BENCHMARKING BATCH-EFFECT CORRECTION METHODS TOWARDS THE CONSTRUCTION OF A TRIPLE-NEGATIVE BREAST CANCER CELL ATLAS

Peter Scheible
Computer Science
Old Dominion University
Norfolk, VA, United States
psche004@odu.edu

Jing He, Ph. D.
Computer Science
Old Dominion University
Norfolk, VA, United States
jhe@cs.odu.edu

Amy H. Tang, Ph. D.
Biomedical and Translational Sciences
Macon & Joan Brock Virginia Health Sciences
Old Dominion University
Norfolk, VA, United States
TangAH@odu.edu

Jiangwen Sun, Ph. D.
The Corresponding Author
Computer Science
old Dominion University
Norfolk, VA, United States
jsun@cs.odu.edu

Abstract—**Triple-negative breast cancer (TNBC) requires detailed cellular mapping given its aggressive nature, immense tumor heterogeneity and genetic diversity. We integrated 156,794 cells from six scRNA-seq datasets—including tumors, metastases, and cell lines—to build a TNBC scRNA cell atlas, focusing on batch effect mitigation while maintaining biological and molecular details. Preprocessing filters noise, normalizes data, and leverages PCA for integration readiness. We utilized scANVI, a semi-supervised tool, to align datasets, preserving TNBC's complex tumor heterogeneity via marker annotations [1]. UMAPs demonstrate biological clustering in integrated data, contrasted with dataset-driven unintegrated patterns. Assessments verifying effective batch correction. This method aligns with NASA's GeneLab supported multi-omics studies under space stressors. Our progress advances personalized TNBC medicine by revealing cellular insights and charts a path toward a comprehensive, multimodal TNBC cell atlas, promising broad impact in oncology and NASA health research.**

Index Terms—**RNA sequencing, Breast cancer, Machine learning, Data integration, Bioinformatics, Biomarkers, Data preprocessing, Computational biology**

I. Introduction

Triple-negative breast cancer (TNBC) is a highly aggressive molecular subtype of breast cancer defined by the absence of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) expression, representing $\sim 15\%$ of all breast cancer cases. This subtype disproportionately affects young women and Black women, who experience higher incidence rates and poorer survival outcomes compared to other racial groups [2]. Lacking targeted therapeutic options, TNBC relies

The authors gratefully acknowledge funding support from the Virginia Space Grant Consortium (VSGC) and the Hampton Roads Biomedical Research Consortium (HRBRC)

heavily on chemotherapy and/or chemo + pembrolizumab, yet 20–30% of patients face early relapse and metastatic progression, underscoring the urgent need for novel approaches to improve prognosis [3].

The molecular heterogeneity of TNBC complicates its management, with six distinct subtypes identified, each exhibiting unique sensitivities to therapeutic agents [4]. Traditional bulk RNA sequencing, which averages gene expression across cell populations, fails to capture the cellular and genetic diversity of TNBC, limiting insights into its complex biology [5]. Alternatively, single-cell RNA sequencing (scRNA-seq) provides unprecedented high resolution, profiling transcriptional states of individual cells to reveal tumor heterogeneity and tumor/tumor microenvironment (TME) interactions [6]. This capability makes scRNA-seq a powerful transformative tool for advancing TNBC research at a single cell resolution.

A single-cell atlas for TNBC, mapping cell types and states, offers a robust framework to address these clinical challenges. By integrating scRNA-seq data to build a TNBC atlas, we may be able to identify subet-specific rare biomarkers and novel therapeutic targets, enabling personalized treatment strategies tailored to individual patients [3]. It also holds promise for unraveling subtype-specific variability and addressing racial disparities in TNBC outcomes, as evidenced by studies linking gene expression heterogeneity to aggressive disease states [2], [7]. If successful, this study may help to build an interactive TNBC atlas as a critical step toward personalized oncology therapeutics.

However, constructing a robust TNBC atlas requires overcoming significant hurdles in data integration. Heterogeneous scRNA-seq datasets, sourced from public repositories like NCBI GEO, vary in experimental protocols, sequencing platforms, and sample conditions, introducing batch effects, technical variations that may mask true biological signals [8], [9]. Existing atlases, often built from one or two datasets, fail to capture the full spectrum of TNBC diversity due to these scalability limitations [3]. Effective batch correction and automated data integration methods are thus essential to unify these diverse TNBC datasets while preserving biological fidelity.

