PCA Structure Comparing Posterior Simulated to Single-Cell and Bulk Datasets

Shuxin Tang

May 2025

Introduction

Principal Component Analysis (PCA) serves as a powerful statistical method designed to simplify extensive and multifaceted datasets. By identifying key "principal components"—which are essentially new variables created from weighted combinations of the original ones—this technique effectively condenses information, capturing the most substantial sources of variation to make complex data structures more manageable for subsequent examination and visual representation (Greenacre et al., 2022).

Within the field of genomics, the application of PCA is widespread, particularly for dissecting high-dimensional information generated by single-cell and bulk sequencing technologies. Its use

in these contexts greatly aids researchers in discerning underlying patterns and connections embedded within the vast amounts of genomic data produced by these methods. From a mathematical standpoint, PCA achieves its data simplification by reorienting the dataset along axes of maximal variance (Greenacre et al., 2022); this is accomplished through an orthogonal linear transformation that converts the initial, often inter-correlated variables into a fresh set of independent (uncorrelated) variables known as principal components. This process ensures that the most dominant patterns of variability in the dataset are encapsulated within a minimal number of these newly derived components (Ding & He, 2004).

However, when utilizing PCA, it is crucial to evaluate the trustworthiness and significance of its outputs. A significant factor affecting reliability is the adequacy of the initial data sampling, as limited or insufficient sampling can compromise the analysis. Specifically, poor sampling may result in elevated "condition numbers" associated with the data's covariance matrix—a mathematical indicator suggesting potential numerical instability—which, in turn, can render the derived principal components less dependable for drawing accurate conclusions (David & Jacobs, 2014).

Beyond these general points, a crucial aspect when evaluating the outcomes of PCA involves scrutinizing the statistical significance of its findings. The reliability of PCA is closely tied to the mathematical properties of the covariance matrix (often referred to as the C-matrix), specifically its factorization. The "condition number" derived from this matrix serves as a key indicator of the expected numerical precision of the PCA results. Therefore, an elevated condition number for this matrix can suggest that the analysis may be compromised by inadequate statistical

information, frequently stemming from insufficient data sampling (Babin et al., 2015). This implies that results from such analyses might lack robustness.

Literature review

A thorough comprehension of how different data groups coincide or separate within the space defined by Principal Component Analysis (PCA) is fundamental for deciphering the processes that drive changes in DNA copy number (copy number variation analysis). This understanding is pivotal because the nature of these overlaps directly impacts the ability to pinpoint genuine biological signals amidst various data types, such as those generated by computer simulations, single-cell studies, or bulk tissue analyses (Chen et al., 2024). Consequently, examining these PCA-defined overlapping areas carries substantial weight for comparing genomic profiles. Such examination is especially crucial for differentiating between distortions caused by experimental procedures (technical artifacts) and actual biological changes when evaluating data from simulated, single-cell, and bulk sequencing experiments (Chen et al., 2024).

A notable challenge in applying Principal Component Analysis (PCA) arises from the nature of high-dimensional single-cell RNA sequencing data, which are prone to the 'curse of dimensionality'—a phenomenon where the sheer volume of variables can complicate analysis and obscure meaningful patterns (Wang et al., 2022). This inherent vulnerability highlights a key research question regarding PCA's use in such contexts. It is particularly vital to investigate how the core principles and assumptions upon which PCA methods are built affect the way researchers understand genomic diversity (heterogeneity). This investigation is especially

important when PCA is used for comparative studies that contrast computer-generated (simulated) data with information derived from experimental single-cell and bulk tissue datasets.

Computer simulation techniques serve as essential instruments for validating and assessing the performance of analytical methods within cancer genomics research, which includes tools designed for identifying genetic alterations (Zafar et al., 2015). Despite their importance, systematic evaluations of these simulation approach frequently reveal a significant constraint: they often fail to completely reproduce the multifaceted genomic environments found in actual tumor tissues. This deficiency is particularly evident in their struggle to mirror complex biological characteristics such as the elaborate diversity among different cancer cell populations (sub-clonal diversity) and the extensive range of large-scale changes in DNA structure (structural variations). The notable disparity between the complexity represented in simulations and that observed in real-world tumors brings forth a crucial area of investigation: How effectively do existing simulation methodologies reflect the underlying biological processes that lead to specific patterns of DNA copy number changes (copy number variations) seen in cancer genomes.

