Bioinformatics, 2026, **42**(5), btag301
https://doi.org/10.1093/bioinformatics/btag301
Published: 14 May 2026
Applications Note | Genome analysis

OXFORD

Yomix: an interactive tool for the exploration of low-dimensional embeddings in omics data

Nicolas Perrin-Gilbert[1,*], Nisma Amjad[1], Pierre Fumeron[2,3], Silvia Tulli[1], Joshua J. Waterfall[2,3,*]

[1]Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique (ISIR), Paris, F-75005, France
[2]Inserm U1330, Institut Curie Centre de Recherche, PSL University, Paris, France
[3]Translational Research Department, Institut Curie Centre de Recherche, PSL University, Paris, France
*Corresponding authors. Nicolas Perrin-Gilbert, Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique (ISIR), Paris, F-75005, France. E-mail: nicolas.perrin@sorbonne-universite.fr; Joshua J. Waterfall, Inserm U1330, Institut Curie Centre de Recherche, PSL University, Paris, France; Translational Research Department, Institut Curie Centre de Recherche, PSL University, Paris, France. E-mail: joshua.waterfall@curie.fr
Associate Editor: Pier Luigi Martelli

Abstract

Summary: **In the analysis of diverse omics data, a common and important preliminary step involves computing low-dimensional embeddings using techniques such as PCA, UMAP, t-SNE, or variational autoencoders. These embeddings provide a global overview of sample distributions and their relationships, often serving as the basis for formulating biological hypotheses. To facilitate rapid and intuitive exploration of such low-dimensional embeddings, we developed Yomix, an interactive omics-agnostic visualization and data exploration tool. Yomix enables users to flexibly define subsets of interest using a lasso selection tool, instantly compute their feature signatures, and compare their distributions. Yomix is a fast and efficient tool for interactive exploration of diverse omics datasets.**

Availability and implementation: **Yomix and its documentation are publicly available at https://github.com/perrin-isir/yomix.**

1 Introduction

The high-dimensional nature of omics datasets presents substantial challenges in data interpretation. To facilitate visualization and exploration of such datasets, a widely adopted preliminary step is the computation of low-dimensional embeddings, typically in two or three dimensions (Becht et al. 2019, Moon et al. 2019). Dimensionality reduction techniques, such as Principal Component analysis (PCA), Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018), t-distributed Stochastic Neighbour Embedding (t-SNE) (van der Maaten 2009), and variational autoencoders (VAEs) (Lopez et al. 2018), are frequently employed for this purpose, offering advantages in capturing both global and local data features. These low-dimensional representations serve as a basis for deriving biological insights, enabling researchers to identify clusters (Kiselev et al. 2019), infer developmental trajectories (Saelens et al. 2019), and formulate hypotheses regarding underlying biological processes (Trapnell et al. 2014).

Despite the widespread adoption of dimensionality reduction techniques in omics workflows, effectively interpreting the resulting low-dimensional embeddings remains a significant challenge. In practice, researchers must frequently explore these embeddings iteratively, zooming into local regions, selecting subpopulations of interest, and querying their molecular signatures. This type of visual exploration is essential for identifying novel cell types, validating marker features (genes, CpG sites, k-mers, etc.). However, most existing analysis frameworks (e.g. Seurat (Hao et al. 2021), Scanpy (Wolf et al. 2018)) rely on static visualizations that are not directly responsive to user interaction or dynamic filtering. This restricts the user's ability to adjust views, subset groups, or overlay metadata in real time, often requiring time-consuming re-computation or manual scripting for even basic exploratory tasks. While recent tools have been developed to address some of these issues (Li et al. 2020, Kotliar and Colubri 2021, Keller et al. 2025) there remains a significant need for fast, flexible approaches to interact with diverse types of omics datasets.

Yomix addresses these challenges through several capabilities not combined in existing tools: a lasso selection tool enabling free-form cluster definition directly on any embedding, with advanced options such as intersection, union and difference, native support for 3D embeddings and user-controlled axis selection across higher-dimensional spaces, interactive annotation capability, and sub-second signature computation across diverse omics data types. A detailed comparison of Yomix with existing interactive visualization frameworks: including scX (Waichman et al. 2024), iSEE (Rue-Albrecht et al. 2018),

Received: 5 December 2025. **Revised:** 15 April 2026. **Accepted:** 1 May 2026

CELLxGENE (Abdulla et al. 2025), ShinyCell (Ouyang et al. 2021), ASAP (David et al. 2020), SEQUIN (Weber et al. 2023), and SCHNAPPS (Jagla et al. 2021), is provided in Supplementary Fig. 1, available as supplementary data at *Bioinformatics* online, highlighting the specific exploratory workflows enabled by Yomix.