Construction of a comprehensive single-cell atlas for triple-negative breast cancer (TNBC) represents the overarching goal of this project, aiming to integrate multimodal datasets spanning genomics, transcriptomics, epigenomics, proteomics and multi-omics approaches into a unified resource for delineating and understanding TNBC's tumor heterogeneity and cellular complexity [10].

At this milestone, the focus narrows to a critical aspect of this vision, mitigating batch effects across diverse scRNA-seq datasets while preserving biological features essential for TNBC characterization. We propose a scalable framework for Six publicly available scRNA TNBC datasets to undergo preprocessing, filtering noise and normalizing mRNA expression, followed by integration and batch-effect correction using scANVI, a semi-supervised variational autoencoder that leverages cell type annotations to align datasets [1]. This alignment enhances biomarker discovery and patient-specific treatment mapping by ensuring data quality and integrity, addressing clinical and scientific gaps in TNBC understanding.

Beyond clinical oncology, this work carries broader significance. Batch correction and integration strategies mirror NASA's GeneLab efforts to harmonize multi-omics data, providing insights into cellular responses under space stressors like microgravity and radiation [11]. By refining these methods, the research supports NASA's mission to protect astronaut health during long-duration spaceflight, while laying groundwork for the full Human Cancer Atlas's future multimodal expansion.

II. Literature Review

Single-cell RNA sequencing (scRNA-seq) has transformed cancer research by enabling high-resolution profiling of cellular heterogeneity, overcoming the limitations of bulk RNA-seq, which averages gene expression across multiple tumor subpopulations [5]. In breast cancer, Chung et al. utilized scRNA-seq to characterize tumor and immune cell diversity in primary tumors, revealing immunosuppressive TMEs [6]. For triple-negative breast cancer (TNBC), Karaayvaz et al. applied scRNA-seq to uncover subclonal heterogeneity and identify useful gene expression signatures linked to treatment resistance and metastasis [7]. These studies underscore scRNA-seq's power to dissect the immense complexity underlining tumor relapse, chemo-resistance, and poor survival.

Integrating diverse and variable scRNA-seq datasets poses significant technical challenges due to their different clinicopathological classification, diverse treatment responses, a multitude of tumor characteristics and dynamic tumor landscape and changing tumor/TME interactions and constant micro-evolutions. Data sparsity, technical noise, and batch effects arising from differences in sequencing platforms, laboratory-to-laboratory variations, scRNA data capturing under different experimental conditions complicate joint analysis [12]. Luecken et al. note that these technical variations are known to obscure and mask biological signals, particularly in atlas-scale large dataset and multiomics projects spanning multiple data sources [8]. Uncorrected batch effects hinder the construction of comprehensive cell atlases, necessitating robust integration methods [9].

A. Anchor-Based Integration

Anchor-based methods align datasets by identifying mutual nearest neighbors (MNN) across batches. Haghverdi et al. pioneered MNN correction, later refined in Seurat for scalable integration [9], [10]. These approaches excel when cell types overlap, preserving biological variation, but scale poorly with large datasets and falter without sufficient shared populations.

B. Variance Decomposition

Variance decomposition methods model and remove technical effects statistically. ComBat-seq adjusts RNA-seq counts using a negative binomial model, while Harmony iteratively corrects batch effects with clustering [13], [14]. Harmony's speed and sensitivity make it widely used, though both methods risk overcorrection when cell frequencies differ significantly across different batches of data.

C. Deep Learning Approaches

Deep learning leverages neural networks to capture non-linear batch effects. scVI uses variational autoencoders for unsupervised integration, while scANVI incorporates cell type labels for improved accuracy [1], [15]. scGPT, a generative transformer, scales to millions of cells, offering a foundation model to capture multiomics datasets [16]. These methods excel in complexity but require careful hyperparameter tuning.