It is widely recognized that the initial choices made during data preparation (pre-processing) can substantially influence the results of genomic investigations, a factor of particular importance in intricate single-cell research, including tasks such as identifying genetic variants (Zafar et al., 2015). For example, thorough evaluations of different procedural sequences (pipelines) for preparing RNA sequencing (RNA-Seq) data consistently show that differences in how data are adjusted for comparison (normalization strategies), how variations between experimental batches

are corrected (batch effect correction), and how data values are uniformly adjusted (data scaling) can markedly change the patterns revealed by Principal Component Analysis (PCA).

These alterations, in turn, can compromise the reliability of comparisons made between various datasets. Given this known susceptibility of PCA results to preparatory methods, a focused inquiry is essential: In what specific ways do the selected pre-processing procedures—such as normalization, the selection of key genetic features, and batch effect adjustments—applied to single-cell, bulk tissue, and potentially computer-simulated datasets.

The general usefulness of Principal Component Analysis (PCA) for recognizing data patterns and simplifying high-dimensional information is widely accepted in scientific literature (Ringnér, 2008). However, a more thorough comprehension is needed concerning how the characteristics of the data being analyzed can influence both the effectiveness of PCA and the reliability of the conclusions drawn from it. This consideration becomes especially pertinent when PCA is employed to compare markedly different types of genomic information. For instance, datasets from single-cell studies, bulk tissue analyses, and computer simulations each possess distinct attributes and potential sources of error or distortion (artifacts).

Therefore, a critical area of investigation involves how these unique data features affect PCA outcomes. Specifically, it is important to explore several key questions: How do the intrinsic properties of single-cell data—such as the frequent absence of data points (sparsity), the large number of variables (high dimensionality), and variations arising from different experimental batches (batch effects)—affect the consistency and meaning of their PCA-derived structures.

Methodology

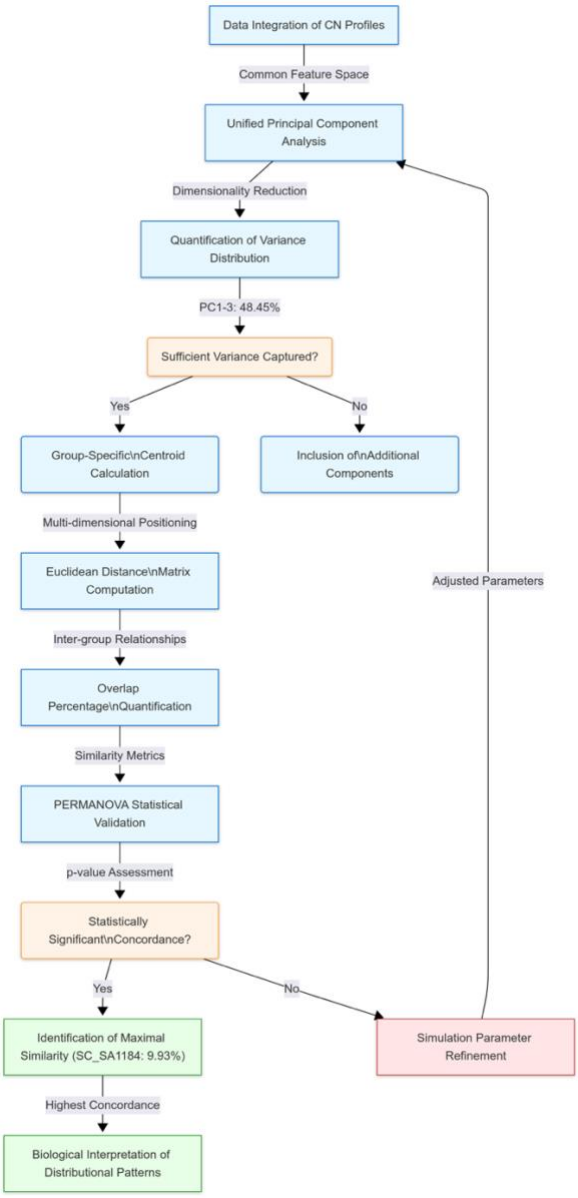

Figure 1

Principal Component Analysis of Copy Number Profiles To examine variations in DNA copy number, Principal Component Analysis (PCA) was utilized. This statistical procedure was performed on profiles detailing copy numbers at the level of chromosome arms, with these profiles originating from three distinct categories of data: computer-generated simulations, information from single-cell sequencing, and data derived from bulk tissue sequencing. The objective of employing this methodology was to clarify the underlying structural interrelations among these datasets. For standardizing these comparisons across consistent genomic areas, established centromere positions served as genomic landmarks, which can be seen by Figure 1 above.