To address these challenges, we developed Yomix, a lightweight yet powerful visualization framework that supports interactive projection views, metadata integration, and feature ranking across a variety of omics data types.

Materials and methods

Yomix is a lightweight tool used to identify biologically relevant structures and discriminative features across multiple datasets. It is developed in Python and uses the Bokeh framework, which generates a JavaScript-based graphical interface and callbacks that run in the web browser. It requires an anndata object (Virshup et al. 2024) saved in .h5ad format as input which contains a data matrix and at least one low-dimensional embedding.

By launching Yomix from the Python session, it directly opens an interactive plot of the embedding chosen by the user (Fig. 1A–E). Users can switch between different embeddings, change plot sizes and labels, and define specific subsets (Fig. 1A). These subsets can be selected using the legend and predefined annotations (Fig. 1B) or manually with a lasso tool (Fig. 1C), making it highly interactive. Yomix is both omics agnostic and handles three-dimensional (3D) embeddings. It also supports embeddings with more than 3 dimensions by letting users pick amongst dimensions for every axis.

Importantly, Yomix is not restricted to t-SNE or UMAP: it supports any low-dimensional embedding method, such as PCA, VAE (Lopez et al. 2018) or diffusion maps (Moon et al. 2019). Users should be aware that low-dimensional embeddings inevitably distort geometries, and structures visible in reduced dimensions, although often relevant, can sometimes be misleading. But there are tools that specifically try to avoid as much as possible those distortions, such as Γ-VAEs (Kim et al. 2024) which regularize geometric curvature. Lasso selections reflect the displayed layout rather than the full high-dimensional structure. The higher the embedding dimensionality, the lower the geometry distortion, so when possible we recommend relying on 3D embeddings. Yomix is fully functional with 3D and its original 3D lasso selection enables users to inspect data just as they would with 2D embeddings. Yomix also supports higher-dimensional embeddings, allowing users to interactively select 3 axes at a time to observe the data from many different perspectives.

Yomix can identify signatures for a user-defined cluster against the remaining data or between two user-defined subsets (Fig. 1D). These signatures, or marker features, are computed using a rapid, two-step filtering process that identifies the most discriminative features.

The algorithm first computes group statistics for each feature, calculating mean and standard deviation for Group A and Group B (either a specific subset or the remaining samples). It then calculates the Wasserstein distance between distributions for each feature, ranks all features in descending order by Wasserstein distance, and selects the features with the highest distributional differences. Wasserstein distances are computed using the closed-form solution for Gaussian distributions parameterized by the empirical mean and standard deviation of each group, and all operations are vectorized using NumPy, enabling sub-second runtimes even on large datasets. The second step identifies the features among the pre-filtered set that most reliably distinguish between groups using optimal binary classification thresholds. Specifically, for each feature, the algorithm tests multiple thresholds across the range of feature values, computes the confusion matrix for each threshold, calculates the Matthews Correlation Coefficient (MCC) at each threshold, and selects the one that maximizes MCC. The computation is vectorized across all features and thresholds simultaneously for efficient computation. A key advantage of this approach is its speed and accuracy, as filtering reduces the search space, and vectorized operations accelerate the search of optimal thresholds. More details on this implementation are available in Yomix documentation https://perrin-isir.github.io/yomix/

In addition, Yomix can compute oriented signatures by letting the user select a subset and draw an arrow in the embedding space. Each element of the subset is projected onto the arrow, creating a vector of positions along the arrow. For every feature, the algorithm computes the Pearson correlation between the feature values and the projection vector. It then returns the top features sorted by highest correlation. Selected features, either from signatures or input by the user, are visualized separately as either violin plots or heatmaps (Fig. 1E). Signatures display the top 100 features, providing a sufficiently large feature set for downstream analyses such as pathway enrichment or transcription factor activity inference. Analyses were run on a Dell desktop computer equipped with 32.0 GiB of RAM and a 13th Gen Intel® Core™ i9-13950HX processor.

