A new statistical approach for the simultaneous clustering of genes and cells in spatial transcriptomic experiments

Andrea Sottosanti

Davide Risso

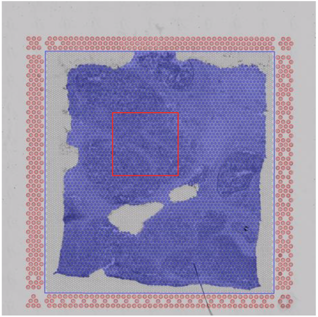
December 14^{th}, 2020

University of Padova - Department of Statistical Sciences
andrea.sottosanti@unipd.it

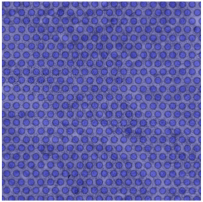
Università
degli Studi
di Padova

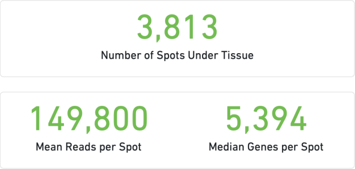
WORK IN PROGRESS

The 10x Genomics Visium technology



$\Downarrow$





3,813
Number of Spots Under Tissue

149,800
Mean Reads per Spot

5,394
Median Genes per Spot

- $j = 1, \ldots, p$ spots, each of which is spatially located;
- Number of spots $\approx$ number of cells;
- for each spot, $i = 1, \ldots, n$ gene expressions are available.

Spatially expressed genes and research motivations

- The rise of such advanced technology has increased the interest for the so-called spatially expressed ($s.e.$) genes.

Spatially expressed genes and research motivations

- The rise of such advanced technology has increased the interest for the so-called spatially expressed (*s.e.*) genes.
- There are methods for discovering *s.e.* genes: `spatialDE` [Svensson et al., 2018], `Trendsceek` [Edsgärd et al., 2018], `SPARK` [Sun et al., 2020].

Spatially expressed genes and research motivations

- The rise of such advanced technology has increased the interest for the so-called spatially expressed (*s.e.*) genes.
- There are methods for discovering *s.e.* genes: `spatialDE` [Svensson et al., 2018], `Trendsceek` [Edsgärd et al., 2018], `SPARK` [Sun et al., 2020].

however...

- These methods do not account for the presence of different cell types.
- Some (clusters of) genes might be *s.e.* just in some specific cell types.

Some aspects to consider

- Let

$$\underset{n\times p}{\mathbf{X}} : \; x_{ij} = \text{measure of expression of the } i\text{-th gene in the } j\text{-th spot.}$$

Some aspects to consider

- Let

$$\underset{n \times p}{\mathbf{X}} : \; x_{ij} = \text{measure of expression of the } i\text{-th gene in the } j\text{-th spot.}$$

- The spatial coordinates of the spots $(\mathbf{s}_1, \ldots, \mathbf{s}_p)$ are known.

Some aspects to consider

- Let

$$\underset{n\times p}{\mathbf{X}}:\ x_{ij} = \text{measure of expression of the } i\text{-th gene}$$
$$\text{in the } j\text{-th spot.}$$

- The spatial coordinates of the spots $(\mathbf{s}_1, \ldots, \mathbf{s}_p)$ are known.
- Just for now, we assume there is only one type of cells.

Some aspects to consider

- Let

$$\underset{n\times p}{\mathbf{X}} : \; x_{ij} = \text{measure of expression of the } i\text{-th gene in the } j\text{-th spot.}$$

- The spatial coordinates of the spots $(\mathbf{s}_1, \ldots, \mathbf{s}_p)$ are known.
- Just for now, we assume there is only one type of cells.

Some aspects to consider:

1. correlation of the genes → $Cor(x_{ij}, x_{i'j})$,
2. (spatial) correlation of the spots → $Cor(x_{ij}, x_{ij'})$.

A statistical model

We assume that the experiment matrix $\mathbf{X}$ distributes as

$$\mathbf{X} \sim \mathcal{MVN}_{n,p}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Delta}),$$

where $\mathcal{MVN}$ denotes the Matrix Variate Normal distribution [Gupta and Nagar, 2018]:

- $\boldsymbol{\mu} = \mu \cdot \mathbf{1}_{n \times p}$ is the mean matrix;
- $\boldsymbol{\Sigma}$ is an $n \times n$ matrix which express the **correlation of the genes** (rows);
- $\boldsymbol{\Delta}$ is an $p \times p$ matrix which express the **correlation of the cells** (columns).

