STAT 35510
Lecture 13

Spring, 2024

Jingshu Wang

Outline

- Identify spatially variable genes
- Cell type deconvolution
- Imputation
  - Impute missing genes for image-based data
  - Increase the resolution of sequencing-based data

Moran's I score in Squidpy

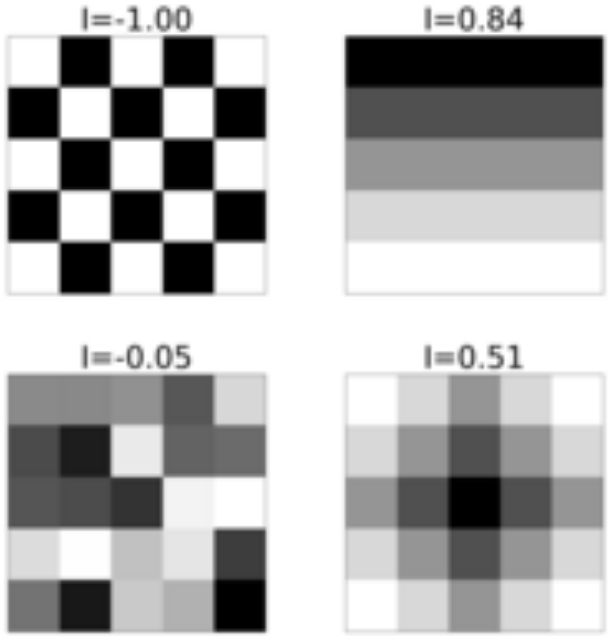
Moran's I score

- Measurement of spatial autocorrelation

$$I = \frac{N}{W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

where

- N is the number of spatial units indexed by i and j;
- x is the variable of interest;
- $\bar{x}$ is the mean of x;
- w_{ij} are the elements of a matrix of spatial weights with zeroes on the diagonal (i.e., $w_{ii} = 0$);
- and W is the sum of all w_{ij} (i.e. $W = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}$).



- Construct p-value for each gene
  - Under the null of no spatial autocorrelation (for specific definitions see Wikipedia)

$$E(I) = \frac{-1}{N-1} \qquad \mathrm{Var}(I) = \frac{NS_4 - S_3 S_5}{(N-1)(N-2)(N-3)W^2} - (E(I))^2$$

  - Convert into z-scores and compute p-value

SpatialDE (Svensson et. al. Nature Methods 2018)

- Main idea: for each gene assume a Gaussian process model

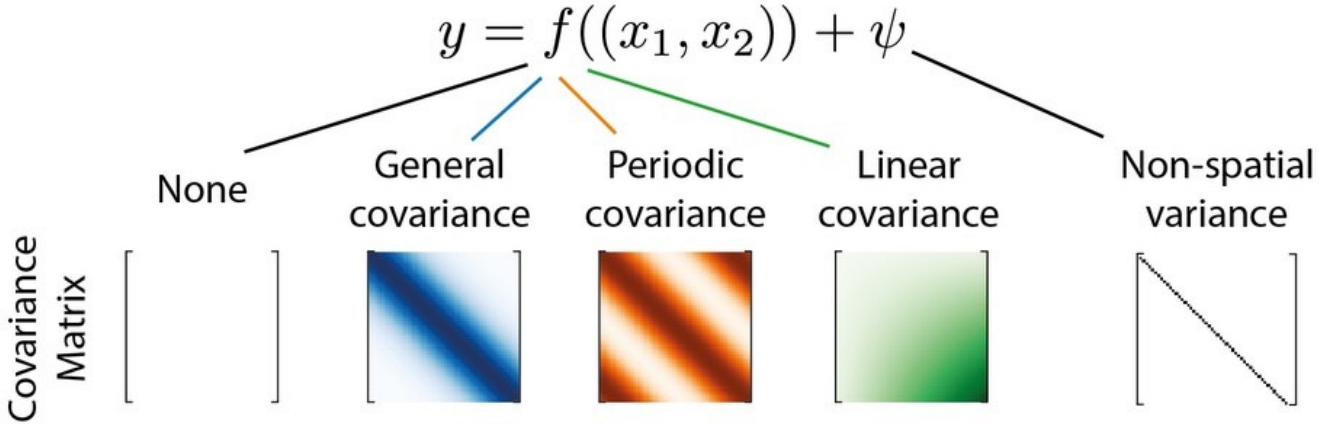
$$P(y \mid \mu, \sigma_s^2, \delta, \Sigma) = N(y \mid \mu \cdot 1, \sigma_s^2 \cdot (\Sigma + \delta \cdot I))$$

  - Common mean across all spots / cells
  - Covariance depend on spatial locations

$$\Sigma_{i,j} = k(x_i, x_j) = \exp\left(-\frac{|x_i - x_j|^2}{2 \cdot l^2} \right)$$

  - Construct p-values: likelihood ratio test testing whether $\Sigma = 0$

- Model selection assuming periodic covariance and linear covariance



SpatialDE (Svensson et. al. Nature Methods 2018)

Expression histology

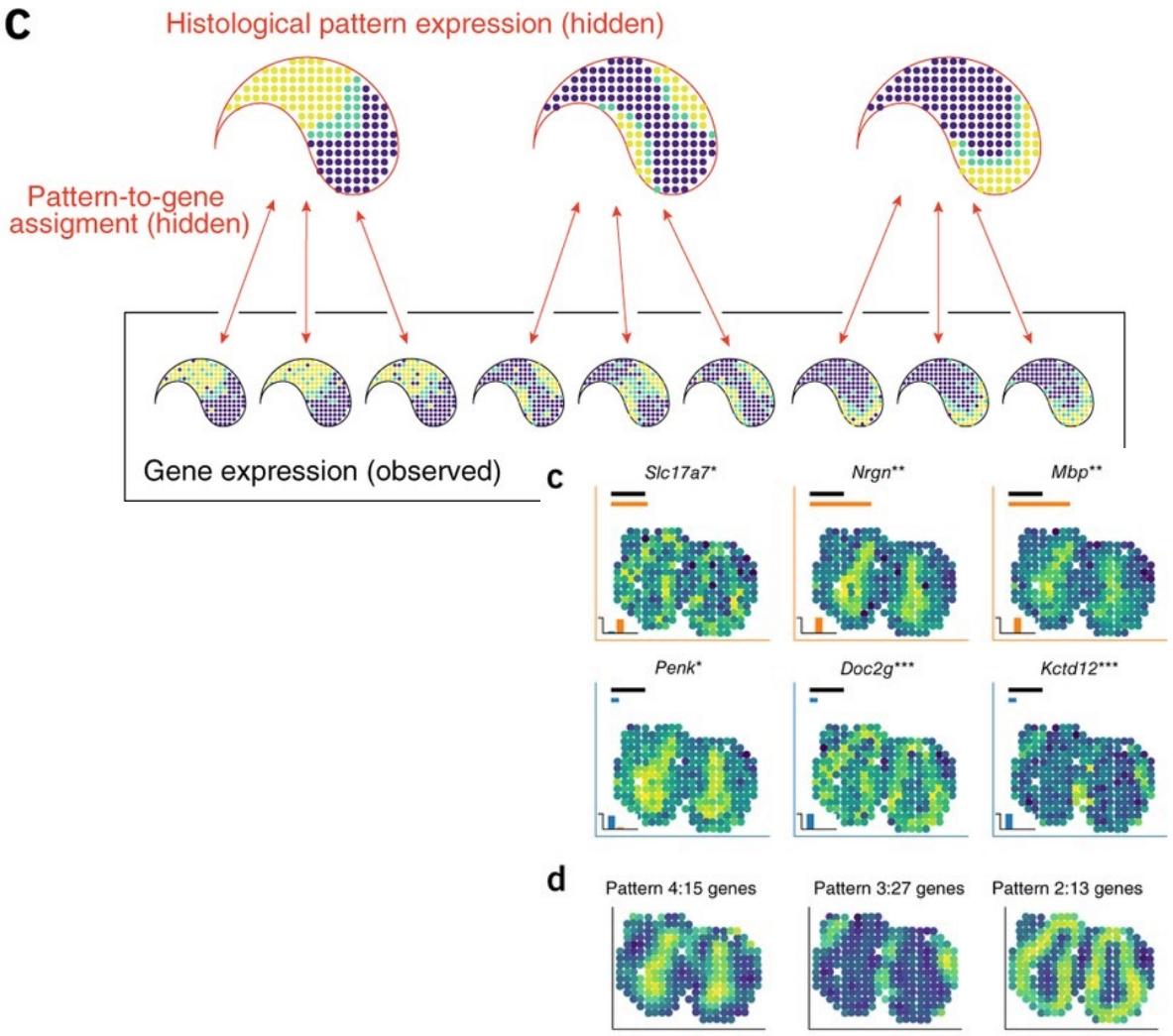
- Perform gene clustering using a hierarchical mixture model on features

