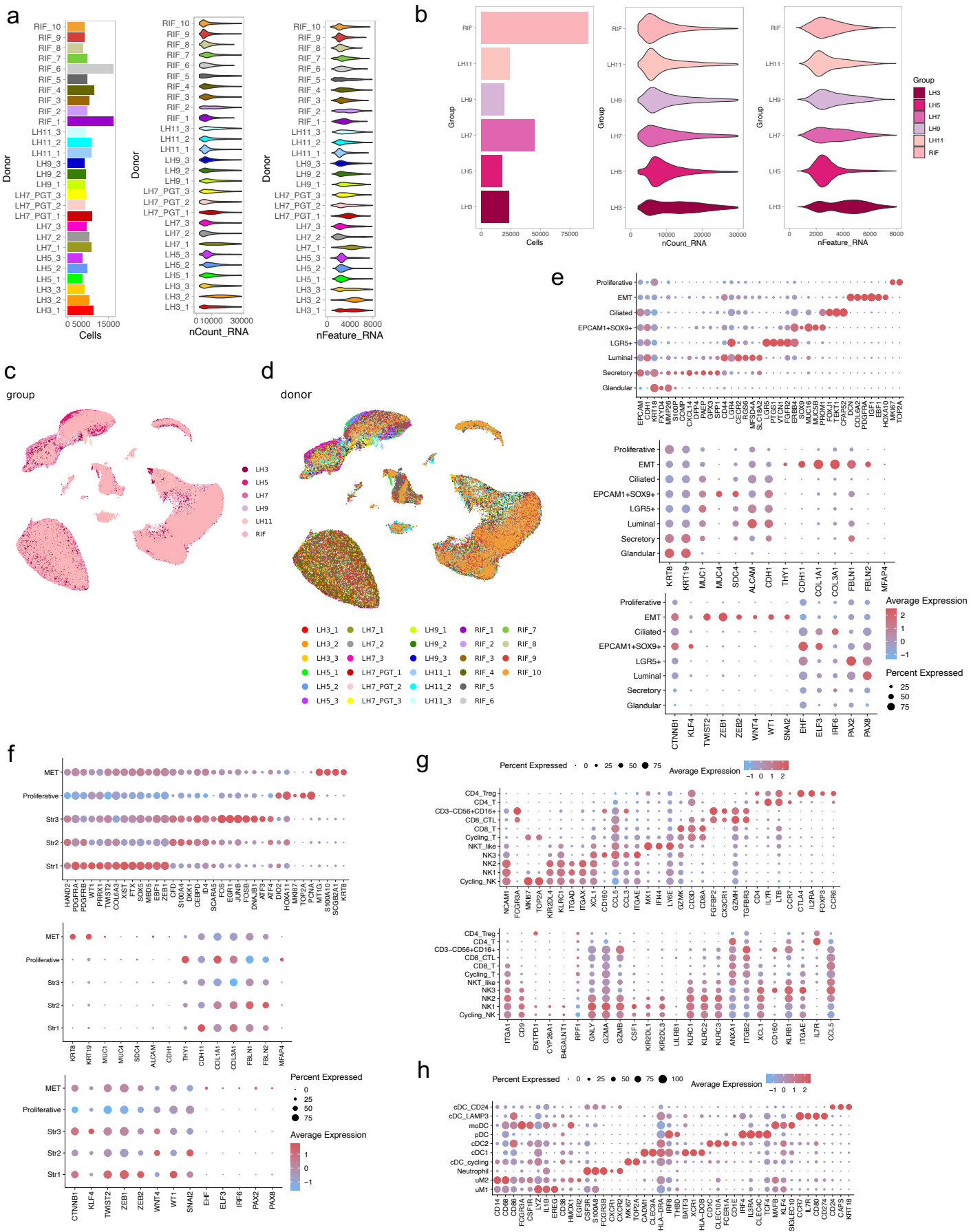
¹Shenzhen Key Laboratory of Fertility Regulation, Reproductive Medicine Center, the University of Hong Kong-Shenzhen Hospital, Shenzhen, China. ²School of Biomedical Sciences, the University of Hong Kong, Hong Kong SAR, China. ³School of Artificial Intelligence, Jilin University, Jilin, China. ⁴School of Medicine, Nankai University, Tianjin, China. ⁵Genomics Institute, Geneplus-Shenzhen, Shenzhen, China. ⁶Department of Obstetrics and Gynaecology, the University of Hong Kong, Hong Kong SAR, China. ⁷Centre for Translational Stem Cell Biology, Building 17 W, The Hong Kong Science and Technology Park, Hong Kong SAR, China. ⁸Department of Statistics and Actuarial Science, the University of Hong Kong, Hong Kong SAR, China. ⁹Department of Gynecology and Obstetrics, Chinese PLA General Hospital, Beijing, China

Correspondence: Yuanqing Yao, Shenzhen Key Laboratory of Fertility Regulation, Reproductive Medicine Center, the University of Hong Kong-Shenzhen Hospital, Shenzhen, China. E-mail: yaoyq_hkusz@126.com

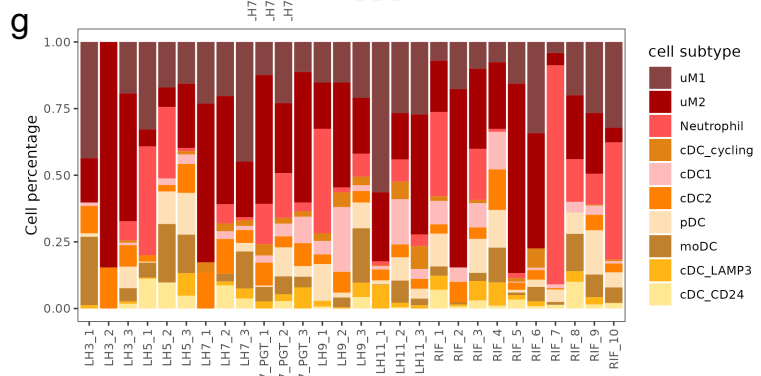
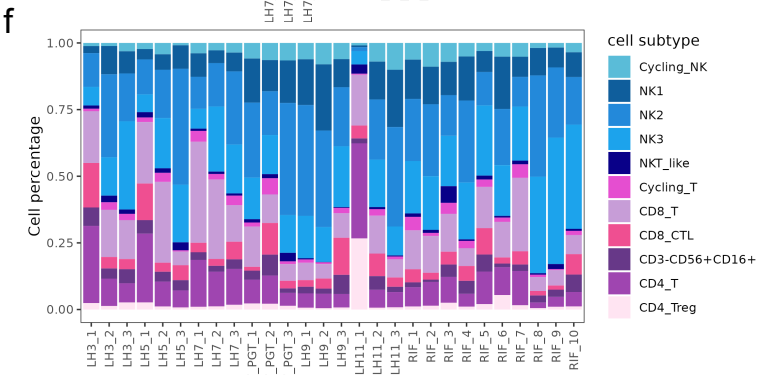
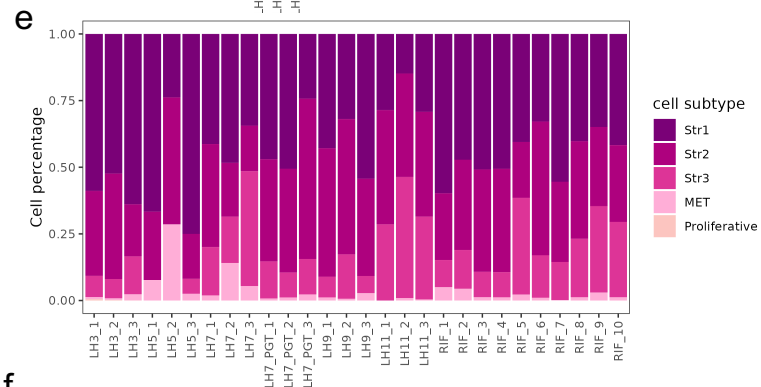
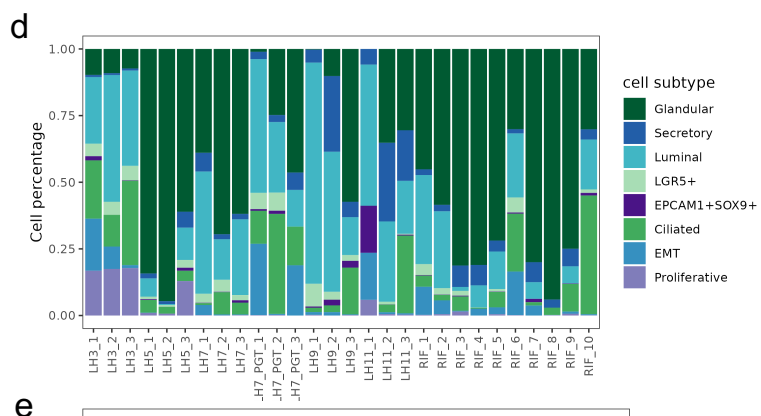
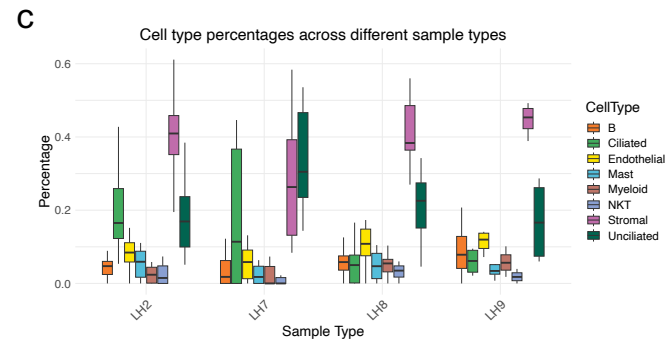
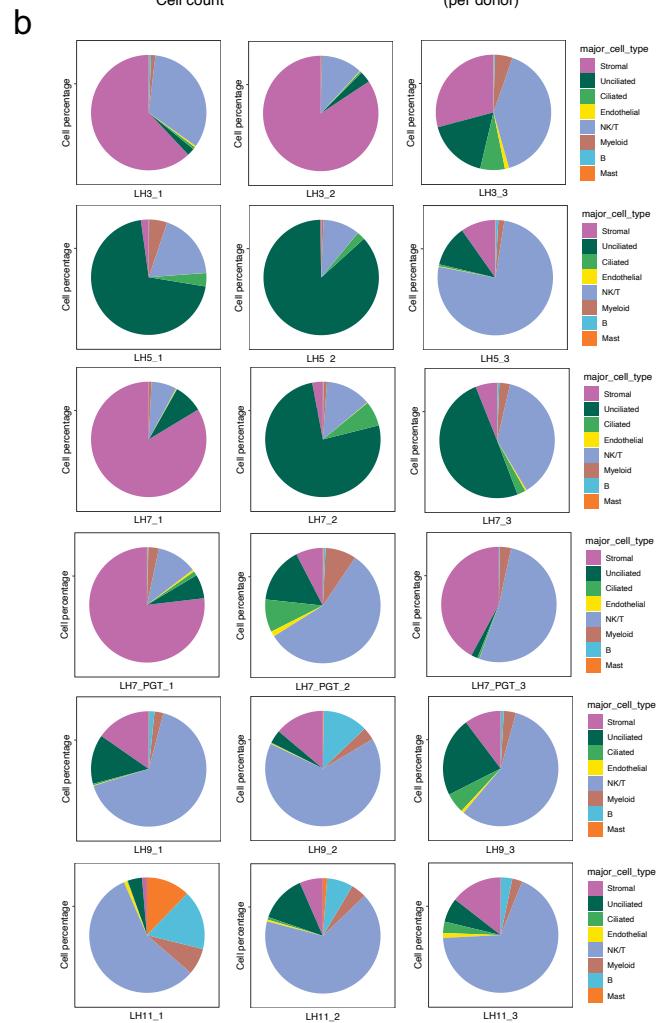
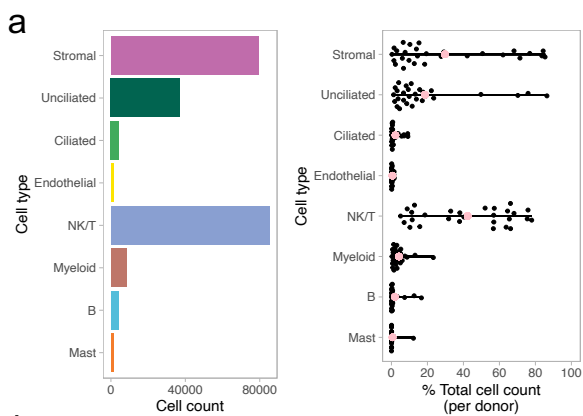
Table of Contents