Results

To evaluate the predictive and computational performance of Yomix, we compared it with three widely used feature selection methods: *t* test, Wilcoxon test, and COSG, which was shown to perform well in a recent benchmarking study (Pullin and McCarthy 2024). COSG is a feature selection method that evaluates features by their geometry in sample space (Dai et al. 2022). For each group, it constructs an ideal feature, one that is expressed exclusively in the cluster of interest and not in any other sample. It then computes the cosine similarity between every feature vector and this ideal vector, returning features with the smallest angular distance to the ideal gene vector of the target group and the largest angular distance to the ideal vectors of other groups.

Benchmarks were performed on five diverse omics datasets that include bulk RNAseq from The Cancer Genome Atlas (TCGA) pan-cancer analysis (Hoadley et al. 2018) (10 541 samples, 8000 features), DNA methylation array profiling of over 1000 sarcomas (Koelsche et al. 2021), single cell multiomics of cord blood (Stoeckius et al. 2017), scRNAseq of peripheral blood mononuclear cells (PBMCs) (Hansen et al. 2023) and proteomics from the National Cancer Institute's Clinical Proteomic Tumor Analysis

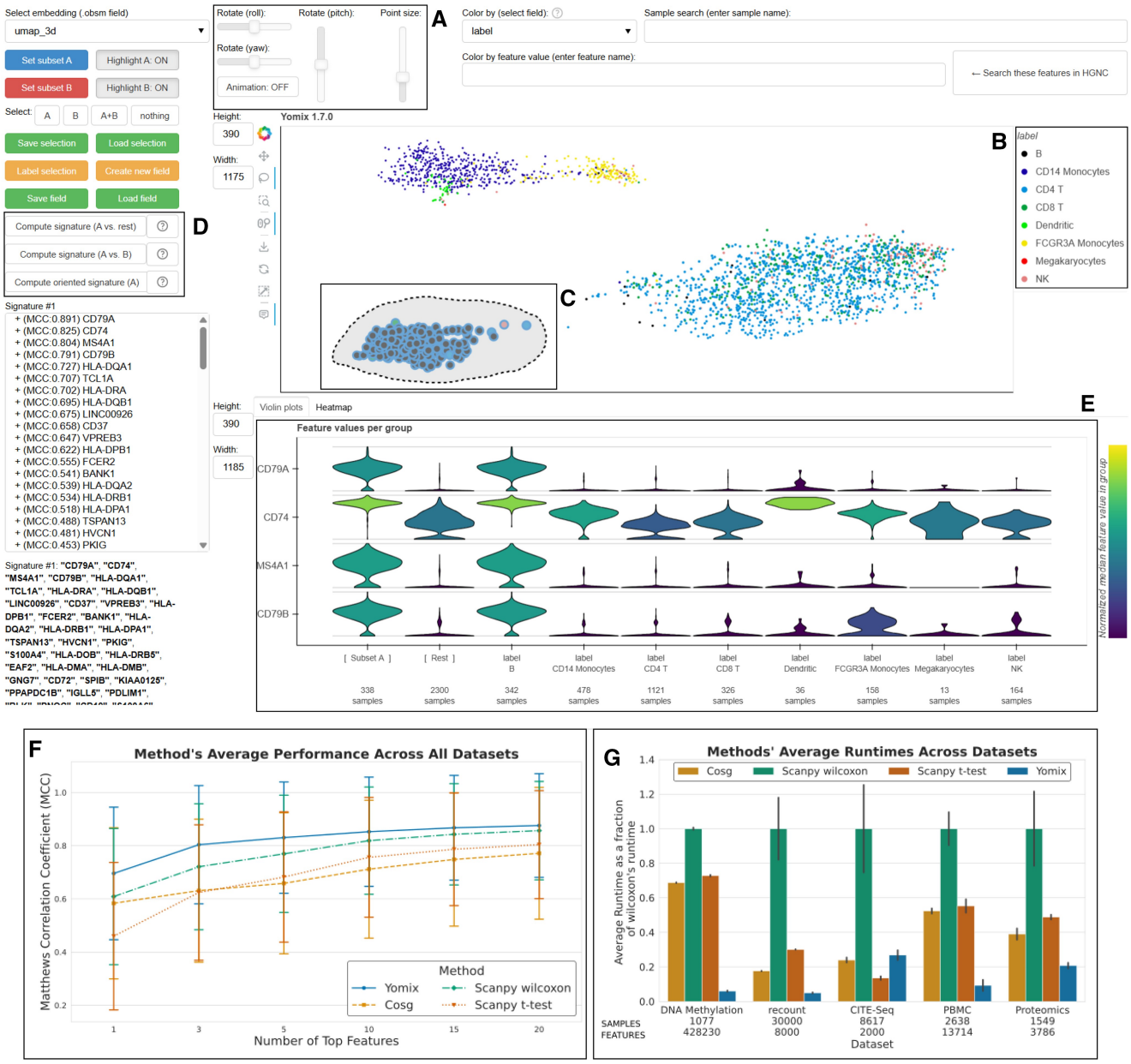

Figure 1 Screenshot of Yomix's interface displaying the PBMC dataset after computing the signature of the bottom-left cluster, mostly composed of B cells. (A) Buttons to control real-time 3D rotation, and to interact in 3D with the scatter plot. (B) Users can click on legend to select data. (C) Lasso tool for manual selection of data points. (D) Users can compute different types of signatures. (E) Violin plots to visualize the distribution of selected features for selected subgroups. Users can switch to a heat map format. (F) Average performance comparison between SVM trained on features selected by different methods, error bars represent standard deviation. G) Runtime comparison between the different methods, error bars represent standard deviation.

Consortium (CPTAC) (Li et al. 2023). Dataset sizes and feature counts are detailed in Fig. 1.

Preprocessing

We used the raw proteomics data. Recount and CITE-seq were log-normalized and subset to the 8000 and 2000 highly variable genes respectively. PBMC was log-normalized as well, but all the available genes were kept for analysis. Regarding the DNA methylation dataset, we followed the procedure described in the original publication (Koelsche et al. 2021).

To evaluate the utility of features selected by the various methods, we computed signatures for one cluster against the remaining samples (A vs Rest) using the 4 different methods (Yomix, *t* test, Wilcoxon test, and COSG). This task was repeated for every cluster for every dataset. In Recount, the clusters are the projects if the samples came from TCGA or the tissues of origin if they came from GTEX. For PBMC and CITE-seq, the goal is to predict the cell type and activation state. In DNA methylation and proteomics datasets, the aim is to predict the cancer type. We selected an increasing number of top features (from 1 to 20) from each method and measured the average MCC using

