Spatial omics feature representation using graph Fourier transform

Yuzhou Chang[1,2,†], Jixin Liu[3,†], Anjun Ma[1,2], Jordan Krull[1,2], Yang Liu[4,5], Rong Fan[4,5,6], Dong Xu[7], Zihai Li[2], Bingqiang Liu[3], and Qin Ma[1,2]

[1] Department of Biomedical Informatics, College of Medicine, Ohio State University, Columbus, OH 43210, USA. [2] Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA. [3] School of Mathematics, Shandong University, Jinan 250100, China. [4] Department of Biomedical Engineering, Yale University, New Haven, CT 06520, USA. [5] Yale Stem Cell Center and Yale Cancer Center, Yale School of Medicine, New Haven, CT 06520, USA. [6] Human and Translational Immunology Program, Yale School of Medicine, New Haven, CT 06520, USA. [7] Department of Electrical Engineering and Computer Science, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA. † These authors contributed equally. Corresponding author: Qin Ma: qin.ma@osumc.edu

Introduction

Background:

- Spatial omics (e.g., spatial transcriptome, proteome, epigenome, and histology) are transforming our understanding of cell or tissue biology in health and disease.
- Cell-centric analysis enables us to investigate the organization of spatial domain, relations of the cell-neighborhood, cell-cell communication, etc.
- Gene-centric analysis enables the discovery of spatially variable features, such as spatial variable genes (SVG).
- However, a quantitative and qualitative representation method of organized spatial pattern presented by diverse spatial omics features is still a gap for further gene-centric analysis.

Challenge:

- The complexity of construction due to multiple spatial omics platforms.
- Interpretability of representation embeddings.
- Scalability issue due to the pixel-level platform (e.g., Stereo-seq and CODEX).
- Non-trivial aggregation method of embedding graph topological structure and graph signal (e.g., gene expression).

Solution: A hypothesis-free graph Fourier transform framework, Spatial Graph Fourier Transform (**SpaGFT**), for spatial omics data feature representation.