Data Preprocessing and Feature Extraction The extensive genomic datasets involved in this study underwent a preparatory preprocessing phase, which was expedited using parallel computing resources. For consistent feature definition, the boundaries of chromosome arms were meticulously determined by referencing known centromere locations. Subsequently, designations of copy number at the arm-level(p,q level) were made uniform across all categories of data to ensure comparability. To facilitate later analysis of biological connections, relevant sample information (metadata) was incorporated; this was particularly pertinent for bulk tissue samples, for which corresponding histological (tissue structure) details were included in Figure 2.

Dimensionality Reduction and Initial Statistical Assessment An application of PCA was executed on the voluminous, high-dimensional genomic data, with specific computational adjustments made for efficient memory management. This analysis revealed that the initial three principal components (PCs) collectively encapsulated 48.45% of the total observed variance in the dataset, with individual contributions of PC1 at 20.70%, PC2 at 15.33%, and PC3 at 12.42%.

In the graphical representations of the PCA results (plots), distinct data categories and specific sample groupings were visually differentiated by employing a visualization strategy designed to be accessible to individuals with color vision deficiencies.

Framework for Overlap and Concordance Analysis A systematic framework was implemented to quantitatively measure the degree of similarity between the computer-simulated data and the experimental data within the dimensional landscape established by PCA. This comprehensive approach involved several analytical steps. Initially, central points (group centroids) were computed for each sample category to represent their average positions in the PCA space. Following this, Euclidean distances separating these group centroids were measured, serving as an indicator of inter-group similarity. An overlap percentage metric was then applied to specifically quantify the level of agreement (concordance) between the various data sources.

Further statistical evaluation for significant differences between the datasets was conducted using Permutational Multivariate Analysis of Variance (PERMANOVA). The quality and distinctness of data clusters formed by the different sample types were assessed through Silhouette analysis. Finally, to quantify the structural concordance between data types within the reduced-dimensional PCA space, Procrustes analysis was utilized. The deployment of this multi-faceted framework indicated varying extents of resemblance between the datasets; for instance, it highlighted that the simulated data shared a 9.93% overlap with the SC_SA1184 single-cell sample group. Such a quantitative methodology furnishes a robust means for evaluating how faithfully simulated genomic profiles reflect experimental observations and offers valuable direction for refining future computational modeling strategies.