$$P(Y,\mu,Z,\sigma_e^2,\Sigma) = P(Y \mid \mu,Z,\sigma_e^2) \cdot P(\mu \mid \Sigma) \cdot P(Z)$$

$$P(Y \mid \mu,Z,\sigma_e^2) = \prod_{k=1}^{K}\prod_{g=1}^{G} N(y_g \mid \mu_k,\sigma_e^2)^{(z_{g,k})}$$

$$P(\mu \mid \Sigma) = \prod_{k=1}^{K} N(\mu_k \mid 0,\Sigma)$$

$$P(Z) = \prod_{k=1}^{K}\prod_{g=1}^{G}\left(\frac{1}{K}\right)^{(z_{g,k})}$$



SpatialDE2 (Kats et. al. BioRXiv, 2021)

- Use Poisson model for the data
- Superior computational speed

- Core steps:
  - Tissue region segmentation using HMRF
    - Assume that gene expression counts follow Poisson distribution within each hidden state / cluster

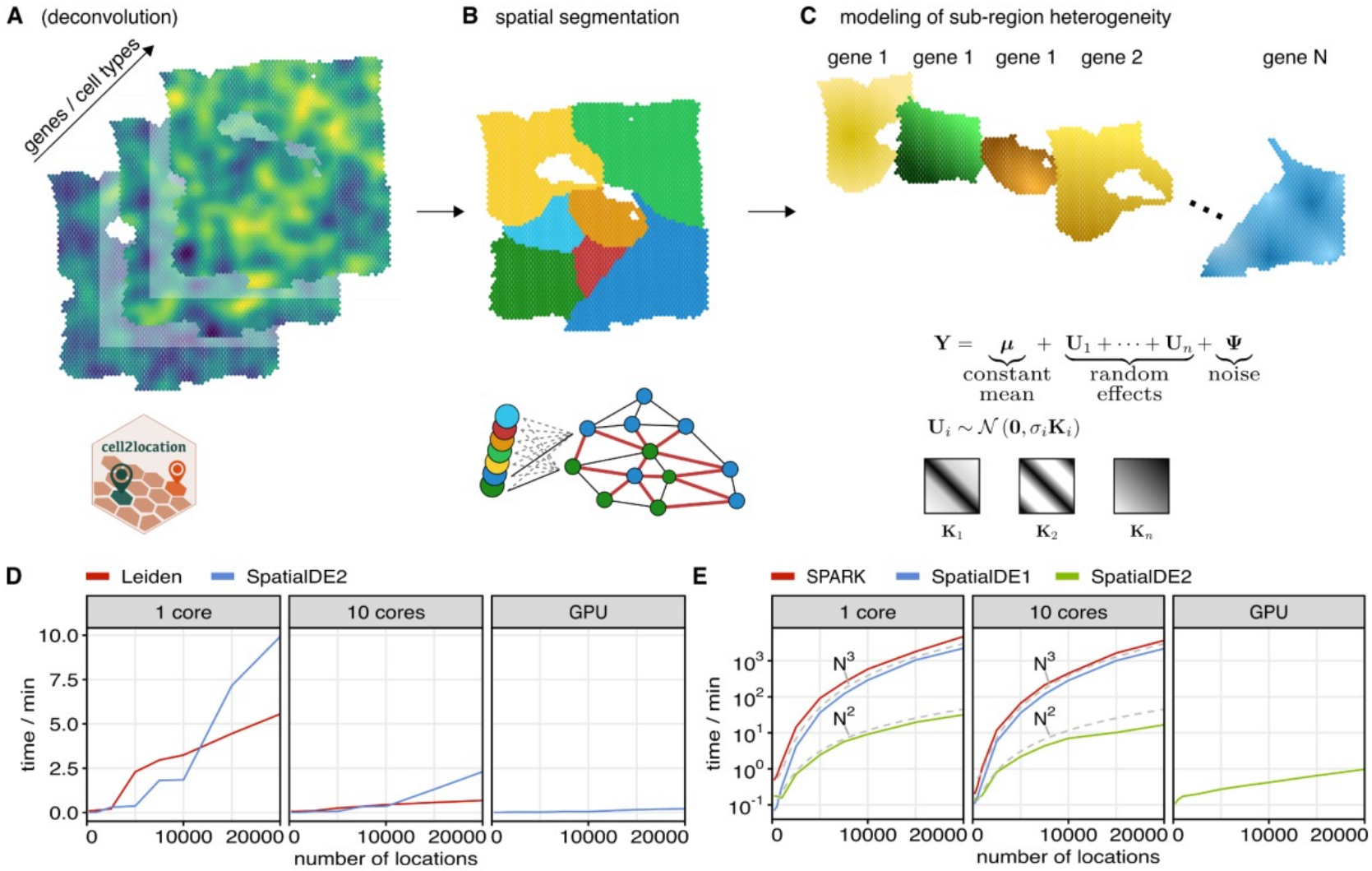
$$\lambda_{gc} \sim \mathcal{G}(\gamma_1, \gamma_2)$$
$$y_{gn} \mid x_n = c, \boldsymbol{\lambda}_g \sim \text{Pois}(S_n \lambda_{gc})$$

