ARTICLE IN PRESS

Enhanced Analytical Framework for Complex Biological Data: Beyond Principal Component Analysis in Cancer Research

Dear Editors:

Ding et al[1] investigated the role of AVL9 in chemoresistance of pancreatic ductal adenocarcinoma under hypoxic and acidic tumor microenvironment conditions. Their analytical approach employed Principal Component Analysis (PCA) for 3 crucial purposes: feature reduction, clustering analysis, and feature importance assessment. For feature reduction, they screened 2000 genes with the highest variation, subsequently reducing dimensionality through PCA. The transformed PCA space enabled cell clustering, revealing distinct cell populations and their relationships. Finally, they determined significant principal components using enrichment scores and *P* values, identifying key gene combinations contributing to biological variation. This multipurpose PCA approach facilitated efficient dimensionality reduction while preserving essential biological signals in their AVL9-IκBα-SKP1 complex investigation.[1]

However, the application of PCA in biological research requires careful consideration. As a linear dimensionality-reduction technique, PCA assumes linear relationships among variables, potentially oversimplifying the inherently nonlinear nature of biological systems. This limitation can lead to incomplete or misleading representations of complex biological interactions, resulting in potentially erroneous conclusions.[2–6]

To address these methodologic limitations, this paper proposes a comprehensive analytical framework enhancing biological data analysis reliability. The framework initiates with thorough exploratory data analysis to evaluate distributions, identify outliers, and understand variable relationships. This is followed by context-specific preprocessing including appropriate normalization techniques, scaling methods, and systematic handling of missing values.

For datasets violating linear assumptions—common in biological systems—this paper recommends advanced nonlinear nonparametric clustering methods. Specifically, HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) and OPTICS (Ordering Points To Identify the Clustering Structure) offer robust solutions for analyzing datasets with varying densities and complex structures, without requiring predefined cluster numbers.[7]

To ensure analytical rigor, this paper advocates for a multimetric clustering validation approach using 3 complementary measures: Silhouette Score for cluster cohesion and separation assessment, Davies-Bouldin Index for evaluating intracluster similarity against intercluster differences, and Gap Statistic for optimal cluster number determination. This comprehensive validation strategy enables robust quality assessment and informed parameter selection.

In addition, this paper recommends implementing multiple nonparametric correlation analyses: Spearman's rank correlation for monotonic relationships, Kendall's tau for ordinal associations, Goodman-Kruskal gamma for tied rankings, Somers' D for asymmetric relationships, and Hoeffding's D for general dependency detection. This comprehensive approach ensures both statistical validity and biological relevance, addressing the limitations of PCA in capturing complex biological relationships. Although the study by Ding et al provided valuable insights, incorporating these additional analytical methods could reveal previously undetected patterns and relationships in their data. Q1

YOSHIYASU TAKEFUJI
Faculty of Data Science
Musashino University
Tokyo, Japan

References

1. Ding J, et al. Gastroenterology 2025;168:539–555.e5.
2. Dyer EL, et al. Proc Natl Acad Sci U S A 2023;120:e2319169120.
3. Cristian PM, et al. Biology (Basel) 2024;13:512.
4. Yao Y, et al. eLife 2023;12:e79238.
5. Elhaik E. Sci Rep 2022;12:14683.
6. Lenz M, et al. Sci Rep 2016;6:25696.
7. Lee T, et al. Int J Precis Eng Manuf 2024;25:51–63.

Conflicts of interest
The author discloses no conflicts.

https://doi.org/10.1053/j.gastro.2025.03.053

UNCORRECTED PROOF