Supplementary Figures 1-22

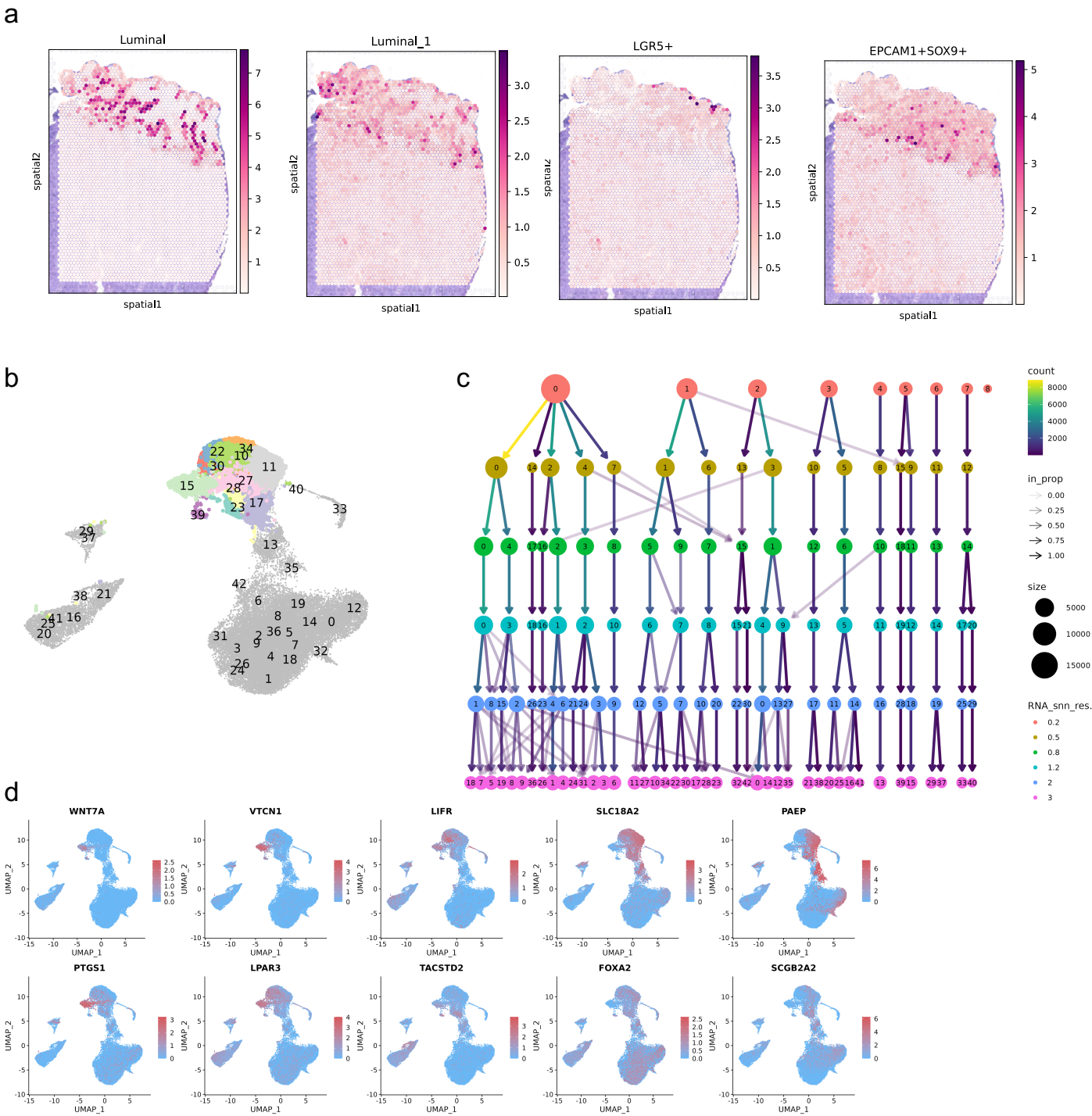
Supplementary Note for StemVAE

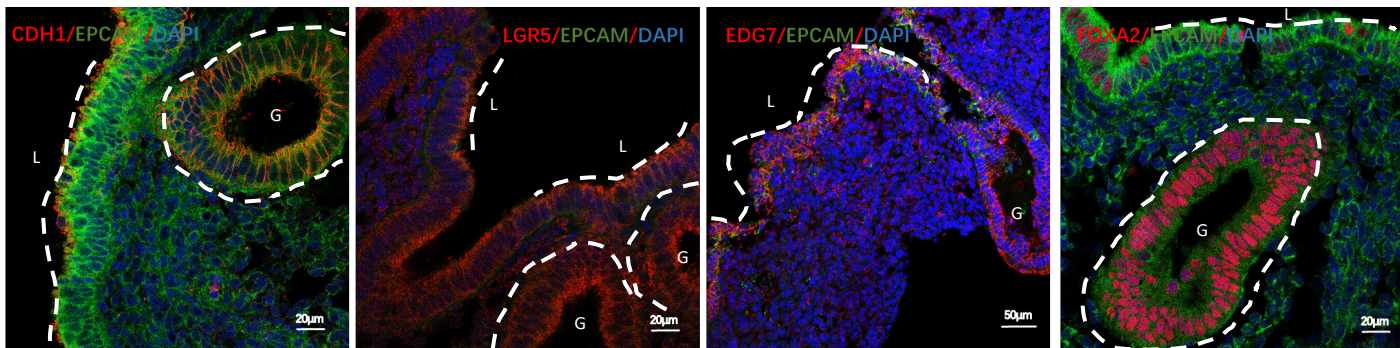


Supplementary Fig. 1 | Quality control matrices for the 28 donors. **a,b**, Distribution of the number of single cells, number of molecules, and number of genes in each donor (**a**) and in each group (**b**). **c,d**, UMAP plots showing the cell distribution colored by group (**c**) and donor (**d**). **e**, Dot plots showing expression of marker genes (top), epithelial and stromal genes (middle), and EMT/MET regulators (bottom) in epithelial subpopulations. ‘Proliferative’ epithelial subpopulation refers to cycling cluster highly expressed MKI67 and TOP2A. **f**, Dot plots showing expression of marker genes (top), epithelial and stromal genes (middle), and EMT/MET regulators (bottom) in stromal subpopulations. **g**, Dot plots showing expression of marker genes (top), and NK cell subtype genes (bottom) in NK/T subpopulations. **h**, Dot plots showing expression of marker genes in myeloid subpopulations.

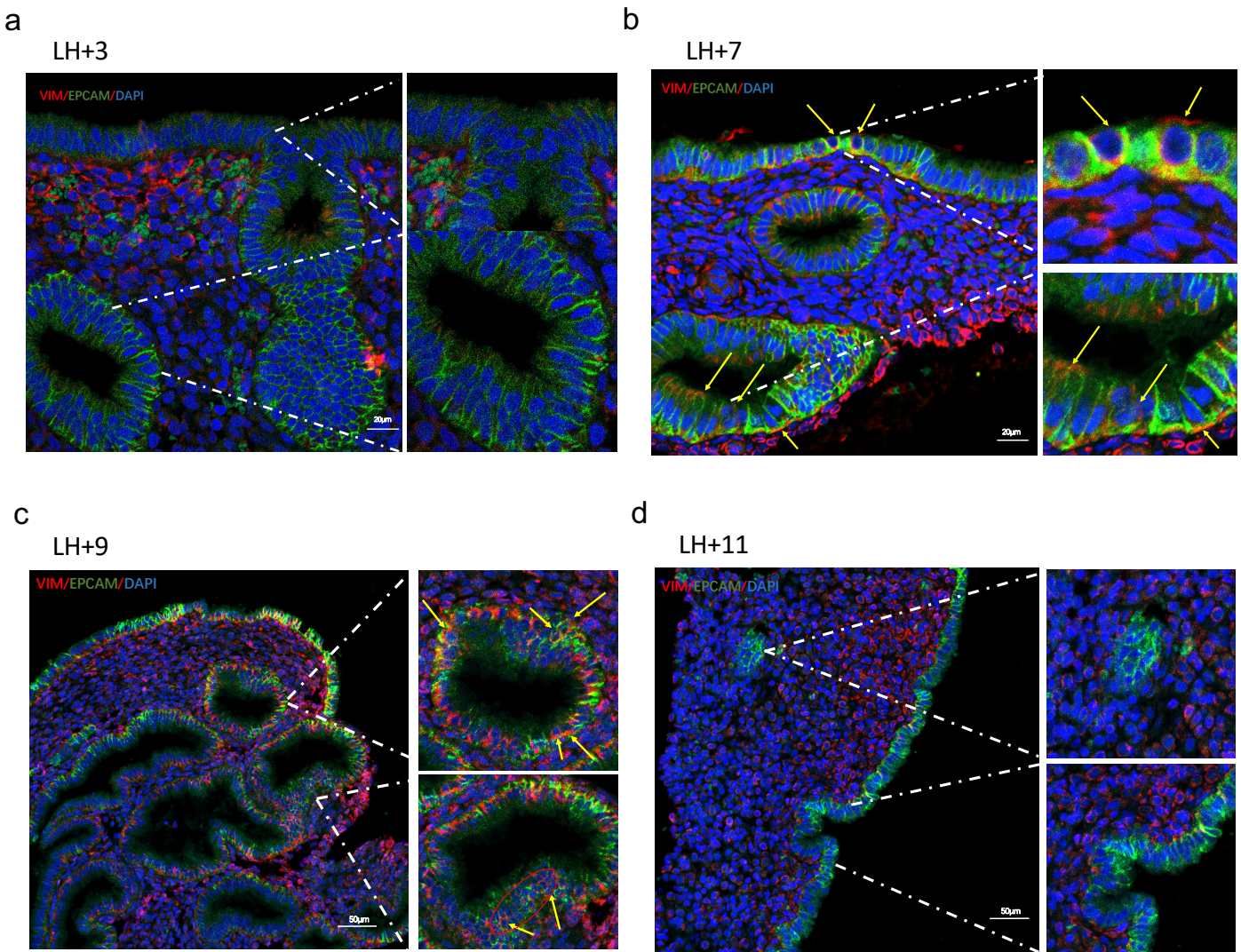


Supplementary Fig. 2| Cell percentage distribution for major and subset cell types. **a**, Cell count and cell percentage distribution for each major cell type. **b**, Individual cellular composition of major cell types for the 18 fertile controls across WOI. **c**, Major cell type percentages in public endometrial bulk samples collected at different time points using deconvolution analysis. **d-g**, Cell composition of epithelial subpopulation (**d**), stromal subpopulation (**e**), NK/T subpopulation (**f**) and myeloid subpopulation (**g**) for each donor.

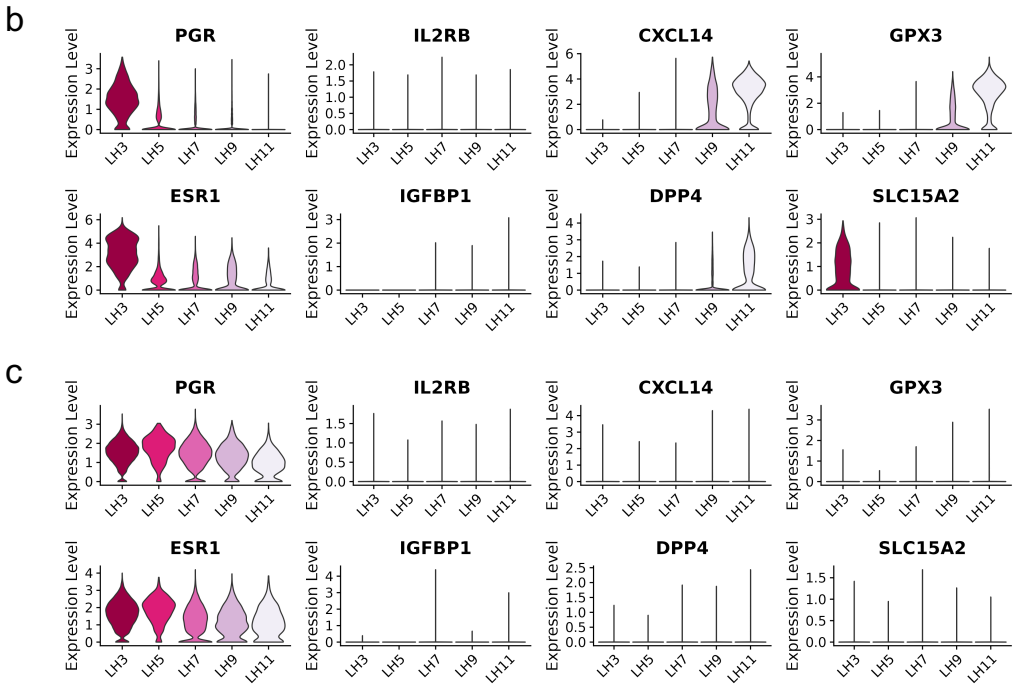
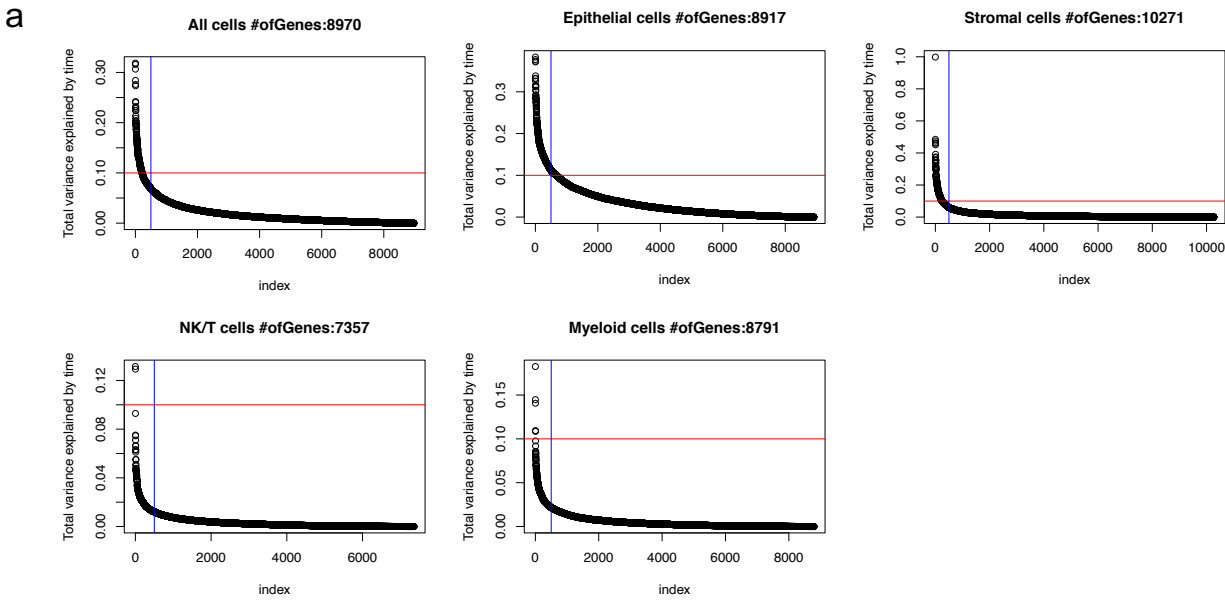




Supplementary Fig. 4| Immunostaining of secretory endometrium. Immunostaining of luminal marker genes (CDH1, LGR5, EDG7) and glandular marker gene (FOXA2) in endometrium of secretory phase (n = 3). G: gland. L: lumen.

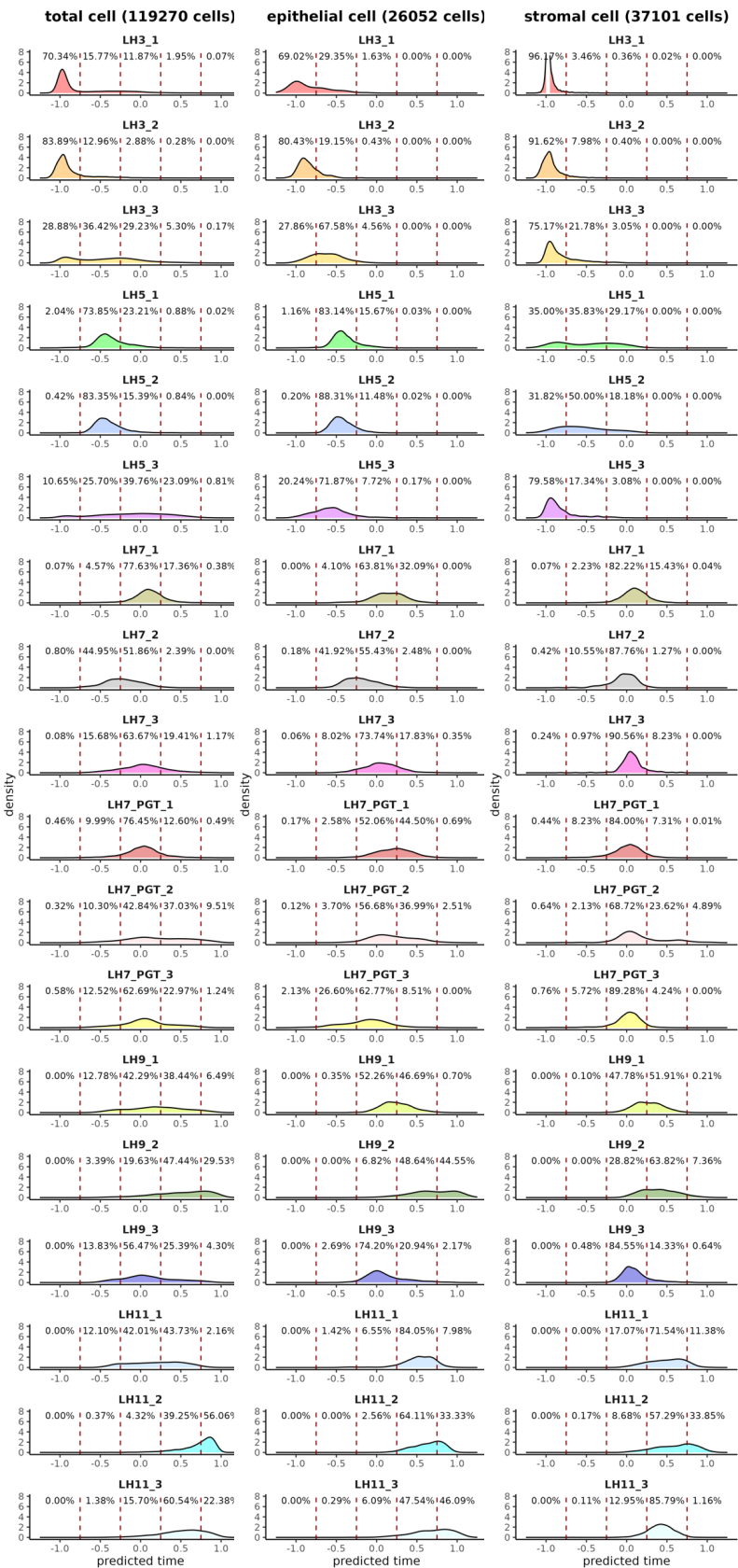


Supplementary Fig. 5 Immunostaining of epithelial cell marker EPCAM and stromal cell marker VIM in **(a)** LH+3 (n = 1), **(b)** LH+7 (n = 1), **(c)** LH+9 (n = 1), **(d)** LH+11 (n = 1) endometrial tissues.

