Supplementary Note

1 Detailed description of ESCO modeling

Modeling the extrinsic variation

$\mathbb{A}$. **Discrete cell groups:** Particularly, denote the set of DE genes as G^{DE}, and the marker gene set $\{G^i\}_{i=1}^K$ for k cell groups such that $G^1 \cup G^2 \cdots \cup G^k \cup \ldots G^K = G^{\text{DE}}$, we let the DE factor for each DE gene g in cell group k be

$$f_g^k = \begin{cases} h_g^k & \text{if } g \in G^k; \\ 1 & \text{otherwise,} \end{cases} \tag{1}$$
$$\text{where } \log h_g^k \overset{iid}{\sim} N(\mu_k, \sigma_k).$$

$\mathbb{B}$. **Tree-structured cell groups:** Specifically, given the similarity between cell groups by a $K \times K$ correlation matrix Σ generated from the tree structure, and a set of DE genes G^{DE}, we firstly select a small proportion of G^{DE} and split them into the marker genes for each group $G^1, G^2, \ldots, G^K$. We let the DE factor for each DE gene g in cell group k be

$$f_g^k = \begin{cases} h_g^k m^k; & \text{if } g \in G^k \\ h_k^g; & \text{otherwise} \end{cases} \tag{2}$$
$$\text{where } (\log h_g^1, \ldots, \log h_g^K) \overset{iid}{\sim} N(\boldsymbol{z}, \text{diag}\{\sigma_1, \ldots, \sigma_K\}),$$
$$\text{with } \boldsymbol{z} := (z_g^1, \ldots, z_g^K) \overset{iid}{\sim} N(\boldsymbol{\mu}, \Sigma),$$

and $m^k > 1$ is a scalar parameter controlling the level of the additional heterogeneity for each group.

$\mathbb{C}$. **Continuous cell trajectories:** Particularly, for each gene in the DE gene set G^{DE}, we simulate the DE factor at each step t in branch b with length T_b as

$$\begin{aligned} &\text{for } t = 1, \ldots, T_b: \\ &\quad f_g^{(t,b)} = \exp\left(w_g^{(t,b)}\right), \\ &\qquad \text{where } w_g^{(t,b)} = w_g^{(t-1,b)} + v_g^{(t-1,b)} \\ &\qquad \text{with } v_g^{(t,b)} = v_g^{(t-1,b)} + N(0,\ 2/T_b). \end{aligned} \tag{3}$$

In particular, we initialize

$$v_g^{(0,b)} \sim N(0, \sigma_b);$$
$$w_g^{(0,b)} = \begin{cases} 0, & \text{if } p(b) = \varnothing; \\ w_g^{(T_{p(b)}, p(b))}, & \text{otherwise.} \end{cases}$$

Then, for each branch b, we randomly sample several time points to generate the final cell samples, and let the "group" identity of cell sample c be $k(c) = (t, b)$.

Finally, we generate the base expression with an adjustment of library size for each gene g in cell c as

$$\begin{aligned} \lambda_{gc} &= L_c \frac{\widetilde{\lambda}_{gc}}{\sum_g \widetilde{\lambda}_{gc}} \quad \text{for each cell } c, \\ &\text{where } \widetilde{\lambda}_{gc} \overset{iid}{\sim} \begin{cases} \lambda_g f_g^{k(c)}, & \text{if } g \in G^{\text{DE}}, \\ \lambda_g, & \text{otherwise;} \end{cases} \\ &\text{and } \log L_c \overset{iid}{\sim} F_L, \end{aligned} \tag{4}$$

where $k(c)$ denotes the group identity of cell c.

Modeling the intrinsic variation

$\mathbb{A}$. **Marginal:** Particularly, we generate the marginal counts $\widetilde{Y}_{gc}$ as:

$$\begin{aligned} \widetilde{Y}_{gc} &\sim NB\left(\frac{1}{B_{gc}}, \frac{1}{\lambda_{gc} B_{gc}^2 + 1}\right) \\ &\text{where } B_{gc} \sim \left(\phi + \frac{1}{\lambda_{gc}}\right) \sqrt{df / \mathcal{X}^2(df)}; \end{aligned} \tag{5}$$

where ϕ is the common dispersion parameter, and df represents the degree of freedom of the $\mathcal{X}^2$, and NB represents the Negative Binomial distribution.

$\mathbb{B}$. **Co-expression:** Recall a copula is defined by a joint cumulative distribution function (CDF), $C(u) : [0,1]^p \to [0,1]$ with uniform marginal distributions. One of the most popular copula models is the Gaussian copula, which is defined simply as:

$$C_{\Sigma}^{\text{Gauss}} = N_{\Sigma}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_p)) \tag{6}$$

where Φ^{-1} denotes the inverse function of standard normal CDF, and N_{Σ} denotes the joint CDF of a multivariate normal random vector with zero means and correlation matrix Σ.