D. Graph-Based Integration

Graph-based methods construct cell similarity networks for data alignment, comparison, and integration. BBKNN balances batch effects efficiently, and Scanorama combines MNN with graph techniques for scalability [17], [18]. While fast and flexible, they often prioritize batch removal over conserving subtle biological differences and preserving the minor molecular signatures.

Advanced batch correction methods push integration further. CLAIRE employs contrastive learning to balance batch mixing and heterogeneity [19], while BERMUDA uses transfer learning to reveal hidden molecular signature and cellular subtypes [20]. BERMAD's multilayer autoencoder addresses under- and overcorrection [21], and adversarial approaches like ABC and scDREAMER optimize biological retention [22], [23]. These innovations highlight ongoing efforts to refine integration.

Existing cancer atlases provide context for TNBC efforts. Dave et al.'s Breast Cancer Single-Cell Atlas maps cell lines and tumors but relies on limited datasets [3]. Chen et al. offer a TNBC genomic-transcriptomic dataset, yet lack broad integration and general application [24]. Most tumor/cancer atlases fail to scale to multimodal, multi-source data [8].

Integration success is gauged by metrics balancing batch removal and biological conservation. kBET and silhouette scores assess batch mixing [25], [26], while cell type ASW and label transfer accuracy measure biological fidelity [8]. Trajectory preservation, per Trapnell et al. and Saelens et al., evaluates dynamic processes [27], [28].

Despite progress, gaps remain in scalability, multimodal data integration accuracy, data generation, data extrapolation, and TNBC-specific preclinical and clinical applications [8]. This work advances the field with a scalable, AI-driven pipeline for a TNBC atlas, building on Harmony and targeting comprehensive dataset integration [14].

III. Methodology

Six publicly available scRNA-seq datasets underpin the construction of a single-cell TNBC atlas, integrated through a pipeline of preprocessing, batch correction, cell type annotation, and evaluation. Computational tools address batch effects and enable biomarker discovery across diverse TNBC profiles.

A. Data Collection

Publicly accessible datasets from the National Center for Biotechnology Information Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) repositories provide a diverse mix of 156,794 tumor cells to build a TNBC atlas, primarily from single-cell RNA sequencing (scRNA-seq). Table I breaks it down. GSE75688 contributes 549 cells from primary breast cancer and metastases across 11 patients, including TNBC cases [6]. GSE176078 presents 100,078 cells from 26 primary tumors, with 10 TNBC samples showing subtype variety [29]. GSE182694 adds over 30,000 cells from breast cancer cell lines, capturing 49 subclusters [3]. SRP157974 includes 15,633 cells come from 401 TNBC patients and 23 white blood cell samples at Fudan University, but uses whole-exome sequencing (WES) data rather than scRNA-seq, offering broader genomic insights at a coarser resolution [24]. GSE118390 includes 1,534 cells from six fresh TNBC tumors, paired with WES data for added investigative depth [7]. SRP114962 adds 9,000 cells from 20 TNBC patients under neoadjuvant chemotherapy. [30].

This mix including primary tissue from untreated tumors and cell line data creates a rich but tricky dataset. Batch effects arise due to differences in sequencing platforms (e.g., 10x Genomics vs. SMART-seq) and sample types (fresh tumors vs. cultured tumor cells).

TABLE I
TNBC Datasets Used in Atlas Construction

Accession	Sample Size	Description	Source
GSE75688	549 cells	Primary and metastatic cells from 11 patients, includes TNBC	GEO
GSE176078	100,078 cells (26 tumors)	10 TNBC cases, subtype heterogeneity	GEO
GSE182694	30,000+ cells	Breast cancer cell lines, 49 subclusters	GEO
SRP157974	15,633 cells (401 patients + 23 WBC)	Mostly WES, some genomic TNBC data	SRA
GSE118390	1,534 cells	6 fresh TNBC tumors, paired with WES data	GEO
SRP114962	9,000 cells (20 patients)	scRNA-seq from neoadjuvant chemotherapy	SRA

B. Preprocessing

Quality control eliminates low-quality cells and genes from the six TNBC scRNA-seq datasets to minimize technical noise. Cells expressing fewer than 200 genes or exceeding 10% mitochondrial gene content are filtered out, as high mitochondrial fractions often indicate cellular stress or cell death [12]. Genes detected in fewer than three cells are similarly removed, ensuring focus on biologically relevant, detectable, and dominant features. Scanpy implements these filters efficiently, leveraging sparse matrix representations to handle the scale of the data, such as the 30,000+ cells in GSE182694 alone [5].