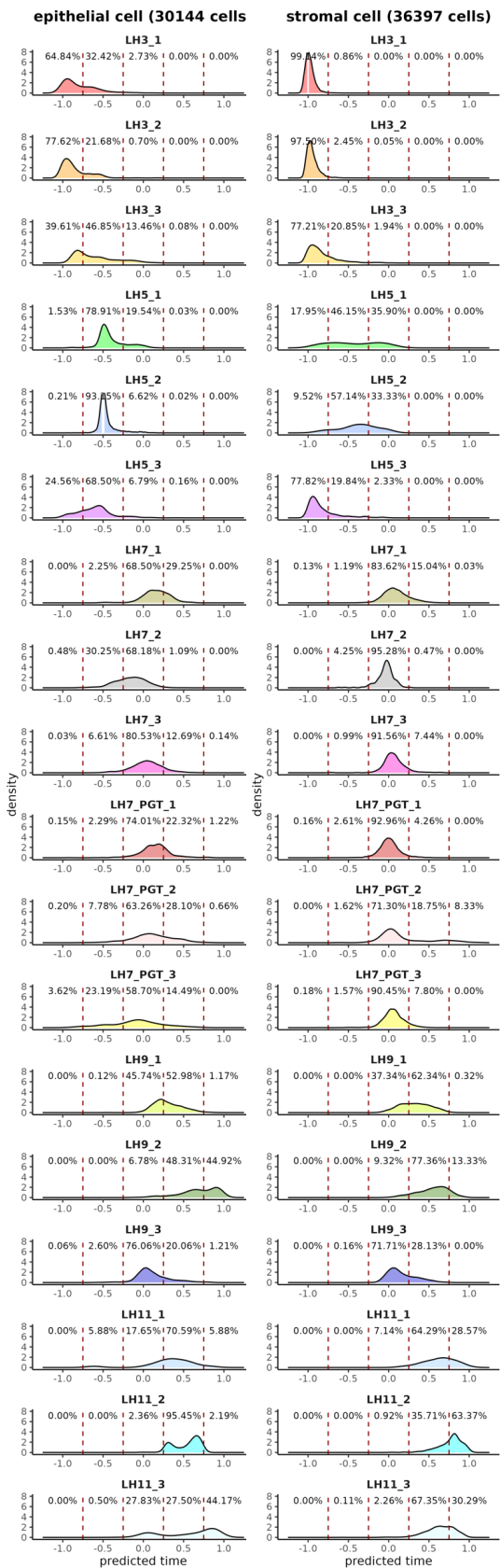


Supplementary Fig. 6| Overall dynamic features of the generated dataset. a, Distribution of variance explained by time for each gene identified using total cells, major stromal, epithelial, NK/T, and myeloid cells. Genes were ordered according to the fraction of the variance. **b,c**, Violin plots showing the expression of reported receptivity regulatory genes in epithelial cells (**b**) and stromal cells (**c**).

a

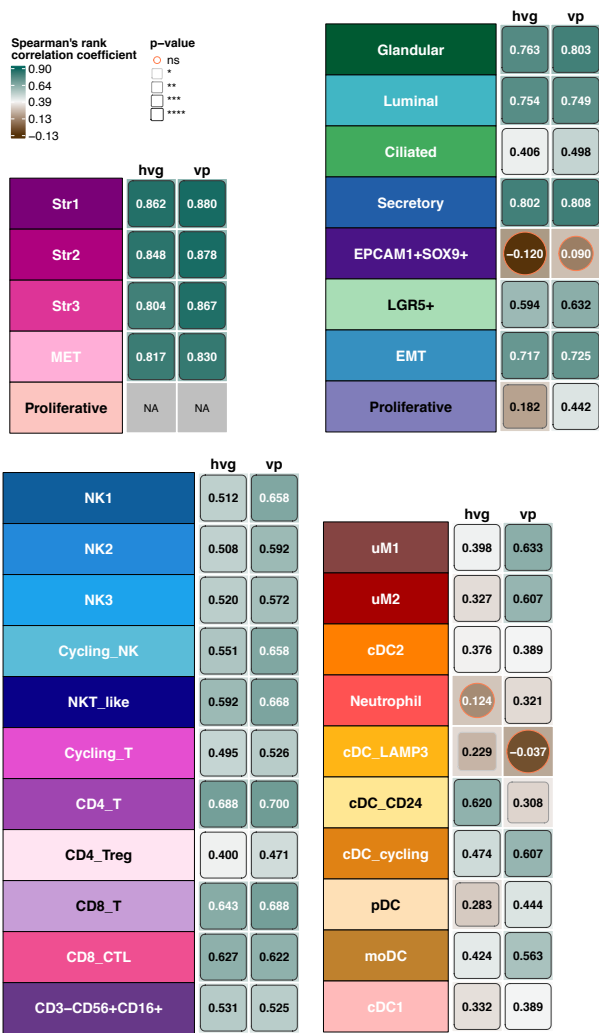


b

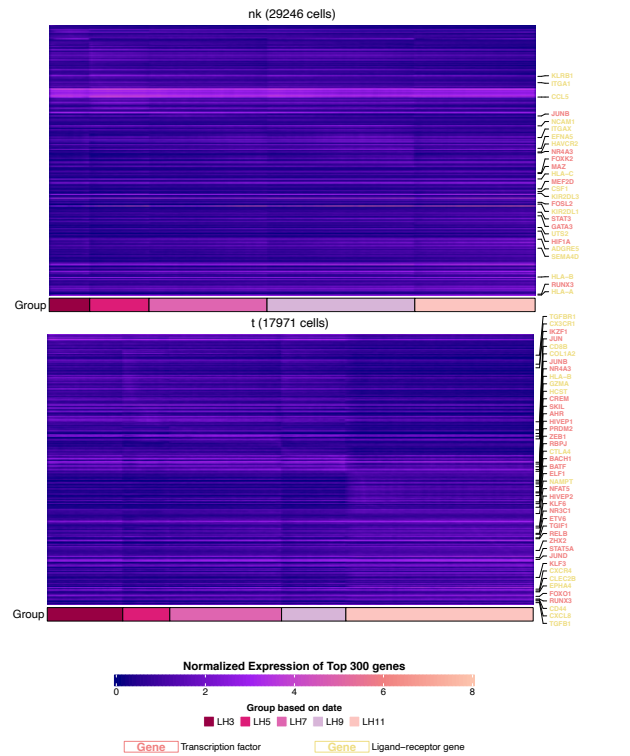


Supplementary Fig. 7| Single-cell predicted time distribution for the fertile controls using leave-one-out mode. **a**, Predicted time distribution of total cells, epithelial cells, and stromal cells when StemVAE was trained regardless of cell types. **b**, Predicted time distribution of epithelial cells and stromal cells when StemVAE was trained separately for epithelial cells or stromal cells.

