stats

MDPI

Article

Refining scRNA-Seq Clusters: The Power of Feature Selection

Ching-Hsuan Chen and Chen-An Tsai *

Department of Agronomy, National Taiwan University, Taipei 106216, Taiwan; miamia.girl1111@gmail.com
* Correspondence: catsai@ntu.edu.tw

Abstract

Feature selection is critical for resolving cell-type heterogeneity in single-cell RNA sequencing (scRNA-seq). DUBStepR (Determining the Underlying Basis using Stepwise Regression) is a widely used gene selection method for scRNA-seq designed to identify feature genes that maximize cell-type separation. DUBStepR has been reported to perform effectively in this domain; however, its reliance on linear Pearson correlation and rigid thresholding limits its effectiveness on complex, high-dimensional datasets. Three enhancements are presented in this study: RFCell-DUBStepR, which uses random forests to capture expression-level importance; Copula-DUBStepR, which models non-linear correlations via Gaussian Copulas; and Zqt-DUBStepR, which utilizes quantile-based selection for improved gene retention. Using both simulated and real-world datasets (scRNA-seq), these modifications are shown to resolve the biases of the original algorithm. The modified methods consistently select a more representative gene set and yield higher clustering accuracy across varying levels of biological complexity. These findings establish the modified DUBStepR frameworks as more reliable tools for high-fidelity subpopulation identification in downstream single-cell analysis.

Keywords: single-cell RNA-seq; DUBStepR; random forest; gaussian copula; clustering

1. Introduction

Single-cell RNA sequencing (scRNA-seq) has emerged as a transformative technique for the quantitative analysis of cellular phenotypes, enabling researchers to explore cellular heterogeneity and identify complex cell types with unprecedented resolution [1]. The typical scRNA-seq workflow involves sophisticated library preparation to quantify transcripts at the single-cell level [2], a process that has become a cornerstone of modern biological research. A critical step in this pipeline is the identification of feature gene sets—subsets of genes selected from raw data to drive cell clustering. The accuracy of these sets is instrumental for the precise identification and annotation of distinct cell subpopulations [3].

Effective unsupervised clustering is the primary vehicle for exploring these cellular differences [4]. The standard pipeline includes quality control [5], normalization [6], dimensionality reduction [7], and similarity assessment [8]. However, the success of this entire process is heavily contingent upon robust feature gene selection. By eliminating "noise" genes and reducing dimensionality, feature selection enhances the signal-to-noise ratio and improves the performance of downstream clustering models. Furthermore, identifying key differentially expressed (DE) genes—those showing significant variation across conditions—provides vital insights into biological pathways, gene regulatory networks, and biomarkers for precision medicine [8–10].

Check for updates

Academic Editors: Tianming Zhu and Jin-Ting Zhang

Received: 7 July 2026
Revised: 29 July 2026
Accepted: 6 August 2026
Published: 11 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1.1. Existing Methodologies and Limitations

Several methods have been developed to address the challenges of feature selection in scRNA-seq data. RFCell [11] integrates permutation and random forest techniques, using Mean Decrease Accuracy (MDA) to rank gene significance. However, RFCell relies on stochastic permutation to generate negative samples and neglects gene-level correlations, a fundamental aspect of genomic structure. Conversely, DUBStepR (Determining the Underlying Basis using Stepwise Regression) [12] was proposed as a correlation-based method to optimize cell-type separation. While DUBStepR has been reported to outperform several existing tools [12], it relies on Pearson correlation coefficients. This presents a notable limitation: Pearson correlation only captures linear relationships [13], whereas scRNA-seq data is characterized by complex, non-linear associations and high sensitivity to outliers [8,14]. Furthermore, DUBStepR often selects a very small number of genes (e.g., 200–500), which may be insufficient compared to the 2000 features typically utilized in the standard single-cell literature [12].

To address non-linearity, tools like scDesign2 [15] have utilized Gaussian copulas to fit correlation matrices and generate high-fidelity synthetic data. By capturing non-linear gene–gene dependencies, copula-based methods have been reported to enhance the reliability of cell clustering [15].

Despite the strengths of DUBStepR, three notable limitations persist:

1. Neglect of Expression Levels: It relies exclusively on correlation, failing to capture the complex patterns inherent in real-world datasets.
2. Scalability Issues: In high-dimensional datasets (e.g., the Ting or Petropoulos datasets), DUBStepR's selection criteria can be overly restrictive, sometimes selecting as few as six genes, leading to unstable clustering.
3. Uneven Selection: The uniform Z-score thresholding used across gene groups results in a skewed distribution, often failing to retain a sufficient number of informative features.

1.2. Proposed Modifications

In this paper, three modifications to the DUBStepR framework are presented to address these challenges:

- RFCell-DUBStepR: Stepwise regression is replaced with a random forest-based ranking method. By generating negative samples through permutation and assessing Mean Decrease Accuracy, this approach considers both gene-wise correlation and raw expression counts.
- Copula-DUBStepR: Linear Pearson coefficients are replaced with Copula Correlation after the initial filtering step, allowing the model to capture non-linear structures and dependencies between genes.
- Zqt-DUBStepR: An adaptive selection strategy is introduced by replacing the uniform Z-score threshold with a quantile-based retention method (top 5%), ensuring a more even and representative selection of genes across different correlation groups.

By evaluating these methods on both simulated and real-world datasets, the proposed modifications are shown to enhance the robustness and accuracy of cell clustering in single-cell analysis.

The remainder of this paper is structured as follows: Section 2 details the materials and our modified DUBStepR frameworks. Section 3 presents the results of the simulation studies. Section 4 discusses the application of our proposed methods to real-life scRNA-Seq datasets. The conclusion and discussion are provided in Section 5.

2. Materials and Methods

In this study, three modifications to DUBStepR: RFCell-DUBStepR, Copula-DUBStepR, and Zqt-DUBStepR are introduced. These methods were evaluated through simulation studies and real single-cell RNA-Seq datasets. Quality control involved removing cells with fewer than 10% of genes expressed and genes expressed in fewer than 10% of cells. The three modified methods were compared against the original DUBStepR, applying a log-normal transformation to the feature-selected matrices. Clustering performance was assessed using SC3 (single-cell consensus clustering) [16] and Consensus Clustering, with evaluation based on the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). Finally, the DE ratio was utilized to assess the efficacy of each method in identifying true differentially expressed (DE) genes.

2.1. The DUBStepR Algorithm

The original DUBStepR [12] process is a multi-step feature selection framework designed to identify an optimal gene set for cell-type separation. The workflow consists of four primary stages: initial filtering, correlation-based feature selection, stepwise regression, and size optimization via the Density Index.

- Initial Filtering and Mean-Based Grouping

DUBStepR first performs a basic quality control step by removing genes expressed in fewer than 5% of the total cell population. For the remaining genes, the mean expression across all cells is calculated. To account for the relationship between mean expression and variance, genes are partitioned into 20 equal-width intervals based on their mean expression values. This grouping strategy ensures that subsequent selection criteria are applied locally, accounting for the inherent heterogeneity in gene expression levels.

