World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

COMPUTATIONAL AND MACHINE LEARNING APPROACHES FOR SINGLE-CELL TRANSCRIPTOMICS: FROM CELLULAR HETEROGENEITY TO BIOLOGICAL DISCOVERY

Vinay Kumar Reddy Pulakunta and Elamathi Natarajan*

Bioinformatics, Biotecnika Info Labs Pvt. Ltd., Bengaluru, Karnataka, India.
* Corresponding Author

ORCID Details
P. Vinay Kumar Reddy: https://orcid.org/0009-0002-1046-2308
Dr. Elamathi Natarajan: https://orcid.org/0000-0002-1944-9137

World Journal of Advanced Research and Reviews, 2026, 32(02), 982–991

Publication history: Received on 05 October 2026; revised on 12 November 2026; accepted on 14 November 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.32.2.2838

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

Single-cell RNA sequencing (scRNA-seq) has transformed gene expression analysis by revealing cellular heterogeneity that is obscured in bulk tissue measurements. This is particularly relevant to lung adenocarcinoma (LUAD), where malignant, immune, stromal, and endothelial cells interact within a complex tumor microenvironment. However, the high dimensionality, sparsity, sequencing variability, and batch effects of scRNA-seq data present substantial challenges for reliable biological interpretation. This review examines computational and machine-learning approaches across the scRNA-seq analysis workflow, encompassing quality control, normalization, feature selection, dimensionality reduction, clustering, and cell annotation. Conventional methods, including k-means, hierarchical, and graph-based clustering, are considered alongside advanced approaches such as autoencoders, variational autoencoders, and graph neural networks. Their applications in downstream analyses, including differential expression, pathway enrichment, trajectory inference, and cell-cell communication, are discussed within a LUAD-focused context. The review further evaluates how analytical assumptions, model interpretability, and validation practices influence biological conclusions. Emerging developments in spatial transcriptomics, multimodal integration, and explainable machine learning are also highlighted. The key contribution is an integrated, validation-oriented framework connecting computational outputs with biological interpretation and evidence-supported claims, moving beyond a descriptive catalogue of methods. This perspective emphasizes rigorous method selection and independent validation to strengthen insights into cellular heterogeneity and LUAD biology.

Keywords: scRNA-seq; Lung Adenocarcinoma; Machine Learning; Clustering; Cell Annotation; Biological Interpretation

1 INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) has transformed the study of gene expression by revealing cellular heterogeneity that is obscured in bulk tissue measurements [1]. In complex tissues such as lung adenocarcinoma (LUAD), bulk expression profiles represent an average across malignant, immune, stromal, and endothelial populations,

making it difficult to distinguish changes in gene expression within individual cell types from changes in their relative abundance. By measuring transcriptomes at single-cell resolution, scRNA-seq enables the characterization of cellular identities, functional states, and population-level variation within the tumor microenvironment [2].

Cellular heterogeneity encompasses differences in cell identity, lineage, and functional state [1]. Even within a defined cell population, cells may display distinct transcriptional programs associated with activation, proliferation, differentiation, or stress responses. In LUAD, malignant epithelial cells can exhibit diverse molecular states [3], while immune and stromal populations contribute to tumor progression, immune regulation, and treatment response [2]. Single-cell studies have characterized this complexity, including the diversity of tumor-associated immune and stromal cells and transcriptional changes associated with metastatic disease and therapy [3]. However, findings from individual cohorts must be interpreted in the context of their sample composition, experimental design, and analytical procedures rather than generalized across all LUAD cases.