A statistical model

Regarding the rows,

$$\mathbf{\Sigma} := \begin{cases} \sigma_i^2 \text{ in position } (i,i); \\ 0 \text{ elsewhere}; \end{cases} \qquad \sigma_i^2 \sim \mathcal{IG}(\alpha, \beta).$$

A statistical model

Regarding the rows,

$$\mathbf{\Sigma} := \begin{cases} \sigma_i^2 \text{ in position } (i,i); \\ 0 \text{ elsewhere}; \end{cases} \qquad \sigma_i^2 \sim \mathcal{IG}(\alpha, \beta).$$

Regarding the columns,

$$\mathbf{\Delta} = \tau \cdot \mathbf{K}(\phi) + \xi \cdot \mathbb{1}_{p \times p}.$$

- $\tau \in \mathbb{R}^+$ is the amount of spatial expression;
- $\mathbf{K}(\cdot)$ is the spatial kernel matrix: example,

$$\mathbf{K}_{j,j'} = \exp\{-||\mathbf{s}_j - \mathbf{s}_{j'}||^2/(2\phi^2)\};$$

- $\phi \in \mathbb{R}^+$ is the spatial scale;
- $\xi \in \mathbb{R}^+$ is the nugget effect (variance not imputable to the spatial structure);

The Co-clustering problem

- K gene clusters → $\mathcal{C}_i = k$ means that gene i belongs to the k-th gene cluster;
- R cell clusters → $\mathcal{D}_j = r$ means that cell j belongs to the r-th cell type.

The Co-clustering problem

- K gene clusters → $\mathcal{C}_i = k$ means that gene i belongs to the k-th gene cluster;
- R cell clusters → $\mathcal{D}_j = r$ means that cell j belongs to the r-th cell type.

$$\mathbf{X} = \begin{array}{c|c|c|c|c|}
 & r=1 & r=2 & \dots & r=R \\ \cline{2-5}
k=1 & \mathbf{X}_{11} & \mathbf{X}_{12} & \dots & \mathbf{X}_{1R} \\ \cline{2-5}
k=2 & \mathbf{X}_{21} & \ddots & \dots & \vdots \\ \cline{2-5}
\dots & \vdots & \dots & \ddots & \vdots \\ \cline{2-5}
k=K & \mathbf{X}_{K1} & \dots & \dots & \mathbf{X}_{KR} \\ \cline{2-5}
\end{array}$$

The Co-clustering problem

- K gene clusters → $\mathcal{C}_i = k$ means that gene i belongs to the k-th gene cluster;
- R cell clusters → $\mathcal{D}_j = r$ means that cell j belongs to the r-th cell type.

$$\mathbf{X} = \begin{array}{c|c|c|c|c|} & r=1 & r=2 & \dots & r=R \\ \cline{2-5} k=1 & \mathbf{X}_{11} & \mathbf{X}_{12} & \dots & \mathbf{X}_{1R} \\ \cline{2-5} k=2 & \mathbf{X}_{21} & \ddots & \dots & \vdots \\ \cline{2-5} \dots & \vdots & \dots & \ddots & \vdots \\ \cline{2-5} k=K & \mathbf{X}_{K1} & \dots & \dots & \mathbf{X}_{KR} \\ \cline{2-5} \end{array}$$

$$\Downarrow$$

$$\mathbf{X}_{kr} \sim \mathcal{MVN}_{n_k, p_r}(\boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}_{kr}, \tau_{kr} \cdot \mathbf{K}(\phi_r) + \xi_{kr} \cdot \mathbb{1}_{p_r \times p_r}),$$

$$\sigma^2_{kr,i} \sim \mathcal{IG}(\alpha_{kr}, \beta_{kr})$$

for $k = 1, \dots, K$ and $r = 1, \dots, R$.

spatialLIBD data

- We exploit the human dorsolateral prefrontal cortex (DLPFC) spatial transcriptomics data generated with the 10x Genomics Visium technology by [Maynard et al., 2020] and contained in the R package spatialLIBD [Collado-Torres et al., 2020].
- We reduced the dataset size, using the first **1000** most variable **genes** measured in **1585 spots**.
- We run our model on $\log$-counts data using $K = 1$ and $R = 4$.
- The estimation procedure is initialized using the results from k-means.