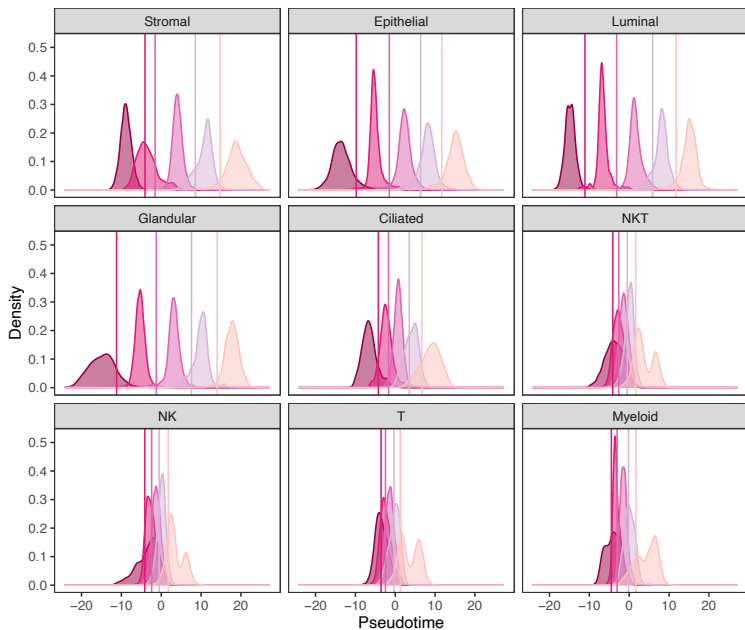
a



b



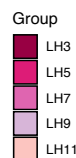
c

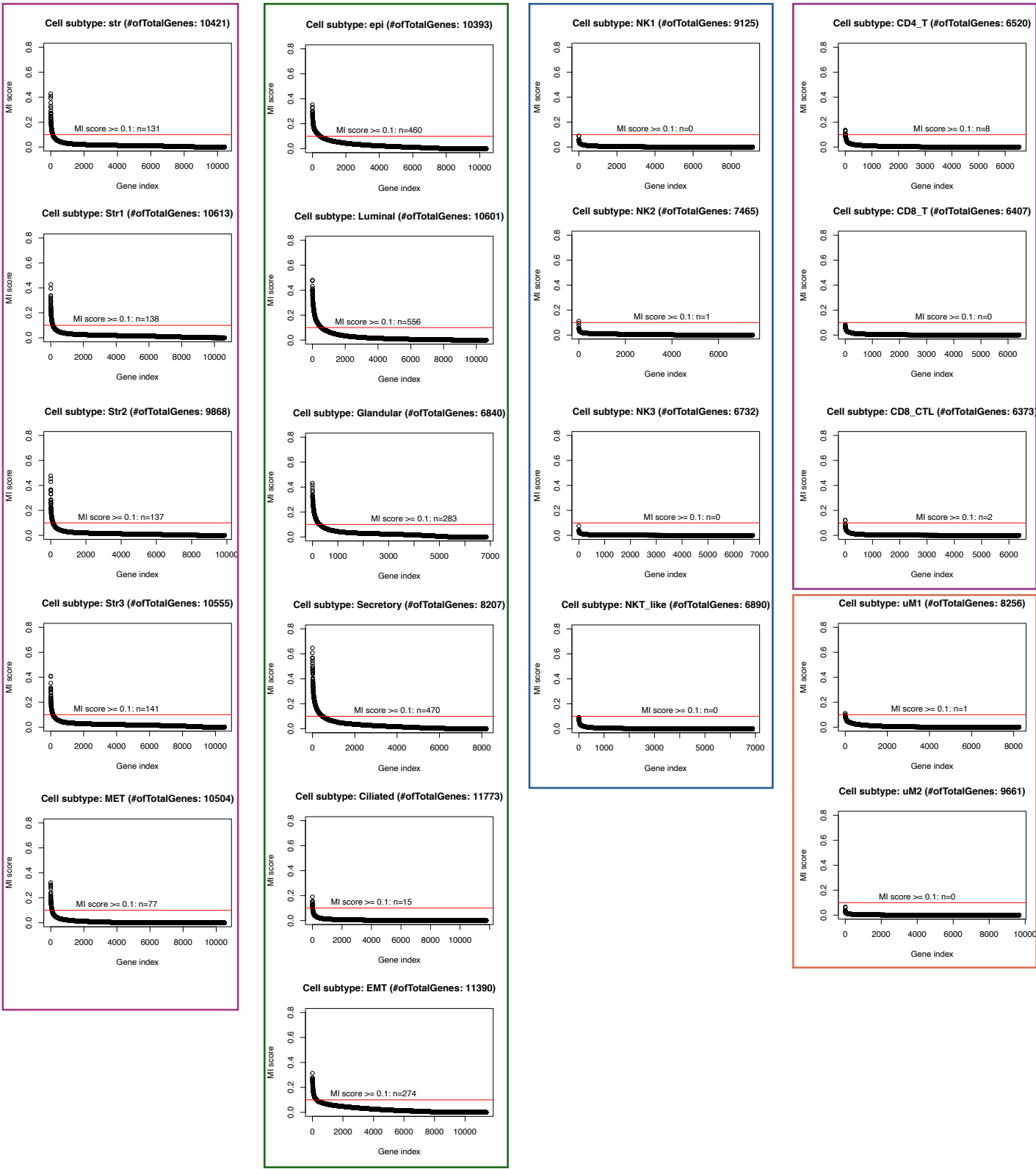


Supplementary Fig. 8|

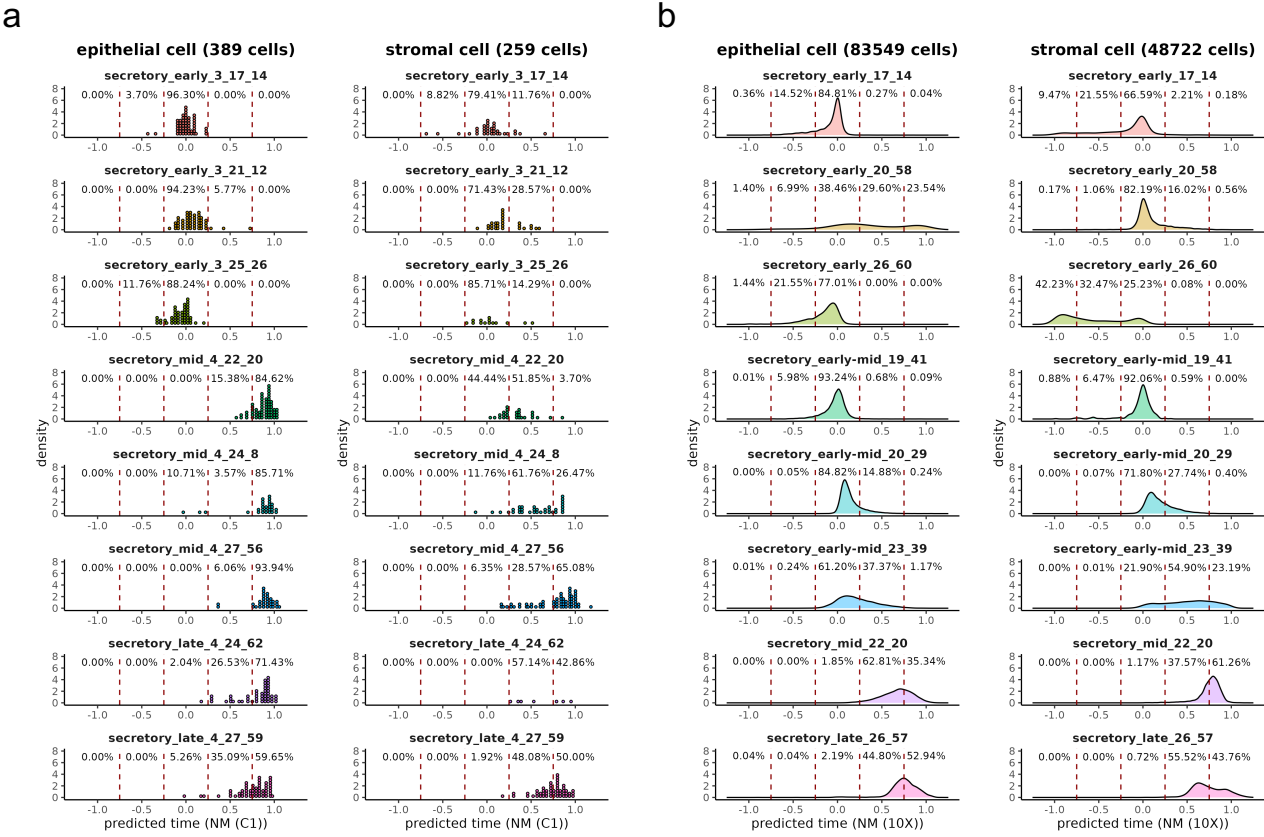
Differential time association for different cell types. a,

Spearman's rank correlation coefficient of predicted time and sampling time for each cell subtype under different training modes of StemVAE. hvg: highly variable genes; vp: time-associated genes from variancePartition. b, Heatmap showing the dynamic time-associated gene expression along the time axis for NK (top) and T cells (bottom). Cells were ordered according to the predicted time within each group. c, Pseudotime distribution for different cell types in different groups across WOI using the method psupertime.

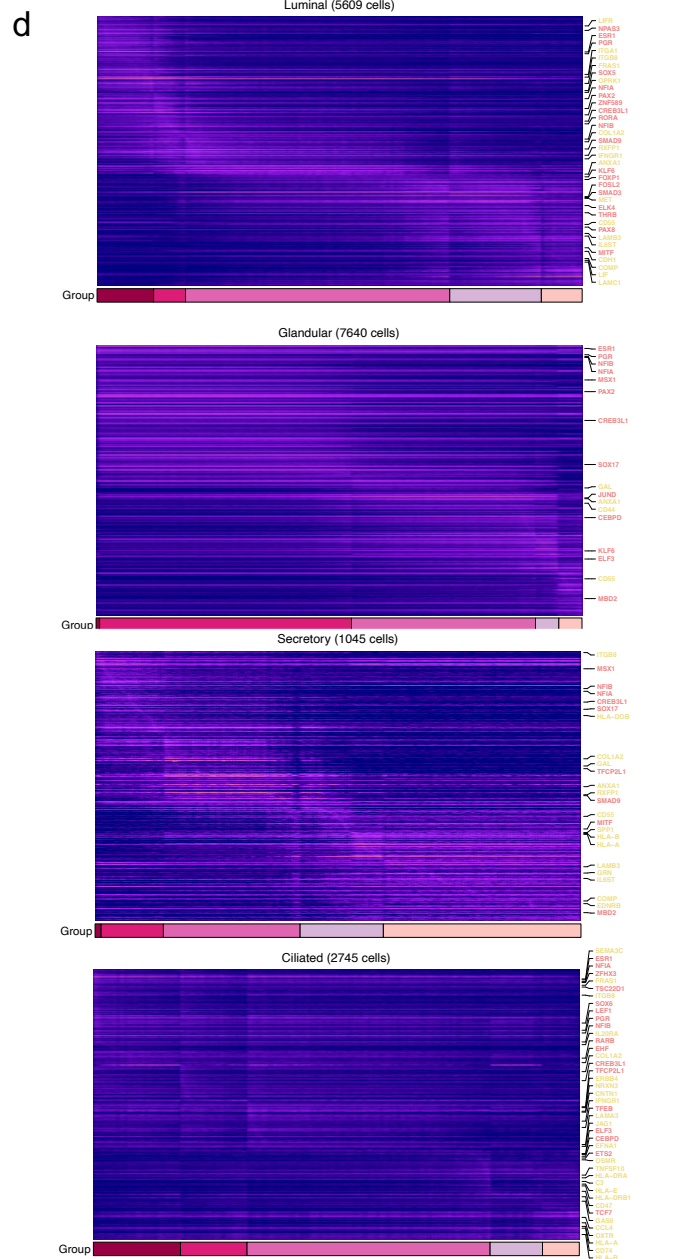
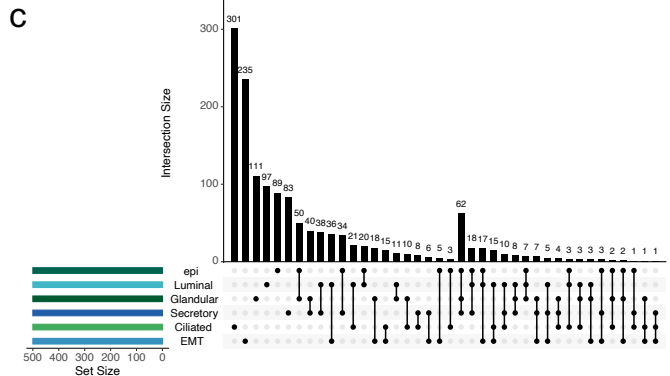
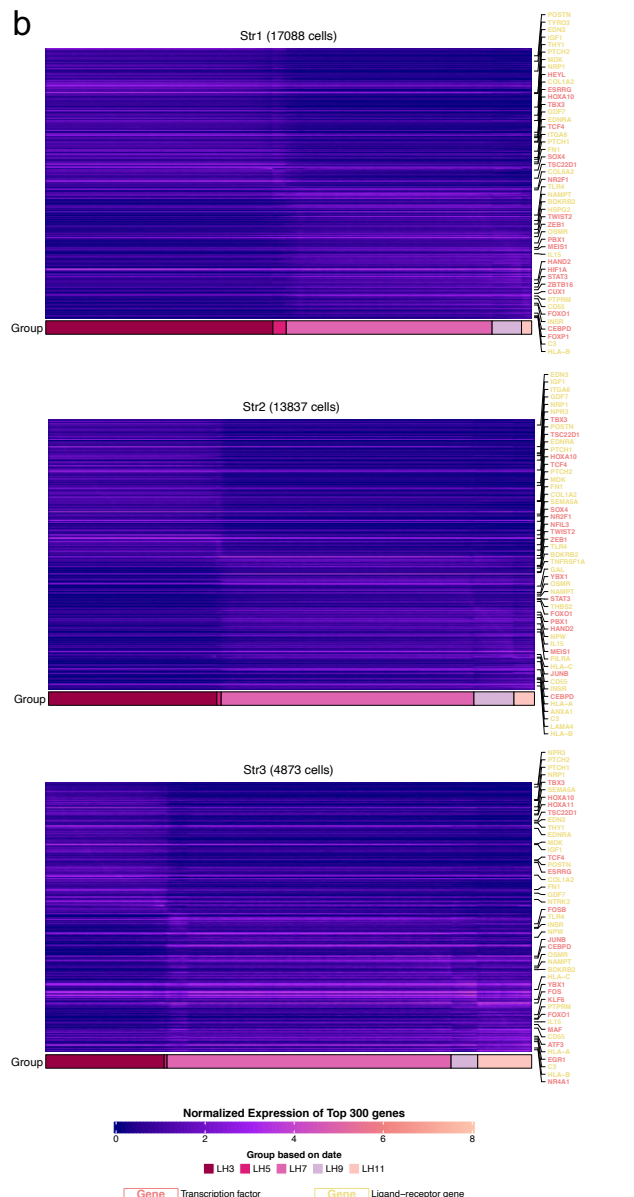
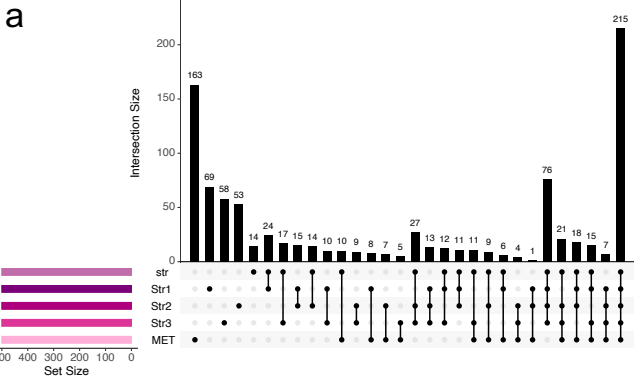




Supplementary Fig. 9| Distribution of MI score of genes in each cell subpopulation. Genes were ordered according to the value of MI score.

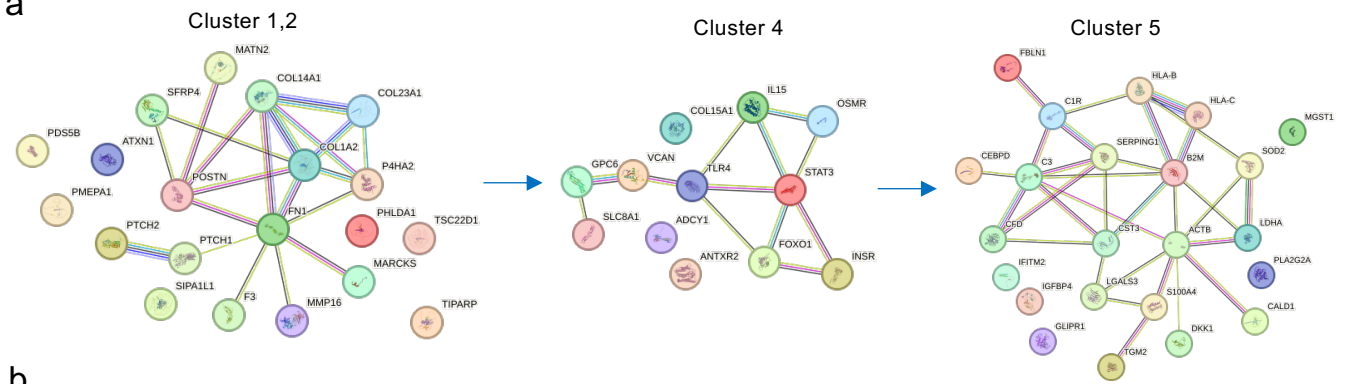


Supplementary Fig. 10| Predicted time distribution of endometrial epithelial and stromal single cells in public database (GSE111976). a, Single-cell dataset generated by C1 platform. b, Single-cell dataset generated by 10X Genomics platform.

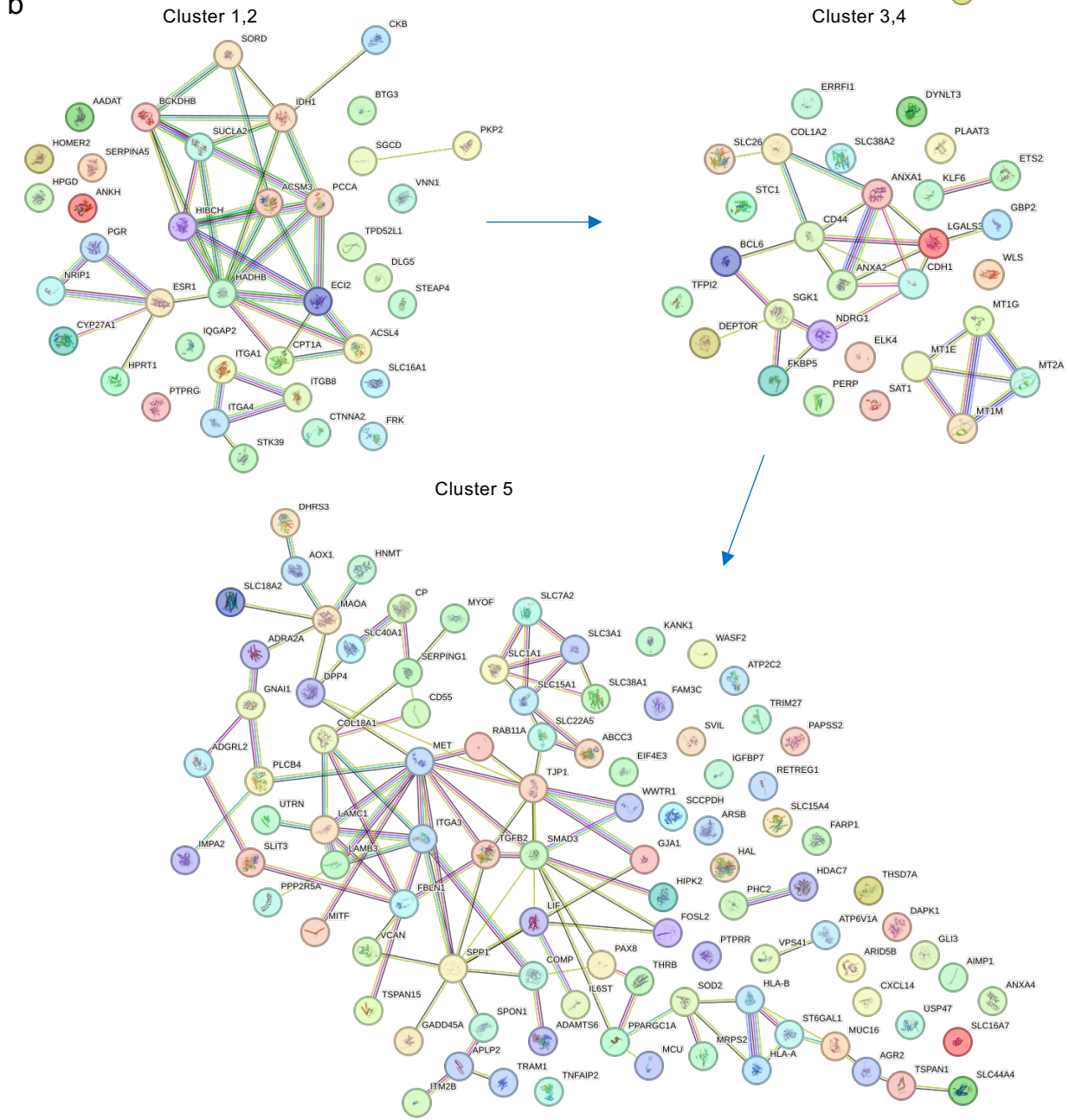


Supplementary Fig. 11| Different dynamics for stromal and epithelial subpopulations. **a**, Upset plot showing the high overlap of time-associated genes identified for each stromal subpopulation. **b**, Heatmap showing the similar dynamic expression pattern of time-associated genes in each stromal subpopulation. **c**, Upset plot showing the minimal overlap of time-associated genes identified for each epithelial subpopulation. **d**, Heatmap showing the differential dynamic expression pattern of time-associated genes in each epithelial subpopulation.