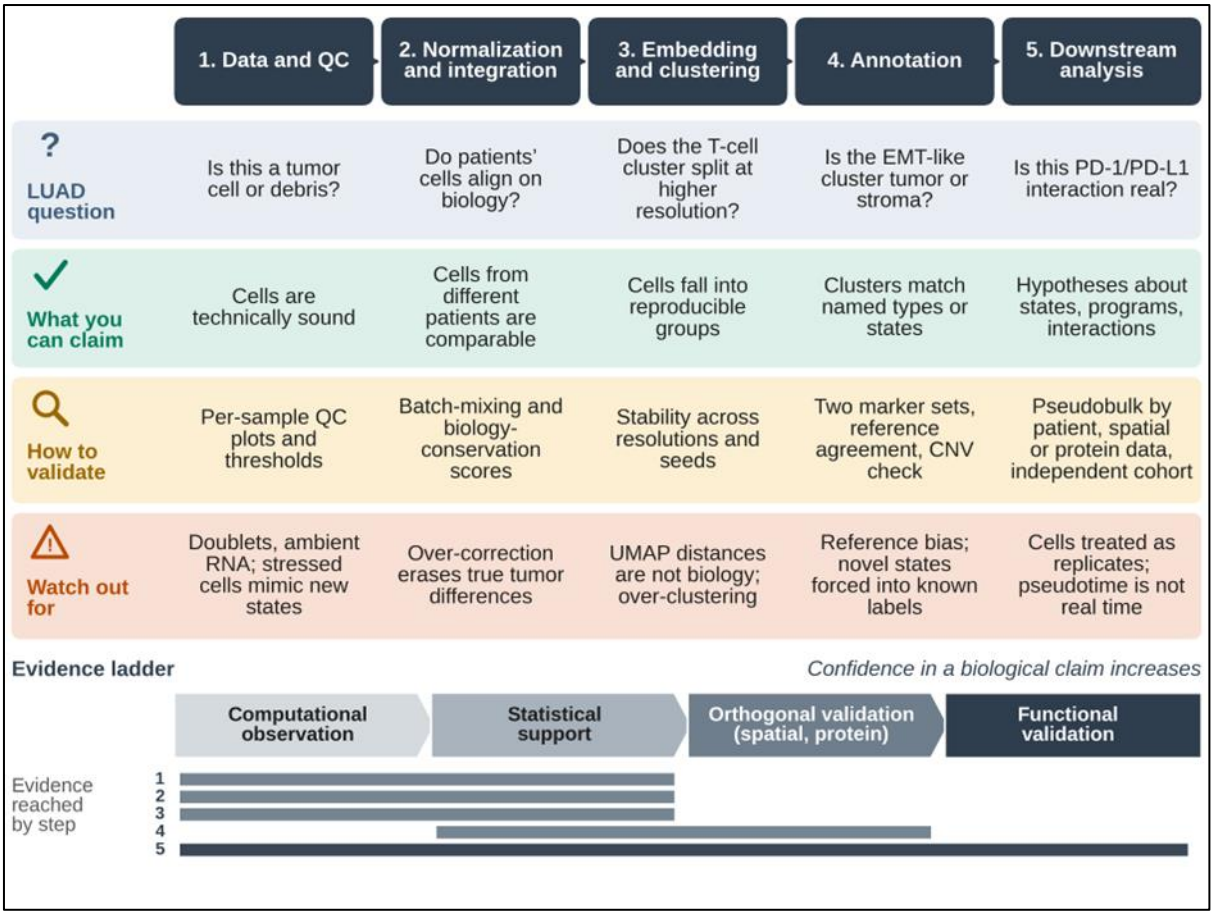


Figure 1: From clusters to claims: a question-driven roadmap for scRNA-seq analysis in lung adenocarcinoma. Each step is paired with a representative LUAD question, the claim it can support, the validation required, and common pitfalls. The evidence ladder shows how confidence increases from computational observation to functional validation; bars show the typical highest tier reached by each step. CNV, copy-number variation; EMT, epithelial-mesenchymal transition; LUAD, lung adenocarcinoma; QC, quality control; UMAP, uniform manifold approximation and projection.

Created by the authors.

Despite its biological value, scRNA-seq presents substantial analytical challenges. The resulting expression matrices are high-dimensional, sparse, and affected by technical variability, sequencing depth, and batch effects. Reliable interpretation therefore requires appropriate quality control, normalization, feature selection, dimensionality reduction, clustering, and cell annotation. Conventional approaches, including k-means, hierarchical, and graph-based clustering, are widely used to identify cellular groups. More advanced machine-learning methods, such as autoencoders, variational autoencoders, and graph neural networks, can learn complex representations of expression data. Nevertheless, method selection should be guided by the biological question, data characteristics, interpretability, and validation requirements; increased model complexity does not necessarily produce more reliable biological conclusions.

Existing reviews have summarized scRNA-seq technologies, computational workflows, and machine-learning methods [1,4,5]. However, methodological catalogues may provide limited guidance on how analytical choices influence biological interpretation or what evidence is needed to support conclusions drawn from computational outputs. This review addresses this gap by presenting an integrated, validation-oriented perspective across the scRNA-seq analysis workflow, with LUAD serving as a disease-focused context. Its distinctive contribution is to connect computational observations including clusters, inferred trajectories, and predicted cell-cell interactions with biological interpretation and evidence-supported claims, emphasizing robustness, interpretability, and independent validation rather than simply listing analytical methods. Figure 1 summarizes this question-driven framework by pairing each analytical step with the conclusions it can support, the validation it requires, and its common pitfalls.