spatialLIBD data - clustering

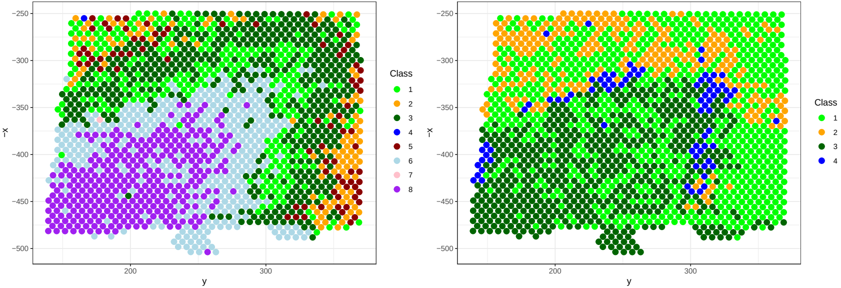


Figure 1: Data: subject 151673. Left: clustering provided by spatialLIBD. Right: Clustering from our method.

cell cluster	$\hat{\mu}$	$\hat{\tau}/\hat{\xi}$	$\hat{\phi}$
1	0.863	0.479	19.159
2	0.451	0.304	21.232
3	0.501	0.357	19.283
4	0.198	0.200	31.440

`spatialLIBD` data - genes variance

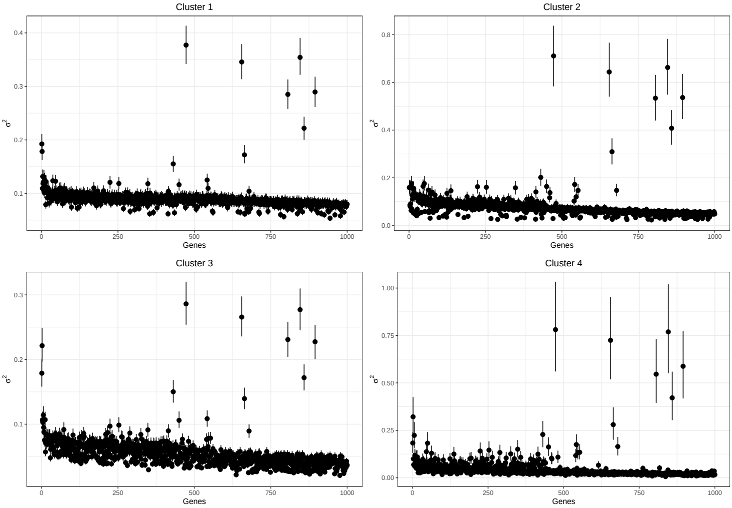


Figure 2: Expected value and 95% interval of σ_i^2 in every cell cluster, given the data the parameter estimates. The first two highly variable genes are $\mathrm{ENSG00000123560}$ and $\mathrm{ENSG00000197971}$.

Acknowledgements

NIH NATIONAL CANCER INSTITUTE

- Dario Righelli
- Martin Morgan, Vince Carey, Levi Waldron
- Giovanna Menardi

Thank you for the attention!

andrea.sottosanti@unipd.it

References I

Collado-Torres, L., Maynard, K. R., and Jaffe, A. E. (2020).
***LIBD Visium spatial transcriptomics human pilot data inspector*.**
https://github.com/LieberInstitute/spatialLIBD - R package version 1.2.0.

Edsgärd, D., Johnsson, P., and Sandberg, R. (2018).
Identification of spatial expression trends in single-cell gene expression data.
Nature methods, 15(5):339–342.

References II

- Gupta, A. K. and Nagar, D. K. (2018).
  ***Matrix variate distributions*, volume 104.**
  CRC Press.

- Maynard, K. E., Collado-Torres, L., Weber, L. M., Uytingco, C., Barry, B. K., Williams, S. R., et al. (2020).
  Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex.
  bioRxiv.

- Sun, S., Zhu, J., and Zhou, X. (2020).
  Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies.
  Nature Methods, 17(2):193–200.

References III

Svensson, V., Teichmann, S. A., and Stegle, O. (2018).
SpatialDE: identification of spatially variable genes.
Nature methods, 15(5):343–346.