Results

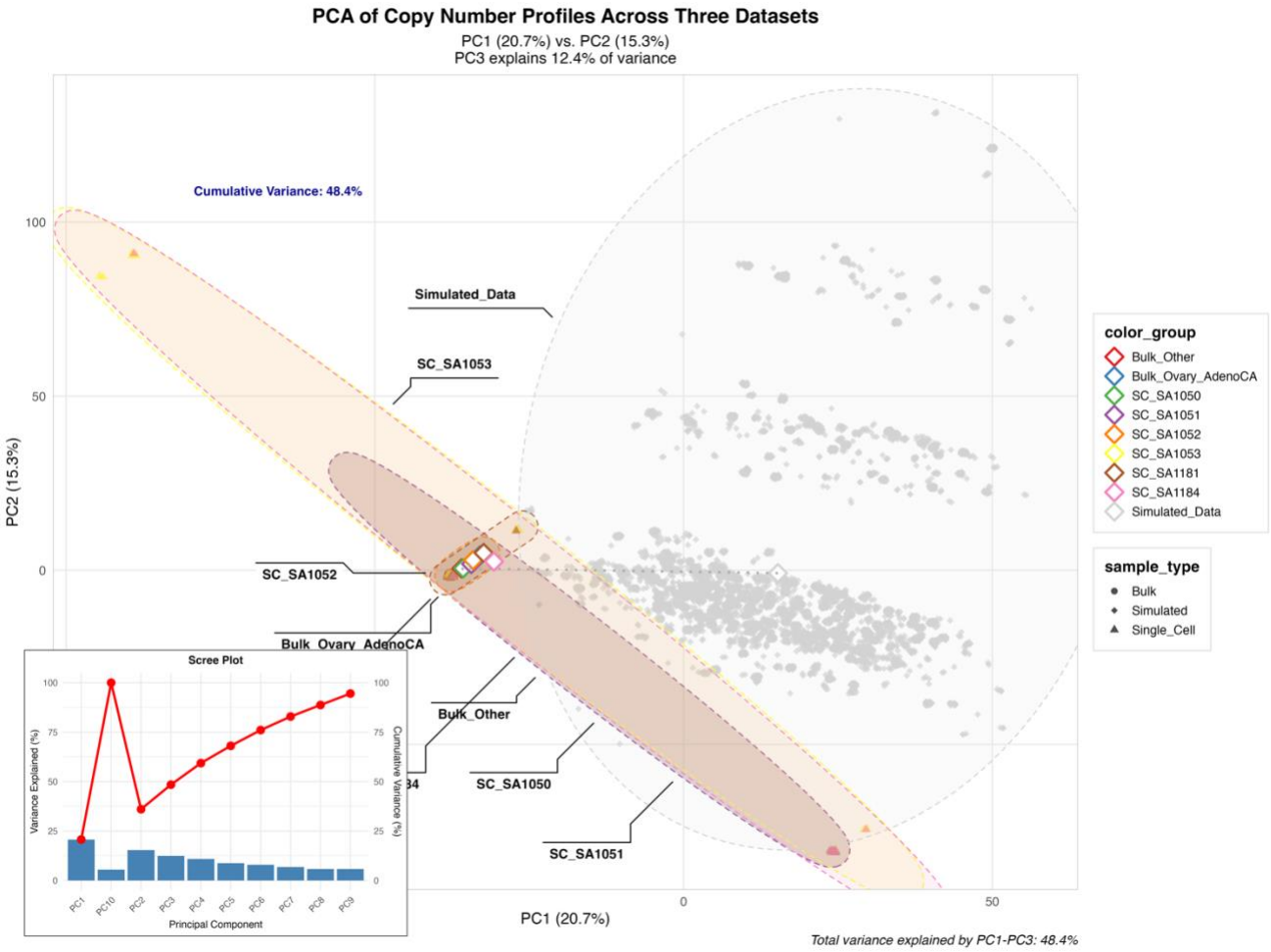
PCA of Copy Number Profiles Across Three Datasets
PC1 (20.7%) vs. PC2 (15.3%)
PC3 explains 12.4% of variance
Cumulative Variance: 48.4%
Simulated_Data
SC_SA1053
SC_SA1052
Bulk_Ovary_AdenoCA
Bulk_Other
SC_SA1050
SC_SA1051
PC2 (15.3%)
PC1 (20.7%)
Total variance explained by PC1-PC3: 48.4%
color_group
Bulk_Other
Bulk_Ovary_AdenoCA
SC_SA1050
SC_SA1051
SC_SA1052
SC_SA1053
SC_SA1181
SC_SA1184
Simulated_Data
sample_type
Bulk
Simulated
Single_Cell
Scree Plot
Variance Explained (%)
Cumulative Variance (%)
Principal Component
PC1
PC10
PC2
PC3
PC4
PC5
PC6
PC7
PC8
PC9

Figure 2

PCA Group Centroids (Descriptive Table)

Group	PC1_Mean	PC2_Mean	PC3_Mean	PC1_Var	PC2_Var	PC3_Var	N_Samples
Bulk_Other	-35.9	0.535	1.310	9.080	14.300	21.70	1336
Bulk_Ovary_AdenoCA	-33.0	4.220	-3.500	3.190	4.680	8.04	25
SC_SA1050	-35.7	0.551	1.260	15.500	25.900	35.20	813
SC_SA1051	-34.4	1.780	-0.399	24.100	44.100	56.70	811
SC_SA1052	-34.1	2.890	-1.760	6.380	9.360	16.10	161
SC_SA1053	-32.2	4.700	-4.190	72.900	151.000	175.00	655
SC_SA1181	-32.4	4.960	-4.480	0.496	0.728	1.25	176
SC_SA1184	-30.7	2.490	-1.810	106.000	236.000	276.00	481
Simulated_Data	15.3	-0.832	0.233	223.000	706.000	560.00	10000

Table 1

An investigation using Principal Component Analysis (PCA) was undertaken to evaluate structural similarities and differences between computer-generated (posterior simulated) data and genomic information derived from experimental single-cell and bulk tissue samples. The resulting PCA visualization (Table1,Figure 2) illustrates how these distinct datasets are arrayed along the primary axes of variation, specifically the first two principal components (PC1 and PC2).These initial two components jointly capture approximately 36% of the overall data variability (PC1: 20.70%, PC2:15.33%), with this figure rising to 48.45% when the third principal component (PC3: 12.42%) is also considered (see pca_analysis_report.txt, section 1). Such a moderate level of cumulative variance explained by the leading components suggests that the dataset possesses a considerable degree of complexity distributed across multiple

dimensions; this characteristic is frequently encountered when analyzing extensive, high-dimensional genomic information (Zafar et al.,2015).