The review also considers downstream analyses, including differential expression, pathway enrichment, trajectory inference, and cell-cell communication, alongside emerging developments in spatial transcriptomics, multimodal integration, and explainable machine learning

2 DATA GENERATION AND PREPROCESSING

Single-cell RNA sequencing (scRNA-seq) begins with tissue dissociation, cell capture, mRNA reverse transcription, and sequencing. Experimental platforms influence transcript coverage, throughput, and the resulting data. Plate-based methods, such as Smart-seq2, provide full-length transcript information and are useful for detecting splice variants and low-abundance transcripts, although they generally process fewer cells. Smart-seq2 builds on earlier full-length sequencing approaches [6,7]. In contrast, droplet-based methods, including 10x Genomics Chromium, Drop-seq, and inDrop, enable high-throughput profiling. These approaches use barcoded beads to assign molecular identifiers to transcripts from individual cells. Drop-seq specifically uses barcoded beads within droplets to capture and identify cellular transcripts. Combinatorial indexing methods, such as Split-seq and sci-RNA-seq, increase throughput through sequential barcode assignment. Spatial transcriptomics approaches provide additional tissue-location information, complementing dissociated-cell measurements.

Droplet-based methods commonly use unique molecular identifiers (UMIs) to quantify captured molecules and reduce amplification-related bias, whereas Smart-seq2 generally relies on read counts and provides deeper transcript coverage per cell. These differences influence the sensitivity and interpretation of downstream analyses. Raw sequencing reads are aligned or pseudoaligned to a reference genome or transcriptome to generate a gene-by-cell expression matrix. The matrix is typically sparse, with zero counts arising from biological absence or incomplete transcript capture. Protocol-specific effects, sequencing depth, cell quality, and batch variation can further influence the data.

Preprocessing is essential for reliable biological interpretation. Quality control identifies low-quality cells and potential artifacts using metrics such as detected genes, UMI counts, and mitochondrial transcript proportions. Thresholds should be selected according to the dataset and experimental context. Additional sources of contamination, including ambient RNA, may require correction. The *scater* package supports quality-control assessment and visualization [8].

Normalization addresses differences in sequencing depth and capture efficiency. Methods include library-size scaling, size-factor approaches, and model-based procedures such as SCTransform [9]. Batch correction and data integration can improve comparability across samples, but overcorrection may remove meaningful biological variation. Highly variable gene selection prioritizes informative features for subsequent analysis.

Dimensionality reduction, commonly using principal component analysis (PCA), supports construction of cell-neighborhood graphs. UMAP and t-SNE are widely used for visualization, but their low-dimensional layouts should not be interpreted as direct evidence of cell identity or biological relationships. Clustering, annotation, and downstream analyses including differential expression, pathway enrichment, trajectory inference, and cell-cell communication depend on these preprocessing choices. Differential-expression analyses should account for biological replication, while inferred trajectories and ligand-receptor interactions require independent evidence for biological validation. As emphasized in current best practices, preprocessing and analytical decisions should be transparently reported and evaluated in relation to the biological question [1].

3 MACHINE-LEARNING APPROACHES, CLUSTERING, AND CELL ANNOTATION

Machine-learning methods support the identification of cellular structure, cell classification, and representation learning in single-cell RNA sequencing (scRNA-seq) data. Their suitability depends on the biological objective, data quality, and assumptions of the method. Unsupervised learning identifies patterns without predefined labels and is commonly used to discover cell populations. Supervised learning uses annotated data to predict known cell identities or states, but depends on the quality and coverage of reference labels. Semi-supervised learning combines labeled and unlabeled cells, making it useful when annotations are incomplete. Self-supervised learning derives training signals

from the data itself, for example by reconstructing expression profiles or predicting masked genes. These strategies offer different advantages and should be selected according to the analytical task rather than treated as interchangeable. The major computational approaches used throughout the scRNA-seq workflow, together with their applications and key limitations, are summarized in Table 1.