- Correlation-Based Feature Selection

Following grouping, a gene–gene Pearson correlation matrix G is constructed. For each gene i, a correlation range c_i is defined to capture its strongest associations while mitigating the influence of outliers:

$$c_i = max_3(G_i) - 0.75 \cdot mean(G_i),$$

where G_i represents the i-th column of the correlation matrix and $max_3(G_i)$ denotes the third-largest value in that column. Within each of the 20 mean-based groups, the Z-score of the c_i values is calculated. Only genes with a Z-score greater than 0.7 are retained for the primary feature set.

- Stepwise Regression and Seed Set Identification

To reduce redundancy within the feature set, DUBStepR then iteratively identifies a "seed set" of genes through a stepwise regression process. The objective is to identify a "seed set" of genes that explains the maximum variance in the correlation matrix. The regression model is defined as:

$$G = gw^T + E$$

where G is the column-wise zero-centered correlation matrix, g is the specific gene column being regressed out, w is the vector of regression coefficients, and E represents the matrix of residuals. In each iteration, the gene that explains the largest proportion of variance is added to the seed set and removed from the matrix; the residuals E then serve as the correlation matrix for the subsequent step. After 30 iterations, the variance explained for the next 70 steps is extrapolated. An "elbow point" e is determined from the resulting scree plot, and these first e genes form the initial core of the feature set.

- Expansion and Optimization

Because the initial elbow point (e) often yields too few genes, DUBStepR expands the feature set by employing a Guilt-by-Association approach, where it identifies the most strongly correlated gene for each gene in the initial set and adds it to the feature list. This step continues until the entire seed set is sequentially ordered. To determine the final, optimal feature set size, DUBStepR uses the Density Index (DI), which measures the separation of cell clusters. Starting from the initial e genes, the algorithm incrementally adds 25 genes at a time. For each iteration, the data is log-normalized, scaled, and reduced via Principal Component Analysis (PCA) (top 20 principal components). The DI is calculated based on the root-mean-squared distances between cells and their neighbors. The subset size that yields the highest DI value is selected as the final optimal feature set.

2.2. RFCell-DUBStepR

DUBStepR often produces a very small optimal gene set, which was observed to be detrimental to clustering performance in real-world datasets—specifically when the subset contains fewer than 10 genes. This limitation originates from the original method's reliance on stepwise regression to determine the initial seed set. Because the scree plot is derived from these regression steps, the resulting elbow point is effectively constrained, often limiting the initial feature pool. Even with the subsequent Guilt-by-Association and Density Index (DI) optimization, this initial bottleneck restricts the diversity of the final feature set.

To address this limitation, RFCell-DUBStepR is presented, replacing the variance-based stepwise regression with RFCell, a Random Forest-based feature selection method. RFCell enhances gene selection by leveraging supervised learning to identify features that distinguish cell types. The algorithm steps are summarized in Figure 1 and the process is as follows:

- Negative Sample Generation: A negative sample matrix is created by randomly permuting the original gene expression data.
- Classification: A Random Forest classifier is trained on a combined dataset of original and permuted samples.
- Importance Evaluation: Mean Decrease Accuracy (MDA) is utilized to quantify gene importance. MDA measures the decline in model classification accuracy when a specific gene's values are permuted.
- Selection: Genes with an MDA score greater than zero are retained, as they provide statistically significant information for cell-type differentiation.

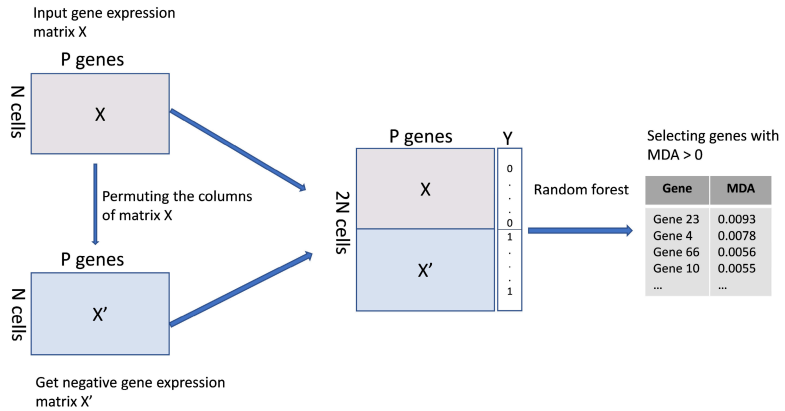


Figure 1. Overview of the full workflow of RFCell. A permutation-based negative control matrix is generated from the original expression matrix, and a random forest classifier trained on the combined real and permuted data is used to rank genes by Mean Decrease Accuracy (MDA); genes with MDA greater than zero are retained as the RFCell-selected feature set used by RFCell-DUBStepR.

Unlike the original variance-based approach, which tends to concentrate importance on a limited number of genes, the distribution of MDA values across the gene set is broader. This produces a more robust scree plot, ensuring the elbow point is not artificially restricted and providing a richer initial pool of candidate genes. By integrating RFCell with DUBStepR, the benefits of correlation-based filtering are combined with the discriminative power of supervised feature selection. The implementation of RFCell-DUBStepR proceeds as follows:

1. Preprocessing: Conduct quality control, calculate the Pearson correlation matrix, and determine the correlation range score for each gene.
2. Filtering: Partition genes into groups by expression level and retain those with a Z-score > 0.7.
3. MDA Ranking: Apply RFCell to the reduced raw count matrix to compute MDA values for all remaining genes.
4. Optimization: Sort genes by MDA and identify the elbow point. Starting from this point, construct candidate sets by iteratively adding 25 genes.
5. Selection: Calculate the Density Index (DI) for each candidate set; the set corresponding to the highest DI value is selected as the optimal feature set.

2.3. Copula-DUBStepR

The original DUBStepR method relies on Pearson correlation coefficients to assess gene–gene relationships. While effective for linear associations, Pearson correlation often fails to capture the complex, non-linear dependencies (e.g., exponential or logarithmic patterns) frequently present in single-cell RNA sequencing (scRNA-seq) data. Furthermore, Pearson correlation is not equipped to account for the heavy zero-inflation characteristic of scRNA-seq counts. To address these limitations, Copula-DUBStepR is introduced. This modification replaces the standard Pearson correlation matrix with a Gaussian Copula correlation matrix. Gaussian Copulas provide a suitable approach in this context, as they model the marginal distribution of each gene separately from the dependence structure, capturing non-linear relationships and providing a more robust measure of correlation [17,18]. The fitting of the Gaussian Copula correlation matrix involves two primary steps:

1. Marginal Distribution Modeling: For each gene, the marginal distribution is modeled based on raw counts across all cells. Four candidate distributions are evaluated: Poisson, Zero-Inflated Poisson (ZIP), Negative Binomial (NB), and Zero-Inflated Negative Binomial (ZINB). The selection of the optimal distribution for each gene is determined by assessing the presence of over-dispersion and the statistical significance of the zero-inflation.
2. Estimation of the Gaussian Copula: After establishing the marginals, the Copula structure is defined by sampling $u_{i,j}$ from a uniform distribution U(0, 1) to handle the discrete nature of the data. The transformed variable $Z_{i,j}$ is calculated as:

$$Z_{i,j} = \Phi^{-1}\big(F_i(X_{i,j}) + u_{i,j}\big(F_i(X_{i,j} - 1) - F_i(X_{i,j})\big)\big),$$

where $X_{i,j}$ is the component of the original count matrix, F_i is the cumulative distribution function (CDF) of the fitted marginal for gene i, and Φ^{-1} is the quantile function of the standard normal distribution. The Gaussian Copula Correlation Matrix R is then estimated using the sample covariance matrix of these transformed variables Z:

$$\hat{R} = corr\big(Z_{i,j}\big),\ i, j = 1, \ldots, n.$$

By replacing the linear Pearson correlation with this copula-based approach, Copula-DUBStepR preserves the underlying biological dependencies that are otherwise obscured

by non-linear noise, leading to more accurate feature selection and improved clustering outcomes in diverse single-cell datasets.

