SpaCoEx: Sparse Gene Selection for Spatially Varying Co-expression in Spatial Transcriptomics

Michael Jian

Mentors: Prof. Gil Alterovitz and Dr. Shaojun Pei

MIT PRIMES Conference

October 24, 2026

Summary

Scientific Question

- Gene expression varies across tissue.
- Co-expression can vary spatially too.
- Example: 36,591 genes imply $>$ 669M gene pairs.
- Goal: preserve that variation with a small gene subset.

SpaCoEx Approach

- Gaussian weighting to estimate local correlations.
- Log-Euclidean transformation for matrix-valued structure.
- L_1-regularized gene weights for structured sparsity.

Key Results

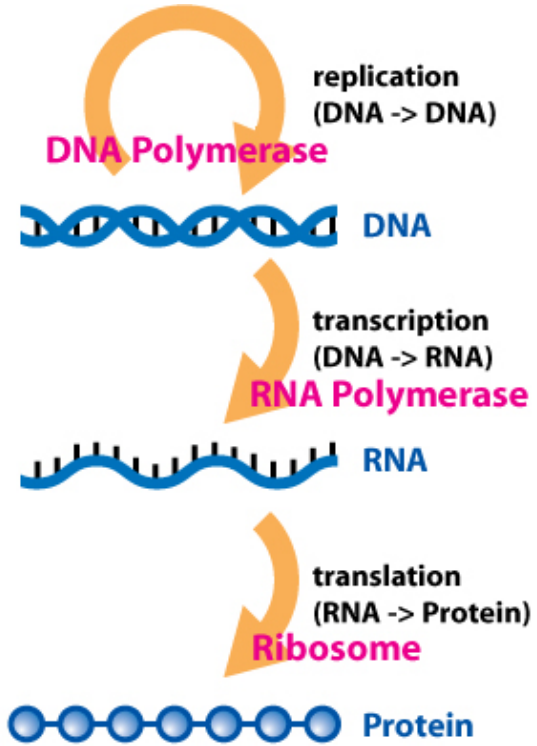
- Keratinocyte: 29 of 45 genes retain 96.74% of variation.
- Joint SpaCoEx: best ARI among non-baseline methods.
- B2M-HLA-C correlation is higher in cancer regions

Conclusion: SpaCoEx identifies a compact set of genes that preserves spatially varying co-expression structure and provides features complementary to conventional expression analysis.

Gene Expression

Gene expression is the process by which the instructions in DNA are converted into the structures and functions of a cell.

- The activity level of genes can vary across cells and tissue regions.
- Gene expression can be quantified via RNA levels.



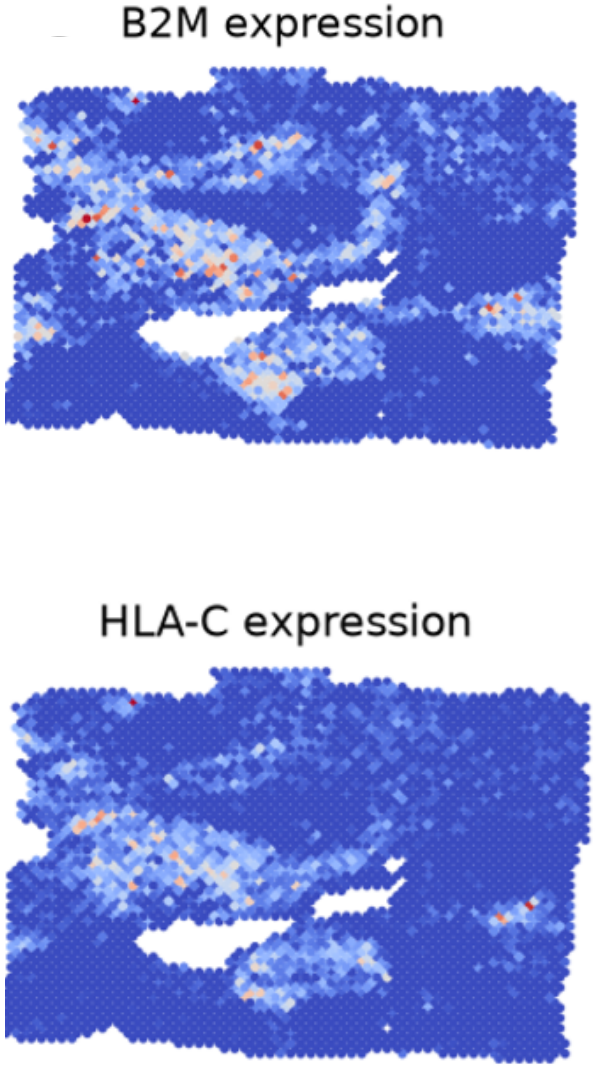
Spatial Transcriptomics

Spatial Transcriptomics is a technique that collects:

- Location of all spots (could be multiple cells)
- gene expression at each spot

Example: a human breast cancer tissue slice with 36,591 genes and 3,798 cells.