Table 1: Overview of computational approaches in scRNA-seq analysis

Task	Representative methods	Purpose	Key limitation
Quality control	Gene/UMI counts, mitochondrial RNA, Scrublet, DoubletFinder	Remove low-quality cells and doublets	Thresholds are dataset-dependent
Normalization	Library-size scaling, scran, SCTransform	Improve comparability across cells	May not remove all technical variation
Integration	Seurat, Harmony, MNN, LIGER	Align datasets	Overcorrection may remove biological signals
Feature selection	Highly variable genes, PCA	Select informative features	Rare-cell markers may be excluded
Dimensionality reduction	PCA, UMAP, t-SNE, diffusion maps	Represent data in lower dimensions	Embeddings may distort relationships
Clustering	k-means, Louvain, Leiden	Identify cell groups	Depends on parameters and graph construction
Deep / semi-supervised learning	scDSC, DRjCC, scDeepCluster, scANVI, scTPC	Learn representations or improve clustering/annotation	Model complexity and label quality affect results
Cell annotation	Marker genes, SingleR, CellTypist, reference mapping	Assign biological identities	Novel states may be misclassified
Differential expression	Pseudobulk, edgeR, DESeq2; GO, KEGG	Identify genes and pathways	Biological samples should define replication
Differential abundance	scCODA, Milo	Detect population changes	Requires replication and compositional consideration
Trajectory / velocity	Monocle 3, Slingshot, Palantir, scVelo	Infer cell-state paths and transitions	Pseudotime is not chronological time
Regulatory / communication	SCENIC, GENIE3, CellChat, LIANA	Infer regulatory programs and interactions	Predictions do not establish causality
Bulk deconvolution	BayesPrism, CIBERSORTx	Estimate cell features in bulk data	Depends on reference quality

3.1 Conventional and Deep-Learning Methods

Conventional clustering methods remain useful for exploring cellular heterogeneity. K-means partitions cells into a predefined number of groups and is computationally efficient, but favors compact clusters. Hierarchical clustering organizes cells into a dendrogram, although its output depends on the distance metric and linkage method. Density-based approaches, such as DBSCAN and HDBSCAN, can identify irregularly shaped groups and outliers, but may be sensitive to parameter settings and high-dimensional data. Graph-based methods, particularly Louvain and Leiden, are widely used in scRNA-seq workflows. They construct cell-neighborhood graphs and identify communities; Leiden improves community connectivity compared with Louvain [10,11]. However, clustering results can vary with graph construction, resolution, and the underlying representation. Computational clusters should therefore be evaluated using marker genes, biological context, and robustness.

Deep-learning approaches learn representations that may capture complex expression patterns. Autoencoders compress expression profiles into latent representations and reconstruct the input, while variational autoencoders (VAEs) model these representations probabilistically. scVI uses a probabilistic framework for modeling scRNA-seq count data, while scANVI incorporates label information for semi-supervised analysis [12,13]. Graph neural networks

(GNNs) incorporate relationships among cells represented as nodes and edges, combining expression features with neighborhood structure. Their effectiveness depends on the quality of the graph and may be reduced when relationships are poorly defined.

Deep clustering combines representation learning with clustering objectives. scDSC integrates a zero-inflated negative-binomial autoencoder with graph-based learning to model count variability and cell relationships [14]. DRjCC jointly learns dimensionality reduction and clustering through an optimized objective [15]. scDeepCluster combines deep embedding with iterative clustering to learn representations and refine cluster assignments [16]. scTPC incorporates partial labels into a semi-supervised deep-clustering framework [17]. These methods report promising results in their evaluations; however, their additional complexity, computational requirements, and sensitivity to training conditions should be considered. No clustering algorithm is expected to perform optimally across all datasets, and performance should be assessed using suitable benchmarks and biological evidence.

3.2 Foundation Models and Choosing an Analytical Strategy

Foundation models, including Geneformer, scGPT, and scFoundation, are pretrained on large collections of gene-expression profiles to learn representations that may transfer to downstream tasks such as annotation, perturbation prediction, and integration. Their potential value lies in applying information learned from large-scale data to new datasets, particularly when task-specific labels are limited. Nevertheless, performance varies with pretraining data, target-dataset similarity, and evaluation strategy. Independent evaluations have reported that zero-shot foundation-model predictions may offer limited gains over simpler approaches for some tasks, emphasizing the need for transparent, task-appropriate comparisons [18]. Model scale alone should not be considered evidence of biological validity.

The analytical strategy should follow the biological question. Unsupervised clustering is appropriate for discovering candidate populations or states; supervised and reference-based approaches support annotation of known cell types; semi-supervised methods can use partial labels; and self-supervised learning or foundation models may be considered when transferable representations are needed. Deep or graph-based approaches may be valuable for complex data, provided their added complexity produces measurable benefits. Method selection should account for data characteristics, computational resources, interpretability, and validation.