a

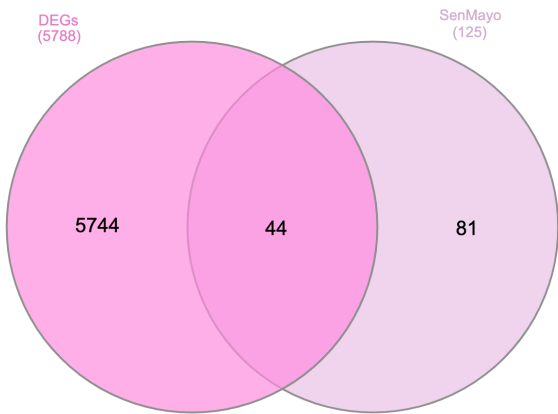


b

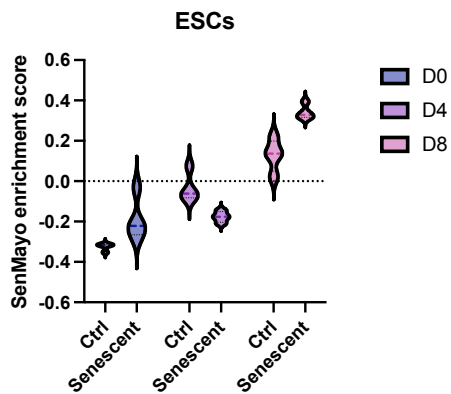


Supplementary Fig. 12| Protein-protein interaction networks for pathway-enriched genes in different stages in stromal cells (a) and in epithelial cells (b) across the WOI.

a



b

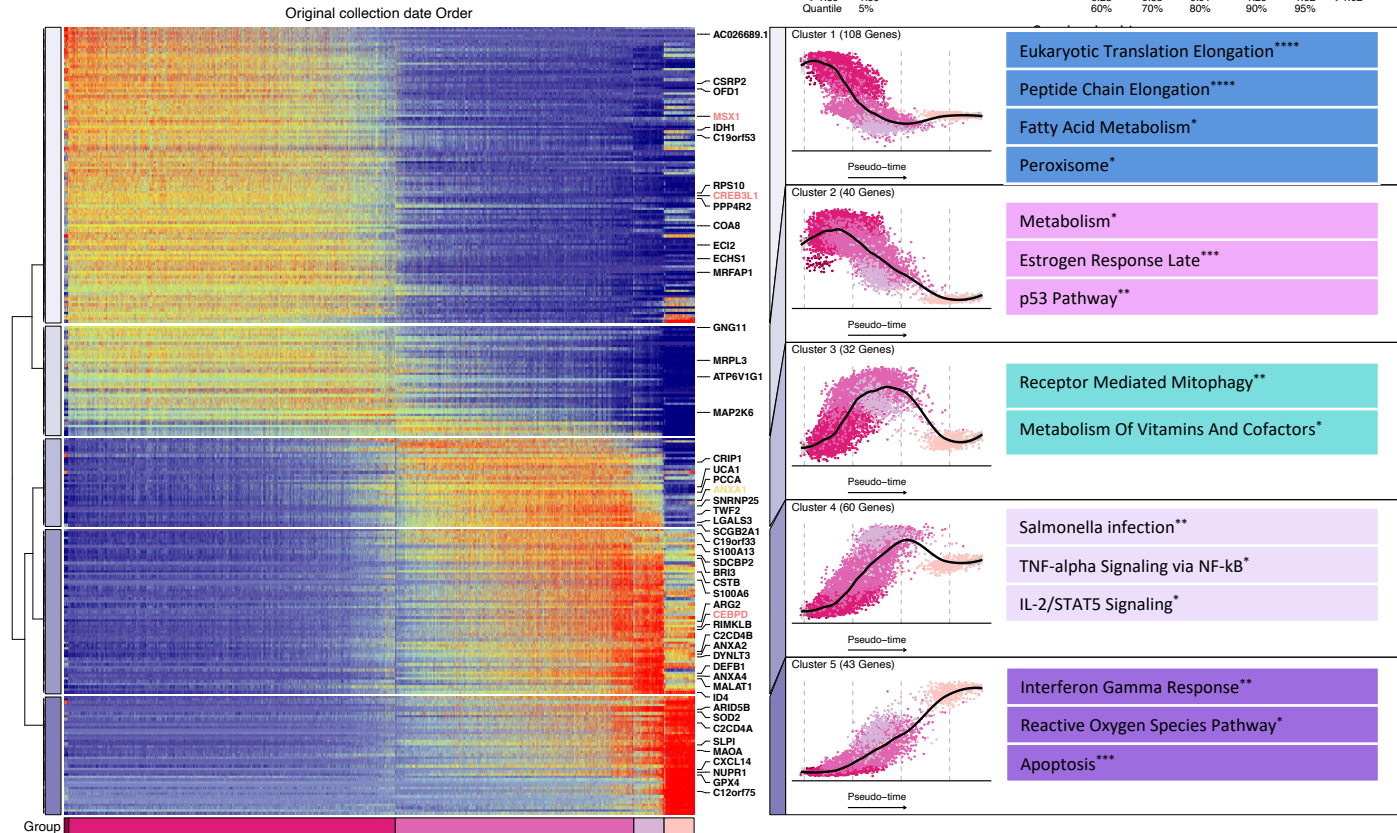


Background gene number: 32463,
target: 5788,
non-target: 32463-5788,
drawing:125,
drawing target:44

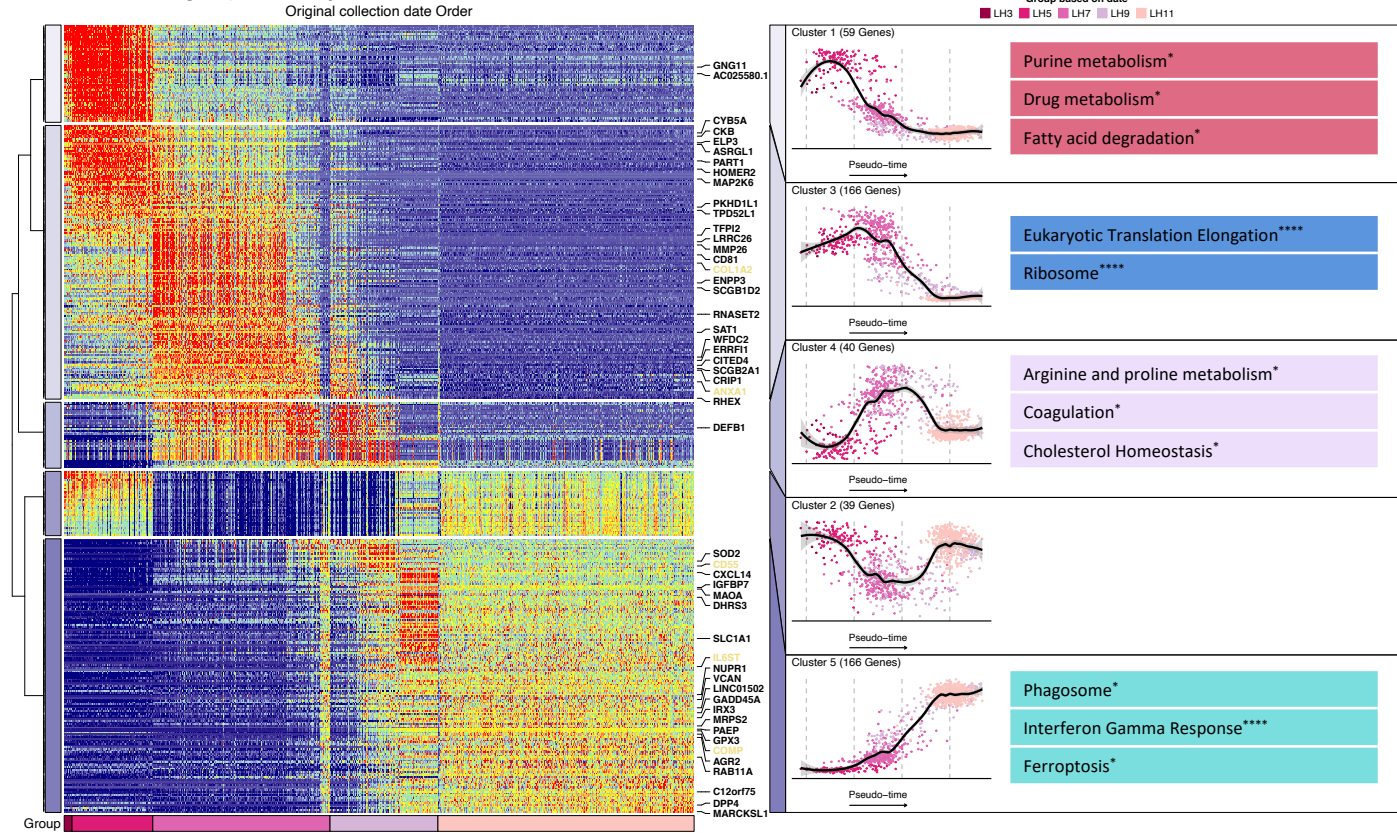
phyper(44-1,5788,32463-5788,125)
p-value = 2.613891e⁻⁰⁶

Supplementary Fig. 13| Validation of SenMayo gene set in human endometrium. a, Intersection between SenMayo gene set and ageing related genes in human endometrium. **b,** GSVA enrichment analysis of human endometrial stromal cells (ESCs) w/o H₂O₂ treatment during the *in vitro* decidualization (D0, D4, D8).