Normalization adjusts raw counts to account for sequencing depth variations across different tumor cells/subtypes. Counts scale to a target sum of 10,000 per cell, followed by log-transformation (log1p) to stabilize variance and mitigate the impact of highly expressed genes [12]. Highly variable genes, identified using a dispersion-based method, capture biological variation while reducing dimensionality from tens of thousands of genes to approximately 2,000, aligning with practices for atlas-scale analysis [10].

Principal component analysis (PCA) further reduces dimensionality to 50 components, retaining principal sources of variance for data integration. This step, executed via Scanpy, balances computational efficiency with biological information preservation, preparing the data for batch correction, automatic integration and data clustering. [5].

C. Cell Type Annotation

Cell identities emerge from the scANVI-integrated latent space through marker gene scoring and clustering, mapping TNBC's diverse cellular landscape. Marker genes, curated from literature, include CD3D, CD3E, and CD3G for T cells, CD19, CD79A, and MS4A1 for B cells, CD68 and CD14 for macrophages, and epithelial markers KRT5 and KRT14 (basal) versus KRT8 and KRT18 (luminal) to distinguish TNBC-relevant subtypes [6], [7]. Scanpy computes scores for each cell type by averaging the expression of valid markers present in the dataset, normalizing against a random gene set to reduce noise [10].

Preliminary annotations assign the highest-scoring cell type per cell, leveraging scANVI's semi-supervised labels for atlas refinement [1]. The Leiden algorithm clusters cells in the integrated space, optimizing resolution (e.g., 0.5-1.0) via silhouette analysis to balance granularity and coherence [26]. Epithelial clusters undergo subtyping: basal (KRT5, KRT14, TP63) and luminal (KRT8, KRT18) scores differentiate TNBC's malignant populations, validated against known profiles from GSE75688 and GSE118390 [6]. Unassigned cells, where scores fall below a threshold (e.g., 0.1), receive an "Unknown" label, addressing sparsity in datasets like SRP157974.

Annotations align with literature benchmarks, ensuring accuracy across immune, stromal, and tumor cells in TNBC [7]. This process, visualized in Figure 1, reveals TNBC's cellular diversity, supporting biomarker identification.

D. Data Integration

Batch effects across the six TNBC scRNA-seq datasets undergo correction using scANVI, a semi-supervised deep learning method leveraging variational autoencoders [1]. Preprocessed PCA embeddings from the datasets feed into scANVI's encoder network, which maps high-dimensional gene expression into a low-dimensional latent space. This latent representation captures both shared biological signals and dataset-specific variations, guided by cell type annotations from marker gene scoring. The decoder reconstructs gene expression, optimizing a loss function that balances reconstruction accuracy with batch alignment, employing adversarial training to minimize batch variation and dataset-specific biases [15].

Training proceeds with labeled and unlabeled cells, utilizing annotations for cell populations to enhance biological fidelity [6].

scANVI's semi-supervised approach refines integration by propagating labels to unannotated cells, addressing the partial annotation common in heterogeneous datasets. Hyperparameters, including learning rate (0.001) and latent dimension (10), tune via cross-validation to optimize convergence, typically requiring 50-100 epochs for larger datasets [1].

Integration quality manifests in UMAP visualizations, with Figure 1 contrasting unintegrated and integrated embeddings. Unintegrated data cluster predominantly by dataset, reflecting batch effects, while scANVI-aligned data group by cell type, indicating effective correction across the atlas's diverse sources [8]. This unified latent space supports downstream TNBC analysis in pilot study, training sets, and large-scale validation studies.

E. Evaluation

We measure integration and annotation quality for the TNBC atlas using metrics that evaluate batch correction and biological fidelity across datasets. The Adjusted Rand Index for batch (ARI_{batch}) assesses cluster alignment with dataset origins, with lower scores from 0 to 1 revealing effective batch mixing. The Adjusted Rand Index for cell type ($ARI_{cell\ type}$) tests consistency with cell type labels, where higher scores highlight preserved biological identity. Graph Connectivity (Graph Connectivity) examines linkage among same-type cells, scored 0 to 1, where higher values confirm successful integration. Normalized Mutual Information (NMI) checks clustering fidelity to annotations, with higher scores up to 1 reflecting accurate mapping [8].