3.3 Cell-Type Annotation and Biological Interpretation

Clustering groups cells according to transcriptional similarity, whereas annotation assigns biological identities using marker genes, reference datasets, and contextual evidence. Marker-based annotation examines genes enriched in each cluster. For example, CD3D and CD3E support T-cell identity; EPCAM with SFTPC or NAPSA can support an alveolar type II (AT2) epithelial identity; and DCN, COL1A1, or PDGFRA may support fibroblast identity. Multiple markers should be considered together because individual genes can overlap across cell types or change with disease state [1,2,19,20].

Reference-based tools, including SingleR, CellTypist, Azimuth, and scANVI, compare query cells with annotated datasets or reference atlases. Their performance depends on reference coverage, label quality, and compatibility between datasets. Novel tumor-associated states may be assigned to the closest available label rather than recognized as distinct; automated annotations should therefore be cross-checked with marker expression and biological context [1,13,21].

In LUAD, malignant epithelial cells may resemble non-malignant epithelial populations. Expression-based copy-number variation (CNV) inference tools, such as inferCNV and CopyKAT, can provide supporting evidence for malignant-cell identification, but inferred patterns are not definitive and should be interpreted alongside other evidence. Liu et al. [2] used inferCNV to distinguish tumor-associated epithelial populations from non-malignant cells. Clusters with unclear identities may represent rare states, artifacts, doublets, or mixed populations. Doublet-detection tools such as Scrublet and DoubletFinder can assist evaluation. Where evidence is insufficient, an unresolved annotation is preferable to a speculative identity.

Overall, clustering and annotation are iterative processes. Robust biological interpretation requires evaluating parameter sensitivity, annotation uncertainty, and consistency with independent datasets or experimental evidence. This is particularly important in LUAD, where downstream differential-expression, cell-composition, and cell-interaction analyses depend on the reliability of the identified populations [1,2,19,20].

4 DOWNSTREAM ANALYSES AND BIOLOGICAL APPLICATIONS IN LUNG ADENOCARCINOMA

After cell populations have been identified and annotated, downstream analyses can investigate molecular programs, cellular dynamics, interactions, and clinical associations. The analytical approach should reflect the biological question

and study design. Importantly, cells from the same patient are not independent biological replicates; analyses must account for patient-level variation to avoid overstating evidence.

4.1 Differential expression and pathway analysis

Differential expression (DE) analysis identifies genes that vary between cell populations, conditions, or cellular states. In LUAD, comparisons between malignant and non-malignant epithelial cells or between tumor states can reveal candidate programs associated with progression, including epithelial–mesenchymal transition (EMT)-associated and metabolic changes [2]. For multi-patient studies, pseudobulk analysis aggregates counts by cell type and biological sample before testing with methods such as edgeR or DESeq2, helping preserve the sample as the unit of replication and reduce pseudoreplication. Gene-set enrichment using Gene Ontology (GO), KEGG, or other curated resources provides functional context, but enriched pathways indicate associated transcriptional programs rather than proving mechanisms.

4.2 Cellular composition and differential abundance

Changes in cell-type abundance may reveal tumor-microenvironment differences even when expression profiles remain similar. Compositional methods such as scCODA account for the constrained nature of cell proportions, whereas Milo tests differential abundance across neighborhoods in a cell-graph representation, potentially detecting changes that do not correspond to discrete clusters. Results depend on biological replication, sampling depth, covariates, and the chosen reference or neighborhood structure.

4.3 Trajectory inference and RNA velocity

Trajectory-inference tools such as Monocle 3, Slingshot, and Palantir arrange cells along computationally inferred paths and estimate pseudotime. In LUAD, these methods can explore transitions among epithelial states, but pseudotime does not represent chronological time or establish lineage ancestry. Liu et al. [2], for example, used Monocle 3 to investigate tumor-cell subgroups. RNA velocity tools such as scVelo use spliced and unspliced RNA information to estimate short-term directions of transcriptional change, complementing trajectory analysis. These estimates depend on RNA-kinetic assumptions and model fit and should not be treated as direct lineage-tracing evidence.