  - Detect spatially variable genes

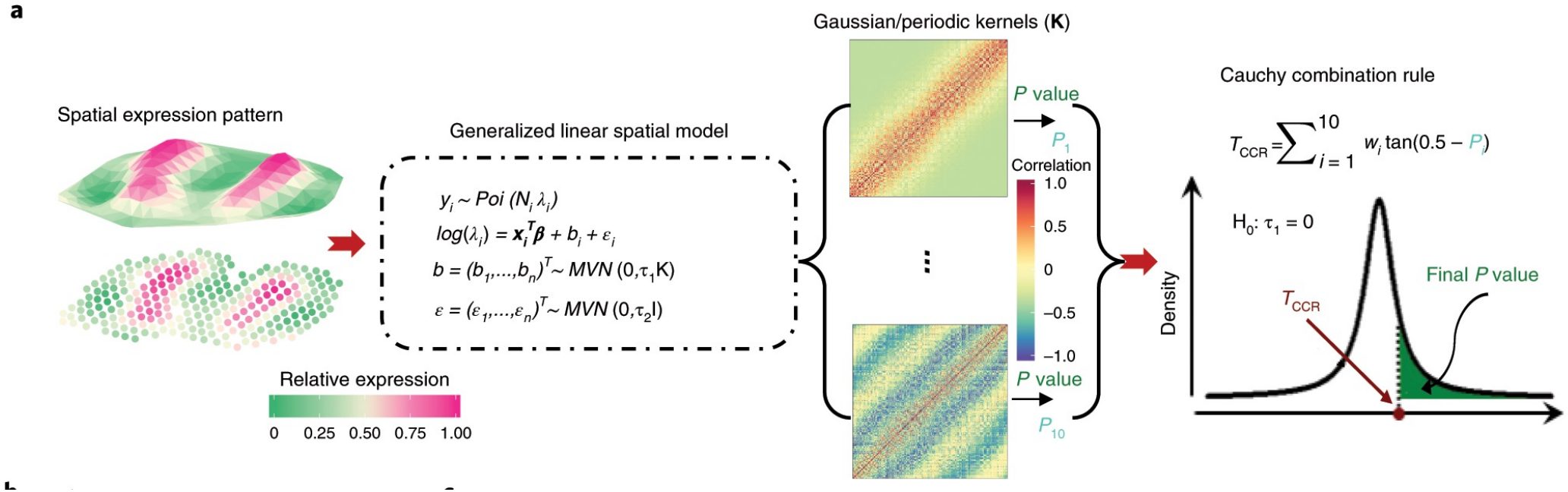
$$\mathbf{e} \sim \mathcal{N}(\mu\mathbf{1}, \sigma_1\mathbf{K}_1 + \cdots + \sigma_k\mathbf{K}_k + \sigma_n\mathbf{I})$$
$$\mathbf{y} \mid \mathbf{e} \sim \text{Pois}(\mathbf{s} \odot \exp(\mathbf{e}))$$

    - Use the GLMM score test and implement in GPU (details omitted)
  - Need to transform data to Gaussian for gene clustering analyses
    - Use variational inference to speed up computation

SpatialDE2 (Kats et. al. BioRXiv, 2021)



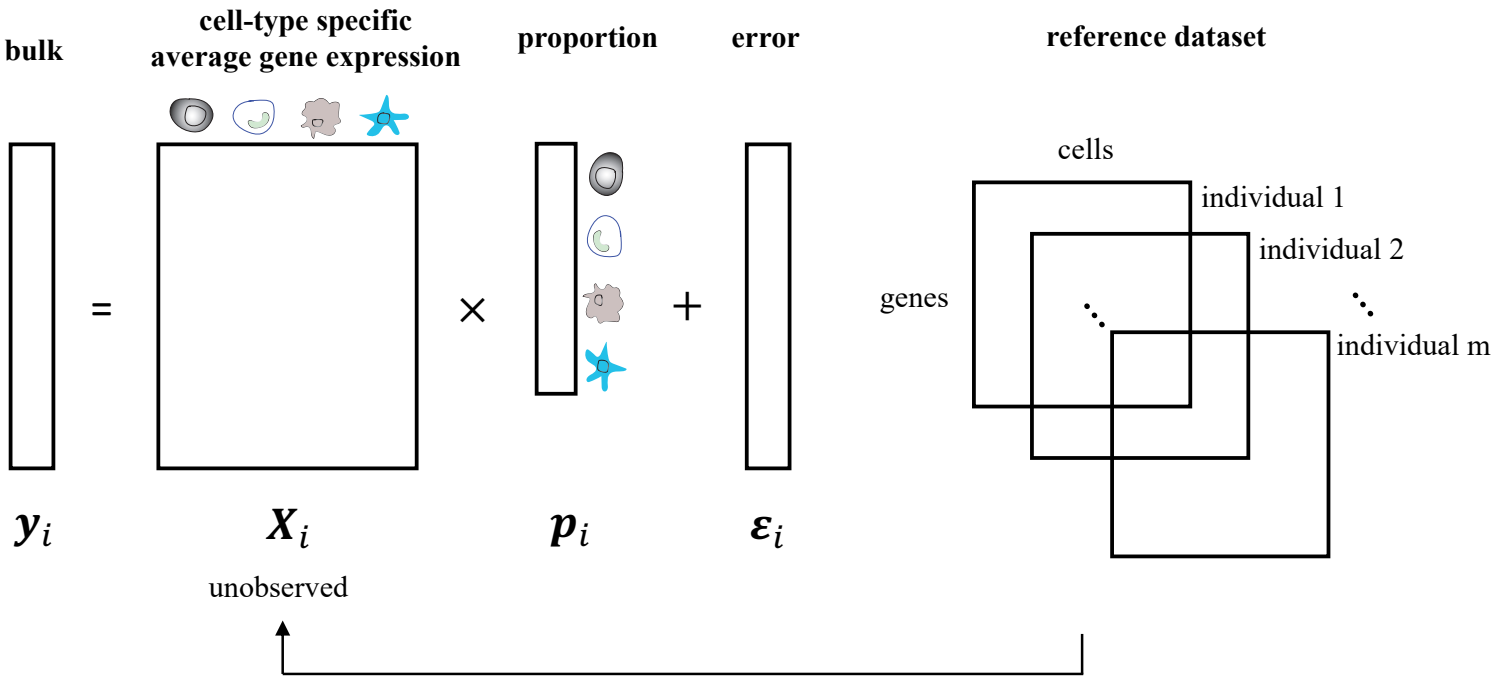
SPARK (Sun et. al., Nature Methods 2020)



- Hierarchical Poisson GP model
  - Can adjust for confounding covariates such as cell types
  - Test for each specific kernel and combine p-values across kernels
  - Testing is challenging -> use a penalized quasi-likelihood algorithm