IV. Results

We constructed a preliminary TNBC single-cell atlas by integrating 156,794 cells across six scRNA-seq datasets using scANVI, targeting batch effect mitigation. Our preprocessing retained high-quality cells, eliminating noise and normalizing expression. We applied scANVI to align these datasets, producing a shared latent space visualized in Figure 1. Our UMAPs show unintegrated data clustering by dataset, while integrated data cluster by cell type, indicating successful batch correction [1].

Metrics reveal a contrast with UMAP results. Unintegrated data produced ARI_{batch} 0.1512, a high score indicating pronounced batch effects, $ARI_{cell\ type}$ 0.5192, showing strong cell type alignment, reflecting cohesive gene clusters, and Graph Connectivity 0.8011, suggesting good connectivity. Integrated data delivered ARI_{batch} 0.0605, $ARI_{cell\ type}$ 0.2365, and Graph Connectivity 0.8769. Unintegrated data outperforms on $ARI_{cell\ type}$ and NMI, with 0.7267 versus 0.6469, highlighting superior cell type fidelity. We trace this to GSE182694's 30,000+ cells, all "cell line," which skews unintegrated gene clustering. scANVI disrupts this to unify datasets, prioritizing batch correction [1].

Our findings prioritize batch mixing (lower ARI_{batch}) and connectivity (higher Graph Connectivity), aligning with UMAP's biological clustering over dataset-specific scores. This atlas reveals resistance signatures in SRP114962 and subtype diversity in SRP157974, enhancing biomarker potential [24].

V. Discussion

Integration of 156,794 cells via scANVI produces a TNBC atlas that corrects batch effects, as UMAPs (Figure 1) show integrated cell type clustering versus unintegrated dataset clustering [1]. Some metrics, however, favor unintegrated data: ARI-celltype (0.5192 vs. 0.2365) and NMI (0.7267 vs. 0.6469) drop post-integration, despite ARI-batch decreasing (0.1512 to 0.0605), confirming batch mixing. We trace this to GSE182694's 30,000+ cells, all labeled "cell line," which cluster tightly pre-integration, boosting unintegrated scores. scANVI's alignment with primary tissue data (e.g., GSE176078) sacrifices this artificial fidelity for signal cohesion, known molecular signatures, biological unity, supported by

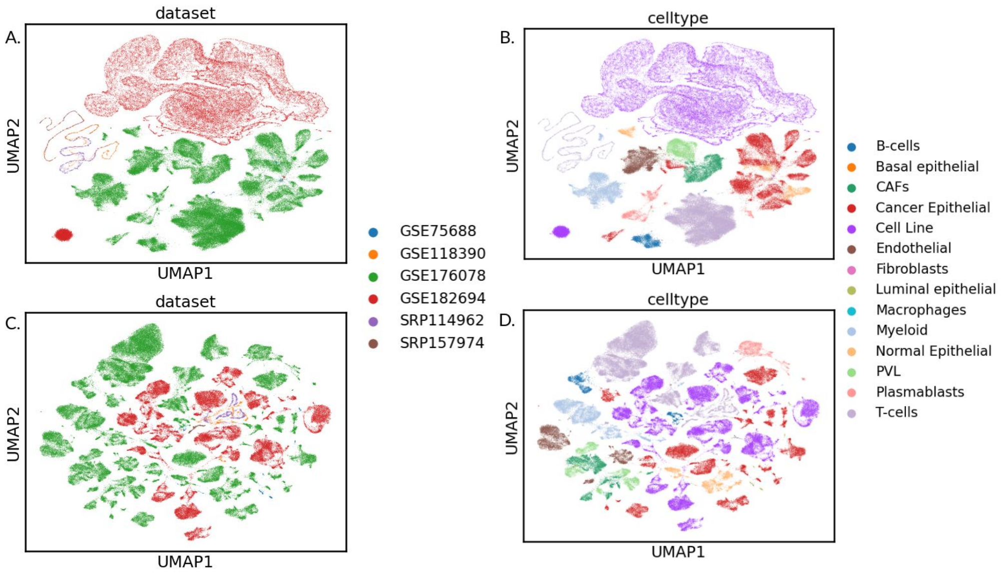

Fig. 1. UMAP embeddings of unintegrated (top) and scANVI-integrated (bottom) TNBC datasets, annotated by dataset (left) and cell type (right).