support vector machines (SVM) as the classifier with default parameters on these features across all datasets. Each classifier is trained on 70% of the dataset, the remaining 30% are used as a test set. The train and test are randomly sampled and the frequencies of the two classes (A and Rest) are preserved. We assessed feature signature quality using MCC scores on the test set from the SVM classification. Benchmarking code is available at https://github.com/PierreFumeron/yomix_benchmark.

The results shown in Fig. 1 demonstrate that Yomix consistently selects feature signatures with the highest predictive power (Fig. 1F). Yomix achieved average MCC scores comparable to or exceeding those of other methods across nearly all feature set sizes, underscoring its ability to select discriminative and biologically informative markers. Wilcoxon was the second best performing method, while COSG and *t* test yielded feature sets with lower MCC scores. All methods showed improved performance when more features were included. Similar results were obtained using KNN and Random Forest classifiers as shown in Supplementary Fig. 2, available as supplementary data at *Bioinformatics* online, confirming that these findings are not specific to the choice of classification model.

As a concrete biological illustration, Fig. 1 displays the PBMC scRNAseq dataset with Yomix's signature computed for the B cell cluster (lasso-selected in panel C) against all remaining cells. The top-ranked features include well-established B cell markers such as CD79A, CD74, MS4A1, and CD79B, whose preferential expression in B cells relative to all other cell populations is clearly visible in the violin plots (Fig. 1E).

In addition to predictive performance, speed was also computed for each method. The shortest runtimes in all datasets were also attained by Yomix, with execution times remaining below 0.7 seconds even for the largest datasets (Fig. 1G). When analyzing the TCGA bulk RNAseq dataset, which has 10 541 samples and 8000 features, the signature computation was completed in approximately 0.165 seconds. In contrast, the Wilcoxon method exhibited the longest runtime, 5.89 seconds on methylation and 2.99 seconds on TCGA datasets. COSG and *t* test methods displayed intermediate performance as they outperformed Wilcoxon but remained slower than Yomix across all datasets. The results demonstrate the significant computational efficiency of Yomix. These findings highlight the scalability and speed of Yomix's two-step algorithm, making it particularly well-suited for the rapid, iterative analysis required in large-scale omics exploration.

This analysis confirms that Yomix not only offers superior computational speed and interactivity but also excels at identifying feature sets with notable biological relevance and predictive capability, regardless of the type of dataset used and the size of samples required.

