bioRχiv

# sotk2: Interpretable module integration across bulk and spatial transcriptomics via program correlations

**Heewon Seo**
Snyder Institute for Chronic Diseases, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada
Correspondence should be addressed to H.S. (Heewon.Seo@ucalgary.ca)

**Gene expression programs (GEPs) inferred by deconvolution provide a compact representation of biological processes. However, comparing these programs across heterogeneous cohorts remains challenging due to platform effects, rank dependence, and cohort-specific signal. We present sotk2 (Spatial Omics Toolkit 2), a reproducible workflow that quantifies cross-cohort similarity among GEPs via correlation structure and a thresholded correlation network, outlines communities of related programs (putative biological modules), and provides diagnostic visualizations spanning GEP- and community-level summaries. sotk2 facilitates cross-dataset interpretation by organizing heterogeneous programs into shared, interpretable modules while distinguishing cohort-shared structure from cohort-specific patterns. By mapping user-derived programs onto these modules, sotk2 further enables annotation of new datasets against established signatures, supporting consistent biological interpretation across platforms and studies. Collectively, sotk2 offers an interpretable and portable framework for integrating deconvolution-derived programs across heterogeneous cohorts, enabling robust module-level comparison and signature-based annotation of new datasets.**

**Availability and implementation: sotk2 is open-source software written in R and distributed as an R package with an accompanying Shiny application for interactive exploration. Source code is freely available on GitHub at https://github.com/Snyder-Institute/sotk2, and the hosted Shiny app is available at https://shinyapps.ucalgary.ca/sotk2/.**

## Introduction

Non-negative matrix factorization[1,2] is widely used to extract interpretable transcriptional programs from gene expression profiles[3], subject to a non-negativity constraint that aligns with biological interpretations of expression magnitudes. However, selecting an appropriate rank and reconciling latent programs across ranks or datasets remains a persistent obstacle, particularly when integrating cohorts generated on different platforms, with varying tissue contexts, and with heterogeneous preprocessing pipelines. Building on the Spatial Omics Toolkit framework for rank selection and metagene network analysis, we previously enabled a systematic, transcriptome-wide evaluation of program structure in a neurodegenerative disease context, supporting the discovery of candidate biomarkers and vulnerability-associated signals[4].

Here, we extend this approach to the increasingly common setting in which modern transcriptomic and multi-omic studies generate multiple cohorts that differ in platform and experimental design. While deconvolution can recover coherent gene expression programs (GEPs; also known as metagenes) within each dataset, cross-cohort alignment is often performed ad hoc, making it difficult to determine whether putatively shared programs are robust or instead cohort-driven. sotk2 addresses this gap by representing all cohort-derived programs as nodes in a program correlation network and using community structure to summarize higher-order modules that are shared across cohorts or specific to individual cohorts. By consolidating deconvolution-derived programs into community networks and providing diagnostic visualizations and community-level summaries, sotk2 supports interpretable integration across heterogeneous datasets while explicitly separating shared structure from cohort-specific signal.

## Example datasets and reproducibility

To illustrate typical usage, we present a case study centered around three cohorts that exemplify common integration challenges: two bulk RNA-seq datasets (GLASS[5] and IVYGAP[6]) and one Visium v1 spatial transcriptomics dataset (HEILAND[7]) (**Fig. 1**). These cohorts vary in platform characteristics and data sparsity, which drives an approach that emphasizes cross-program similarity over direct sample-level harmonization. Programs were inferred separately within each cohort using consensus non-negative matrix factorization (cNMF) implemented in the *cnmf* Python package (v1.4)[8]. We applied the following filtering and regularization parameters: `removeAbove` = 10, `removeBelow` = 0, and `alpha` = 1; the cNMF `density_threshold` parameter was set to 2. For GLASS and IVYGAP expression profiles, we evaluated a rank grid spanning $k = 2$–60, whereas for HEILAND Visium v1 spot-level profiles—characterized by increased sparsity—we used a coarser grid ($k = 5, 10, 15, 20, 25, 30$). All remaining cNMF settings (including `n_iter` and `beta_loss`) were left at package defaults, and a fixed random seed (`seed` = 14) was used for reproducibility.

To facilitate reproducible exploration, the Shiny app distributes precomputed cNMF outputs across the defined rank grids and a bundled, precomputed sotk2 object for default rendering of reference figures. It is implemented to

avoid unnecessary recomputation by loading the precomputed object unless the user changes parameters that alter the analytical state (e.g., correlation method, correlation threshold, community detection algorithm, or layout rewiring weights). When recomputation is triggered, the app rebuilds only the required intermediate objects deterministically and reports status messages that distinguish precomputed views from results generated on demand.

vThe outputs of the workflow are encapsulated within an R object that preserves intermediate representations, such as correlation structures, metagene and community networks, community assignments, and mappings of cohort-related metadata. This object can be exported directly from the application in the form of an RDS file, thereby facilitating the archiving, sharing, and programmatic reanalysis of the exact analytical state utilized to generate figures. This design fosters a methodology-driven approach, where interactive exploration is employed to identify defensible parameter regimes. Subsequently, downstream analyses and reporting can be conducted based on the exported object, all while adhering to version-controlled practices for software and data snapshots.