- **B2M** expression.
- **HLA-C** expression.



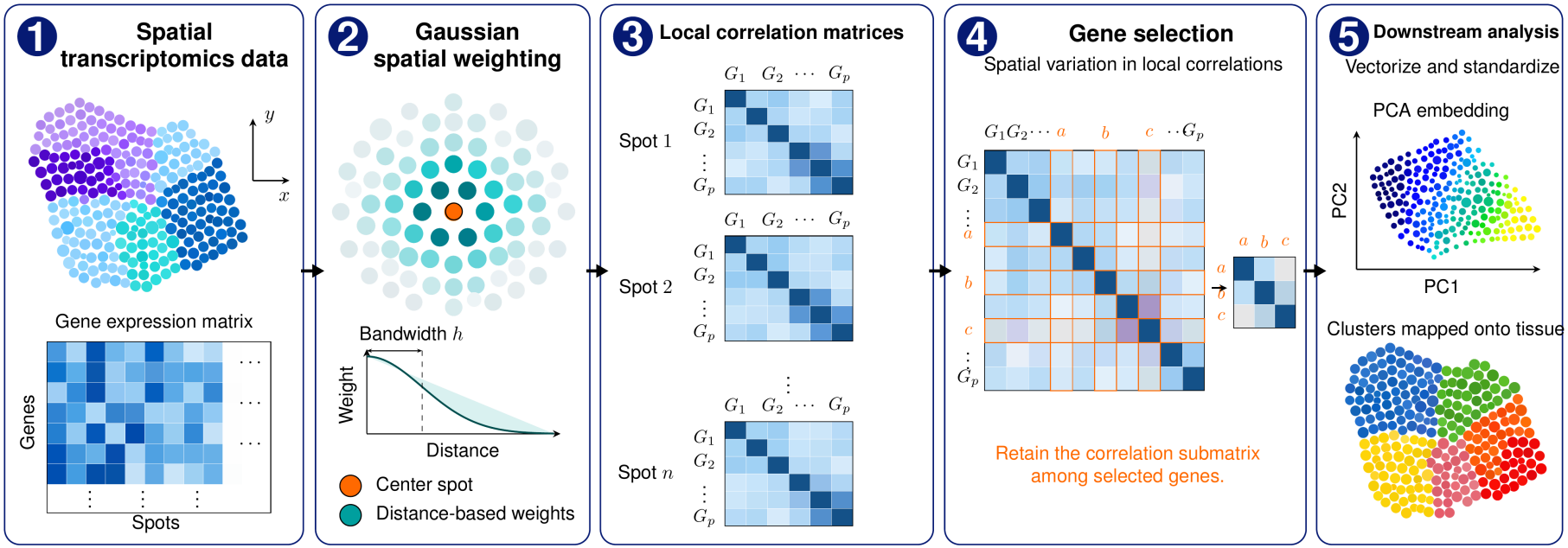
Scientific Question:

- Existing studies mainly focus on spatially varying genes, but genes rarely work alone.
- Their co-expression is also expected to vary spatially.
- For p genes, the number of gene pairs is $\binom{p}{2}$.
- For the breast cancer tissue slice, there are more than 669 million of gene pairs!

Question

How can we select **a small set of genes** that preserves spatial variation in local co-expression?

SpaCoEx Workflow



Spatial coordinates and gene expression measurements.

Weight surrounding spots by Gaussian distance from each center spot.

Estimate a Gaussian-weighted gene–gene correlation matrix at each spot.

Remove genes with the smallest marginal contribution; choose a subset preserving the desired variation.