Then we generate true counts Y_{gc} via the following model:

$$\begin{aligned} Y_{gc} &= NB_{gc}^{-1}\left(\Phi^{-1}(X_{gc})\right) \quad \text{for } g = 1, 2, \ldots, p, \\ &\text{where } (X_{1c}, X_{2c}, \ldots, X_{pc}) \sim N(\mathbf{0}, \Sigma); \end{aligned} \tag{7}$$

and NB_{gc}^{-1} is the quantile function of the Negative Binomial distribution with parameters indexed by cell c and gene g in equation (5), and Σ is the target correlation matrix.

Modeling the technical noise

Particularly, as for the empirical approach from SymSim, one may resort to Zhang *et al.* (2019) for details. While as for the parametric approach from Splat, the observed counts Z_{gc} from the data is generated via the following

$$\begin{aligned} Z_{gc} &= Y_{gc}(1 - D_{gc}) \\ &\text{where } D_{gc} \sim Ber(\pi_{gc}) \\ &\text{with } \pi_{gc} = \frac{1}{1 + \exp\{-k(\log(\lambda_{gc}) - x_0)\}}, \end{aligned} \tag{8}$$

where π_{gc} denotes the probability of zero-inflation, given the expression mean λ_{gc}, Ber denotes the Bernoulli distribution, and Z_{gc} denotes the final observed counts.

2 Estimating the technical noise

ESCO also allows estimation of the median zero-inflation and shape parameters in equation (8). Though Splat already includes the corresponding estimation via fitting a logistic regression between the log-transformed gene mean and their observed zeros proportions, it is biased towards inflating the probability of excess zeros, as can be understood via the following reasoning:

Given a real scRNA-seq data set $\mathcal{Z} \in \mathbb{R}^{p\times n}$, where each element Z_{gc} is the observed count of the expression of gene g in cell c, let

$$\pi'_{gc} := \Pr\{Z_{gc} = 0\}\,. \tag{9}$$

Splat estimates π'_{gc} via fitting a logistic function to model the relationship between the log means of the normalized counts and the proportion of cell samples that are zero for each gene. Then Splat plugs the estimation $\widehat{\pi}_{gc}$ in place of π_{gc} in equation (8) to simulate $\widehat{Z}_{gc}$,

$$\widehat{Z}_{gc} = \widehat{Y}_{gc}(1 - \widehat{D}_{gc}), \quad \text{where } \widehat{D}_{gc} \sim Ber(\widehat{\pi}_{gc}). \tag{10}$$

and $\widehat{Y}_{gc}$ is the imitation of the true counts Y_{gc} for gene g in cell c simulated in the previous steps.

Assuming the estimation of π'_{gc} is accurate and the simulated true counts $\widehat{Y}_{gc}$ well mimics the real truth Y_{gc}, then this approach would cause more sparsity than expected, since the proportion of zeros in the simulated observation will be

$$\begin{aligned}\Pr\Big\{\widehat{Z}_{gc} = 0\Big\} &= \Pr\Big\{\widehat{Y}_{gc} = 0\Big\} + \Pr\Big\{\widehat{Y}_{gc} \neq 0, \widehat{D}_{gc} = 1\Big\} \\ &\overset{(*)}{=} \Pr\Big\{\widehat{Y}_{gc} = 0\Big\} + \Pr\Big\{\widehat{Y}_{gc} \neq 0\Big\} \Pr\Big\{\widehat{D}_{gc} = 1\Big\}\,,\end{aligned} \tag{11}$$

where (*) is true since $\widehat{Y}_{gc}$ and $\widehat{D}_{gc}$ are independent once condition on λ_{gc}. Therefore,