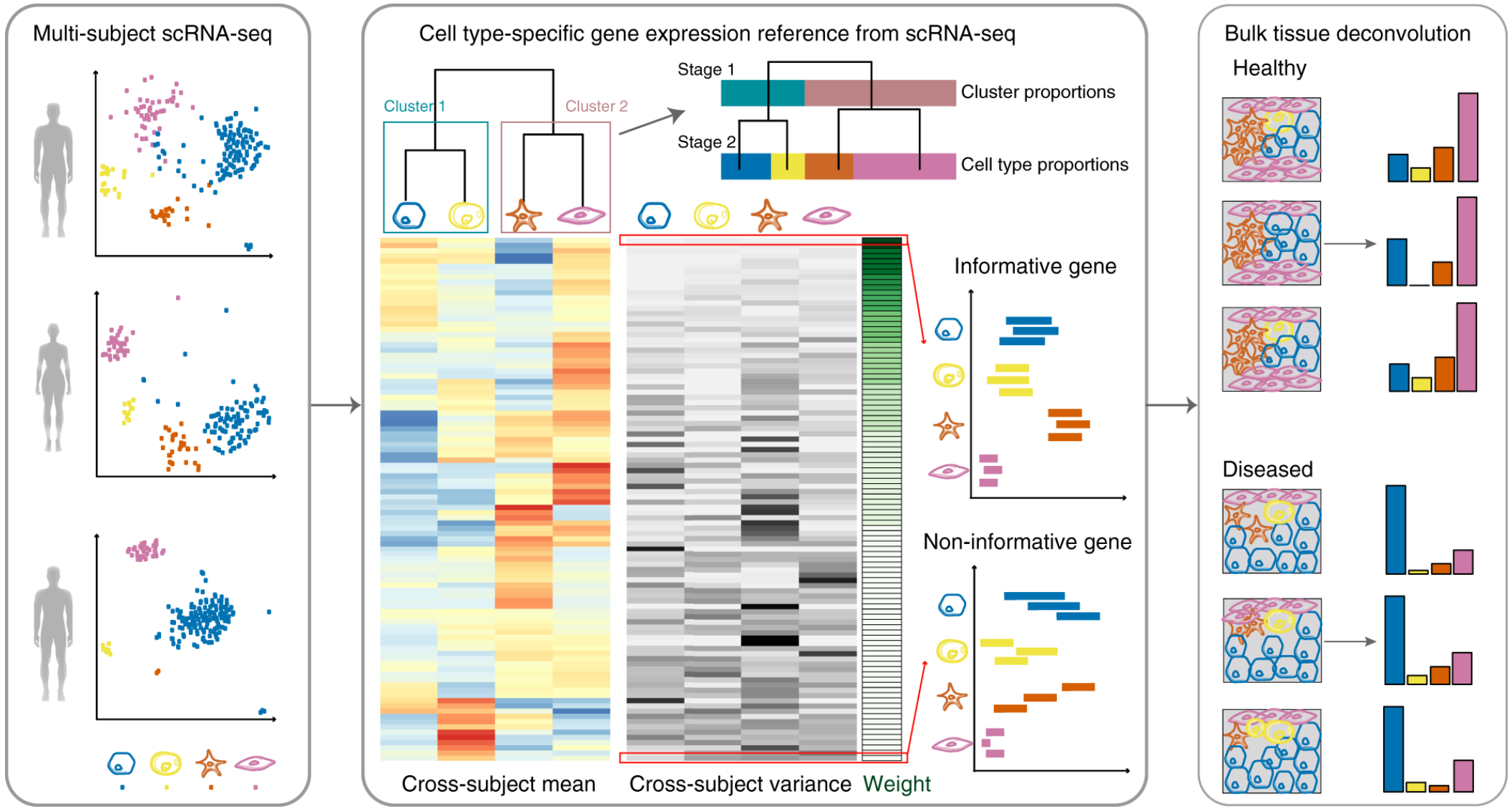
- As other GP models, computational cost is high
- The authors have later developed SPARK-X (Zhu et. al., Genome Biology 2021) to speed up
  - based on similarity test between gene covariance matrix and spatial similarity

Cell type deconvolution for bulk RNA-seq



- Main challenges (Xie and Wang, ArXiv, 2022):
  - Cell-type specific gene expressions from reference datasets may not be reliable
    - Variability of gene expression across individuals
    - Platform specific biases between bulk and single-cell RNA sequencing data
  - Missing cell types in the reference data
  - Genes are not independent across each other
  - How does the uncertainty in estimated cell types affect downstream analyses?

MuSiC (Wang et. al., Nature Comm, 2019)



MuSiC (Wang et. al., Nature Comm, 2019)

- Key steps
  - Normalize reference scRNA-seq data
    - Normalize by the library size but no transformations (why?)
  - A linear regression model:

$$Y_{jg} = C_j \cdot \left(\sum_{k=1}^{K} p_{jk} \, S_k \, \theta_{jg}^k + \epsilon_{jg} \right)$$

    - Two constraints:

$$\text{(C1) Non-negativity: } p_j^k \geq 0 \text{ for all } j, k; \text{ (C2) Sum-to-one: } \sum_{k=1}^{K} p_j^k = 1 \text{ for all } j$$

    - $S_k \theta_{jg}^k$: absolute gene expression profiles for each cell type
      - scRNA-seq only provides relative abundance θ_{jg}^k
      - Assume S_k (cell size) is the same across all cell types
  - Solve the model by Weighted non-negative least squares
    - Intuitively, marker genes should have higher weights
      - That intuition is wrong by Gauss-Markov theorem (Xie and Wang, ArXiv, 2022)
    - Give higher weights to genes that can be estimated and measured more accurately
      - Genes with less variability across samples for cell-type specific expressions
      - Genes has less technical noise

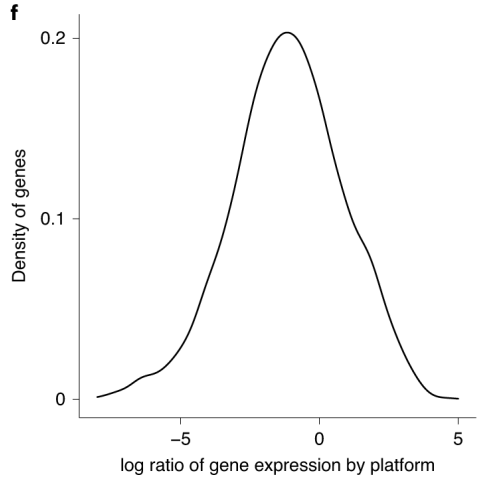
MuSiC (Wang et. al., Nature Comm, 2019)

- Key steps
  - Normalize reference scRNA-seq data
  - A linear regression model:

$$Y_{jg} = C_j \cdot \left(\sum_{k=1}^{K} p_{jk} \, S_k \, \theta_{jg}^{k} + \epsilon_{jg} \right)$$

  - Solve the model by Weighted non-negative least squares
    - Intuitively, marker genes should have higher weights
      - That intuition is wrong by Gauss-Markov theorem (Xie and Wang, ArXiv, 2022)
    - Give higher weights to genes that can be estimated and measured more accurately
      - Genes with less variability across samples for cell-type specific expressions
      - Genes has less technical noise
    - Iteratively reweighting in MuSiC
      - Estimate the variance of each gene given the current estimated p_{jk}
      - Inverse variance weighting to update the estimate of p_{jk}
  - Recursive tree-guided deconvolution
    - Deconvolute major cell types first

RCTD (Cable et. al, Nature Biotech 2022)



- Cell type deconvolution for spatial transcriptomics
  - Consider gene-specific biases across platforms
  - Does not smooth across spatial locations
  - Need to decompose many spots simultaneously