2.4. Zqt-DUBStepR

The original DUBStepR algorithm partitions genes into 20 equidistant groups based on their average expression and applies a fixed Z-score threshold of 0.7 to filter "highly correlated" genes within each group. However, because gene expression levels are not uniformly distributed across these groups, a static threshold is often inappropriate. When the distribution of correlation scores varies significantly between groups, a rigid threshold leads to inconsistencies: some groups may be over-represented with noisy features, while others may lose biologically meaningful genes. This variability often prevents DUBStepR from selecting a sufficient number of representative features, ultimately undermining the accuracy of downstream cell clustering. To address this limitation, Zqt-DUBStepR is introduced, which replaces the static Z-score cutoff with a dynamic, quantile-based selection criterion. Instead of using a universal threshold of 0.7, Zqt-DUBStepR retains the top 0.05 quantile (top 5%) of genes with the highest Z-scores within each group. The benefits of the modified quantile approach are threefold. (1) Adaptive Filtering: By using a quantile threshold, the selection process automatically adapts to the varying density of gene correlations within each expression group. (2) Balanced Feature Set: This approach ensures that highly informative genes are consistently captured from every expression level, rather than being biased toward groups where correlations naturally skew higher. (3) Improved Clustering: By maintaining a more consistent and representative pool of genes, Zqt-DUBStepR preserves the biological signal necessary for the Density Index (DI) to effectively identify optimal cell clusters.

2.5. Clustering Methods

This study employs two primary unsupervised clustering methods, SC3 [16] and Consensus Clustering (CC) [19], to evaluate the performance of the proposed gene selection techniques. Both methods utilize consensus-based strategies to improve stability and reliability.

SC3 is an unsupervised tool specifically optimized for the unique characteristics of single-cell RNA sequencing (scRNA-seq) data. The process involves:

- Distance Computation: It calculates cell–cell distances using three metrics: Euclidean, Pearson, and Spearman.
- Transformation: Each distance matrix is transformed using both Laplacian and PCA methods, creating six distinct representations of the data.
- K-means & Consensus: It performs k-means clustering on each transformed matrix (with k set to the true number of groups) and aggregates these results into a single consensus matrix.
- Hierarchical Labeling: It calculates squared Euclidean distances from the consensus matrix and applies hierarchical clustering to generate the final cluster assignments.

Consensus clustering is implemented via the ConsensusClusterPlus [19] R package. This method assesses cluster stability through repeated subsampling:

- Data Preparation: The input data is log-normalized.
- Repeated Subsampling: The algorithm performs 10 rounds of random sampling, each selecting 80% of the total cells.
- Stability Analysis: In each round, it calculates Euclidean distances and performs k-means clustering. It then tracks how frequently pairs of cells are assigned to the same cluster across these iterations.

- Matrix Scaling: A consensus matrix is constructed based on these frequencies and scaled between 0 (complete disagreement) and 1 (complete agreement).
- Final Labeling: As with SC3, final cell labels are determined by performing hierarchical clustering on the scaled consensus matrix.

2.6. Performance Evaluation

To evaluate the effectiveness of the proposed methods in clustering single-cell data, three primary metrics are utilized: the Adjusted Rand Index (ARI), Normalized Mutual Information (NMI), and the DE ratio. The DE ratio was computed only for the simulated datasets, since ground-truth differentially expressed genes are not available for the real datasets used in this study.

The Adjusted Rand Index (ARI) quantifies the similarity between two partitions (the clustering results and the ground truth) by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted versus true labels. ARI is bounded between -1 and 1, where a score of 1 indicates perfect agreement, 0 suggests agreement expected by chance, and negative values indicate performance worse than random. The ARI is calculated as:

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right]/\binom{n}{2}},$$

where n_{ij} represents the elements of the contingency table; a_i and b_j are the row and column sums of the contingency table, respectively.

Normalized Mutual Information (NMI) is an information-theoretic metric that measures the amount of information shared between the clustering results (A) and the true cell types (B). NMI is normalized to range from 0 to 1, where 1 represents identical clustering and 0 indicates no shared information. The NMI is calculated using the mutual information $I(A,B)$ and the entropy $H(A)$ and $H(B)$:

$$\text{NMI}(A,B) = \frac{2I(A,B)}{H(A) + H(B)}$$

where:

- $P(i)$ and $P(j)$ are the marginal probabilities of clusters in A and B, respectively.
- $P(i,j)$ is the joint probability of cells belonging to cluster i in A and cluster j in B.

$$I(A,B) = \sum_{i,j} P(i,j) log\left(\frac{P(i,j)}{P(i)P(j)}\right)$$

$$H(A) = -\sum_i P(i) log P(i)$$

To evaluate the biological accuracy of the feature selection methods, the DE ratio is used. Since the simulated datasets contain a known proportion of differentially expressed (DE) genes, this ratio measures the enrichment of true DE genes within the final selected feature set. The DE ratio is defined as:

$$\text{DE-ratio} = \frac{\text{Number of true DE genes in the selected feature set}}{\text{Total number of genes in the selected feature set}}$$

2.7. Software Implementation

RFCell-DUBStepR, Copula-DUBStepR, and Zqt-DUBStepR are not pre-existing software packages but modifications of the original DUBStepR algorithm [12], implemented in R 4.0.2 by the authors, that each replace one stage of the DUBStepR pipeline described in Section 2.1. RFCell-DUBStepR replaces the stepwise-regression seed-set step with the

RFCell random forest ranking procedure [11] (Section 2.2); Copula-DUBStepR replaces the Pearson correlation matrix with a Gaussian copula correlation matrix (Section 2.3); and Zqt-DUBStepR replaces the fixed Z-score threshold of 0.7 with a dynamic top-5% quantile threshold applied within each expression group (Section 2.4). For Copula-DUBStepR, the marginal count distributions (Poisson, zero-inflated Poisson, negative binomial, and zero-inflated negative binomial) are fitted per gene in R, and the resulting Gaussian copula correlation matrix is constructed using the correlation-matching approach of Avramidis, Channouf, and L'Ecuyer [17] together with the Nataf transformation formalized by Lebrun and Dutfoy [18]; no dedicated external copula-fitting software package was used. All three variants are available as custom R scripts built around the DUBStepR pipeline rather than as standalone or previously published tools, so readers seeking to reproduce the analyses should treat DUBStepR [12] and RFCell [11] as the underlying published methods being extended.