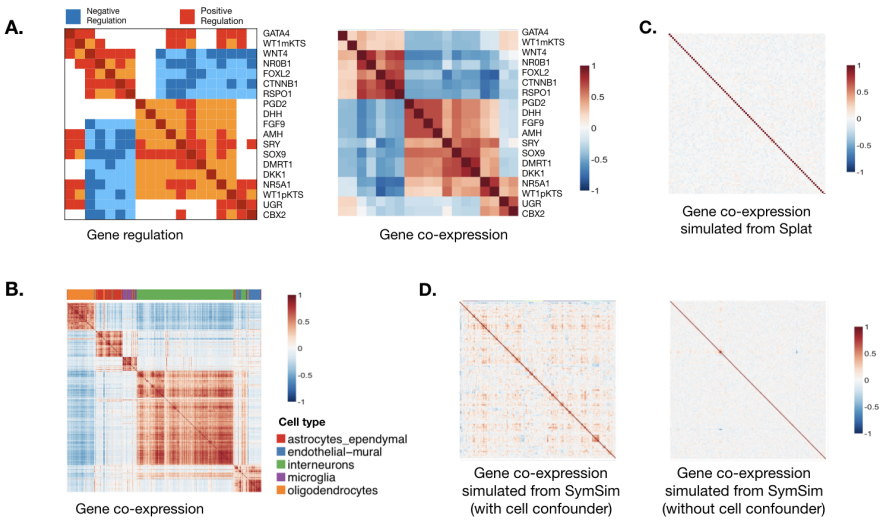
$$\begin{aligned}\Pr\Big\{\widehat{Z}_{gc} = 0\Big\} &= \Pr\Big\{\widehat{Y}_{gc} = 0\Big\} + \Pr\Big\{\widehat{Y}_{gc} \neq 0\Big\}\,\widehat{\pi}_{gc} \\ &\geq \Pr\Big\{\widehat{Y}_{gc} = 0\Big\}\,\widehat{\pi}_{gc} + \Pr\Big\{\widehat{Y}_{gc} \neq 0\Big\}\,\widehat{\pi}_{gc} \\ &= \widehat{\pi}_{gc} = \pi'_{gc} = \Pr\{Z_{gc} = 0\}\,,\end{aligned} \tag{12}$$

From the above calculation, one simple correction for this bias uses:

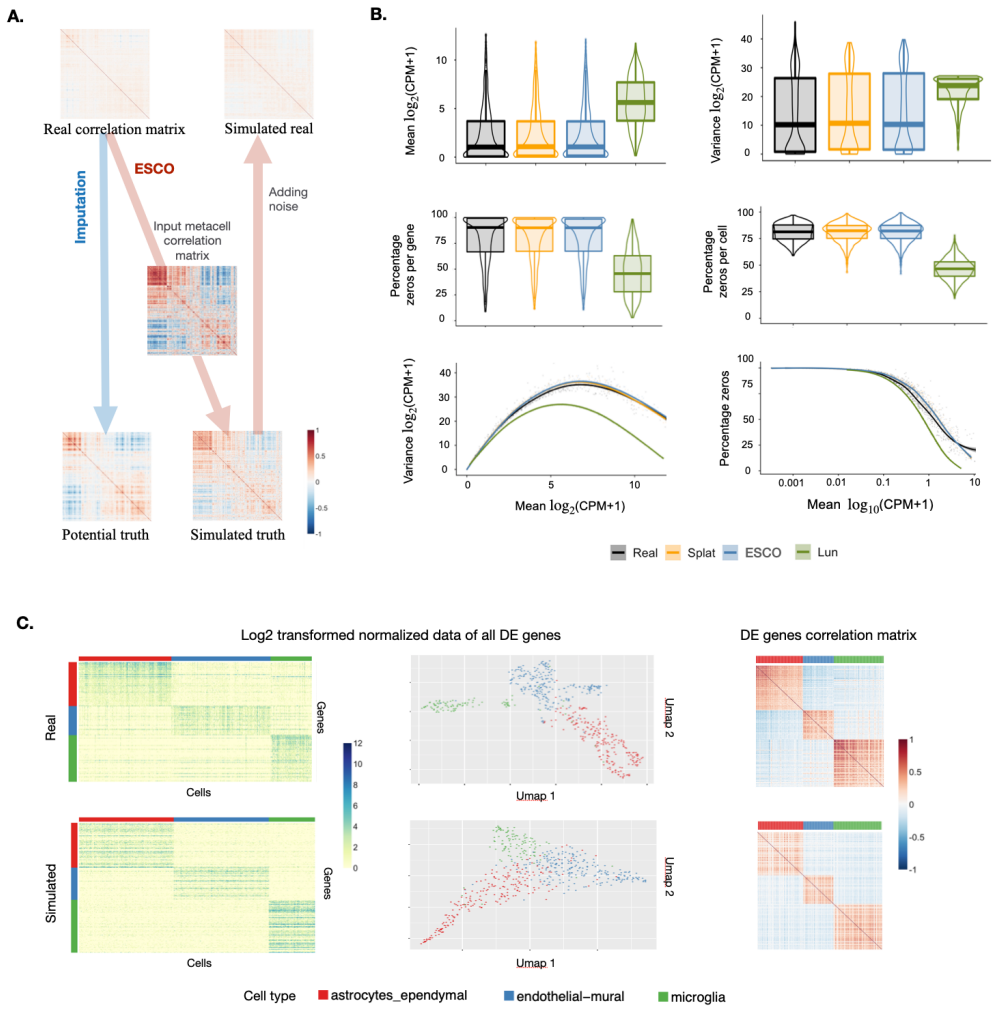
$$\widetilde{\pi}_{gc} = \frac{\widehat{\pi}_{gc} - \Pr\Big\{\widehat{Y}_{gc} = 0\Big\}}{1 - \Pr\Big\{\widehat{Y}_{gc} = 0\Big\}} \tag{13}$$

as the plug-in for equation (8). Particularly, ESCO approximates $\Pr\Big\{\widehat{Y}_{gc} = 0\Big\}$ using the CDF of Poisson with mean λ_{gc} at zero.