## Features and implementation

sotk2 implements an interpretable, network-centric strategy for integrating deconvolution-derived GEPs across heterogeneous cohorts. The core design principle involves treating each program as a comparable unit of biological signal and modeling cross-cohort similarity through the correlation structure of program activity[9], rather than attempting direct one-to-one factor matching. Specifically, sotk2 begins by aggregating programs across cohorts and ranks (**Fig. 1A**), as provided by upstream deconvolution runs, and calculates pairwise correlations between program profiles. The resulting correlation-density summaries offer an initial quality-control perspective on within-cohort versus between-cohort similarity, enabling users to predict whether the integration will be influenced more by robust cross-links (indicating shared structure) or fragmented due to cohort-specific signals. This stage also facilitates the calibration of a correlation coefficient threshold utilized downstream to determine which program–program relationships will be retained.

With the thresholded correlations, sotk2 constructs a GEP correlation network wherein nodes represent programs, and edges denote retained correlations (**Fig. 1B**). This network acts as the primary integration framework: cohorts contribute nodes to a shared graph, and cross-cohort

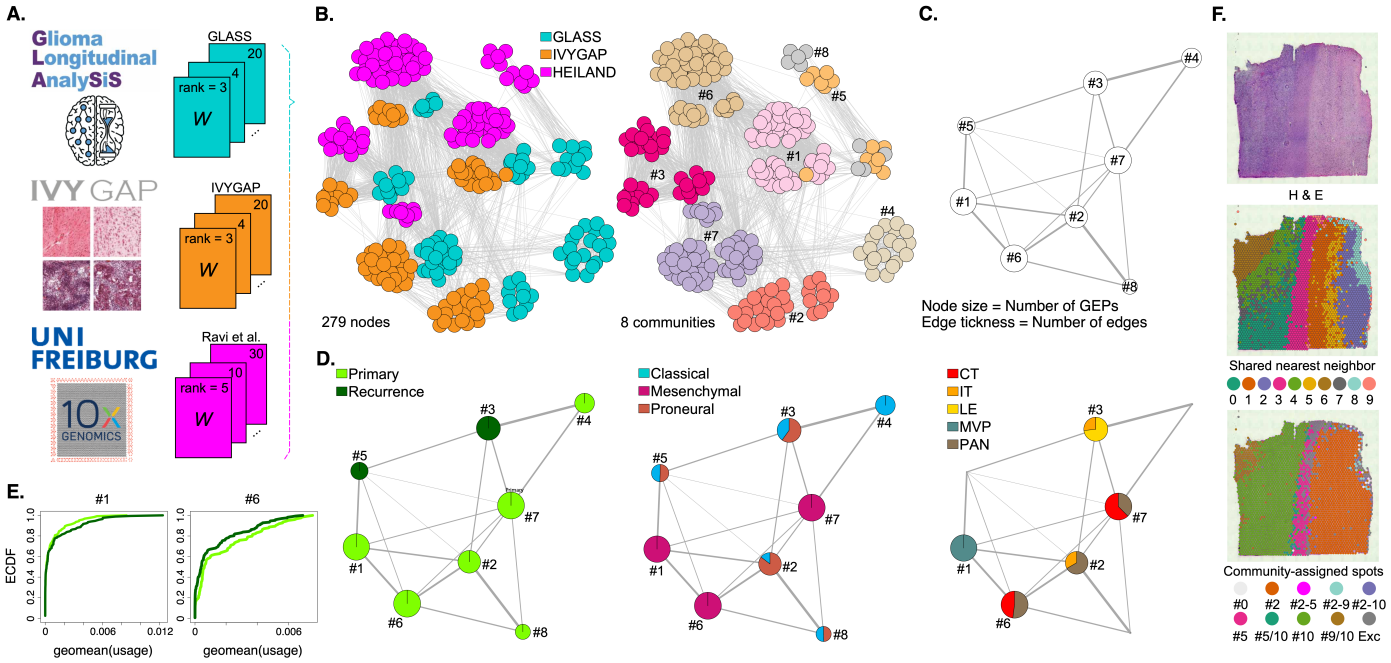


**Figure 1. Conceptual overview and case study for sotk2. (A)** Three cohorts with distinct platforms and resolutions were deconvolved using consensus non-negative matrix factorization to derive gene expression programs (GEPs). Program (*W*) matrices were used to compute cross-program correlations and to construct a GEP correlation network. **(B)** GEP-level correlation networks. Left: applying a correlation threshold of >0.3 retained 279 GEPs and positive correlations were represented as edges; node colors indicate cohort. Right: community detection (*fast greedy*) partitioned the network into eight communities; node colors indicate community membership. **(C)** Community-level abstraction of the GEP network, where each node represents a community and edges summarize inter-community connectivity. Node size is proportional to the number of GEPs assigned to the community, and edge thickness is proportional to the number of inter-community edges in the underlying GEP network. **(D)** Community-level networks with sample-annotation overlays shown as pie charts based on standardized chi-squared residuals (enrichment/depletion relative to expectation). Left: GLASS primary versus recurrence. Center: GLASS molecular subtype (Verhaak signatures). Right: IVYGAP anatomical features (CT, cellular tumor; IT, infiltrating tumor; LE, tumor's leading edge; MVP, microvascular proliferations; PAN, palisading cells around necrosis) **(E)** Community activity summaries for Communities #1 and #6 using the geometric mean of continuous usage values across community-assigned GEPs. Empirical cumulative distribution functions compare activity distributions between sample groups and provide complementary evidence to the annotation enrichments in (D). **(F)** Spatial transcriptomics example (10x Genomics Visium v1) from HEILAND. Top: H&E image of the UKF269_T glioblastoma section. Middle: Seurat shared nearest neighbor clusters overlaid on Visium spots. Bottom: sotk2 community labels projected onto spots; spots may map to multiple communities. Hashmarks (e.g., "#2") indicate the community identifier (here, Community 2); "#0" denotes spots with memberships spanning more than two communities, and "Exc" denotes spots that could not be assigned under the selected criteria.

integration is represented as edges connecting programs across datasets. Given that the threshold influences network sparsity and component structure, sotk2 regards this decision as critical to analysis. The Shiny application explicitly supports iterative exploration to identify conditions where the network is sufficiently connected for module discovery, while avoiding overly dense graphs that can be challenging to interpret and may reflect weak or noisy associations. Detailed guidance on parameter tuning and key diagnostic features is provided in the Shiny application's Interpretation tab.