Graph Connectivity's improvement (0.8011 to 0.8769) [8].

This dedicated balance and controlled trade-off advances our goal: batch-corrected data enables cross-dataset insights and biomarker discovery for personalized treatment mapping. Beyond oncology, the methodology parallels NASA's GeneLab, where multi-omics integration informs cellular responses to space stressors like radiation.

Limitations include partial annotations and computational data integration and data accuracy demands. Sparse or incorrect labels restrict scANVI's semi-supervised potential, while training on 156,794 cells requires significant resources, scaling poorly beyond millions of cells [1]. Multimodal integration (e.g., scATAC-seq) remains unaddressed, limiting epigenetic insights [24]. Future work will incorporate additional datasets, automate annotation using Foundational Models such as scGPT, and extend to multi-omics, enhancing scalability and depth [16]. This prototype TNBC atlas lays a foundation for precision medicine and space biology, bridging terrestrial and extraterrestrial health challenges.

VI. Conclusion

Our team achieved a significant milestone by constructing a preliminary single-cell atlas for triple-negative breast cancer (TNBC), integrating 156,794 cells from six diverse scRNA-seq datasets using scANVI [1]. We actively processed datasets spanning primary tumors, metastases, cell lines, patient cohorts, fresh primary tumors, and chemotherapy-treated residual tumors, addressing batch effects while preserving biological integrity. Our application of scANVI aligned these heterogeneous sources into a shared latent space, as demonstrated by UMAP embeddings shifting from dataset-specific clusters to cell type-driven groupings (Figure 1).

Our integration framework parallels NASA's GeneLab mission, harmonizing multi-omics data to study cellular responses under microgravity and radiation, directly supporting astronaut health for long-duration spaceflight. The approach's scalability and adaptability enhance its relevance to high-risk and high-grade TNBC patients facing early relapse, treatment-resistance, poor outcome and reduced survival in the clinic

We acknowledge limitations in annotation coverage and computational efficiency. Sparse and potentially erroneous labels constrain scANVI's semi-supervised potential, while processing 156,794 cells demands significant resources, hinting at challenges for million-cell scales [1]. Multimodal integration, such as scATAC-seq, remains unexplored. [24].

In the next year, we plan to expand the atlas by incorporating additional TNBC datasets, integrating multi-omics data such as scATAC-seq and proteomics, and enhancing annotation and integration with foundational models such as scGPT to streamline analysis [16]. These steps will transform the preliminary atlas into a comprehensive resource, amplifying its impact on personalized medicine and biological research.

REFERENCES

[1] C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, and N. Yosef, "Probabilistic Harmonization and Annotation of Single-cell Transcriptomics Data with Deep Generative Models," Jan. 2019. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/532895

[2] S. Siddharth and D. Sharma, "Racial Disparity and Triple-Negative Breast Cancer in African-American Women: A Multifaceted Affair between Obesity, Biology, and Socioeconomic Determinants," *Cancers*, vol. 10, no. 12, p. 514, Dec. 2018. [Online]. Available: https://www.mdpi.com/2072-6694/10/12/514

[3] A. Dave, D. Charytonowicz, N. J. Francoeur, M. Beaumont, K. Beaumont, H. Schmidt, T. Zeleke, J. Silva, and R. Sebra, "The Breast Cancer Single-Cell Atlas: Defining cellular heterogeneity within model cell lines and primary tumors to inform disease subtype, stemness, and treatment options," *Cellular Oncology*, vol. 46, no. 3, pp. 603–628, Jun. 2023. [Online]. Available: https://link.springer.com/10.1007/s13402-022-00765-7