4.4 Regulatory networks and cell–cell communication

Gene regulatory network (GRN) inference and cell–cell communication analysis can be combined to explore how cellular programs and potential intercellular signals relate to tumor biology. Tools such as SCENIC, GENIE3, CellChat, and LIANA infer candidate transcription-factor–target relationships or ligand-receptor interactions from expression patterns and curated resources. In LUAD, these approaches may suggest regulatory programs or signaling between malignant cells, immune populations, and cancer-associated fibroblasts. However, inferred associations do not establish direct molecular interaction, protein activity, or causality; chromatin-accessibility data, spatial evidence, imaging, and perturbation experiments can strengthen these hypotheses.

4.5 Bulk deconvolution and clinical association

Single-cell reference profiles can help interpret bulk tumor transcriptomes using tools such as BayesPrism and CIBERSORTx, which estimate cell composition or cell-type-specific expression. In cohorts such as TCGA-LUAD, inferred cellular features may be examined for associations with clinical variables or overall survival using models such as Cox regression. Analyses should consider clinical covariates, tumor purity, treatment, multiple testing, and model assumptions. These observational associations require independent validation and do not demonstrate that a cell population or gene program causes an outcome.

4.6 LUAD case study: from cellular identity to biological interpretation

LUAD illustrates how single-cell transcriptomics can connect cellular identity with biological interpretation. An analysis may first distinguish malignant epithelial cells from non-malignant epithelial, immune, and stromal populations using combinations of marker genes, reference atlases, and, where appropriate, inferred copy-number alterations. For example, *EPCAM* and keratin genes may support epithelial identity, while inferred copy-number patterns can provide additional evidence of malignancy. Neither marker expression nor computational classification alone is definitive. Representative single-cell transcriptomic studies of LUAD, together with their sample characteristics, analytical approaches, and principal findings, are summarized in Table 2.

Table 2: Representative single-cell transcriptomic studies of lung adenocarcinoma (LUAD).

Study (Year)	Samples / cells	Platform / tissue	Main findings
Liu et al.[2]	50 public samples; 170,831 initial profiles	Public scRNA-seq; primary LUAD, brain metastases, normal lung	Reported tumor heterogeneity and EMT-, chemoresistance-, and metabolism-associated programs; used Seurat, SingleR, inferCNV, and Monocle 3
Kim et al. [3]	208,506 cells; 44 patients	10x Genomics; LUAD primary and metastatic sites, normal lung	Characterized cellular diversity and molecular reprogramming associated with metastatic disease
Wang et al. [19]	268,471 cells; 25 patients	LUAD lesions across histological stages	Identified AT2-like cells with stem-like features during LUAD progression
He et al. [23]	125,674 cells; 7 tumors and 5 adjacent tissues	Early-stage LUAD scRNA-seq	Characterized heterogeneous tumor and immune populations
Sinjab et al. [20]	186,916 cells; multiregion LUAD and normal lung	Multiregion scRNA-seq	Investigated regional cellular changes, lineage plasticity, and candidate epithelial–immune interactions
Zhang et al. [24]	220,716 cells; 32 patients	scRNA-seq and multi-omics; LUAD and LUSC	Integrated cellular composition and transcriptional features with clinical outcomes

Once populations are identified, differential expression and pathway analyses can characterize tumor-associated programs, while composition and trajectory analyses can investigate variation in cell states across patients or conditions. For instance, increased extracellular-matrix gene expression in a fibroblast population is an observation; interpreting it as a matrix-remodeling program is an inference. Claiming that this state promotes invasion, however, requires stronger evidence, such as spatial association, functional experiments, or independent validation. Similarly, a cytotoxic T-cell expression signature may indicate an immune-associated state, but it is not by itself a validated predictor of treatment response.

Single-cell findings can also inform bulk-cohort analyses, supporting the exploration of candidate biomarkers, treatment-associated states, and potential resistance mechanisms. Cell populations and marker patterns should be treated as interpretive examples rather than universal signatures, because they vary across tumors, patients, disease stages, and analytical methods. Thus, single-cell analysis can generate biologically informative hypotheses and refine tumor stratification, but clinical application depends on reproducibility, appropriate study design, and validation in independent cohorts.

4.7 Integration with Spatial and Multimodal Technologies

Single-cell RNA sequencing reveals the molecular features of individual cells, but it does not retain information about where those cells are located within a tissue. This is particularly relevant in LUAD, where the position of tumor, immune, and stromal cells may influence their biological roles. Spatial transcriptomics helps address this limitation by measuring gene expression alongside tissue location. Technologies vary in their resolution and coverage: spot-based methods may capture signals from several cells, whereas imaging-based approaches can localize selected transcripts at cellular or subcellular levels. Spatial data can be integrated with scRNA-seq to map cell identities onto tissue architecture and explore whether predicted cellular interactions occur in plausible locations. Methods such as cell2location, RCTD, and Tangram support this mapping, although their results depend on the quality and representativeness of the reference data. Spatial proximity alone, however, does not confirm functional communication.