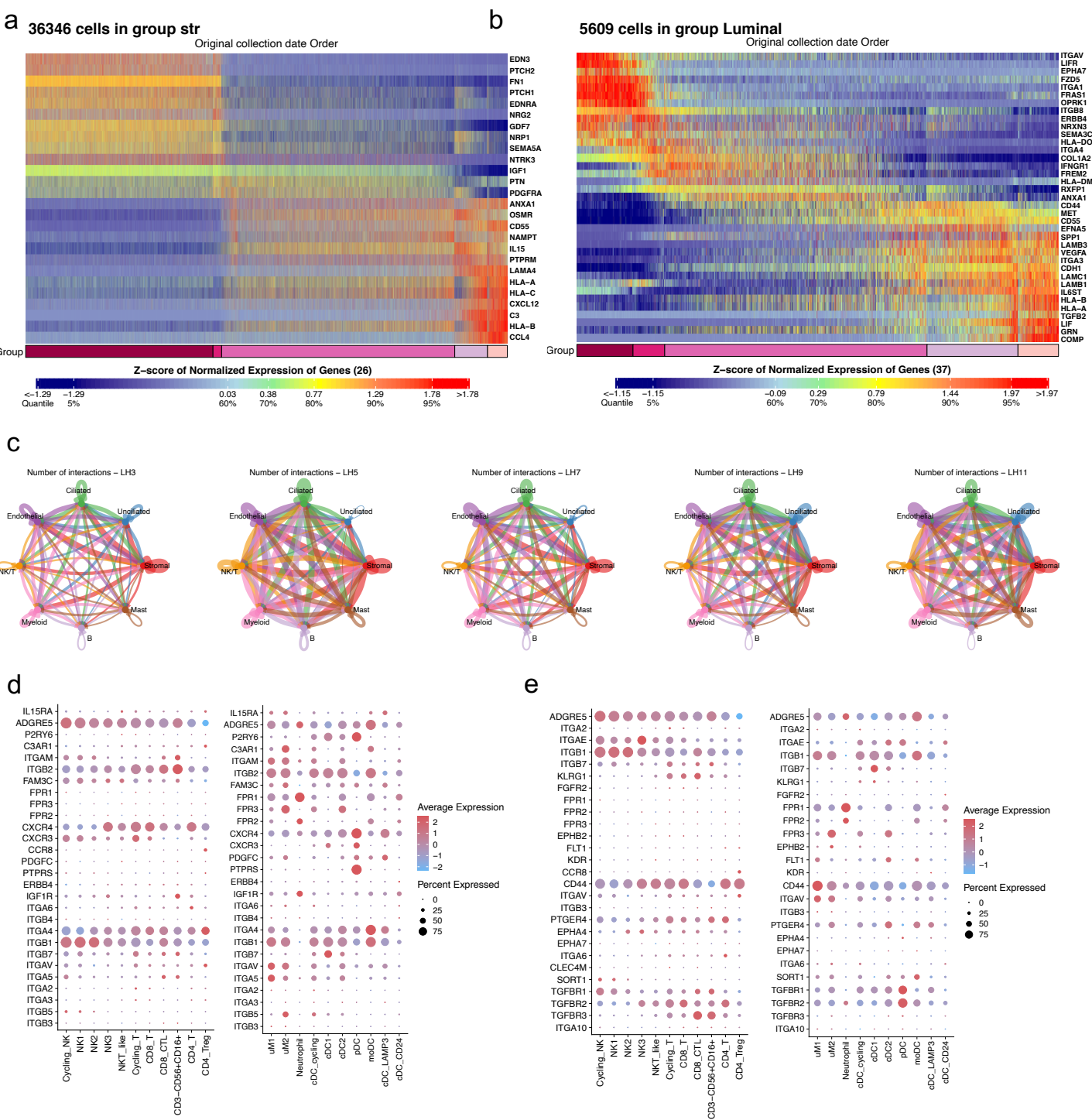
a 7640 cells in group Glandular



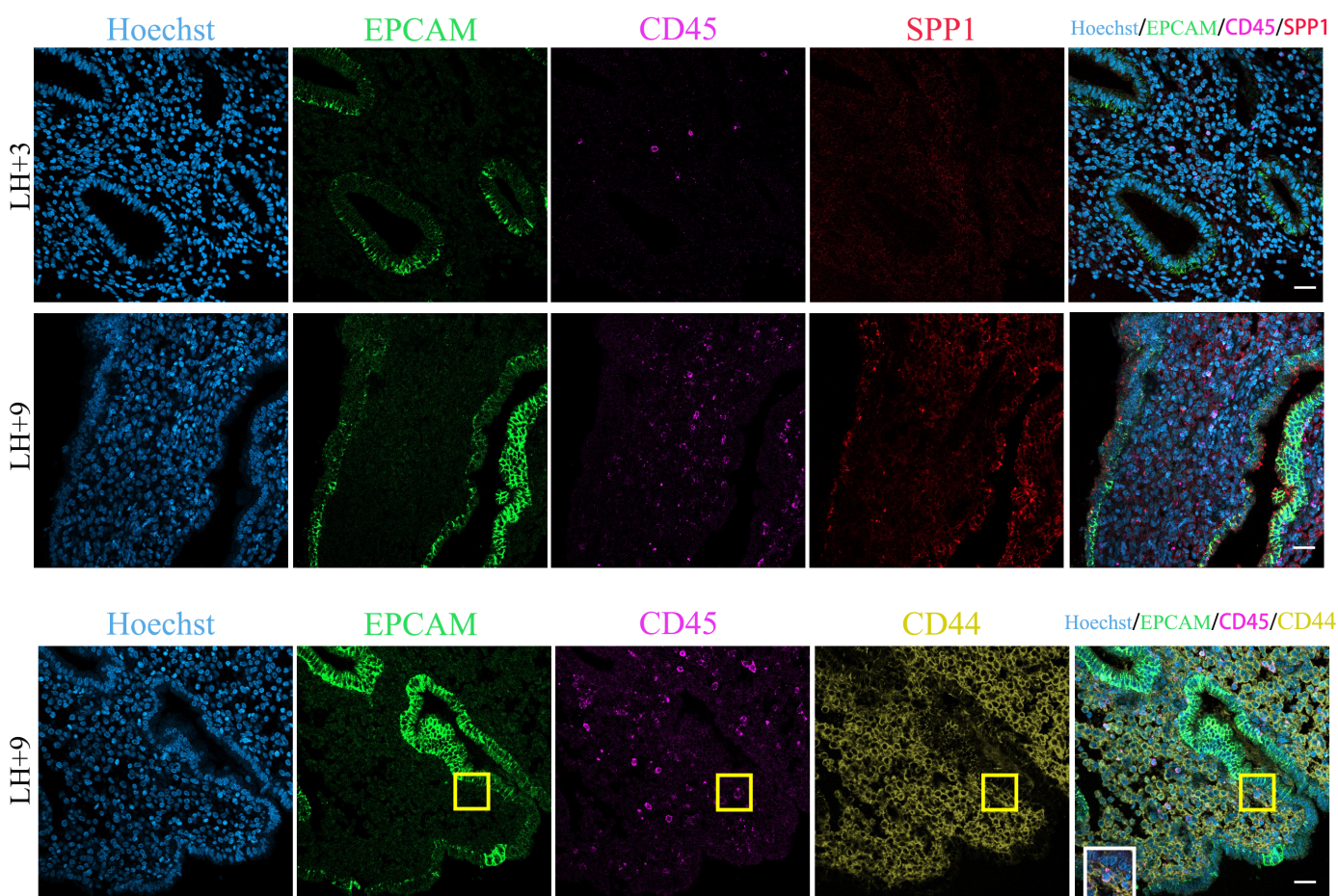
b 1045 cells in group Secretory



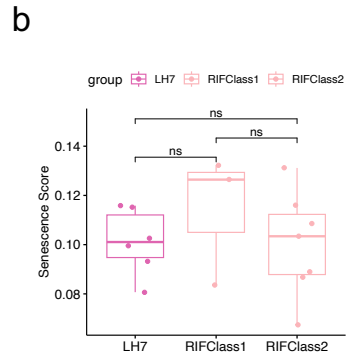
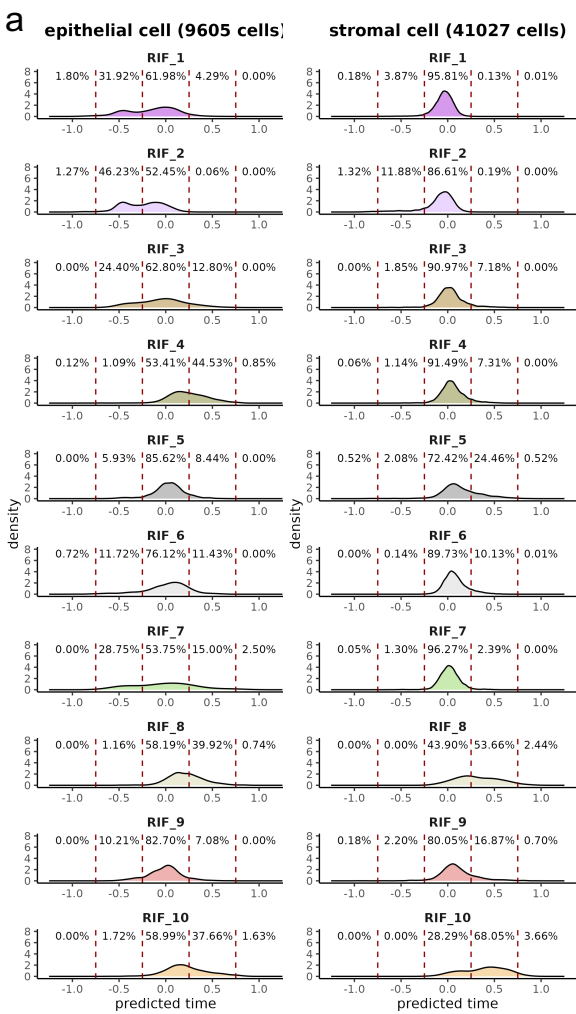
Supplementary Fig. 14| Characteristics of glandular and secretory epithelial dynamics across the window of implantation. a,b An integrated scheme showing the heatmap of denoised and scaled expression of time-associated genes (left), expression trend of the identified gene clusters (middle), and significantly enriched biological terms/pathways from KEGG, Reactome and MSigDB Hallmark (right) for glandular (**a**) and secretory (**b**) epithelial cells. P values ≤ 0.05 were represented as *, ≤ 0.01 as **, ≤ 0.001 as ***, ≤ 0.0001 as ****.



Supplementary Fig. 15| Cell-cell interaction analysis between stromal/epithelial and immune cells across WOI. a,b, Heatmap showing the dynamic expression of potential interacting genes that are time-associated across WOI in stromal (a) and epithelial (b) cells. The ligand/receptor was determined according to the CellChat database. **c,** Potential interactions among the major cell types at each time point. **d,** Dot plot showing the expression of significant stromal interacting genes in NK/T and myeloid subpopulations. **e,** Dot plot showing the expression of significant epithelial interacting genes in NK/T and myeloid subpopulations.

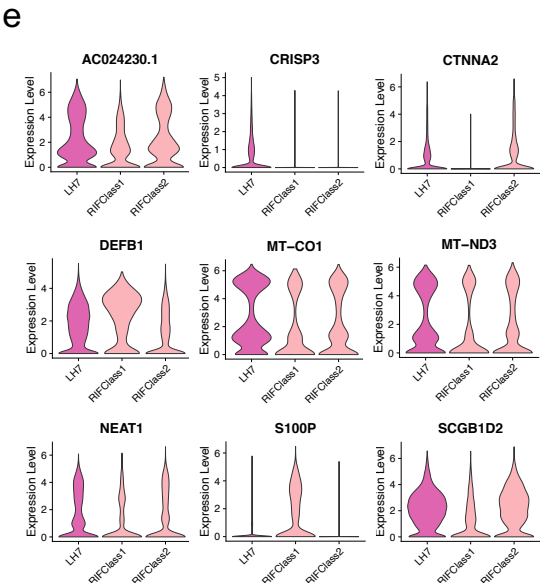
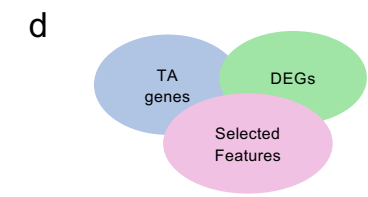


Supplementary Fig. 16| Immunostaining of EPCAM, CD45, SPP1, CD44 in human endometrium. The yellow rectangle regions shows the co-expression of CD44 and CDD45 as well as their proximal location to epithelial cells. The white rectangle region enlarges the corresponding yellow rectangle region. (Scale bar, 20 μ m; n = 2)

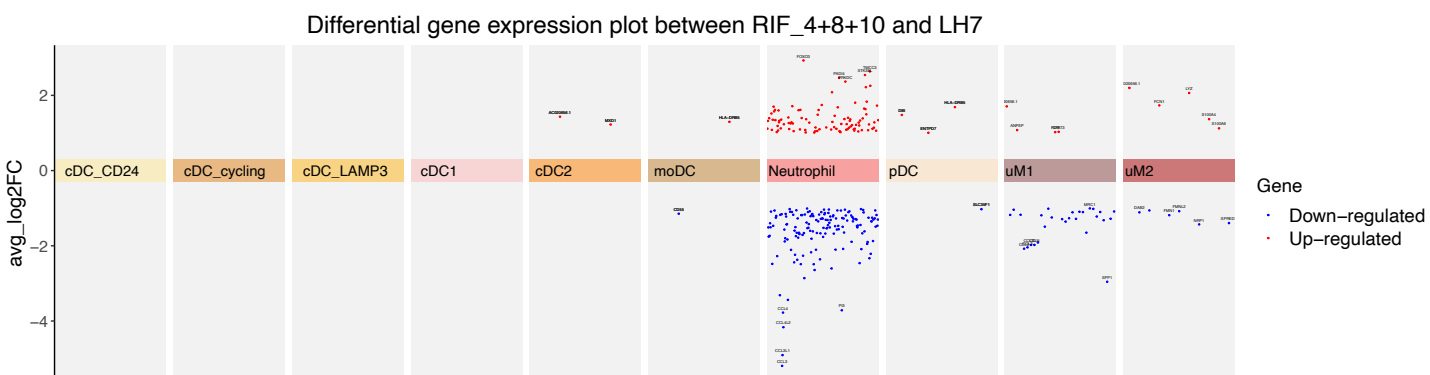
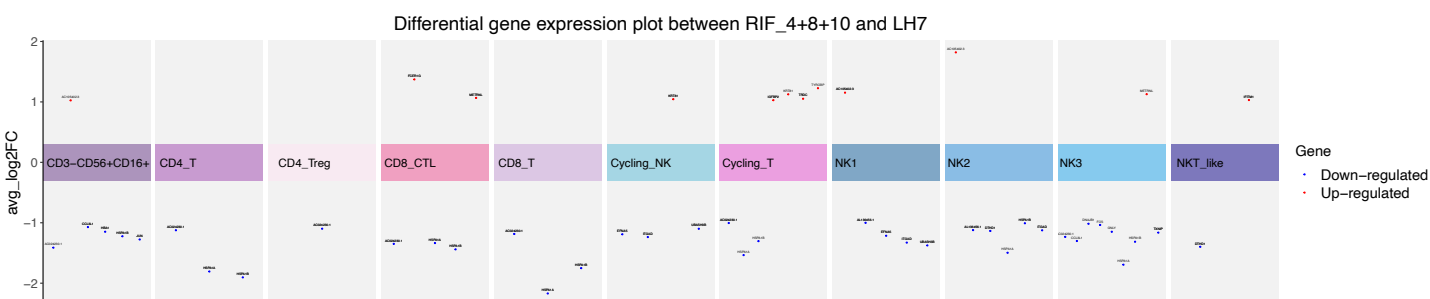
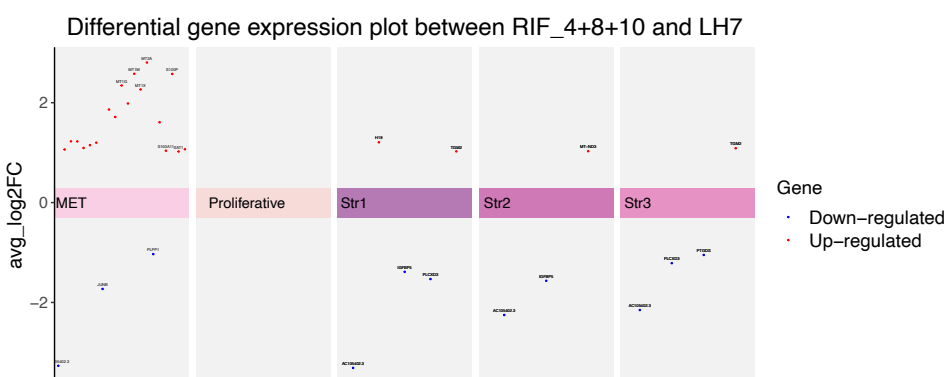
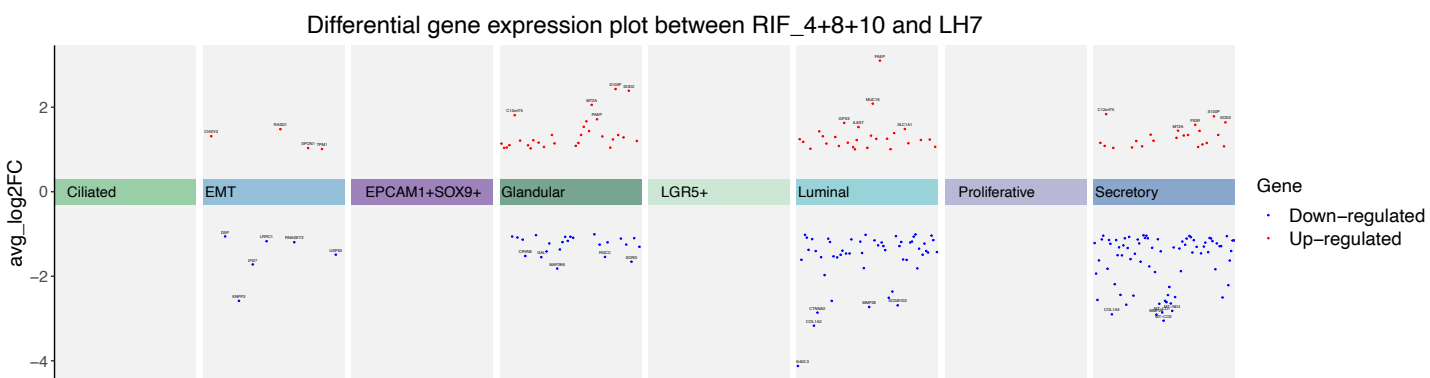