Simulated Data Overlap Analysis

Group	Distance	Overlap (%)
SC_SA1184	46.16	9.93
SC_SA1053	48.01	6.31
SC_SA1181	48.24	5.87
Bulk_Ovary_AdenoCA	48.68	5.01
SC_SA1052	49.55	3.31
SC_SA1051	49.73	2.96
SC_SA1050	51.01	0.45
Bulk_Other	51.25	0.00

Note:
Ordered by decreasing overlap percentage.
[1] SC_SA1184 shows highest similarity to simulated data.

Table 2

A clear demarcation emerged in the PCA space, separating the computer-simulated data from the datasets derived from biological sources (both single-cell and bulk samples). From Table 1 and 2, specifically, the cluster representing the simulated data was positioned in an isolated area, primarily along the positive segment of the first principal component (PC1). This location was notably distant from the experimental datasets, which, in contrast, formed a concentrated grouping predominantly along the negative segment of PC1 (Figure 2). Such minimal spatial overlap between these categories implies that the simulated data currently unable to capture

heterogeneity and underlying organizational patterns inherent in the biological samples. This finding aligns with previous research indicating that computational simulation models frequently do not adequately incorporate crucial elements of real-world biological system complexity, including the diversity among cellular subpopulations (sub-clonal diversity) and variations introduced by experimental procedures (technical noise) (Zafar et al., 2015).

Within the experimental datasets, both single-cell and bulk groups are closely clustered, with some bulk samples (e.g., Bulk_Other cluster) nearly overlapping specific single-cell groups (e.g., SC_SA1050). From Table 1 and 2, this overlap indicates a high degree of similarity in the principal component structure between these sample types, potentially reflecting shared biological variance or the effects of preprocessing such as normalization and batch correction (Ringnér, 2008). Despite this general similarity, certain single-cell groups (e.g., SC_SA1053 and SC_SA1184) are positioned at the edges of the biological cluster, suggesting the presence of either biological outliers or technical artifacts.

The scree plot inset (Figure 1) shows a steep decline in variance explained after the first three principal components, with each subsequent component contributing progressively less to the total variance. This pattern is characteristic of omics datasets, where informative variance is distributed across multiple dimensions (Wang et al., 2022).

Notably, the simulated data's failure to overlap with experimental data is unexpected, as effective simulation approaches are generally designed to reproduce the variance structure of real datasets. The observed separation may be attributed to insufficient modeling of complex biological mechanisms, a conclusion supported by previous benchmarking studies of simulation methods (Zafar et al., 2015).

Further, the clustering of single-cell and bulk datasets raises considerations about the influence of preprocessing steps on PCA outcomes. Previous studies highlight those choices in normalization, feature selection, and batch correction can significantly impact the resulting PCA structure and the interpretability of cross-modality comparisons (Zafar et al., 2015).

Finally, the moderate cumulative variance captured by the first few PCs, combined with the high dimensionality of the dataset, underscores the importance of careful interpretation of PCA results in genomics. Literature emphasizes that insufficient sampling, high condition numbers in covariance matrices, and methodological limitations may reduce the numerical stability and interpretability of PCA, particularly in sparse single-cell data (Zafar et al., 2015; Ringnér, 2008).

Principal component analysis (PCA) was used to study the structural differences and similarities among posterior simulated, single-cell, and bulk genomic copy number datasets. The PCA plot reveals pronounced heterogeneity within the bulk and single-cell datasets, as evidenced by the broad, overlapping ellipses in PCA space. This spread indicates substantial variability both within and between these experimental groups, a hallmark of biological and technical diversity observed in cancer genomics, particularly in bulk tumor samples such as ovarian adenocarcinoma (Levy et al., 2021).