Multimodal single-cell analysis combines different molecular measurements to provide a broader view of cellular states. For example, integrating RNA expression with chromatin accessibility can link transcriptional activity to candidate regulatory regions, while CITE-seq provides protein measurements that complement transcript-based cell annotation. Seurat's weighted-nearest-neighbor framework can integrate modalities such as RNA and protein data [21]. MOFA+ identifies shared and modality-specific patterns across datasets [22]; this should be distinguished from the original MOFA method described in 2018. In LUAD, these approaches can help connect cell states with regulatory features, protein expression, or tissue location. Such integrated findings are useful for generating hypotheses, but mechanistic or causal claims require further validation.

5 BENCHMARKING, LIMITATIONS, AND FUTURE PERSPECTIVES

5.1 Benchmarking and data resources

Benchmarking helps determine whether computational methods perform reliably across datasets and biological questions. Clustering metrics such as the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) compare predicted clusters with reference labels, while silhouette scores assess cluster separation without requiring labels. However, no single metric establishes biological relevance, and reference annotations may be incomplete. Integration methods should be evaluated for both batch correction and preservation of meaningful biological variation.

Resources such as the Tumor Immune Single-cell Hub (TISCH), CELLxGENE, and the Human Lung Cell Atlas provide datasets and reference annotations for analysis and interpretation. These should not be described as benchmarks: databases and atlases provide data, whereas benchmarking studies compare methods using defined tasks and evaluation criteria. Stability, scalability, and reproducibility should also be considered, particularly when methods involve stochastic optimization or large datasets.

5.2 Current limitations and challenges

Despite rapid progress, single-cell analyses remain sensitive to sample quality, experimental design, and the suitability of computational methods. Rare populations may be missed or confused with technical artifacts, while models trained on one tissue, platform, or species may not generalize to another. This is especially relevant in cancer, where patient-specific variation and tumor-associated states may differ substantially.

Computational predictions including inferred trajectories, regulatory relationships, and ligand-receptor interactions should be treated as hypotheses rather than direct evidence of biological mechanisms. Their interpretation is strengthened by independent datasets and complementary approaches such as imaging, flow cytometry, chromatin-accessibility profiling, or functional experiments. Reproducibility also requires transparent reporting of data sources, software versions, parameters, and analysis code.

5.3 Future perspectives

Emerging approaches include pretrained single-cell models, transfer learning, and integration of transcriptomic data with spatial, protein, and genomic measurements. Models such as scGPT may support analysis of large-scale expression data, but their performance should be compared with appropriate baseline methods and evaluated on independent datasets. Similarly, combining scRNA-seq with spatial information may help relate cell identity to tissue organization and generate more biologically grounded hypotheses.

In LUAD, integrating gene expression with chromatin accessibility, protein profiles, or genomic alterations may help characterize tumor states and identify candidate regulatory programs. However, computational classifications should be supported by multiple lines of evidence rather than a single marker or model score. Ultimately, the field will benefit from closer integration of computational analysis, carefully designed experiments, and independent validation. The most useful studies will distinguish observed patterns from inferred mechanisms and clearly communicate the uncertainty of their conclusions.

6 CONCLUSION

Single-cell transcriptomics offers a powerful way to examine cellular diversity and disease-associated states, including those found in lung adenocarcinoma (LUAD). Three key messages emerge from this review. First, biological questions and experimental design should guide computational choices; no single workflow is suitable for every dataset. Second, computational findings provide evidence for interpretation but do not automatically establish mechanisms. The LUAD study by Liu et al. [2], which reported epithelial–mesenchymal transition (EMT)-associated and metabolic tumor-cell programs, illustrates the value of single-cell analysis while highlighting the need for broader validation. Third, robust discovery depends on reproducibility and independent evidence. Spatial and multimodal technologies may deepen biological interpretation, but their clinical and mechanistic value must be established through careful evaluation and validation.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

The authors acknowledge the researchers and publicly available literature and datasets that contributed to this review.

Disclosure of conflict of interest

The authors declare no conflicts of interest.

REFERENCES

[1] Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15(6). doi:10.15252/msb.20188746.