3. Simulation Studies

3.1. Setup

To evaluate the performance of the proposed methods, comprehensive simulation studies were conducted using the GeneScape R package [20]. This simulator allows for the generation of synthetic scRNA-seq datasets with customizable gene-cell counts, group structures, and differential expression (DE) patterns. GeneScape generates a realistic count matrix by modeling biological mean expression, sequencing depth, and technical dropout. The generation process follows these key steps:

- Baseline Expression: A basic gene mean count matrix (μ) is generated using a $\Gamma(0.3, 0.15)$ distribution.
- Differential Expression: Fold changes for DE genes are applied using a lognormal distribution ($\mu = \alpha_k, \sigma = 0.2$). A distinction is made between "core" genes (specific to a group) and "shared" genes (common across multiple groups).
- Depth Normalization: Sequencing depth is modeled using a lognormal distribution ($\mu = 9.3$, $\sigma = 0.25$). The count matrix is normalized by multiplying the gene mean matrix by the ratio of the sequencing depth to the total sum of the matrix columns.
- Poisson Modeling: The initial count matrix N is generated via a Poisson distribution based on the normalized mean.
- Correlation Structure: To simulate group-specific gene co-expression, multivariate normal vectors are generated for each group. The covariance matrix is calculated by raising the correlation parameter ρ to the power of the number of DE genes (n). These vectors are then converted to counts using the Poisson quantile function.
- Dropout Incorporation: To simulate technical noise, the logistic dropout model defines how dropout rates correlate with mean expression levels with the dropout probability is calculated as:

$$\mathrm{P(dropout)} = [1 + \exp(-\mathrm{slope}\cdot(\mathrm{mean} - \mathrm{location}))]^{-1}.$$

The final observed count matrix X is then adjusted using a binomial distribution $\mathrm{B(N, 1 - P(dropout))}$.

Ten distinct simulation scenarios were configured to assess robustness across varying group sizes and DE proportions (Table 1). A fixed size of 1000 cells and 10,000 genes was maintained for all scenarios, with each simulation repeated 50 times to ensure statistical reliability. By varying the number of groups from 2 to 6 and increasing the correlation coefficients incrementally from 0.4 to 0.9, increasingly complex biological environments are simulated. This design rigorously evaluates how well the modifications—RFCell, Copula,

and Zqt—maintain clustering performance under both low-noise and high-complexity conditions.

Table 1. Summary of experimental simulation settings.

Scenario	Groups (k)	Core Genes (α_k)	Share Genes (α_k)	Correlation (ρ)	DE Proportion
1/2	2	(0.1, 0.5)	(0.3, 0.7)	(0.4, 0.5)	0.03/0.06
3/4	3	(0.1, 0.5, 0.9)	(0.3, 0.7, 1.1)	(0.4, 0.5, 0.6)	0.03/0.06
5/6	4	(0.1, 0.5, 0.9, 1.3)	(0.3, 0.7, 1.1, 1.5)	(0.4, 0.5, 0.6, 0.7)	0.03/0.06
7/8	5	(0.1, 0.5, 0.9, 1.3, 1.7)	(0.3, 0.7, 1.1, 1.5, 1.9)	(0.4, 0.5, 0.6, 0.7, 0.8)	0.03/0.06
9/10	6	(0.1, 0.5, 0.9, 1.3, 1.7, 2.1)	(0.3, 0.7, 1.1, 1.5, 1.9, 2.3)	(0.4, 0.5, 0.6, 0.7, 0.8, 0.9)	0.03/0.06

3.2. Results

The performance of the three proposed modifications was evaluated—RFCell-DUBStepR, Copula-DUBStepR, and Zqt-DUBStepR—against the original DUBStepR method across 10 simulated scenarios. These scenarios were categorized by DE gene proportions (low: 3% vs. high: 6%) and complexity of cell group structures (low: 2–4 groups vs. high: 5–6 groups). Clustering performance was assessed using SC3 and Consensus Clustering, with metrics focused on the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI).

In Figure 2, the SC3 clustering results indicate that the modifications provide distinct advantages depending on dataset complexity:

- Low Group Complexity (2–4 groups): RFCell-DUBStepR achieved higher ARI and NMI values than the other methods. As shown in Figure 2, it achieved the highest median ARI and NMI values with minimal variance, indicating both high accuracy and stability of the resulting clusters.
- High Group Complexity/Low DE Proportion (5–6 groups): Copula-DUBStepR emerged as the most effective method. It provided clearer cluster distinction, achieving the highest ARI for both 5 and 6 groups and the highest NMI for 6 groups, suggesting its robustness in capturing non-linear dependencies when biological signals are subtle.
- High Group Complexity/High DE Proportion (5–6 groups): Performance remained competitive across all methods. While variation increased, RFCell-DUBStepR and Copula-DUBStepR frequently attained the best metrics, reflecting their ability to handle larger pools of informative genes.

In Figure 3, it was observed that the Consensus clustering results reinforced the patterns observed in the SC3 analysis:

- Higher Performance of RFCell-DUBStepR: Across almost all scenarios, particularly those with low to moderate group numbers, RFCell-DUBStepR maintained the highest median ARI and NMI, with consistently low variance across these scenarios.
- High Complexity Resilience: In high-group scenarios with low DE proportions, RFCell-DUBStepR remained the top performer in NMI. While performance was more uniform across all four methods when DE proportions were high, RFCell-DUBStepR retained an advantage in cluster identification for 5-group scenarios.

Overall, the results indicate that RFCell-DUBStepR achieves the highest stability among the modifications, with higher ARI and NMI values than the original DUBStepR across the majority of simulation conditions (Table 2). Copula-DUBStepR provides an advantage in high-complexity environments where capturing non-linear relationships is critical. These findings suggest that the proposed methods improve the accuracy and stability of feature selection for single-cell clustering.