[4] B. D. Lehmann, J. A. Bauer, X. Chen, M. E. Sanders, A. B. Chakravarthy, Y. Shyr, and J. A. Pietenpol, "Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies," *Journal of Clinical Investigation*, vol. 121, no. 7, pp. 2750–2767, Jul. 2011. [Online]. Available: http://www.jci.org/articles/view/45014

[5] V. Svensson, R. Vento-Tormo, and S. A. Teichmann, "Exponential scaling of single-cell RNA-seq in the past decade," *Nature Protocols*, vol. 13, no. 4, pp. 599–604, Apr. 2018. [Online]. Available: https://www.nature.com/articles/nprot.2017.149

[6] W. Chung, H. H. Eum, H.-O. Lee, K.-M. Lee, H.-B. Lee, K.-T. Kim, H. S. Ryu, S. Kim, J. E. Lee, Y. H. Park, Z. Kan, W. Han, and W.-Y. Park, "Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer," *Nature Communications*, vol. 8, no. 1, p. 15081, May 2017. [Online]. Available: https://www.nature.com/articles/ncomms15081

[7] M. Karaayvaz, S. Cristea, S. M. Gillespie, A. P. Patel, R. Mylvaganam, C. C. Luo, M. C. Specht, B. E. Bernstein, F. Michor, and L. W. Ellisen, "Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq," *Nature Communications*, vol. 9, no. 1, p. 3588, Sep. 2018. [Online]. Available: https://www.nature.com/articles/s41467-018-06052-0

[8] M. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. Mueller, D. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and F. Theis, "Benchmarking atlas-level data integration in single-cell genomics," May 2020. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2020.05.22.111161

[9] L. Haghverdi, A. T. L. Lun, M. D. Morgan, and J. C. Marioni, "Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors," *Nature Biotechnology*, vol. 36, no. 5, pp. 421–427, May 2018. [Online]. Available: https://www.nature.com/articles/nbt.4091

[10] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, and R. Satija, "Comprehensive Integration of Single-Cell Data," *Cell*, vol. 177, no. 7, pp. 1888–1902.e21, Jun. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0092867419305598

[11] J. D. Welch, V. Kozareva, A. Ferreira, C. Vanderburg, C. Martin, and E. Z. Macosko, "Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity," *Cell*, vol. 177, no. 7, pp. 1873–1887.e17, Jun. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0092867419305045

[12] A. T. L. Lun, K. Bach, and J. C. Marioni, "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts," *Genome Biology*, vol. 17, no. 1, p. 75, Dec. 2016. [Online]. Available: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0947-7

[13] Y. Zhang, G. Parmigiani, and W. E. Johnson, "ComBat-seq: batch effect adjustment for RNA-seq count data," *NAR Genomics and Bioinformatics*, vol. 2, no. 3, p. lqaa078, Sep. 2020. [Online]. Available: https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqaa078/590951

[14] I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P.-r. Loh, and S. Raychaudhuri, "Fast, sensitive and accurate integration of single-cell data with Harmony," *Nature Methods*, vol. 16, no. 12, pp. 1289–1296, Dec. 2019. [Online]. Available: https://www.nature.com/articles/s41592-019-0619-0

[15] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nature Methods*, vol. 15, no. 12, pp. 1053–1058, Dec. 2018. [Online]. Available: https://www.nature.com/articles/s41592-018-0229-2

[16] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo,

N. Duan, and B. Wang, “scGPT: toward building a foundation model for single-cell multi-omics using generative AI,” *Nature Methods*, vol. 21, no. 8, pp. 1470–1480, Aug. 2024. [Online]. Available: https://www.nature.com/articles/s41592-024-02201-0

[17] K. Polański, M. D. Young, Z. Miao, K. B. Meyer, S. A. Teichmann, and J.-E. Park, “BBKNN: fast batch alignment of single cell transcriptomes,” *Bioinformatics*, vol. 36, no. 3, pp. 964–965, Feb. 2020. [Online]. Available: https://academic.oup.com/bioinformatics/article/36/3/964/5545955

[18] B. Hie, B. Bryson, and B. Berger, “Efficient integration of heterogeneous single-cell transcriptomes using Scanorama,” *Nature Biotechnology*, vol. 37, no. 6, pp. 685–691, Jun. 2019. [Online]. Available: https://www.nature.com/articles/s41587-019-0113-3