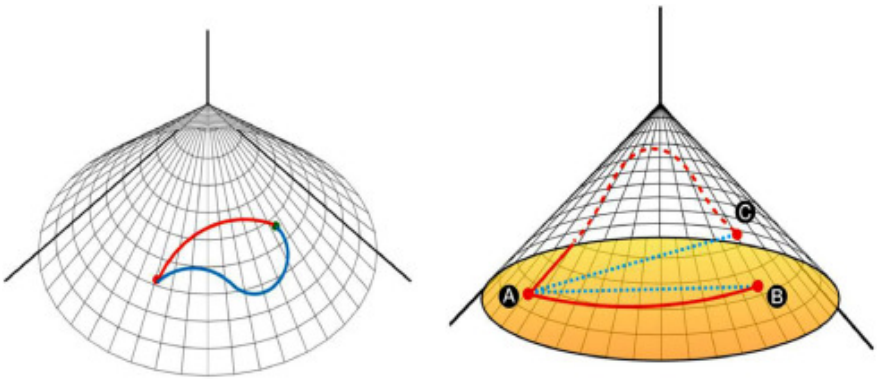
Apply PCA and K-means clustering, then map clusters onto the tissue.

Two Challenges

Challenge 1

- The measurement of interest is not univariate or a standard vector.
- Solution: Log-Euclidean (matrix-log).

$$A(s) = V(s)log[D(s)]V(s)^{-1}$$



Challenge 2

- Including / removing a gene operates on the associated row and column.
- Solution: structured gene selection through gene weights.

$$A(s) = \begin{array}{c} \\ j \end{array} \overset{\overset{j}{\downarrow}}{\begin{bmatrix} a_{11}(s) & \cdots & a_{1j}(s) & \cdots & & a_{1p}(s) \\ \vdots & \ddots & \vdots & & & \vdots \\ a_{j1}(s) & \cdots & a_{jj}(s) & \cdots & & a_{jp}(s) \\ \vdots & & \vdots & \ddots & & \vdots \\ & & & & & \\ a_{p1}(s) & \cdots & a_{pj}(s) & \cdots & & a_{pp}(s) \end{bmatrix}}$$

L_1 Regularized Objective Function and Algorithm

- Structured gene-selection through gene weights:

$$\mathbf{w} = (w_1, \ldots, w_p)^T, \qquad w_i \in [0,1]. \quad \mathbf{W} = \text{diag}(w).$$

- Objective function:

$$\max_{\mathbf{w}\in[0,1]^p} \left[\sum_{s=1}^{n} \left\| W^{1/2}(A(s) - \overline{A}) W^{1/2} \right\|_F^2 - \lambda \|w\|_1 \right].$$

- Algorithm: projected proximal gradient ascent, which alternates three steps:
  1. **gradient ascent**: update $\mathbf{w}$ to increase objective.
  2. **proximal** for L_1 penalty to encourage sparsity.
  3. **project** to keep each w_i between 0 and 1.

L_1–L_0 Equivalence

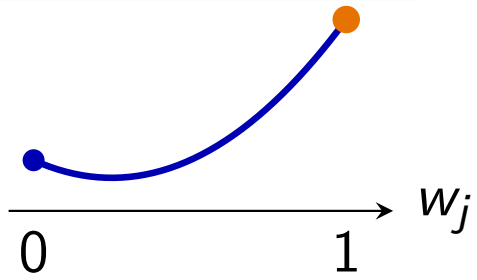
Theorem

The continuous L_1-penalized formulation has the same global maximum with the L_0-penalized subset-selection problem:

$$\max_{\mathbf{w}\in[0,1]^p} \left[\sum_{s=1}^{n} \left\| W^{1/2}(A(s)-\overline{A})W^{1/2} \right\|_F^2 - \lambda \|w\|_1 \right]$$

$$= \max_{\mathbf{w}\in\{0,1\}^p} \left[\sum_{s=1}^{n} \left\| W^{1/2}(A(s)-\overline{A})W^{1/2} \right\|_F^2 - \lambda \|w\|_0 \right]$$

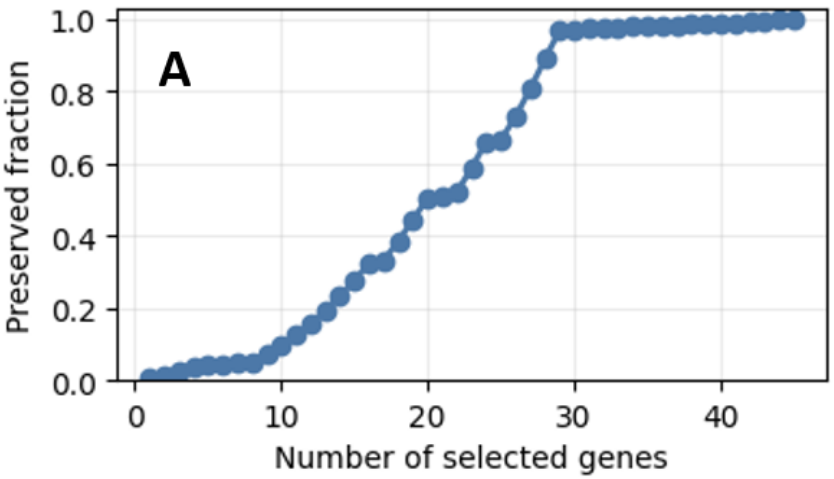
- The L_1 formulation has an exact subset-selection interpretation at the global optimum.