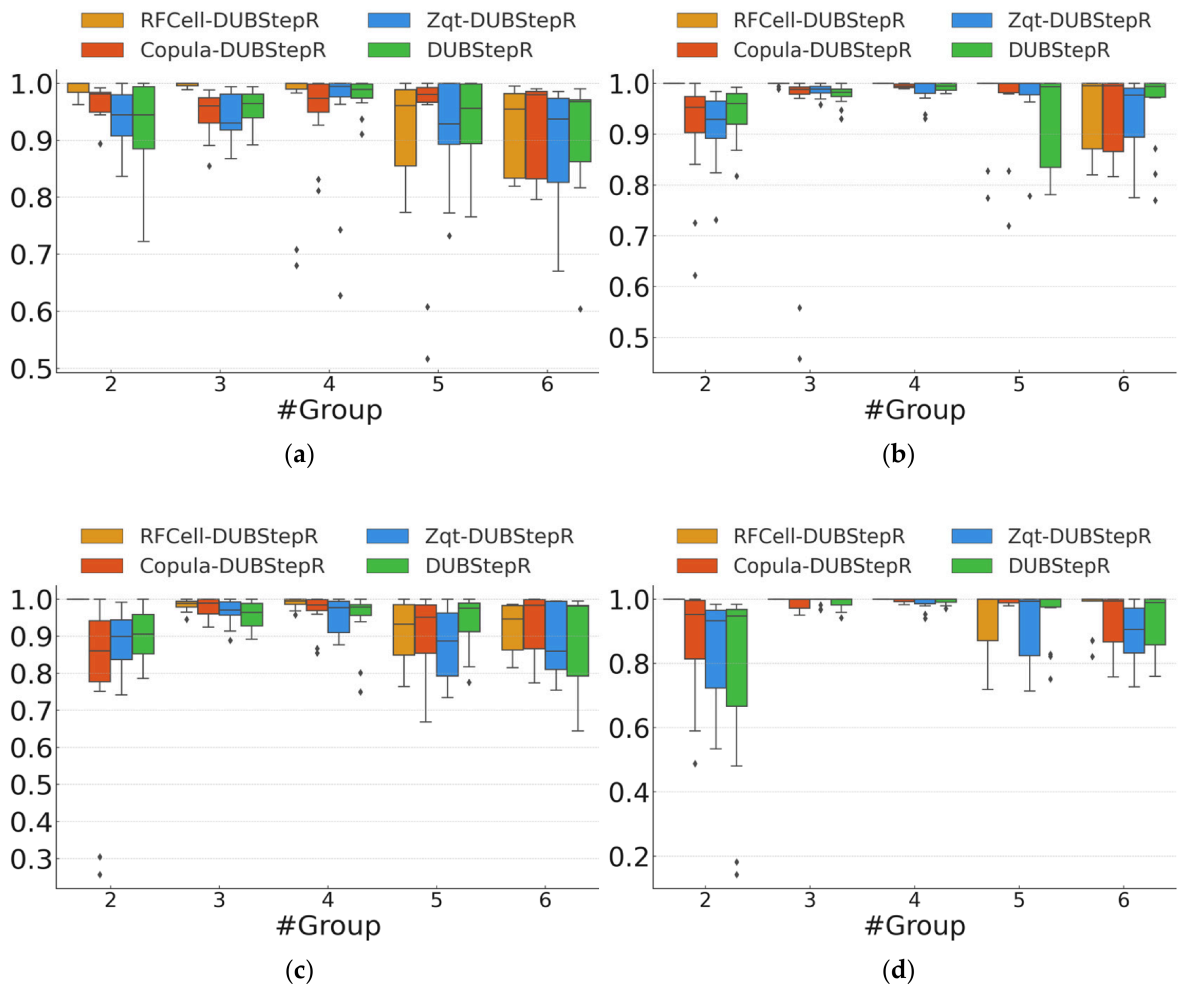


Figure 2. Boxplots of SC3 clustering performance on simulated data. Results are evaluated using the Adjusted Rand Index (ARI) for differentially expressed (DE) gene proportions of 3% (**a**) and 6% (**b**), and using Normalized Mutual Information (NMI) for DE gene proportions of 3% (**c**) and 6% (**d**). At low group complexity, RFCell-DUBStepR attains the highest median ARI and NMI with minimal variance, whereas Copula-DUBStepR becomes the strongest performer as group complexity increases, particularly under low DE-gene proportions.

In addition, the efficacy of RFCell-DUBStepR, Copula-DUBStepR, and Zqt-DUBStepR in selecting biologically meaningful DE genes is evaluated, comparing their performance against the original DUBStepR method. Under a 3% DE-gene proportion (see Figure 4a), RFCell-DUBStepR consistently achieves the highest median DE ratio, maintaining values above 0.3 across all scenarios and peaking above 0.4 in the 6-group configuration. While Zqt-DUBStepR, Copula-DUBStepR, and DUBStepR exhibit similar trends with performance gains after the fourth group, they remain consistently lower than the Random Forest-based approach in complex biological settings. Similarly, with a 6% DE-gene proportion (see Figure 4b), RFCell-DUBStepR shows steady improvement as group complexity increases, reaching a peak near 0.40 at the 6-group level, while Zqt-DUBStepR follows as the second-best performer, leaving Copula-DUBStepR and DUBStepR with notably lower DE ratios below 0.025. Across all group complexities and DE proportions, RFCell-DUBStepR shows higher DE ratios than the alternative methods, consistent with its higher ARI and NMI values reported above, which together identify RFCell-DUBStepR as the method with the highest DE ratio for distinguishing cell types and selecting biologically relevant features among those evaluated.

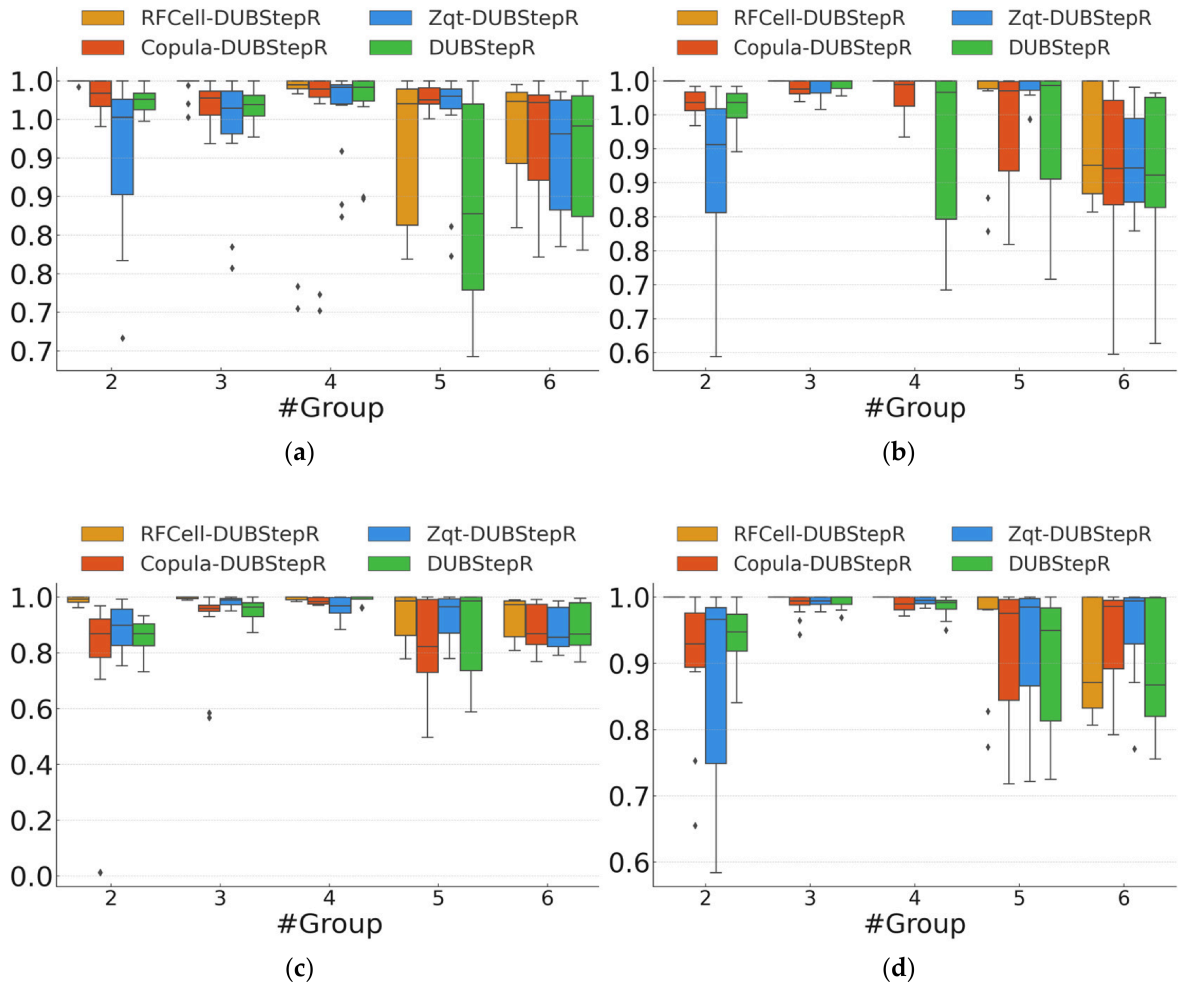


Figure 3. Boxplots of Consensus clustering performance on simulated data. Results are evaluated using the Adjusted Rand Index (ARI) for differentially expressed (DE) gene proportions of 3% (**a**) and 6% (**b**), and using Normalized Mutual Information (NMI) for DE gene proportions of 3% (**c**) and 6% (**d**). These results reinforce the patterns observed in the SC3 analysis: RFCell-DUBStepR maintains the highest median ARI and NMI with consistently low variance across most scenarios, retaining an advantage in 5-group configurations even as performance across methods becomes more uniform under high DE proportions.