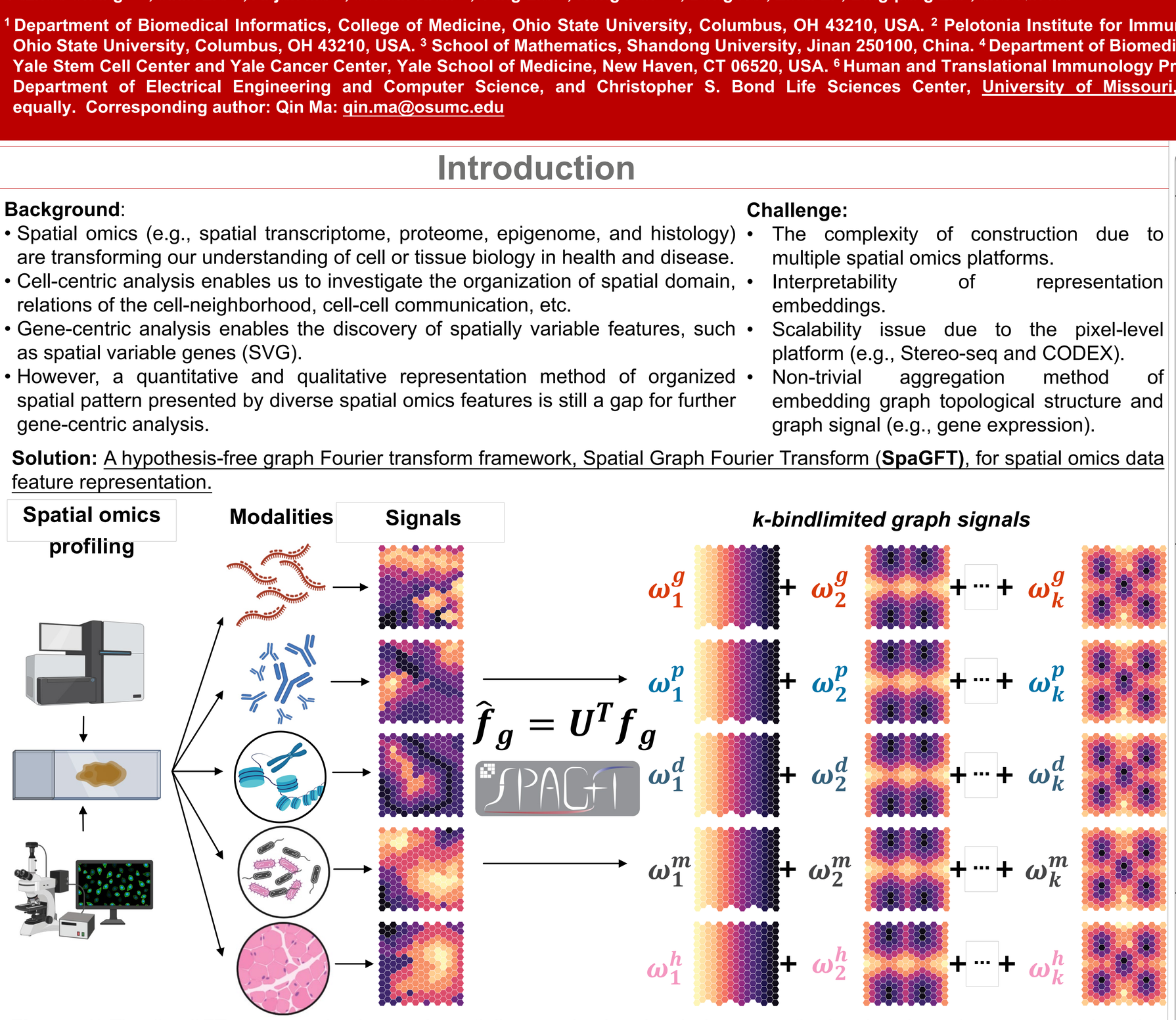


Figure. 1| The SpaGFT conceptual schema. A spatially organized molecule is a smooth signal and can be represented as the linear combination of k low-frequency Fourier mode (FMs), where a low-frequency FM contributes to a slow and smooth graph signal variation. Fourier coefficient (FC) can measure FMs contribution.

Results-SVG identification performance

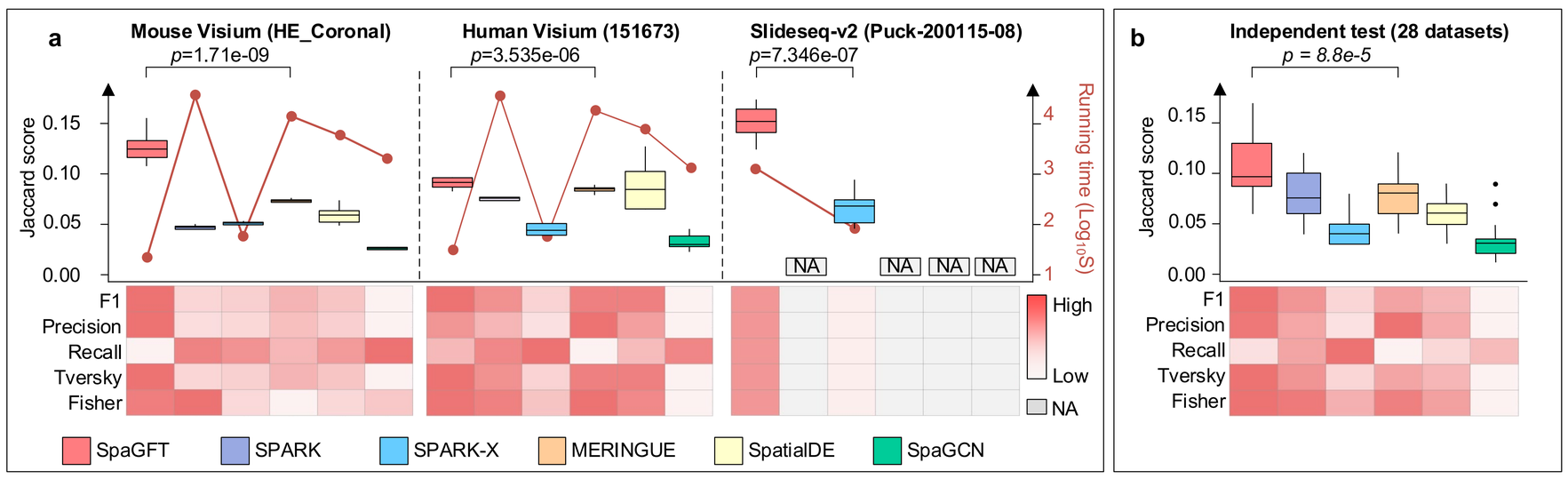


Figure. 2| a. The SVG prediction evaluation was compared to five benchmarking tools. The running time (log transformation seconds) of each tool is represented as red lines. **b.** After parameter selection (three high-quality datasets), the SVG prediction performance of SpaGFT on additional 28 independent datasets was compared to those of the five benchmark tools. **Conclusion: SpaGFT identifies SVG more accurately and faster.**

