Received: 8 July 2025 | Accepted: 9 July 2025

DOI: 10.1097/HEP.0000000000001479

CORRESPONDENCE

Letter to the Editor: Reevaluating PCA and SCTransform usage in single-cell intrahepatic cholangiocarcinoma analysis

To the editor,

Fan et al[1] investigated the co-localization of MARCO + tumor-associated macrophages and CTSE+ tumor cells, identifying it as a determinant of poor prognosis in intrahepatic cholangiocarcinoma. For their analysis, gene expression data were normalized using the Seurat SCT normalization algorithm. Principal component analysis (PCA) was employed for dimensionality reduction, extracting the first 20 principal components. Subsequently, Louvain clustering was applied at a resolution of 0.8 to define distinct cell types.[1]

Nevertheless, this study raises significant theoretical and empirical concerns regarding the exclusive reliance on PCA, a linear method that can yield misleading interpretations when confronted with nonlinear biological data.[2–5] PCA relies on several assumptions, such as linear inter-variable correlations, continuous and standardized data, an adequate sample size, homoscedasticity, and minimal outliers, while ensuring the orthogonality of principal components. Applying linear techniques like PCA to nonlinear datasets, or using parametric models on nonparametric data, can distort results related to variance and feature importance, potentially leading to erroneous conclusions. This highlights the critical need for a thoughtful consideration of the data's inherent characteristics when selecting analytical methods in biological research, advocating for a more nuanced approach that better accommodates the complexities of biological data.

However, even widely accepted preprocessing steps such as normalization and transformation can introduce artifacts. Methods like Seurat's SCTransform fit and remove technical effects (for example, sequencing depth and mitochondrial read percentage) using a regularized negative binomial model. While this stabilizes variance across genes, it also reshapes the original count distribution, potentially decorrelating true biological signals or amplifying noise in low-abundance transcripts. Such distortions can profoundly affect downstream analyses, including estimates of variance and feature importance.

These distortions fall into 1 of 3 major categories of machine learning misapplications. First is the violation of algorithmic assumptions: PCA assumes linear relationships, orthogonal components, homoscedastic residuals, and minimal outlier conditions, rarely satisfied in single-cell gene expression data. Second are ground truth challenges in model interpretation: without an independent "gold standard" or orthogonal validation (for example, protein-level confirmation of cell types), it is difficult to assess whether clustering or component loadings truly reflect underlying biology. Third are critical preprocessing errors: normalization, batch correction, or filtering steps that inadvertently warp the feature space before any modeling takes place.

In their study, Fan and colleagues confront both the first and the third issues. They rely solely on PCA despite its linear assumptions, and they employ SCTransform without demonstrating that the transformed expression values faithfully preserve the original biological variation. By overlooking these concerns, the authors risk drawing conclusions from distorted representations of the data.

Moving forward, researchers should validate preprocessing pipelines against untransformed data or external benchmarks, complement linear methods with nonlinear unsupervised approaches such as feature agglomeration and highly variable gene selection to capture complex cellular relationships, and explicitly test whether key findings hold under alternative normalization schemes. Only by aligning analytical choices with the data's intrinsic characteristics can we ensure robust, biologically meaningful insights.

CONFLICTS OF INTEREST

The author has no conflicts to report.

Yoshiyasu Takefuji

Faculty of Data Science, Musashino University, Tokyo, Japan

Correspondence
Yoshiyasu Takefuji, Faculty of Data Science, Musashino University, 3-3-3 Ariake, Koto-ku, Tokyo 135-8181, Japan.
E-mail: takefuji@keio.jp

ORCID
Yoshiyasu Takefuji https://orcid.org/0000–0002–1826–742X

REFERENCES

1. Fan G, Tao C, Li L, Xie T, Tang L, Han X, et al. The co-location of MARCO+ tumor-associated macrophages and CTSE + tumor cells determined the poor prognosis in intrahepatic cholangiocarcinoma. Hepatology. 2025;82:25–41.
2. Dyer EL, Kording K. Why the simplest explanation isn't always the best. Proc Natl Acad Sci USA. 2023;120:e2319169120.
3. Cristian PM, Aarón VJ, Armando EHD, Estrella MLY, Daniel NR, David GV, et al. Diffusion on PCA-UMAP manifold: The impact of data structure preservation to denoise high-dimensional single-cell RNA sequencing data. Biology. 2024;13:512.
4. Yao Y, Ochoa A. Limitations of principal components in quantitative genetic association models for human studies. eLife. 2023;12:e79238.
5. Elhaik E. Principal component analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. Sci Rep. 2022;12:14683.