Community detection is subsequently employed within the GEP network to encapsulate program relationships into cohesive modules. The software tool sotk2 facilitates a variety of community search algorithms, including but not limited to *fast greedy* methods, *Louvain*, *Leiden*, and *random-walk* based approaches. This design acknowledges the variability of community structure contingent upon algorithmic assumptions and the underlying graph topology. Notably, this flexibility is integrated as a user-selectable parameter, allowing for the recalibration of community assignments upon modification while maintaining visual parameters independently. The primary objective is to enable comparative analyses: users can ascertain the stability of modules across different methodologies and evaluate whether the resulting communities correspond to anticipated cohort mixing patterns indicative of shared biological processes.

In terms of visualization and interpretability, sotk2 provides a range of complementary network representations using the *igraph* R package[10]. An unweighted layout allows for the examination of the retained edge set and overall connectivity without the influence of edge weights, which is instrumental in validating that the correlation threshold produces a credible topology. Conversely, a weighted layout view is employed to highlight stronger relationships and enhance module differentiation in two-dimensional layouts, yielding figures suitable for presentation. Within the context of this weighted layout, sotk2 incorporates adjustable rewiring weights expressly designed to augment layout interpretability—such as minimizing edge crossings and stabilizing module separation—without influencing the community detection processes. This intentional distinction between structural decisions (e.g., thresholding and community search methodology) and visualization-specific controls is a carefully considered design choice aimed at mitigating the risk of inadvertently altering biological interpretations during aesthetic adjustments.

Beyond the GEP-level graphical representations, the sotk2 framework comprehensively summarizes results at the community level by developing a community network (**Fig. 1C**). In this network, each node symbolizes a community, while edges encapsulate inter-community connectivity. The size of each node is indicative of the number of programs assigned to a community, and the thickness of the edges represents the degree of inter-community connectivity, such as the number of cross-community edges derived from the foundational metagene network. This abstraction facilitates a compact representation of the relational dynamics among modules, thereby enabling the prioritization of communities for subsequent interpretative analyses.

When cohort-specific sample metadata is available, sotk2 enhances its analytical capacity by incorporating community annotation overlays that draw connections between module structures and biological strata (**Fig. 1D**). For each cohort, programs that are allocated to a community contribute corresponding samples under a maximum-usage assignment heuristic. The category composition is then articulated utilizing standardized chi-squared residuals, which reflect enrichment and depletion relative to expected values. These residual-based pie overlays offer a structured and comparable summary across communities, enabling the identification of phenotype-associated modules (e.g., subtype-specific or condition-specific communities), while maintaining a clear methodological interpretation of "enrichment" as a deviation from a contingency-table expectation. In instances where curated annotations are unavailable, sotk2's community structure still facilitates inference by transferring labels from well-annotated cohorts through shared community membership, thus providing a pragmatic approach to contextualizing programs in new datasets.