Results – graph signal representation and process

Figure. 3| a. SVGs identified by SpaGFT were distinguishably separated from non-SVGs on the FM-based UMAP with a clear boundary, whereas SVGs were irregularly distributed on the PC-based gene UMAP. **b.** SpaGFT can enhance and remove noise and outperformed other gene enhancement tools. **c.** SpaGFT enhanced signal and removed the noisy background for spatial omics platforms. **Conclusion: (1) FC is a transformed simple but informative topological features for representing complex structures with irregular topologies. (2) The FCs of low-frequency FMs will be enhanced and those of high-frequency FMs will be diminished.**

Graph signal transform

Step 1: K-nearest neighbor (KNN) Graph construction

(i) Define an undirected graph and adjacency matrix.

$$G = (V, E)$$

where $V = \{v_1, v_2, \ldots, v_n\}$ is the node set referring to n spots; E is the edge set defined by KNN.

(ii) An adjacent binary matrix $\boldsymbol{A} = (a_{ij})$ with rows and columns as n spots is defined as:

$$a_{ij} = \begin{cases} 1, & e_{ij} \in E \\ 0, & else. \end{cases}$$

(iii) A diagonal matrix $\boldsymbol{D} = diag(d_1, d_2, \ldots, d_n)$, where $d_i = \sum_{j=1}^{n} a_{ij}$ represents the degree of v_i

Step 2: Fourier mode calculation

(i) Using matrices $\boldsymbol{A}$ and $\boldsymbol{D}$, a Laplacian matrix $\boldsymbol{L}$ can be obtained by

$$\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$$

(ii) The Laplacian matrix $\boldsymbol{L}$ can be decomposed using spectral decomposition

$$\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathbf{T}}$$

$$\boldsymbol{\Lambda} = \boldsymbol{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

$$\boldsymbol{U} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_n),$$

where the diagonal elements of $\boldsymbol{\Lambda}$ are the eigenvalues of $\boldsymbol{L}$ with $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$, where λ_1 is always equal to 0 regardless of graph topology.

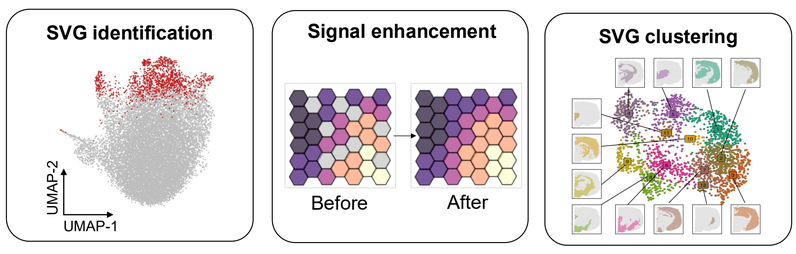
Step 3: Graph Fourier transform

The graph signal of a gene g is defined as $\boldsymbol{f}_g = (f_g^1, f_g^2, \ldots, f_g^n) \in \mathbb{R}^n$, which is a n-dimensional vector and represents the gene expression values across n spots. The graph signal $\boldsymbol{f}_g$ is transformed into a Fourier coefficient $\hat{\boldsymbol{f}}_g$ by

$$\hat{\boldsymbol{f}}_g = \boldsymbol{U}^T \boldsymbol{f}_g, \hat{\boldsymbol{f}}_g = (\hat{f}_g^1, \hat{f}_g^2, \ldots, \hat{f}_g^n)$$

$\hat{f}_g^k$ is the projection of $\boldsymbol{f}_g$ on FM $\boldsymbol{\mu}_k$, representing the contribution of FM $\boldsymbol{\mu}_k$ to graph signal $\boldsymbol{f}_g$, k is the index of $\boldsymbol{f}_g$ (e.g., $k = 1, 2, \ldots, n$). This Fourier transform harmonizes gene expression and its spatial distribution to represent gene g in the frequency domain.

Discussion



- Transformed signals (FCs) compressed graph topology and graph signal (e.g., gene expression).
- Low-frequency FCs represent features' spatial smooth pattern, leading to a new method for SVG identification.
- A low-pass filter enhances signal and removes noise.
- FCs can be used for other downstream tasks (e.g., SVG clustering to identify functional niches).

Link and acknowledgement

Paper QR **SpaGFT GitHub QR**

Funding: This work was supported by an NIH R01 award #R01GM131399 and #R35-GM126985; This work used Ohio Supercomputer Center (OSC) and the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation grant number ACI-1548562. This work was also supported by the Pelotonia Institute of Immuno-Oncology (PIIO).