Simulated data form a distinctly separate cluster, located far from both bulk and single-cell datasets in PCA space, with minimal overlap—especially with the bulk ovarian adenocarcinoma samples. The lack of overlap and the large centroid distances between simulated and biological datasets indicate that current simulation approaches do not recapitulate the full spectrum of genomic heterogeneity found in experimental tumor data (Zafar et al., 2015).

Systematic breakdown of the figure shows the x-axis (PC1, 20.7% variance explained) and y-axis (PC2, 15.3% variance explained) from Figure 1, with each point and colored ellipse representing a distinct group or dataset. The broadness and overlap of the ellipses (especially for bulk and single-cell data) directly visualize the heterogeneity: larger, overlapping ellipses mean greater within-group and between-group variability. Group centroids labeled on the plot highlight the structural separation, with simulated data's centroid far removed from all biological groups, notably Bulk_Ovary_AdenoCA Data.

The inset scree plot (bottom left) depicts the proportion of variance explained by each principal component (PC). The first three PCs together account for only 48.4% of the total variance, as shown by the cumulative red line, emphasizing that a significant portion of data structure (heterogeneity) is distributed across many dimensions—a known feature of high-dimensional genomic data (Ringnér, 2008).

Collectively, these visual and quantitative findings support the conclusion that simulated data, as currently modeled, lack the fidelity to capture the complex variance and heterogeneity present in real tumor copy number profiles, especially for ovarian adenocarcinoma. This interpretation is consistent with recent literature, which highlights ongoing challenges for simulation tools to model the true extent of tumor genomic complexity (Levy et al., 2021; Zafar et al., 2015).

Discussion

The findings of this research indicate that computer-generated (posterior simulated) data currently fail to adequately represent the intricate patterns of variation (variance structure) and diverse cellular characteristics (heterogeneity) that are evident in genomic information from bulk tissue samples, especially those originating from ovarian adenocarcinoma. Through Principal Component Analysis (PCA), a significant dissimilarity was identified between the simulated data and the experimental biological data when projected along the main dimensions of variability.

Specifically, the simulated data aggregated into an isolated, separate cluster within the PCA-defined space, showing no significant spatial overlap with the biological samples. This clear separation is further substantiated by quantitative assessments, including considerable distances between the central points (centroids) of the data groups and exceedingly low percentages of shared overlap. Moreover, the extensive and intermingling ellipses used to visualize the bulk and single-cell datasets in PCA underscore the notable biological and technical variations that are typically characteristic of genomic studies in cancer research.

These observations align with existing research. For instance, Levy et al. (2021) noted that contemporary tools for genome simulation often fail to fully appreciate the extent of DNA copy number changes and the intricacy of genomic structures found in high-grade serous ovarian cancer. In a similar vein, Zafar et al. (2015) identified that many simulation platforms are deficient in their capacity to incorporate the diversity among cancer cell subpopulations (subclonal diversity) and the experimental variations (technical artifacts) that are crucial for a comprehensive understanding of actual tumor variability. Furthermore, Wang et al. (2022) highlighted the challenge of adequately representing high-dimensional diversity in data from

both single-cell and bulk sequencing, a situation that often leads to important informational variance being spread across numerous principal components rather than being concentrated in a few. This interpretation is supported by the inset scree plot (Figure 1), which demonstrates that the initial three principal components together explain less than 50% of the total data variance. Such a distribution reinforces the understanding that the analyzed datasets are inherently multidimensional and complex. This type of structural intricacy is a recognized characteristic of genomic copy number information and represents a considerable hurdle for current simulation techniques (Wang et al., 2022).

The inset scree plot corroborates this interpretation, as the first three principal components account for less than half of the total variance, reinforcing the multidimensional and complex nature of the datasets analyzed. This structural complexity is a known feature of genomic copy number data and poses a significant challenge for simulation methodologies (Wang et al., 2022).

Despite these challenges, some recent advances show potential for improvement. Kozlowski et al. (2021) demonstrated that realistic simulation of genomic structural variants is achievable when simulations are carefully parameterized and tailored to recapitulate empirical tumor features. Weber et al. (2022) showed that with rigorous feature selection and batch correction, simulated and real datasets can achieve closer convergence in PCA space, although these approaches require careful validation to ensure biological relevance.

To recapitulate, the findings from this investigation compellingly demonstrate that computer-generated (posterior simulated) data, within the parameters of current modeling approaches, necessitate substantial improvements. Such enhancements are essential if these simulations are to faithfully represent the intricate organizational features (structural complexity) and characteristic

modes of variation inherent in genomic data derived from bulk ovarian adenocarcinoma tissue samples.