- Model

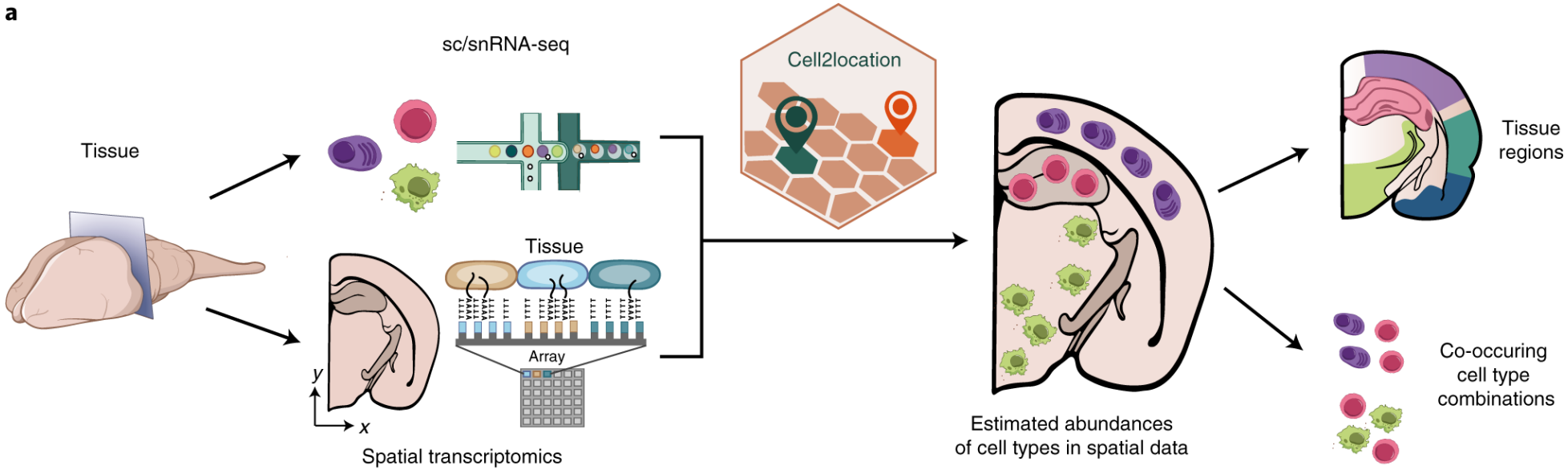
$$Y_{i,j}|\lambda_{i,j} \sim \text{Poisson}(N_i \lambda_{i,j})$$
$$\log(\lambda_{i,j}) = \alpha_i + \log\left(\sum_{k=1}^{K} \beta_{i,k}\mu_{k,j}\right) + \gamma_j + \varepsilon_{i,j}$$

  - $\beta_{i,k}$: cell type proportion per spot
  - γ_j: gene-specific biases, prior $\gamma_j \sim N(0, \sigma_\gamma^2)$

- Model fitting
  - Estimate cell-type specific gene expressions from the reference
  - Select marker genes for each cell type
  - Estimate γ_j by aggregating across spots
    - Identification issues if γ_j are arbitrarily different and only marker genes are selected? (Wang and Xie, Arxiv, 2022)

Cell2location (Kleshchevnikov et. al., Nature Biotech 2022)

- Goal: get spatial distribution of cells types → cell type deconvolution



- Model for a spatial spot

$$d_{s,g} \sim NB\left(\mu_{s,g}, \alpha_{e,g}\right)$$

$$\mu_{s,g} = \left(\underbrace{m_g}_{\text{technology sensitivity}} \cdot \underbrace{\sum_f w_{s,f}\, g_{f,g}}_{\text{cell type contributions}} + \underbrace{s_{e,g}}_{\text{additive shift}} \right) \cdot \underbrace{y_s}_{\text{per-location sensitivity}}$$

Additive shift account for contaminating RNA

Cell2location (Kleshchevnikov et. al., Nature Biotech 2022)

$$\mu_{s,g} = \left(\underbrace{m_g}_{\text{technology sensitivity}} \cdot \underbrace{\sum_f w_{s,f}\, g_{f,g}}_{\text{cell type contributions}} + \underbrace{s_{e,g}}_{\text{additive shift}} \right) \cdot \underbrace{y_s}_{\text{per-location sensitivity}}$$

- Hierarchical model on the proportions $w_{s,f}$ assuming factor models

$$w_{s,f} \sim \texttt{Gamma}(\mu^w_{s,f} v^w, v^w)$$

$$\mu^w_{s,f} = \sum_r z_{s,r}\, x_{r,f}$$

  - Priors on the factors add some regularization?

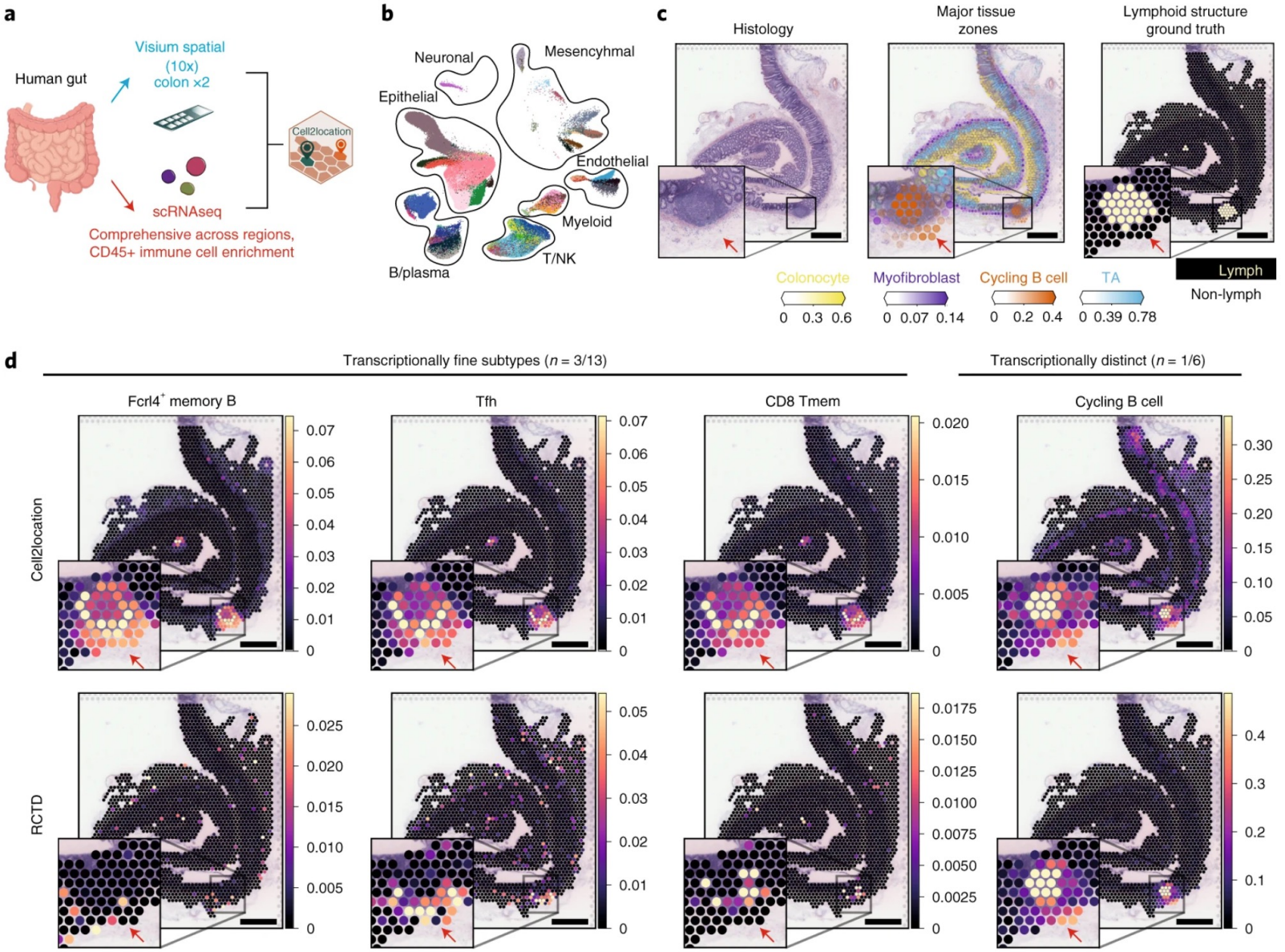
$$z_{s,r} \sim \texttt{Gamma}(B_s/R, 1/(N_s/B_s)) \qquad N_s \sim \texttt{Gamma}(\hat{N} \cdot v^n, v^n) \qquad B_s \sim \texttt{Gamma}(\hat{B}, 1),$$

$$x_{r,f} \sim \texttt{Gamma}(K_r/R, K_r) \qquad K_r \sim \texttt{Gamma}(\hat{A}/\hat{B}, 1)$$