[19] X. Yan, R. Zheng, F. Wu, and M. Li, “CLAIRE: contrastive learning-based batch correction framework for better balance between batch mixing and preservation of cellular heterogeneity,” *Bioinformatics*, vol. 39, no. 3, p. btad099, Mar. 2023. [Online]. Available: https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btad099/7055295

[20] T. Wang, T. S. Johnson, W. Shao, Z. Lu, B. R. Helm, J. Zhang, and K. Huang, “BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes,” *Genome Biology*, vol. 20, no. 1, p. 165, Dec. 2019. [Online]. Available: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1764-6

[21] X. Zhan, Y. Yin, and H. Zhang, “BERMAD: batch effect removal for single-cell RNA-seq data using a multi-layer adaptation autoencoder with dual-channel framework,” *Bioinformatics*, vol. 40, no. 3, p. btae127, Mar. 2024. [Online]. Available: https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btae127/7619104

[22] R. Danino, I. Nachman, and R. Sharan, “Batch correction of single-cell sequencing data via an autoencoder architecture,” *Bioinformatics Advances*, vol. 4, no. 1, p. vbad186, Jan. 2024. [Online]. Available: https://academic.oup.com/bioinformaticsadvances/article/doi/10.1093/bioadv/vbad186/7502962

[23] A. Shree, M. K. Pavan, and H. Zafar, “scDREAMER for atlas-level integration of single-cell datasets using deep generative model paired with adversarial classifier,” *Nature Communications*, vol. 14, no. 1, p. 7781, Nov. 2023. [Online]. Available: https://www.nature.com/articles/s41467-023-43590-8

[24] Q. Chen, Y. Liu, Y. Gao, R. Zhang, W. Hou, Z. Cao, Y.-Z. Jiang, Y. Zheng, L. Shi, D. Ma, J. Yang, Z.-M. Shao, and Y. Yu, “A comprehensive genomic and transcriptomic dataset of triple-negative breast cancers,” *Scientific Data*, vol. 9, no. 1, p. 587, Sep. 2022. [Online]. Available: https://www.nature.com/articles/s41597-022-01681-z

[25] M. Büttner, Z. Miao, F. A. Wolf, S. A. Teichmann, and F. J. Theis, “A test metric for assessing single-cell RNA-seq batch correction,” *Nature Methods*, vol. 16, no. 1, pp. 43–49, Jan. 2019. [Online]. Available: https://www.nature.com/articles/s41592-018-0254-1

[26] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0377042787901257

[27] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells,” *Nature Biotechnology*, vol. 32, no. 4, pp. 381–386, Apr. 2014. [Online]. Available: https://www.nature.com/articles/nbt.2859

[28] W. Saelens, R. Cannoodt, H. Todorov, and Y. Saeys, “A comparison of single-cell trajectory inference methods,” *Nature Biotechnology*, vol. 37, no. 5, pp. 547–554, May 2019. [Online]. Available: https://www.nature.com/articles/s41587-019-0071-9

[29] M. Papanicolaou, A. L. Parker, M. Yam, E. C. Filipe, S. Z. Wu, J. L. Chitty, K. Wyllie, E. Tran, E. Mok, A. Nadalini, J. N. Skhinas, M. C. Lucas, D. Herrmann, M. Nobis, B. A. Pereira, A. M. K. Law, L. Castillo, K. J. Murphy, A. Zaratzian, J. F. Hastings, D. R. Croucher, E. Lim, B. G. Oliver, F. V. Mora, B. L. Parker, D. Gallego-Ortega, A. Swarbrick, S. O’Toole, P. Timpson, and T. R. Cox, “Temporal profiling of the breast tumour microenvironment reveals collagen XII as a driver of metastasis,” *Nature Communications*, vol. 13, no. 1, p. 4587, Aug. 2022. [Online]. Available: https://www.nature.com/articles/s41467-022-32255-7

[30] R. Vishnubalaji and N. M. Alajez, “Transcriptional landscape associated with TNBC resistance to neoadjuvant chemotherapy revealed by single-cell RNA-seq,” *Molecular Therapy - Oncolytics*, vol. 23, pp. 151–162, Dec. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2372770521001297