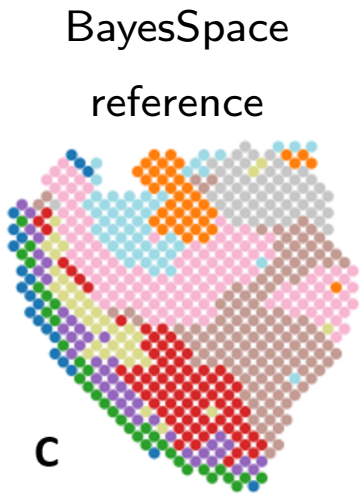


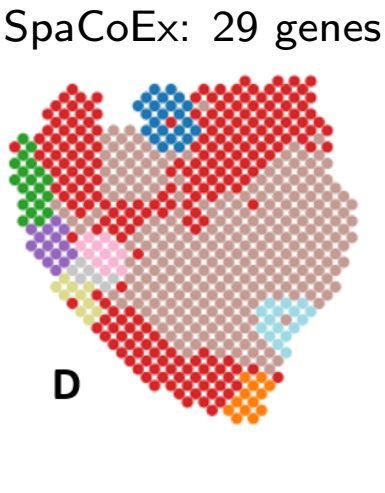
Case Study 1: Keratinocyte Gene Selection

45 candidates **29 retained**

96.74% preserved co-expression variation







Preserved variation refers to the method's representation, not classification accuracy.

Case Study 1: Cluster Analysis (10 clusters)

Representation	Spatial agreement	ARI
Co-expression only	0.746	0.149
Joint ($\delta = 0.25$)	0.549	**0.186**
Expression only	0.310	0.123
SpaGCN	0.387	0.095
stLearn	0.638	0.068
BayesSpace (baseline)	0.600	1.000

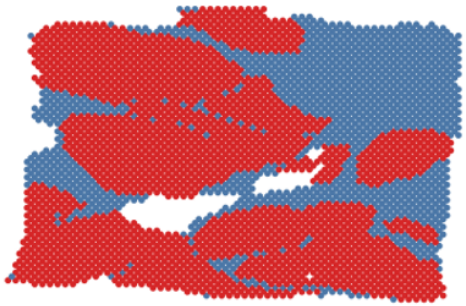
Table: Comparison with expression-only clustering and two existing spatial-domain methods, SpaGCN and stLearn. ARI measures agreement with the BayesSpace annotations; the reference ARI of 1 is a self-comparison.

Main Finding

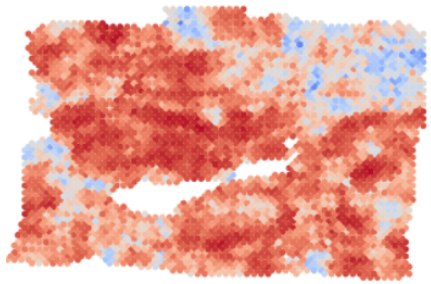
Joint SpaCoEx achieves the highest ARI, while co-expression-only SpaCoEx achieves the highest spatial agreement.

Case Study 2: Breast Cancer: B2M and HLA-C

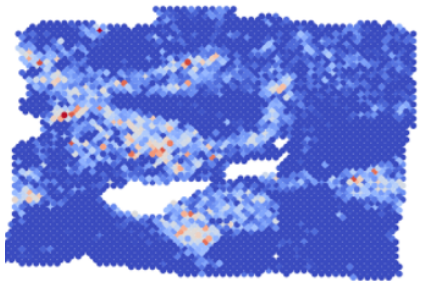
Annotated cancer-associated regions



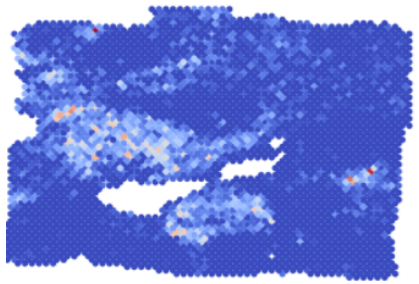
Local correlation



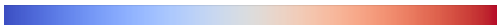
B2M expression



HLA-C expression



Annotation: red = cancer-associated; blue = other.