  - Similar priors on other parameters

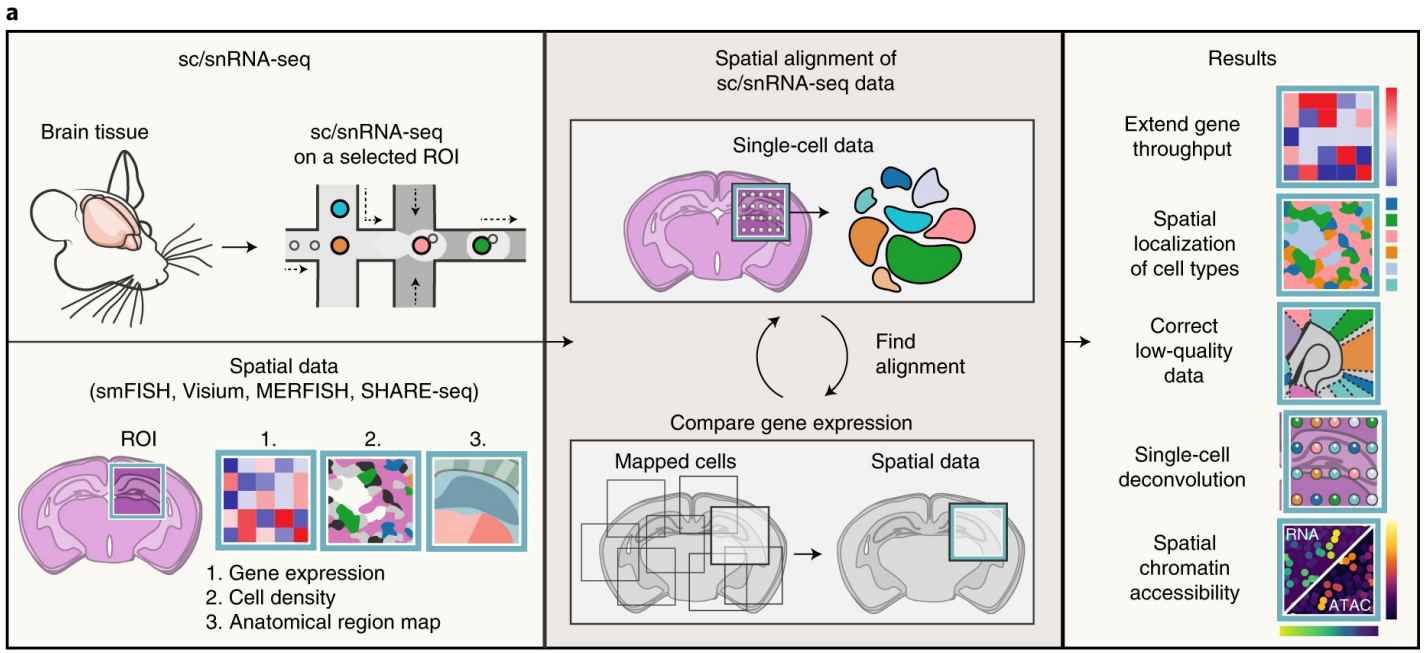
- Use Variational Bayes to solve the model
  - Seems to be challenging to solve

Cell2location (Kleshchevnikov et. al., Nature Biotech 2022)



Tangram (Biancalani et. al., Nature Methods 2021)

- Predict the spatial locations of each cell in sc/snRNA-seq data by leveraging spatial transcriptomics
  - Spatial transcriptomics can be either sequencing-based or image-based
  - Goals:
    - Impute missing gene expressions in image-based spatial transcriptomics
    - Denoising and cell type deconvolution for sequencing-based spatial transcriptomics data
    - Map sc/snRNA-seq data to spatial locations
    - Predict chromatin accessibility for spatial transcriptomics data



Tangram (Biancalani et. al., Nature Methods 2021)

- Cell mapping
  - Input: spatial voxel by gene matrix G, cell density vector d across voxels d, Cell by gene expression matrix S
  - Output: cell by voxel mapping matrix M
  - Loss function

$$\Phi\left(\tilde{M}\right) = KL\left(\vec{\mathbf{m}}, \vec{\mathbf{d}}\right) - \sum_{k}^{n_{genes}} cos_{sim}\left((M^T S)_{*,k}, G_{*,k}\right)$$

$$- \sum_{j}^{n_{voxels}} cos_{sim}\left((M^T S)_{j,*}, G_{j,*}\right),$$

$$m_j = \sum_{i}^{n_{cells}} M_{ij} / n_{cells}$$

$$M_{ij} = softmax(\tilde{M})_{ij} = \frac{e^{\tilde{M}_{ij}}}{\sum_{l}^{n_{voxels}} e^{\widetilde{M_{il}}}}$$

  - Mapping only a subset of genes (mapping with a filter)
    - Include a real-values filtering vector $\tilde{f}$ in training $f_i = \sigma(\tilde{f}_i)$

$$\Phi\left(\tilde{M}, \overrightarrow{\tilde{f}}\right) = KL\left(\overrightarrow{\mathbf{m}^{\mathbf{f}}}, \vec{\mathbf{d}}\right) - \sum_{k}^{n_{genes}} cos_{sim}\left((M^T S^f)_{*,k}, G_{*,k}\right)$$