Supplementary Figure



Supplementary Figure 1: Gene co-expression is informative, but we lack satisfactory methods to simulate it for scRNA-seq data. **A.** Connection between gene regulation and gene co-expression. The left panel shows the regulation relationship between the 19 genes in Gonadal Sex Determination (Ríos *et al.*, 2015), while the right panel shows Pearson's correlation matrix for these 19 genes with inferred expression (Pratapa *et al.*, 2020). **B.** Connection between gene co-expression and cell group clusters. The correlation matrix of the 500 most significant marker genes of the five major cell types from the Zeisel data (Zeisel *et al.*, 2015) with corresponding gene types marked with a color bar on top, clustered using hierarchical clustering. **C.** The correlation matrix for 200 simulated genes from Splat (Zappia *et al.*, 2017), without zero-inflation. **D.** The correlation matrix for 200 simulated genes from SymSim (Zappia *et al.*, 2017), without zero-inflation. The left and right panels show results with and without the cell confounding effect, respectively. Specifically, the confounding effect arise as SymSim generates the gene expression for gene g in cell c via a random product model, that is expression $Y_{gc} = \lambda_g \tau_c$, where $\lambda_g \overset{iid}{\sim} F$, and $\tau_c \overset{iid}{\sim} G$. Once conditioning on the cell confounder τ_c, the correlation between expression of genes g_1 and g_2 disappears.



Supplementary Figure 2: ESCO can learn both the cell heterogeneity and gene co-expression from the data. **A.** The generation process of gene co-expression for one homogeneous cell group from real data using ESCO. Particularly, the example is for 500 randomly selected genes in pyramidal CA1 cell type (911 cells) from Zeisel data. **B.** The comparison of marginal features of real data consist of 500 randomly selected genes in pyramidal CA1 cell type (911 cells) extracted from Zeisel data, and the corresponding simulated data using different simulators. Particularly, Lun (Lun *et al.*, 2016) is one of the earliest scRNA-seq simulators, which has been found to be suboptimal (Zappia *et al.*, 2017). We include it here as a clear contrast with the state-of-art methods. **C.** The comparison of real data consist of 4000 most differential expressed genes in three cell types (astrocytes_ependymal, endothelial_mural, microglia) of 526 cells in total extracted from Zeisel data, and the corresponding simulated data using ESCO. While the UMAP depiction differs somewhat, the expression and co-expression patterns match closely.

Supplementary Table

(a) With gene co-expression

(#genes, #cells)	(1000, 300)	(5000, 500)	(10000, 1000)	(15000, 3000)	(20000, 5000)
One group	10.6	17.2	49.8	343.5	1102.8
Discrete groups	15.8	27.5	89.7	458.9	1365.7
Tree structured groups	17.8	31.2	80.5	454.6	1328.2
Continuous trajectories	16.3	29.5	99.1	452.6	1270.8

(b) Without gene co-expression

(#genes, #cells)	(1000, 300)	(5000, 500)	(10000, 1000)	(15000, 3000)	(20000, 5000)
One group	2.5	8.0	10.4	42.0	94.2
Discrete groups	6.6	12.5	28.0	91.7	184.0
Tree structured groups	7.4	16.0	34.1	112.7	212.6
Continuous trajectories	7.3	12.9	30.2	101.7	196.5

Supplementary Table 1: Time (seconds) spent of simulating large complex data.

References

Lun, A. T. *et al.* (2016). Pooling across cells to normalize single-cell rna sequencing data with many zero counts. *Genome biology*, **17**(1), 75.

Pratapa, A. *et al.* (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, **17**(2), 147–154.

Ríos, O. *et al.* (2015). A boolean network model of human gonadal sex determination. *Theoretical Biology and Medical Modelling*, **12**(1), 26.

Zappia, L. *et al.* (2017). Splatter: simulation of single-cell rna sequencing data. *Genome biology*, **18**(1), 174.

Zeisel, A. *et al.* (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, **347**(6226), 1138–1142.

Zhang, X. *et al.* (2019). Simulating multiple faceted variability in single cell rna sequencing. *Nature communications*, **10**(1), 1–16.