Additionally, sotk2 implements a complementary community activity diagnostic predicated on continuous usage magnitudes. For each selected cohort and community, sotk2 aggregates per-sample program usage values across all programs associated with that community by employing the geometric mean (**Fig. 1E**). This approach is robust against scale disparities and mitigates the influence of single extreme values while accurately reflecting consistent activity across programs. The resulting distribution of community activity is summarized via empirical cumulative distribution functions stratified by available sample annotations. This analytical perspective is designed to complement the residual-based community annotation methodology: a community may exhibit enrichment for a category based on membership counts, yet reveal an inverse activity shift when continuous usage is evaluated, thereby highlighting competition or mixing phenomena that discrete assignments may obscure. The integrated design of both enrichment-based and activity-based summaries is intended to fortify inference by necessitating consistency across independent diagnostics before advancing robust biological claims.

Lastly, the inferred biological modules can be overlayed onto and systematically compared with existing clustering frameworks. For instance, we projected the community assignments generated by sotk2 onto Visium spots (**Fig. 1F**), subsequently analyzing the congruence of these module labels with the Shared Nearest Neighbor (SNN) clusters computed using Seurat[11], a prominent methodology for unsupervised spatial clustering. While SNN primarily delineates cluster architecture, the biological interpretation of such clusters typically necessitates a subsequent annotation step[12] to bestow meaningful labels upon the clusters or spots. In contrast, the sotk2 framework facilitates direct, cohort-informed interpretation by utilizing module definitions

derived from multiple cohorts with well-established annotations. This approach not only enables more consistent annotation of spatial spots but also promotes comparability across diverse platforms.

## Discussion

sotk2 addresses a recurring challenge in integrative transcriptomics: reconciling deconvolution-derived programs across cohorts generated on different platforms, with varying signal-to-noise ratios and cohort-specific biological context. Rather than attempting direct factor matching by ad hoc heuristics, sotk2 places all programs in a shared correlation network and uses community structure to organize them into shared, interpretable modules while preserving the ability to detect cohort-specific structure. This design supports cross-dataset interpretation by enabling users to assess whether apparent similarities reflect broadly shared biology or are driven by cohort-specific connectivity patterns.

A key practical contribution of sotk2 is the availability of complementary community summaries that emphasize different aspects of evidence. Community-level annotation based on standardized residuals can highlight category enrichment using discrete membership counts. In contrast, usage-based community activity summaries capture continuous magnitude differences and can reveal competition or mixing between strata that membership heuristics may obscure. Together, these views encourage robust interpretation by requiring concordance between annotation enrichment and activity shifts, rather than relying on a single diagnostic. sotk2 is therefore well-suited for settings where users seek portable, interpretable modules that generalize across cohorts, as well as settings where identifying cohort-restricted programs is essential for understanding technical limitations or context-specific biology.

In conclusion, sotk2 provides an interpretable and portable framework for integrating deconvolution-derived programs across heterogeneous cohorts via correlation network organization and community-level summaries. By separating cohort-shared structure from cohort-specific signals and supporting module-level comparison and signature-based annotation of new datasets, sotk2 facilitates consistent biological interpretation across platforms and studies.

## References


1. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999)
2. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010)
3. Brunet, J.P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci.* **101**, 4164–4169 (2004)
4. Seo, H. et al. Dual platform spatial transcriptomics reveals metagenes associated with vulnerable parvalbumin interneurons in female 5xFAD mice. *in revision* (2026)
5. Varn, F. S. et al. Glioma progression is shaped by genetic evolution and microenvironment interactions. *Cell* **185**, 2184-2199.e16 (2022)
6. Puchalski, R. B. et al. An anatomic transcriptional atlas of human glioblastoma. *Science* **360**, 660–663 (2018)
7. Ravi, V. M. et al. Spatially resolved multi-omics deciphers bidirectional tumor-host interdependence in glioblastoma. *Cancer Cell* **40**, 639-655.e13 (2022)
8. Kotliar, D. et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* **8**, e43803 (2019)
9. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10**, 1108–1115 (2013)
10. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Systems**, 1695 (2006)
11. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021)
12. Maynard, K. R. et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat. Neurosci.* **24**, 425–436 (2021)


## Acknowledgements

Support from the Snyder Institute for Chronic Diseases at the University of Calgary is gratefully acknowledged. The Institute's interdisciplinary research environment contributed to the development and evaluation of sotk2.

## Author contributions

HS conceived the project and wrote the manuscript.

## Competing interest statement

The author declares no competing interests.

## Preprint Information

Manuscript written on December 29, 2025. It will be submitted to bioRxiv once the citation for Reference 4, which describes the SOTK methodology, becomes available. This version has not been peer reviewed.