Table 2. Summary of DE ratio and clustering performance (ARI/NMI) trends across simulated scenarios.

Method	DE Ratio (3% DE)	DE Ratio (6% DE)	ARI/NMI—Low Complexity (2–4 Groups)	ARI/NMI—High Complexity (5–6 Groups)
RFCell-DUBStepR	>0.3, peak >0.4 (6 groups)	Peak ~0.40 (6 groups)	Highest median, lowest variance	Top NMI (low DE); competitive with Copula-DUBStepR (high DE)
Copula-DUBStepR	Lower than RFCell-DUBStepR; gains after 4-group	<0.025	Lower than RFCell-DUBStepR	Highest ARI (5–6 groups); highest NMI (6 groups) at low DE
Zqt-DUBStepR	Lower than RFCell-DUBStepR; gains after 4-group	Second-best performer	Comparable to Copula-DUBStepR	Competitive at high DE proportion
DUBStepR (original)	Lowest among methods	<0.025	Lowest median ARI/NMI	Lowest performer across most scenarios

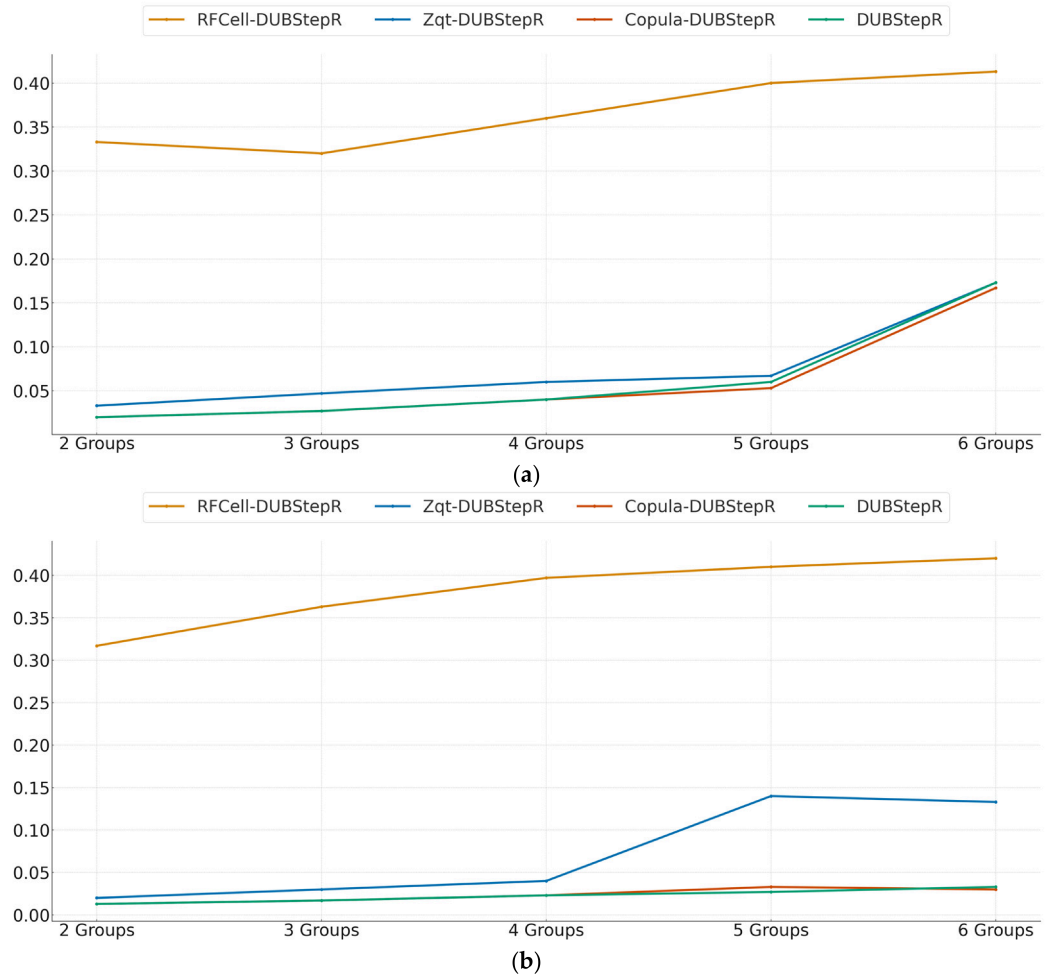


Figure 4. Plots of median DE ratio on simulated data. Results are evaluated using the DE ratio for differentially expressed (DE) gene proportions of 3% (**a**) and 6% (**b**) across varying group complexities. RFCell-DUBStepR consistently attains the highest median DE ratio at both DE proportions, while Copula-DUBStepR and the original DUBStepR yield the lowest DE ratios, particularly under the 6% DE proportion where values fall below 0.025.

4. Application

To evaluate the performance of the proposed gene selection methods, they were tested against the original DUBStepR method using two benchmark datasets: Ting [21] and Petropoulos [22]. Performance was assessed through SC3 and Consensus clustering, with quantitative metrics (ARI and NMI) serving as the primary indicators of clustering accuracy and cell-type resolution.

4.1. The Ting Dataset Analysis

The Ting dataset comprises 29,018 genes and 149 cells across seven cell types, representing circulating tumor cells (CTCs) in a mouse pancreatic cancer model. For this dataset, the cluster number/resolution parameter for both SC3 and Consensus Clustering was set to k = 7, matching the number of known cell types. As summarized in Table 3, the modified methods consistently outperformed the original DUBStepR across all metrics. In SC3 clustering, Copula-DUBStepR achieved the highest performance (ARI: 0.813; NMI: 0.876). In Consensus clustering, Zqt-DUBStepR demonstrated the strongest performance (ARI: 0.664; NMI: 0.754). Notably, the original DUBStepR yielded significantly lower ARI and NMI scores, likely due to its restrictive feature selection process.

Table 3. Comparison of SC3 and consensus clustering performance on the Ting dataset. The respective highest value is highlighted in bold.

Method	SC3		Consensus		Optimal Set Size
	ARI	NMI	ARI	NMI	
RFCell-DUBStepR	0.665	0.769	0.593	0.731	2105
Copula-DUBStepR	**0.813**	**0.876**	0.639	0.735	488
Zqt-DUBStepR	0.681	0.765	**0.664**	**0.754**	529
DUBStepR	0.443	0.546	0.500	0.581	6

In Figure 5, a visualization of the t-distributed stochastic neighbor embedding (t-SNE) [23,24] projections confirms these quantitative findings. While DUBStepR (see Figure 5d) fails to distinguish most cell types—resulting in significant inter-group overlap—the modified methods provide more clearly defined clusters. Specifically, Copula-DUBStepR and Zqt-DUBStepR effectively isolate four distinct cell types (TuMP, nb505, MEF, and MP3/MP27), indicating higher biological resolution than the original method.