Blue to red: correlation -1 to 1; expression 0 to 100% of each gene's maximum.

Neither *B2M* nor *HLA-C* showed a statistically significant difference in mean expression between cancer and non-cancer regions. However, their local correlation was significantly higher in cancer regions ($P < 1.0 \times 10^{-10}$), with spatial dependence accounted for using Conley spatial heteroskedasticity- and autocorrelation-consistent standard errors.

Conclusions and Future Work

- SpaCoEx is a framework to identify genes whose co-expression varies across spatial locations.
- Co-expression features extracted can be combined with expression features for downstream analysis.
- I am currently revising the manuscript for *Briefings in Bioinformatics*. There are multiple updates motivated by the reviewers' constructive comments.
- Future work: validation across independent sections, multiple samples.

Acknowledgments and Questions

Thank you to my mentors:

Prof. Gil Alterovitz
Dr. Shaojun Pei

and to the **MIT PRIMES program**.

Questions?

Manuscript

Extra Slide: Notations

- The spatial variation of the correlation between gene i and gene j is denoted by:

$$S_{ij} = \sum_{s=1}^{n} (A(s)_{ij} - \overline{A}_{ij})^2.$$

- The objective function can be rewritten as

$$\max_{\{0 \leq w_i \leq 1\}} \left[\sum \sum w_i w_j S_{ij} - \lambda \sum w_i \right].$$

- Let $f(w)$ be the smooth part, i.e.,

$$f(w) = \sum \sum w_i w_j S_{ij}.$$

Extra Slide: Projected Proximal Gradient Ascent

- First, take a gradient ascent step on the smooth preservation term:

$$u^{(t+1)} = w^{(t)} + \eta \nabla f(w^{(t)}).$$

- Second, apply the proximal step for the L_1 penalty.

$$v^{(t+1)} = \max\left(0, u^{(t+1)} - \eta\lambda\right).$$

  This step shrinks each weight by an amount proportional to λ. If a weight becomes small enough, it is set exactly to zero, which removes the corresponding gene.

- Third, project the weights back into the allowed interval $[0, 1]$:

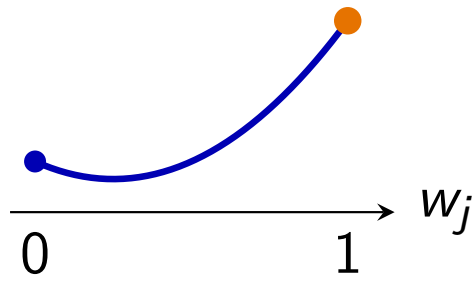
$$w^{(t+1)} = \min\left(1, v^{(t+1)}\right).$$

Equivalence Between the L_0 and L_1: Proof

For symmetric S with $S_{jj} \geq 0$, fixing all weights except $w_j = t$ gives

$$F(t) = S_{jj}t^2 + \left(2\sum_{b \neq j} S_{jb}w_b - \lambda\right) t + C.$$

1. Start from a global optimizer.
2. Since $F(t)$ is convex, choose $t = 0$ or $t = 1$ without decreasing the objective.
3. Repeat for every coordinate to obtain a binary global optimizer.



Conclusion

For binary weights, $\|w\|_1 = \|w\|_0$. Therefore, the two problems have the same global optimal value.

References: Data and Representation

- Ji, A. L., et al. (2020). Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell*, 182(2), 497–514. Sample GSM4284326.
- 10x Genomics (2020). Human Breast Cancer, Block A Section 1. Spatial Gene Expression dataset.
- Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. (2006). Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine*, 56(2), 411–421.
- Breast annotation source: SEDR analyses, BRCA1 metadata.

References: Comparison Methods and more

- Zhao, E., et al. (2021). Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 39(11), 1375–1384.
- Hu, J., et al. (2021). SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature Methods*, 18(11), 1342–1351.
- stLearn project documentation: `https://stlearn.readthedocs.io/`
- Conley, T. G. (1999). GMM estimation with cross sectional dependence. *Journal of Econometrics*, 92(1), 1–45.