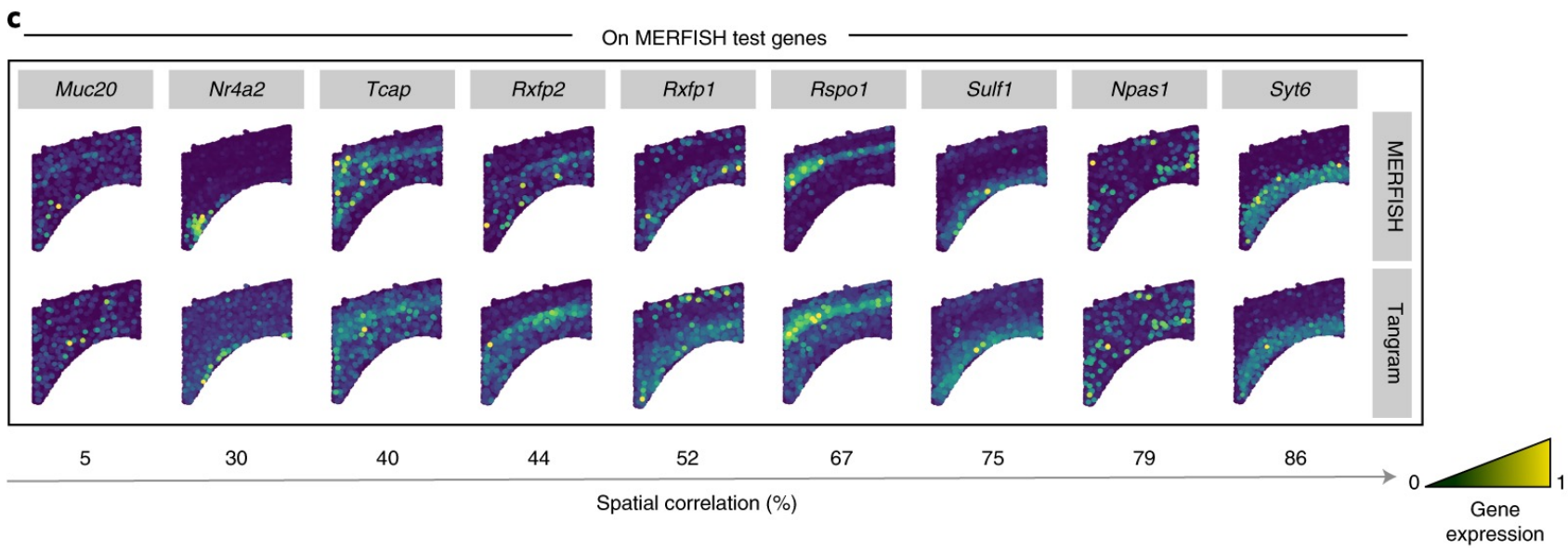
$$- \sum_{j}^{n_{voxels}} cos_{sim}((M^T S^f)_{j,*}, G_{j,*}) - \lambda_{r_1} \sum_{i,j}^{n_{cells}, n_{voxels}} M_{ij} log\left(M_{ij}\right)$$

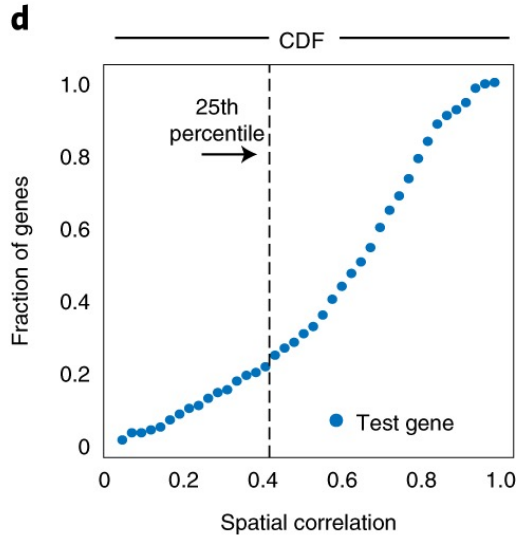
$$+ abs(\sum_{i}^{n_{cells}} f_i - \mathbf{n}_{\text{target\_cells}}) + \sum_{i}^{n_{cells}} (f_i - f_i^2).$$

Tangram (Biancalani et. al., Nature Methods 2021)

- Cell mapping
  - For image-based spatial transcriptomics, n_(target_cells)=n_voxel

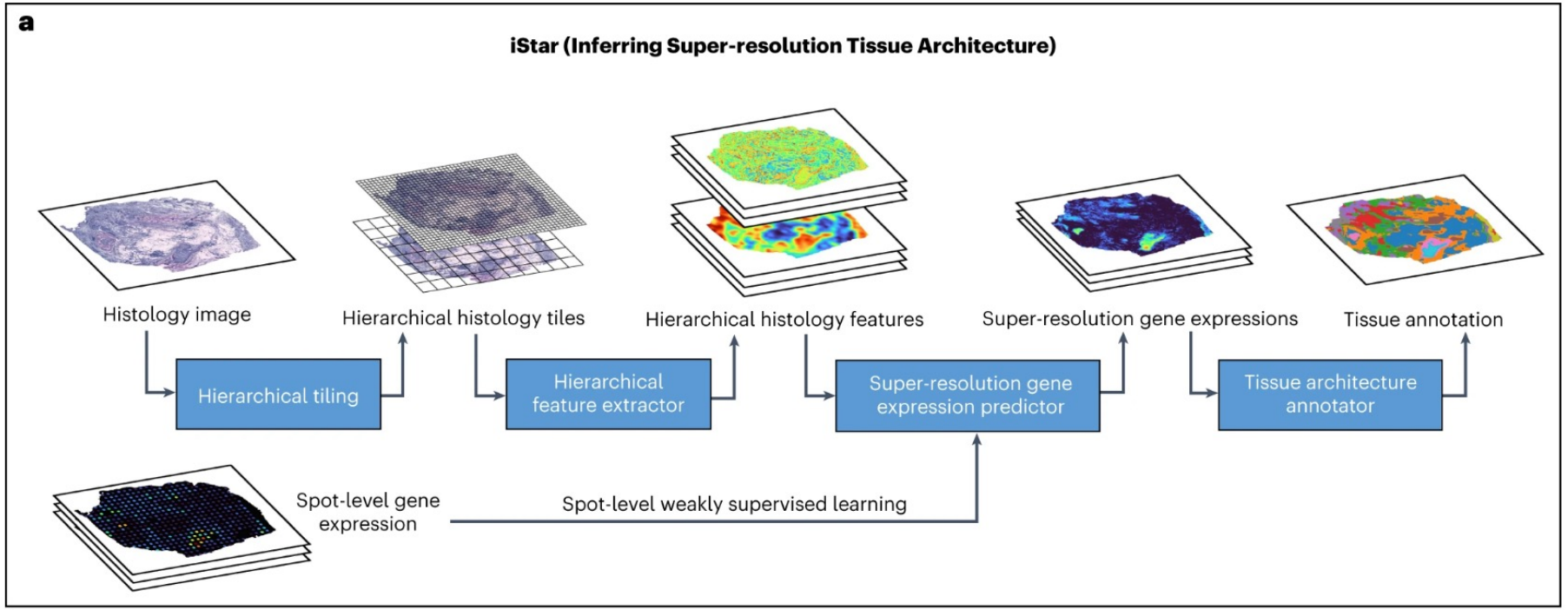
- Transfer cell type annotations in sc/snRNA-seq to spatial data based on M
  - For low resolution spatial transciptomics, assign a probability (cell type proportions)
  - For single-cell resolution data, assign to the cell type with maximum probability





iStar (Zhang et. al., Nature Biotech 2024)

- Making histological image as a guidance of the cell type at a higher resolution to increase the resolution of the sequencing-based data



iStar (Zhang et. al., Nature Biotech 2024)

- Key steps:
  - Extract histological features
    - Partition into image tiles hypercritically with different resolution:
      original pixel, 16*16 pixel blocks, 256 * 256 pixel blocks
    - Use hierarchical vision transformers (HViTs) to extract features from the image tiles
      - Each 16 * 16 block and 25 * 25 block receives a low-dimensional embedding (dimension C_1 and C_2)
      - Pretrain the HViTs model on public histological image
    - Final output:

      $$\text{histology feature image } H = [h_{mn}]_{m=1,n=1}^{M',N'} \text{ of size } M' \times N' \text{ with } C_1 + C_2 + 3 \text{ channels}$$