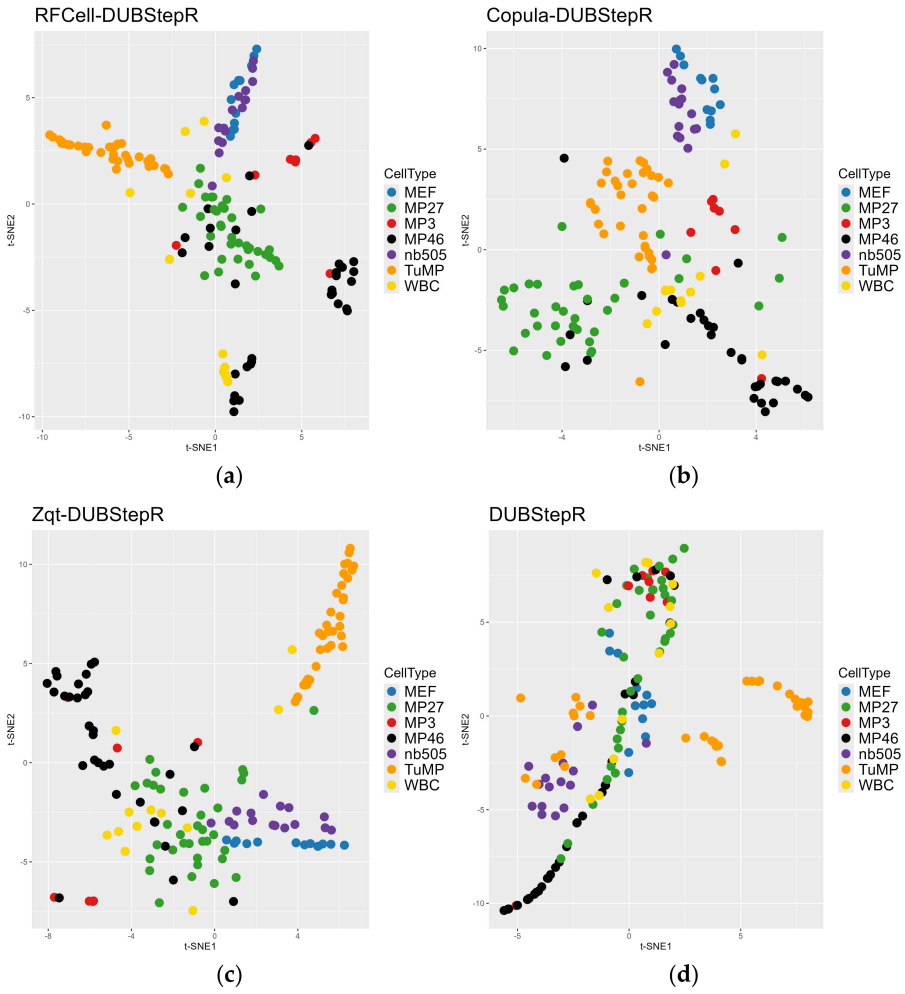
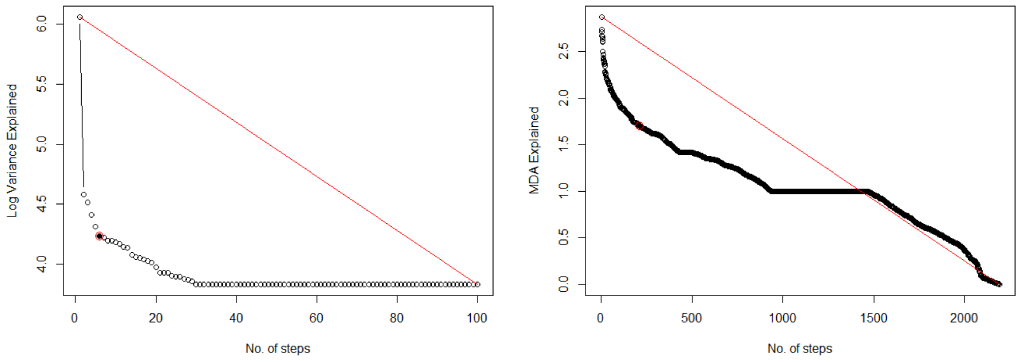


Figure 5. t-SNE visualization of the Ting dataset. (**a**) RFCell-DUBStepR, (**b**) Copula-DUBStepR, (**c**) Zqt-DUBStepR, and (**d**) DUBStepR, showing the gene sets selected by the original and proposed methods. DUBStepR fails to distinguish most cell types, resulting in significant inter-group overlap, whereas Copula-DUBStepR and Zqt-DUBStepR effectively isolate four distinct cell types (TuMP, nb505, MEF, and MP3/MP27), indicating higher biological resolution than the original method.

A critical limitation of DUBStepR is its tendency to select a small number of genes (e.g., only 6 genes in the Ting dataset), which is insufficient to capture the complexity of 29,018 genes. This stems from DUBStepR's reliance on a limited 30-step stepwise re-

gression to determine the "elbow point" in the scree plot (see Figure 6a). In contrast, RFCell-DUBStepR utilizes the Random Forest (RFCell) method to prioritize gene importance, allowing for a more robust identification of the elbow point based on a larger gene pool (see Figure 6b). Consequently, RFCell-DUBStepR identified 2105 optimal genes, leading to substantially higher clustering accuracy and more biologically relevant gene sets.



(**a**) (**b**)

Figure 6. Scree plots for elbow point determination in the Ting dataset. The red line joins the first and last points of the plot and the elbow is the point which is the furthest from the line. (**a**) Elbow point identification based on Log Variance ranking (DUBStepR). (**b**) Elbow point identification based on non-zero MDA ranking (RFCell-DUBStepR). DUBStepR's reliance on a limited 30-step stepwise regression to locate the elbow point yields only 6 selected genes for the Ting dataset, whereas the broader gene pool ranked by RFCell's non-zero Mean Decrease Accuracy (MDA) allows RFCell-DUBStepR to identify 2105 optimal genes.

4.2. The Petropoulos Dataset Analysis

The Petropoulos dataset tracks human embryonic development from day 6 to day 14, containing 21,794 genes across 1529 cells categorized into four developmental groups (TE, PE, EPI, Prelin) across stages E3–E7. For this dataset, the cluster number/resolution parameter for both SC3 and Consensus Clustering was set to k = 4, matching the number of known developmental groups. The results in Table 4 indicate that the proposed methods consistently outperform the original method. In SC3 clustering, Zqt-DUBStepR achieved the highest performance (ARI: 0.455; NMI: 0.521), while in Consensus clustering, Copula-DUBStepR emerged as the top performer. As seen from Figure 7, DUBStepR struggles to separate the later developmental stages (E6 and E7), whereas the modified methods offer clearer differentiation of these dynamic embryonic states.

Table 4. Comparison of SC3 and consensus clustering performance on the Petropoulos dataset. The respective highest value is highlighted in bold.