c

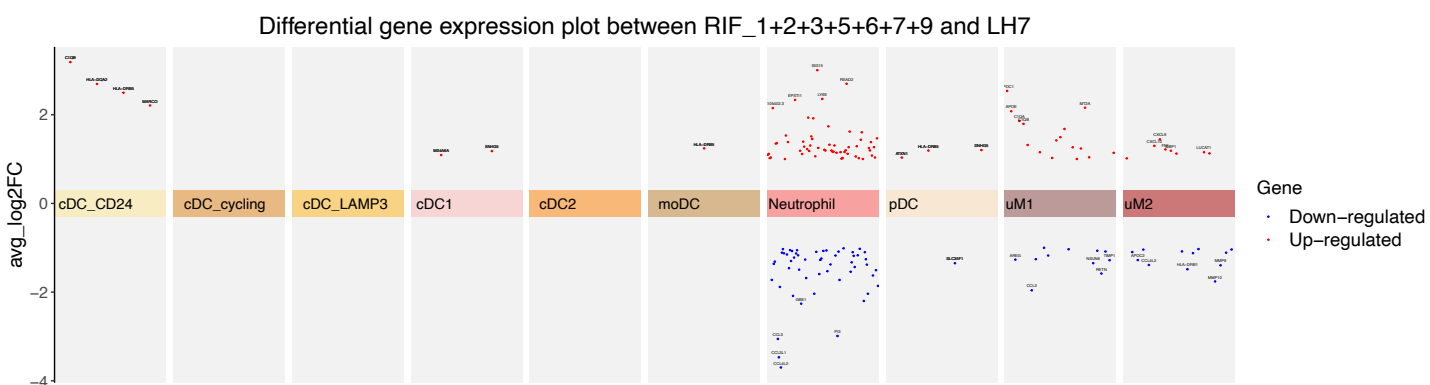
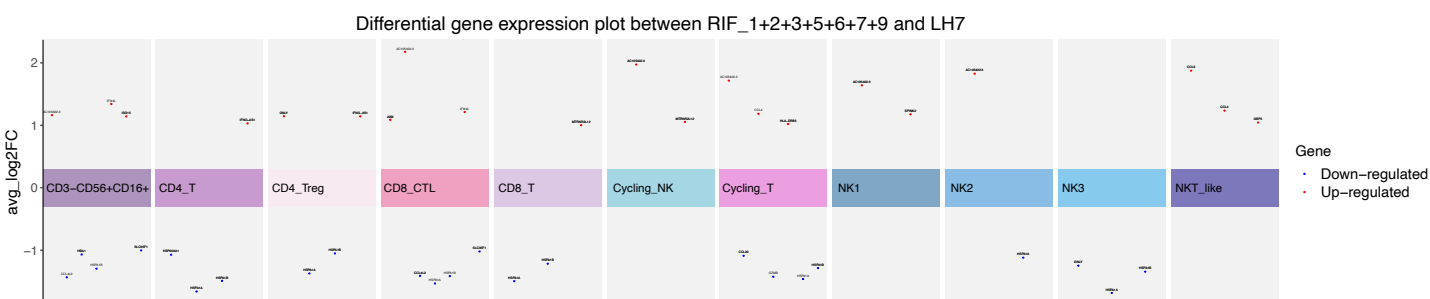
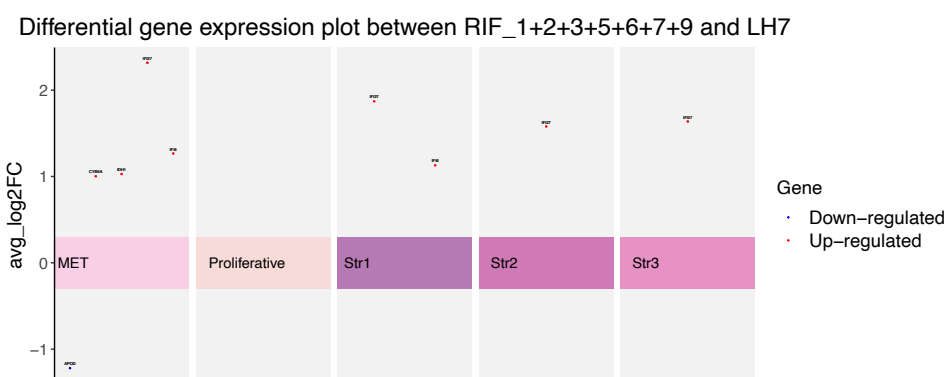
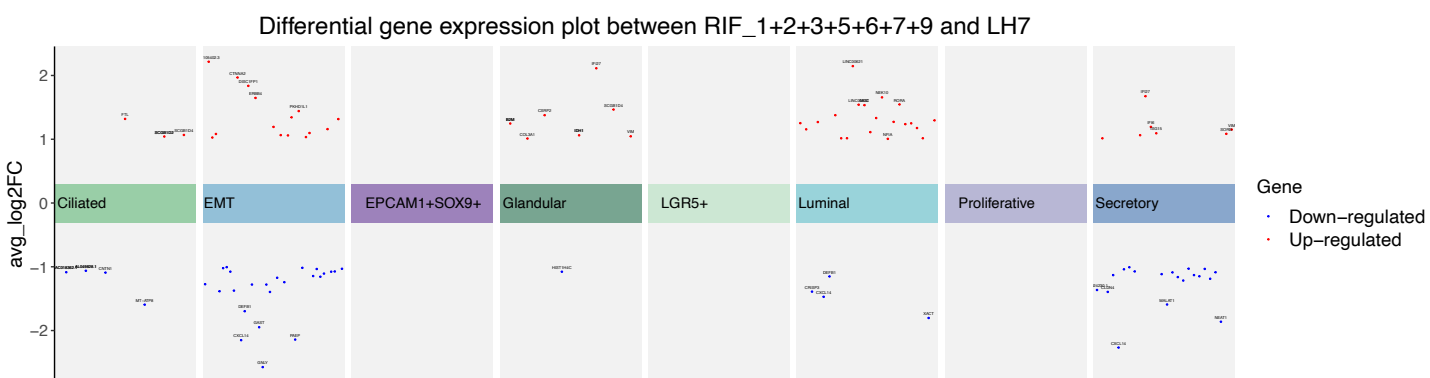
Binary classification	ACC	AUCR OC	AUCP RC	F1 score	GM score	MCC
normalDonor VS. RIFClass1	0.939	0.838	0.590	0.742	0.745	0.710
normalDonor VS. RIFClass2	0.880	0.846	0.718	0.798	0.800	0.716



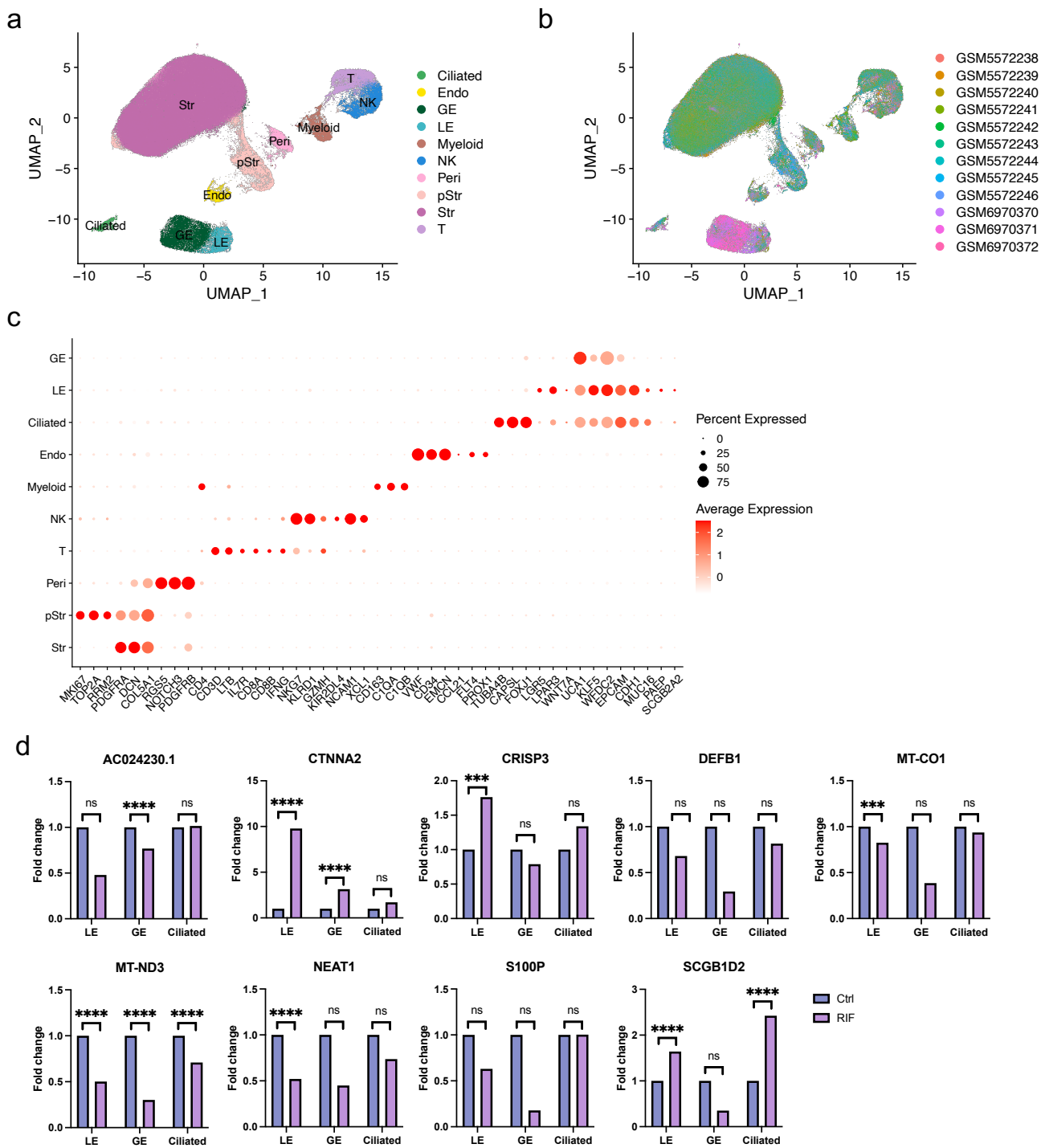
Supplementary Fig. 17 | Characterization of RIF endometrium. **a**, Predicted time distribution of epithelial and stromal single cells by StemVAE built on 18 fertile donors. **b**, Senescence score calculated based on the SenMayo gene set among different groups. **c**, Performance of classifiers built for different scenarios. **d**, Identification of dysregulated genes in RIF endometrium considering different aspects. TA genes: time-associated genes; DEGs: differentially expressed genes; Selected Features: genes identified using random forest selection. **e**, Violin plots showing the identified top-regulated gene in epithelial cells among different groups.



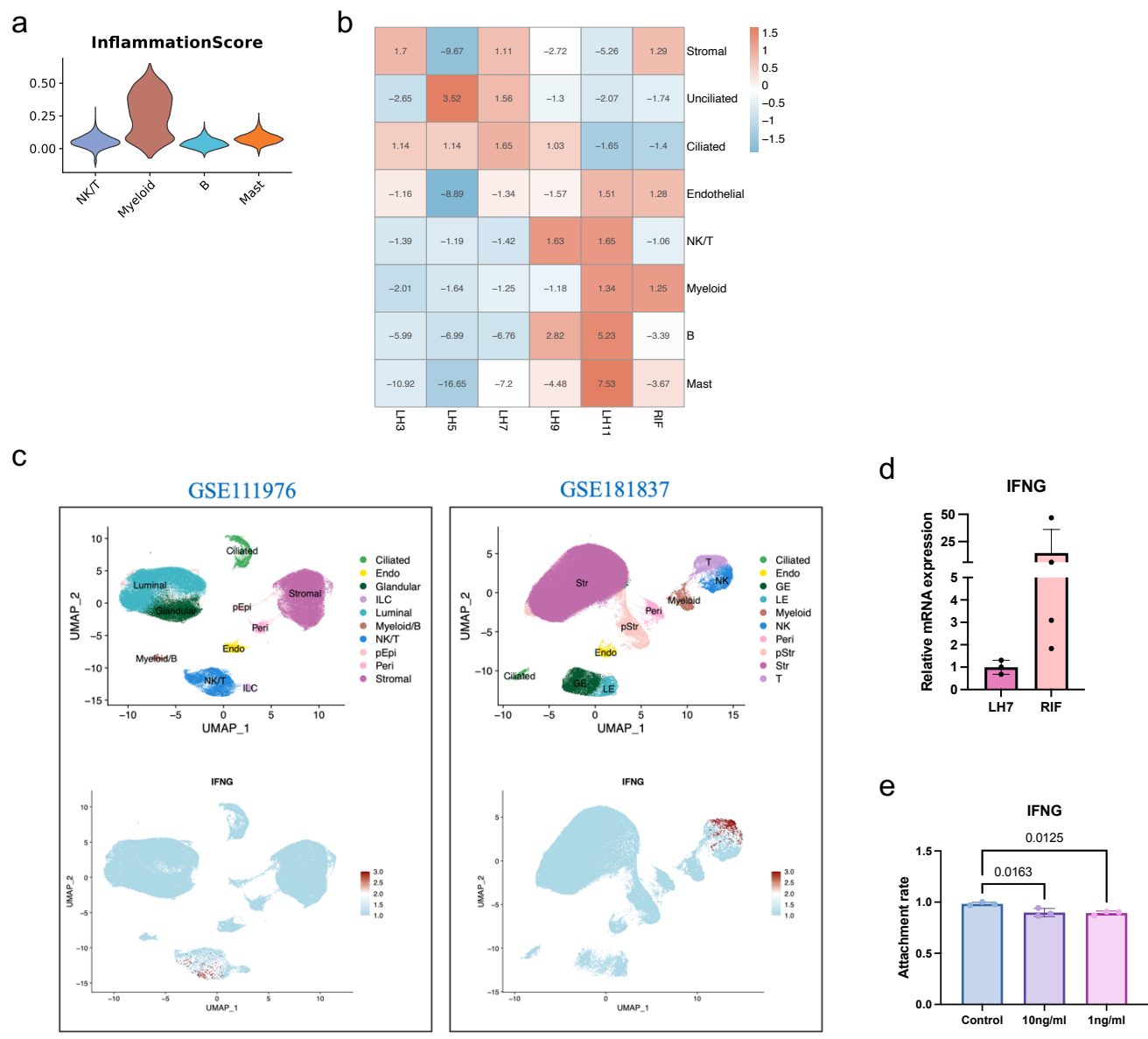
Supplementary Fig. 18| Beeswarm plot showing the distribution of differentially expressed genes in different cell subpopulations detected in displaced RIF samples (RIFClass1).



Supplementary Fig. 19| Beeswarm plot showing the ddistribution of differentially expressed genes in different cell subpopulations detected in in-phase RIF samples (RIFClass2).