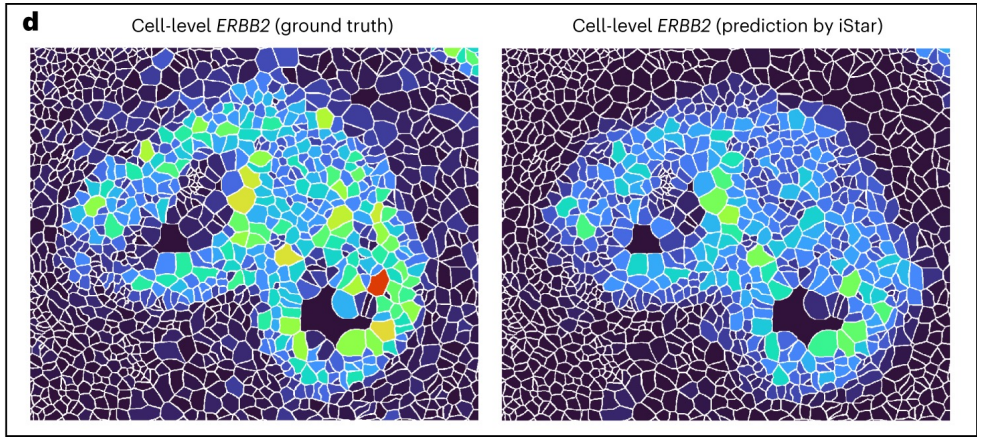
      - $(M', N') = \left(\frac{M}{16}, \frac{N}{16}\right), \left(\frac{M}{32}, \frac{N}{32}\right), \left(\frac{M}{64}, \frac{N}{64}\right), \left(\frac{M}{128}, \frac{N}{128}\right)$

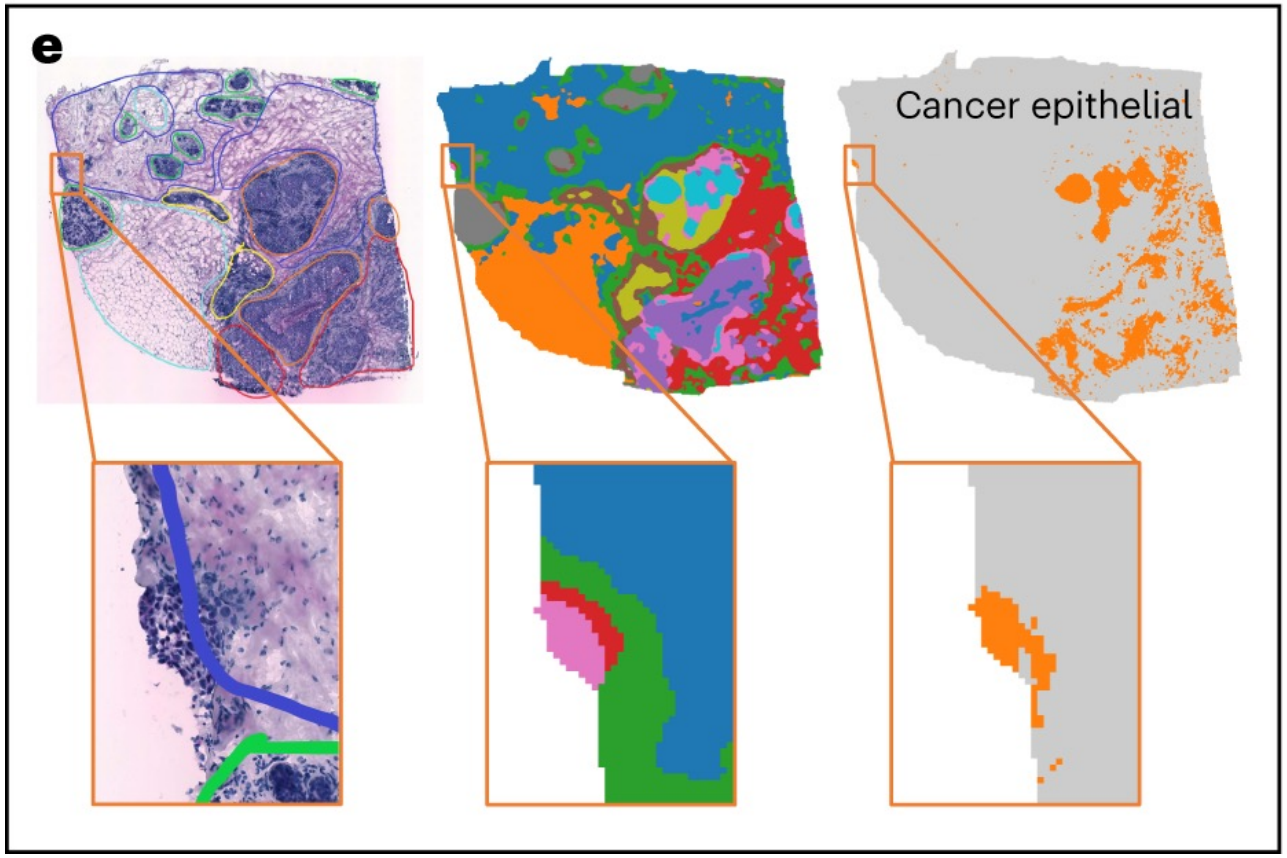
  - Prediction super-resolution gene expressions

$$\mathscr{L} = \sum_{k=1}^{K} \sum_{s=1}^{S} \left(y_{ks} - \sum_{(m,n) \in \mathscr{M}_s} g_k \left(h_{mn} \right) \right)^2$$

    - No use of scRNA-seq data at all for deconvolution
    - If cell segmentation is provided, can obtain single-cell level gene expressions
    - Only predict the top 1000 HVGs
  - Provide cell type annotations of the regions

iStar (Zhang et. al., Nature Biotech 2024)





Related papers

- Svensson, V., Teichmann, S. A., & Stegle, O. (2018). SpatialDE: identification of spatially variable genes. *Nature methods*, *15*(5), 343-346.
- Kats, I., Vento-Tormo, R., & Stegle, O. (2021). SpatialDE2: fast and localized variance component analysis of spatial transcriptomics. *Biorxiv*, 2021-10.
- Sun, S., Zhu, J., & Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nature methods, 17(2), 193-200.
- Zhu, J., Sun, S., & Zhou, X. (2021). SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. Genome biology, 22(1), 184.
- Xie, D., & Wang, J. (2022). Robust Statistical Inference for Cell Type Deconvolution. arXiv preprint arXiv:2202.06420.
- Wang, X., Park, J., Susztak, K., Zhang, N. R., & Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature communications, 10(1), 380.
- Cable, D. M., Murray, E., Zou, L. S., Goeva, A., Macosko, E. Z., Chen, F., & Irizarry, R. A. (2022). Robust decomposition of cell type mixtures in spatial transcriptomics. Nature biotechnology, 40(4), 517-526.
- Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H. W., Li, T., ... & Bayraktar, O. A. (2022). Cell2location maps fine-grained cell types in spatial transcriptomics. Nature biotechnology, 40(5), 661-671.
- Biancalani, T., Scalia, G., Buffoni, L., Avasthi, R., Lu, Z., Sanger, A., ... & Regev, A. (2021). Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. Nature methods, 18(11), 1352-1362.
- Zhang, D., Schroeder, A., Yan, H., Yang, H., Hu, J., Lee, M. Y., ... & Li, M. (2024). Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. Nature Biotechnology, 1-6.