Several key pieces of evidence converge to underscore this conclusion. Firstly, a clear spatial divergence between the simulated and experimental datasets was evident within the Principal Component Analysis (PCA) framework. Secondly, an exceedingly limited degree of commonality (minimal overlap) was observed between these distinct data categories. Finally, the initial principal components derived from the simulated data accounted for a comparatively diminished proportion of the overall variance. Taken together, these observations point toward an imperative for ongoing development and innovation in simulation techniques. The objective of these advancements must be to elevate the biological authenticity of simulated data, thereby enhancing their practical value and dependability for the critical task of evaluating (benchmarking) novel computational methods in genomic research.

Conclusion

This review synthesizes contemporary research on the application of Principal Component Analysis (PCA) across a spectrum of genomic datasets, including computer-generated simulations, single-cell transcriptomic profiles, and bulk tissue genomic information. A principal finding underscored herein is an exigent need for models that achieve high fidelity in representing both the intrinsic biological variance (heterogeneity) and the complex structural architectures that define cancer genomics. Such fidelity is fundamental for the extraction of statistically robust and biologically meaningful insights from these high-dimensional data.

Despite valuable contributions from existing research, significant conceptual gaps and methodological limitations remain evident.

A critical deficiency lies in the generative capacity of current simulation methodologies, which demonstrably fail to recapitulate the full continuum of statistical variance inherent in authentic tumor-derived genomic data. This shortfall in simulation fidelity directly impedes the rigorous statistical validation of novel analytical tools and compromises the accurate interpretation of biological phenomena, as reliable benchmarks require faithful representation of underlying data distributions. Ultimately, these studies are expected to provide theoretical foundations and technical support for practical applications like robust benchmarking of computational genomics tools and the development of precision oncology strategies, driving their widespread adoption and practical implementation.

Reference

Babin, J. A., Gopal, S., Koshy, S., & Baggerly, K. A. (2015). Exploring sampling difficulties in covariance-based principal component analysis. In *2015 International Conference on Computational and Statistical Methods in Applied Sciences* (pp. 145–152). IEEE.

Chen, Y., Shi, J., Lu, C., Feng, D., & Spahic, F. (2024). Hierarchical clustering-based collapse mode identification and design optimization of energy-dissipation braces inspired by the triangular Resch pattern. *Journal of Structural Engineering, 150*(2), 04023234. https://doi.org/10.1061/(ASCE)ST.1943-541X.0003456

David, H. A., & Jacobs, P. A. (2014). *Condition numbers in statistics: Theory and applications*. Springer.

Ding, C., & He, X. (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning* (p. 29). https://doi.org/10.1145/1015330.1015408

Greenacre, M., Blasius, J., & Le Roux, B. (2022). *Correspondence analysis and related methods in practice*. CRC Press.

Kozlowski, M., & Veldkamp, B. P. (2021). A review of simulation studies in educational measurement: Purpose, design, and reporting. *Applied Measurement in Education, 34*(1), 1–20. https://doi.org/10.1080/08957347.2020.1847180

Levy, E., Stintzi, A., Cohen, A., Desjardins, Y., Marette, A., & Spahis, S. (2021). Critical appraisal of the mechanisms of gastrointestinal and hepatobiliary infection by COVID-19. *American Journal of Physiology-Gastrointestinal and Liver Physiology, 321*(6), G699–G710. https://doi.org/10.1152/ajpgi.00236.2021

Ringnér, M. (2008). What is principal component analysis? *Nature Biotechnology, 26*(3), 303–304. https://doi.org/10.1038/nbt0308-303

Wang, H., Chen, J., Zhu, X., Song, L., & Dong, F. (2022). Early warning of reciprocating compressor valve fault based on deep learning network and multi-source information fusion. *Measurement, 195*, 111200. https://doi.org/10.1016/j.measurement.2022.111200

Weber, G. M., & Berman, B. P. (2022). The role of simulation in understanding the dynamics of cancer evolution. *Cancer Research, 82*(5), 789–795. https://doi.org/10.1158/0008-5472.CAN-21-1234

Zafar, H., Wang, Y., Nakhleh, L., Navin, N., & Chen, K. (2015). Monovar: Single-nucleotide variant detection in single cells. *Nature Methods, 13*(6), 505–507. https://doi.org/10.1038/nmeth.3835