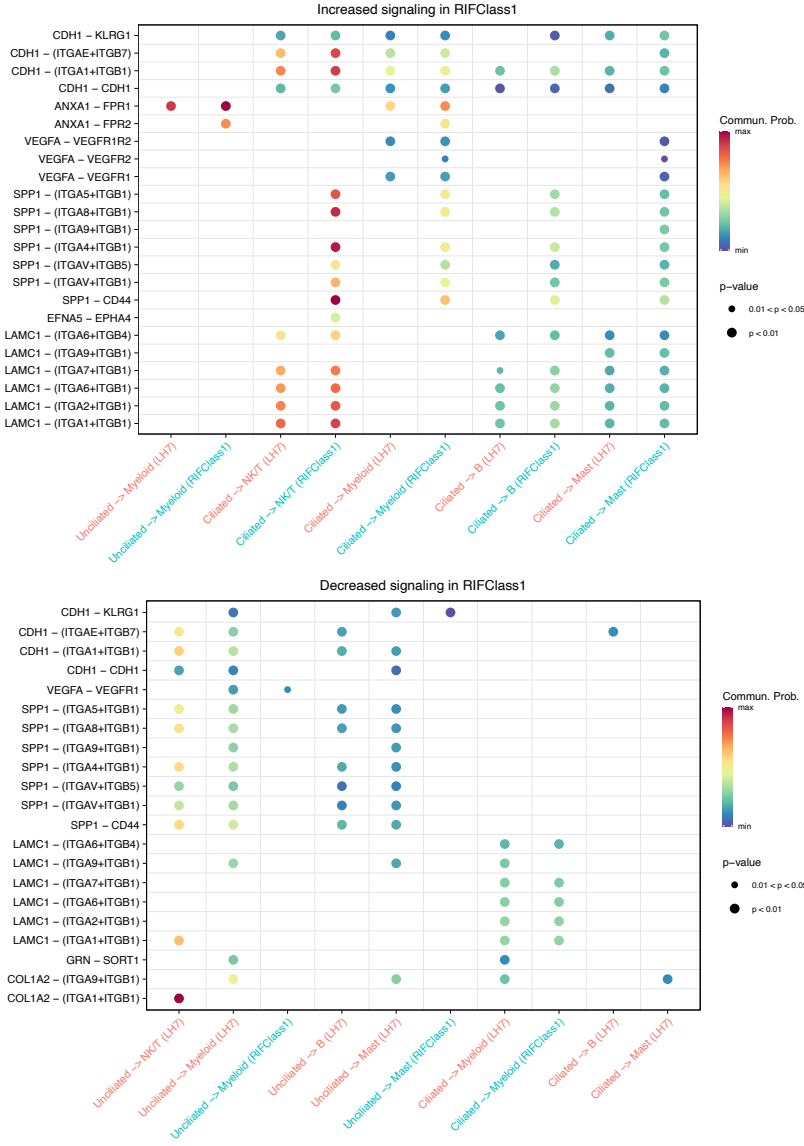


Supplementary Fig. 20| Orthogonal validation of common RIF DEGs. **a**, UMAP plot of integrated cells from single-cell transcriptomic datasets of RIF endometrium (GSE183837, GSE223672) colored by annotated cell types. **b**, UMAP plot of cells colored by donor. **c**, UMAP plot of cell identity marker expression. **d**, Bar plot of average fold change and significance for the 9 common DEGs in the referred datasets. Wilcoxon Rank Sum test was used with bonferroni correction. P values ≤ 0.05 were represented as *, ≤ 0.01 as **, ≤ 0.001 as ***, ≤ 0.0001 as ****.

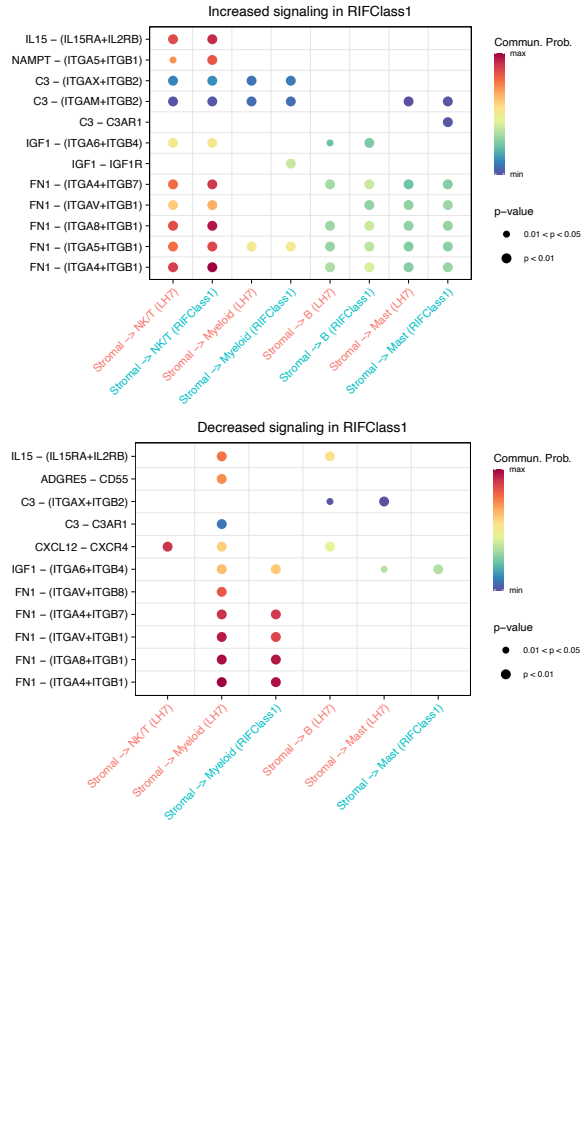


Supplementary Fig. 21| Hyperinflammation characteristics of RIF endometrium. a, Violin plot showing inflammation score distribution in major immune cell types. **b,** Fold enrichment of major cell types in different groups. **c,** IFNG expression in external endometrial single-cell datasets. **d,** qPCR results of IFNG in fertile (n = 3) and RIF (n = 4) endometrial cell mixtures. **e,** IFNG treatment of ISHAKAWA cells decreased the trophoblast spheroid attachment (n = 3 for each group). Data are presented as mean values +/- SD. For statistical analysis, one-way ANOVA was used.

a



b



Supplementary Fig. 22| Differential dynamic changing cell interactions between RIF and fertile endometrium. a. Increased (upper) and decreased (lower) epithelial ligand-receptor signaling in in-phase RIF (RIFClass1) endometrium. **b.** Increased (upper) and decreased (lower) stromal ligand-receptor signaling in in-phase RIF endometrium.

Supplementary Note for “Time-series single-cell transcriptomic profiling of luteal-phase endometrium uncovers dynamic characteristics and its dysregulation in recurrent implantation failures”

1 StemVAE: single-cell temporal variational auto-encoder

1.1 Data Pre-processing workflow

As we set multiple experiments to ensure the validity of our model and the unbiased findings, a detailed pre-processing workflow is provided as shown in Alg. 1.1. Note that the input of the pre-processing workflow should be the raw cell-by-gene expression count matrix. If the user has a custom gene set or donor, the workflow will filter the input matrix accordingly. To mitigate the potential impact of sequencing depth on experimental outcomes, we provide three different downsampling methods in Line 4 for result validation. This step can also add the data augmentation function according to the user’s requests. Next, we filter the cells and genes according to thresholds, followed by total-count normalization of 1,000,000 reads per cell and logarithmize the data with plus one to transform the data. Finally, the normalized log-count values for each gene are standardised to Gaussian distribution $\mathcal{N}(0,1)$. This step is important to prevent the bias caused by genes with higher expression or larger variance, which is also adopted in the preprocessing steps by another method [MGC22].

Algorithm 1 Preprocess single-cell RNA-seq data (SC data)

- 1: Concat SC count data and external SC count data (if exist)
 - 2: If request a constraint of gene set, filter SC data.
 - 3: If request a constraint of dropout donor set, filter SC data.
 - 4: If request downsampling on test donors or cells, downsampling 10 times using a multinomial distribution on element, row, or column
 - 5: **repeat**
 - 6: Dropout cell with less than 100 gene expression
 - 7: Dropout gene with expression in less than 50 cells
 - 8: **until** Each cell and gene satisfy the expression threshold
 - 9: If request different cell subtypes have the same cell number, random select cells from each subtype. While the cell number of a certain cell type is below the minimum requirement, keep all. Update SC data.
 - 10: Normalize counts per cell to 1e6 and update SC data.
 - 11: Logarithmize the data matrix,
 - 12: Re-scale each gene expression as Gaussian-dist (0, 1)
- Output:** Preprocessed SC data and corresponding cell info
-

1.2 The StemVAE model

We developed a probabilistic latent space model named StemVAE (**S**ingle-cell **t**emporal **V**AE) 1 to infer the pseudo-time of cells. The input of StemVAE consists of an mRNA expression UMI count matrix and real-time labels of cells, and the output is the reconstructed expression matrix and estimated pseudo-time. StemVAE, based on a canonical variational auto-encoder (VAE) [KW13] tailored for single-cell temporal analysis, includes an encoder, a cell-decoder, and a time-decoder. The Encoder

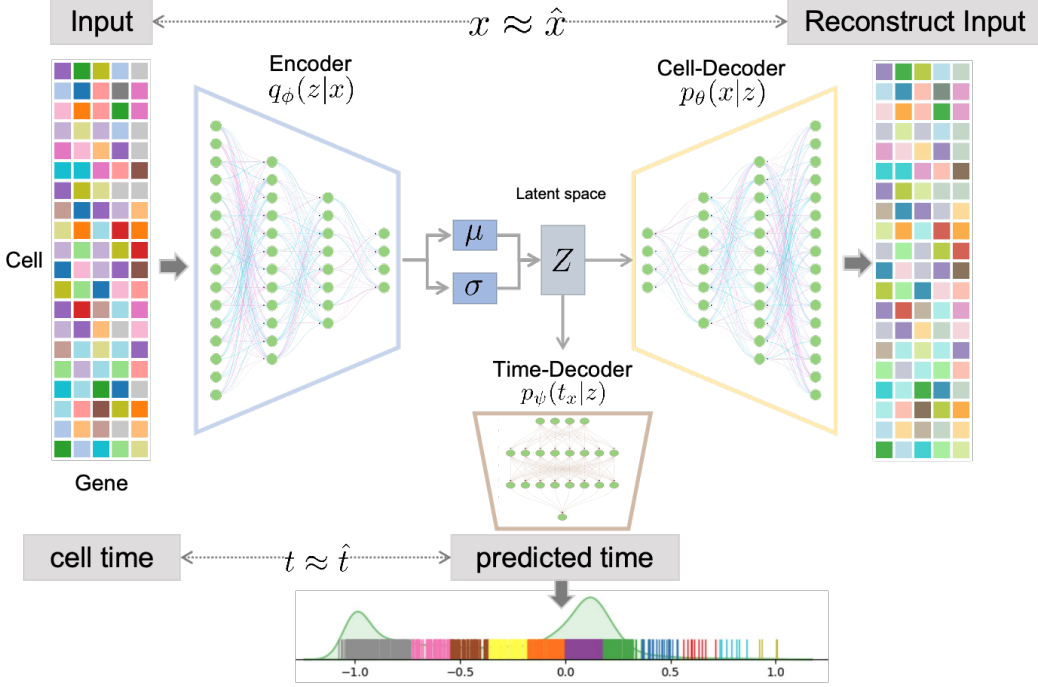


Figure 1: Workflow of **StemVAE**. The input is mapped into a low-dimensional distribution $p_\theta(z)$ by **Encoder** and the latent encoding vector z is used as the input of **Cell-Decoder** and **Time-Decoder**. The **Cell-Decoder** is a generative model to reconstruct the input matrix. The **Time-Decoder** is a supervised regression model to predict the pseudo-time of cells.

parameterized by ϕ (a set of parameters in a feedforward network) encodes the mRNA expression matrix to a low-dimensional distribution. Specifically, the input, whose dimension is the number of genes, passes through three hidden modules with 256, 128, and 64-dimensional linear layers, respectively. Each hidden module includes a Batch-Normalization layer and uses LeakyReLU as the activation function after the linear layer. The final output layer of the encoder is 50 dimension distribution $\mathcal{N}(\mu, \sigma)$. The Cell-Decoder, parameterized by θ , reconstructs input data from latent variable space with the input z which is sampled from a prior distribution $p_\theta(z)$. The detailed neural network structure of the Cell-Decoder mirrors that of the Encoder in a reverse order. The Time-Decoder, parameterized by ψ , extracts the temporal information from latent space with the latent encoding vector z and time labels as input. It includes 2 hidden neural layers with 50 and 10 dimensions. The last layer is only one dimension predicting the biological time of single cells. Here, we treated the biology time via a continuous variable instead of a discrete variable, considering the former can better capture the temporal continuity (both interpolation and extrapolation) of each cell.

Similar to the standard VAE model [KW13], the loss function of StemVAE can be extended by adding a loss term of the time decoder, as follows:

$$\begin{aligned}
 \mathcal{L}_{\text{VAE}}(\theta, \phi, \psi) &= -\log p_\theta(\mathbf{x}) + D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})) - \log p_\psi(\mathbf{t}_\mathbf{x}) \\
 &= -\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x})} \log[p_\theta(\mathbf{x}|\mathbf{z})p_\psi(\mathbf{t}_\mathbf{x}|\mathbf{z})] + D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})), \\
 \theta^*, \phi^*, \psi^* &= \arg \min_{\theta, \phi, \psi} \mathcal{L}_{\text{VAE}},
 \end{aligned} \tag{1}$$

where the symbol definition and meaning are in Tab1. In Eq.1, $D_{\text{KL}}(q_\phi||p_\theta)$ is the Kullback-Leibler divergence to quantify the distance between these two distributions. By minimizing the loss, we can obtain a set of neural network parameters that maximize the lower bound of the joint probability of observing the transcriptomes and the time of the input cells.

1.3 Comparison of model setting to existing method

Besides StemVAE, psupertime is an alternative method recently proposed to perform a similar analysis. In comparison, there are two major differences in model settings, making StemVAE uniquely suitable

Symbol	Mean
$\mathcal{D}$	The dataset contains n cells, $\mathcal{D} = \{(x^{(1)}, t^{(1)}), (x^{(2)}, t^{(2)}), \dots, (x^{(n)}, t^{(n)})\}$, $ \mathcal{D} = n$.
$x^{(i)}$	Each cell expresses on d genes, $x^{(i)} = [x_1^{(i)}, x_2^{(i)}, \dots, x_d^{(i)}]$.
$t^{(i)}$	The observed time of cell i .
x	One cell from dataset $\mathcal{D}$.
$\hat{x}, \hat{t}$	The reconstructed x and predicted t .
μ, σ	The mean and variance parameter of the multivariate Gaussian distribution.
z	The compressed low dimensional representation of x , and $z = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$
$q_\phi(z x)$	Encoder parameterized by ϕ outputs latent variable space distribution (μ, θ) .
$p_\theta(x z)$	Cell-Decoder parameterized by θ represents the likelihood of generating x given z .
$p_\psi(t_x z)$	Time-Decoder parameterized by ψ represents the likelihood of inferring observed time t_x .

Table 1: Symbol notation.

for analysing the time-series data here. First, psupertime is a supervised pseudo-time approach based on an ordinal logistic regression model which is suitable for scenarios where the response variable’s categories have a natural order, but the distances between the categories are not known. However, in biological scenarios, the time matters not only on the order but also the actual difference; namely, cellular differences are more significant with greater temporal differences in general. Second, since psupertime models the categorical variable, it requires seeing all time labels in the training, hindering the prediction of unseen time points. In other words, it is difficult to perform cross-validation by leaving one point out. In contrast, StemVAE directly predicts the physical time of the cell via a multilayer neural network, which can be easily used to predict unseen time labels while preserving the temporal characteristic of the cell. Therefore, when assessing the performance of StemVAE, the prediction is always performed as a test set by leaving the patient out of the training, including in main Figures 2b, 2c, and 2e.

References

- [KW13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [MGC22] Will Macnair, Revant Gupta, and Manfred Claassen. psupertime: supervised pseudotime analysis for time-series single-cell rna-seq data. *Bioinformatics*, 38(Supplement_1):i290–i298, 2022.