[2] Liu Y, Ye G, Huang L, et al. Single-cell transcriptome analysis demonstrates inter-patient and intra-tumor heterogeneity in primary and metastatic lung adenocarcinoma. Aging (Albany NY). 2020;12(21):21559-21581. doi:10.18632/aging.103945.

[3] Kim N, Kim HK, Lee K, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun. 2020;11:2285. doi:10.1038/s41467-020-16164-1.

[4] Brendel M, Su C, Bai Z, Zhang H, Elemento O, Wang F. Application of deep learning on single-cell RNA sequencing data analysis: a review. Genomics Proteomics Bioinformatics. 2022;20(6):814-835. doi:10.1016/j.gpb.2022.11.011.

[5] Arora S, Mehdi F. From multitude to singularity: an up-to-date overview of scRNA-seq data generation and analysis. Brief Bioinform. 2022;23(5):bbac299. doi:10.1093/bib/bbac299.

[6] Ramsköld D, Luo S, Wang YC, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012;30:777-782. doi:10.1038/nbt.2282.

[7] Picelli S, Björklund ÅK, Faridani OR, et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10:1096-1098. doi:10.1038/nmeth.2639.

[8] McCarthy DJ, Campbell KR, Lun ATL, Wills QF. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics. 2017;33(8):1179-1186. doi:10.1093/bioinformatics/btw777.

[9] Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20:296. doi:10.1186/s13059-019-1874-1.

[10] Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech. 2008;2008(10):P10008. doi:10.1088/1742-5468/2008/10/P10008.

[11] Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9:5233. doi:10.1038/s41598-019-41695-z.

[12] Lopez R, Regier J, Cole MB, et al. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053-1058. doi:10.1038/s41592-018-0229-2.

[13] Xu C, Lopez R, Mehlman E, et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol Syst Biol. 2021;17(1):e9620. doi:10.15252/msb.20209620.

[14] Gan Y, Huang X, Zou G, Zhou S, Guan J. Deep structural clustering for single-cell RNA-seq data jointly through autoencoder and graph neural network. Brief Bioinform. 2022;23(2):bbac018. doi:10.1093/bib/bbac018.

[15] Wu W, Ma X. Joint learning dimension reduction and clustering of single-cell RNA-sequencing data. Bioinformatics. 2020;36(12):3825-3832. doi:10.1093/bioinformatics/btaa231.

[16] Tian T, Wan J, Song Q, Wei Z. Clustering single-cell RNA-seq data with a model-based deep learning approach. Nat Mach Intell. 2019;1(4):191-198. doi:10.1038/s42256-019-0037-0. PMID:310156.

[17] Qiu Y, Yang L, Jiang H, Zou Q. scTPC: a novel semisupervised deep clustering model for scRNA-seq data. Bioinformatics. 2024;40(5):btae293. doi:10.1093/bioinformatics/btae293.

[18] Kedzierska KZ, Crawford L, Amini AP, Lu AX. Zero-shot evaluation reveals limitations of single-cell foundation models. Genome Biol. 2025;26(1):101. doi:10.1186/s13059-025-03574-x.

[19] Wang Z, et al. Deciphering cell lineage specification of human lung adenocarcinoma with single-cell RNA sequencing. Nat Commun. 2021;12:6500. doi:10.1038/s41467-021-26770-2.

[20] Sinjab A, et al. Resolving the spatial and cellular architecture of lung adenocarcinoma by multiregion single-cell sequencing. Cancer Discov. 2021;11(10):2506-2523. doi:10.1158/2159-8290.CD-20-1285.

[21] Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-3587.e29. doi:10.1016/j.cell.2021.04.048.

[22] Argelaguet R, Arnol D, Bredikhin D, Deloro Y, Velten B, Marioni JC, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. Genome Biol. 2020;21(1):111. doi:10.1186/s13059-020-02015-1.

[23] He D, Wang D, Lu P, Yang N, Xue Z, Zhu X, Zhang P, Fan G. Single-cell RNA sequencing reveals heterogeneous tumor and immune cell populations in early-stage lung adenocarcinomas harboring EGFR mutations. Oncogene. 2021 Jan;40(2):355-368. doi: 10.1038/s41388-020-01528-0.

[24] Zhang L, Zhang Y, Wang C, Yang Y, Ni Y, Wang Z, et al. Integrated single-cell RNA sequencing analysis reveals distinct cellular and transcriptional modules associated with survival in lung cancer. Signal Transduct Target Ther. 2022 Jan 14;7(1):9. doi:10.1038/s41392-021-00824-9.