To further benchmark Yomix in the context of data types for which specialized methods exist, we performed two additional comparisons. For bulk RNA-seq, we compared Yomix against pyDESeq2 on a randomly subsampled subset of 50% of the TCGA dataset (5272 samples, 8000 highly variable genes), performing differential analysis for each of the 33 TCGA project types against the remaining samples. Results are shown in Supplementary Fig. 3, available as supplementary data at *Bioinformatics* online; while pyDESeq2 required an average of 6 minutes per project comparison, Yomix completed the equivalent task in under one second. For proteomics, we compared Yomix against ProMS (Shi et al. 2021), a method specifically designed for protein biomarker discovery. As shown in Supplementary Fig. 4, available as supplementary data at *Bioinformatics* online, the two methods yield comparable feature selection performance on the CPTAC proteomics dataset.

Taken together, these results demonstrate that Yomix offers strong computational efficiency, seamless interactivity, and feature selection performance that is competitive with established methods across diverse omics data types. While the benchmarking presented here covers five datasets spanning five data types, we consider these representative use cases rather than exhaustive proof of universal superiority; the primary strengths of Yomix lie in its speed and the interactive workflows it uniquely enables.

Author contributions

Nicolas Perrin-Gilbert (Conceptualization [Lead], Formal analysis [Lead], Investigation [Lead], Methodology [Lead], Project administration [Lead], Resources [Lead], Software [Lead], Supervision [Lead], Validation [Lead], Writing—original draft [Equal], Writing—review & editing [Equal]), Nisma Amjad (Investigation [Equal], Software [Equal], Visualization [Equal], Writing—original draft [Equal]), Pierre Fumeron (Investigation [Equal], Software [Equal], Visualization [Equal], Writing—original draft [Equal], Writing—review & editing [Equal]), Silvia Tulli (Software [Equal], Visualization [Equal], Writing—review & editing [Equal]), and Joshua J. Waterfall (Conceptualization [Lead], Formal analysis [Lead], Investigation [Lead], Methodology [Lead], Project administration [Lead], Software [Supporting], Supervision [Lead], Validation [Lead], Writing—original draft [Equal], Writing—review & editing [Equal])

Supplementary data

Supplementary material is available at *Bioinformatics* online.

Conflict of interests

None declared.

Funding

This work was supported by ITMO Cancer of Aviesan within the framework of the 2021–2030 Cancer Control Strategy, on funds administered by Inserm [22CM060-00] and by the Fight Kids Cancer and St Baldrick's Foundation Arceci Innovation Award to JJW.

Data availability

Yomix and its documentation are publicly available at https://github.com/perrin-isir/yomix. Benchmarking code is available at https://github.com/PierreFumeron/yomix_benchmark. scRNAseq

of peripheral blood mononuclear cells (PBMCs) is available at https://bioconductor.org/packages/TENxPBMCData/.

References

Abdulla S, Aevermann B, Assis P *et al.* CZ cellxgene discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res* 2025;**53**: D886–900. https://doi.org/10.1093/nar/gkae1142

Becht E, McInnes L, Healy J *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* 2019;**37**: 38–44. https://doi.org/10.1038/nbt.4314

Dai M, Pei X, Wang X-J. Accurate and fast cell marker gene identification with COSG. *Brief Bioinform* 2022;**23**:bbab579. https://doi.org/10.1093/bib/bbab579

David FPA, Litovchenko M, Deplancke B *et al.* ASAP 2020 update: an open, scalable and interactive web-based portal for (single-cell) omics analyses. *Nucleic Acids Res* 2020;**48**: W403–W414. https://doi.org/10.1093/nar/gkaa412

Hansen KD, Risso D, Hicks S. TENxPBMCData: PBMC data from 10X Genomics, 2023. Bioconductor package.

Hao Y, Hao S, Andersen-Nissen E *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–87.e29. https://doi.org/10.1016/j.cell.2021.04.048

Hoadley KA, Yau C, Hinoue T, *et al.* Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 2018;**173**:291–304.e6. https://doi.org/10.1016/j.cell.2018.03.022

Jagla B, Libri V, Chica C *et al.* SCHNAPPs – single cell sHiNy APPlication(s). *J Immunol Methods* 2021;**499**:113176. https://doi.org/10.1016/j.jim.2021.113176

Keller MS, Gold I, McCallum C *et al.* Vitessce: integrative visualization of multimodal and spatially resolved single-cell data. *Nat Methods* 2025;**22**:63–7. https://doi.org/10.1038/s41592-024-02436-x

Kim JZ, Perrin-Gilbert N, Narmanli E *et al.* γ-VAE: Curvature regularized variational autoencoders for uncovering emergent low dimensional geometric structure in high dimensional data. *arXiv preprint*2024. https://doi.org/10.48550/arXiv.2403.01078, preprint: not peer reviewed.

Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat Rev Genet* 2019;**20**: 273–82. https://doi.org/10.1038/s41576-018-0088-9

Koelsche C, Schrimpf D, Stichel D *et al.* Sarcoma classification by DNA methylation profiling. *Nat Commun* 2021;**12**:498. https://doi.org/10.1038/s41467-020-20603-4

Kotliar D, Colubri A. Sciviewer enables interactive visual interrogation of single-cell RNA-seq data from the python programming environment. *Bioinformatics* 2021;**37**:3961–3. https://doi.org/10.1093/bioinformatics/btab689

Li Y, Dou Y, Da Veiga Leprevost F, *et al.* Proteogenomic data and resources for pan-cancer analysis. *Cancer Cell* 2023;**41**: 1397–406. https://doi.org/10.1016/j.ccell.2023.06.009

Li B, Gould J, Yang Y *et al.* Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nat Methods* 2020;**17**:793–8. https://doi.org/10.1038/s41592-020-0905-x

Lopez R, Regier J, Cole MB *et al.* Deep generative modeling for single-cell transcriptomics. *Nat Methods* 2018;**15**:1053–8. https://doi.org/10.1038/s41592-018-0229-2

McInnes L, Healy J, Saul N *et al.* UMAP: uniform manifold approximation and projection. *JOSS* 2018;**3**:861. https://doi.org/10.21105/joss.00861

Moon KR, van Dijk D, Wang Z *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nat Biotechnol* 2019;**37**:1482–92. https://doi.org/10.1038/s41587-019-0336-3

Ouyang JF, Kamaraj US, Cao EY *et al.* ShinyCell: simple and sharable visualisation of single-cell gene expression data. *Bioinformatics* 2021;**37**:3374–6. https://doi.org/10.1093/bioinformatics/btab209

Pullin JM, McCarthy DJ. A comparison of marker gene selection methods for single-cell RNA sequencing data. *Genome Biol* 2024;**25**:56. https://doi.org/10.1186/s13059-024-03183-0

Rue-Albrecht K, Marini F, Soneson C *et al.* iSEE: interactive SummarizedExperiment explorer. *F1000Res* 2018;**7**:741. https://doi.org/10.12688/f1000research.14966.1

Saelens W, Cannoodt R, Todorov H *et al.* A comparison of single-cell trajectory inference methods. *Nat Biotechnol* 2019;**37**: 547–54. https://doi.org/10.1038/s41587-019-0071-9

Shi Z, Wen B, Gao Q *et al.* Feature selection methods for protein biomarker discovery from proteomics or multiomics data. *Mol Cell Proteomics* 2021;**20**:100083. https://doi.org/10.1016/j.mcpro.2021.100083

Stoeckius M, Hafemeister C, Stephenson W *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017;**14**:865–8. https://doi.org/10.1038/nmeth.4380

Trapnell C, Cacchiarelli D, Grimsby J *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 2014;**32**:381–6. https://doi.org/10.1038/nbt.2859

van der Maaten L. Learning a parametric embedding by preserving local structure. *J Mach Learn Res Proc Track* 2009; **5**:384–91.

Virshup I, Rybakov S, Theis FJ *et al.* Anndata: access and store annotated data matrices. *JOSS* 2024;**9**:4371. https://doi.org/10.21105/joss.04371

Waichman TV, Vercesi ML, Berardino AA *et al.* scX: a user-friendly tool for scRNA-seq exploration. *Bioinform Adv* 2024;**4**:vbae062. https://doi.org/10.1093/bioadv/vbae062

Weber C, Hirst MB, Ernest B *et al.* SEQUIN is an R/shiny framework for rapid and reproducible analysis of RNA-seq data. *Cell Rep Methods* 2023;**3**:100420. https://doi.org/10.1016/j.crmeth.2023.100420

Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:15. https://doi.org/10.1186/s13059-017-1382-0

Bioinformatics, 2026, 42, 1–5
https://doi.org/10.1093/bioinformatics/btag301
Applications Note