Method	SC3		Consensus		Optimal Set Size
	ARI	NMI	ARI	NMI	
RFCell-DUBStepR	0.412	0.456	0.377	0.481	3966
Copula-DUBStepR	0.388	0.453	**0.391**	**0.492**	4602
Zqt-DUBStepR	**0.455**	**0.521**	0.364	0.477	403
DUBStepR	0.368	0.424	0.273	0.402	350

The experimental results indicate that Copula-DUBStepR, Zqt-DUBStepR, and RFCell-DUBStepR consistently surpass the original DUBStepR in both clustering accuracy and cell-type recognition. By replacing standard Pearson correlation with more robust statistical approaches and refining the elbow-point selection logic, the modified methods provide a more accurate and comprehensive representation of single-cell transcriptomic heterogeneity.

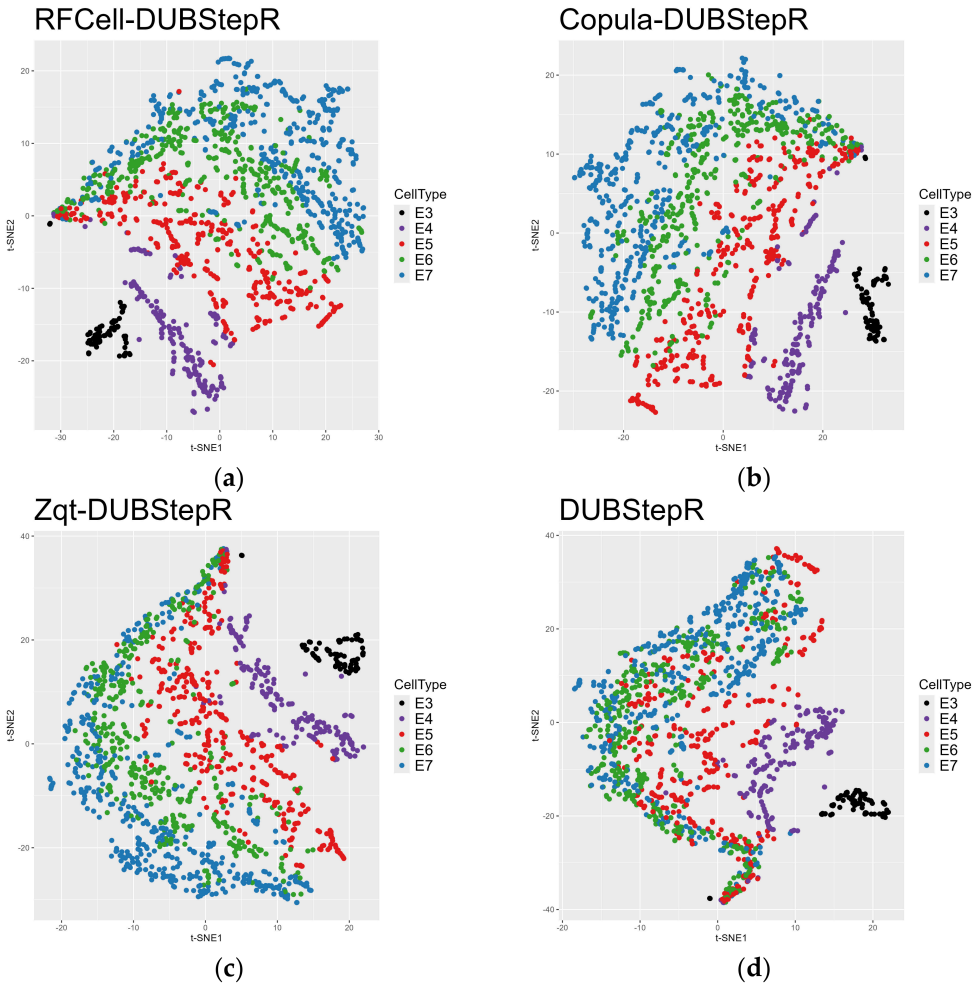


Figure 7. t-SNE visualization of the Petropoulos dataset. (**a**) RFCell-DUBStepR, (**b**) Copula-DUBStepR, (**c**) Zqt-DUBStepR, and (**d**) DUBStepR, showing the gene sets selected by the original and proposed methods. DUBStepR struggles to separate the later developmental stages (E6 and E7), whereas the modified methods offer clearer differentiation of these dynamic embryonic states.

4.3. Biological Relevance of the Improved Cluster Separation

In the Ting dataset, the clearer separation achieved by the proposed variants carries direct biological relevance for circulating tumor cell (CTC) research. CTCs are exceedingly rare and transcriptionally heterogeneous, and their signal is easily masked by contaminating cell populations such as leukocytes and fibroblast- or cell-line-derived controls [21]. By producing feature sets that more cleanly separate the CTC population (TuMP) from contaminating cell groups (e.g., MEF, nb505, and MP3/MP27), the proposed methods make it easier to isolate CTC-specific expression signals that would otherwise be diluted by non-tumor cells in the sample. Because CTCs are considered key mediators of metastatic spread, more reliable computational separation of CTCs from contaminating cell types is directly relevant to downstream efforts to characterize the CTC subpopulations that drive metastasis [21].

In the Petropoulos dataset, the improved resolution of the E6 and E7 developmental stages reflects a biologically meaningful window of human embryonic development. During this period, the human blastocyst undergoes concurrent specification of the trophectoderm (TE), epiblast (EPI), and primitive endoderm (PE) lineages, a process that in humans occurs simultaneously around E5–E7 rather than in the sequential two-step manner reported in mouse embryos [22]. Because these three lineages emerge from a common progenitor population over a narrow developmental window, distinguishing

them computationally is challenging, and the difficulty of the original DUBStepR framework in separating E6 and E7 cells (Figure 7d) is consistent with this biological complexity. The clearer lineage separation achieved by the proposed variants therefore corresponds to a more faithful recovery of the underlying TE/EPI/PE lineage-specification process, which is central to understanding early cell-fate decisions and the loss of pluripotency during human preimplantation development [22].

5. Conclusions

In this study, critical limitations in the DUBStepR framework were identified and addressed—specifically its inability to capture complex non-linear gene dependencies and its tendency to select a small number of features for downstream analysis (e.g., clustering analysis). To address these shortcomings, three enhanced modifications are proposed—RFCell-DUBStepR, Copula-DUBStepR, and Zqt-DUBStepR—each targeting a distinct aspect of the gene selection pipeline. Across both simulated and real single-cell RNA sequencing datasets, all three methods selected a broader set of informative genes and achieved higher cell clustering accuracy than the original DUBStepR.

Specifically, RFCell-DUBStepR integrates both gene correlation structure and expression-level information by leveraging random forest-based gene importance scoring. This approach proved particularly effective in simulated scenarios with a relatively small number of cell types (2–4 groups), achieving the strongest overall clustering performance in those settings. Copula-DUBStepR addresses the non-linearity limitation directly by modelling gene dependencies through a Gaussian Copula framework, enabling the capture of complex correlation structures that linear methods cannot detect. This method outperformed the original DUBStepR in the majority of simulated scenarios involving a larger number of cell types (5–6 groups), where non-linear relationships are more prevalent. Zqt-DUBStepR adopts a quantile-based retention strategy, preserving the top 5% quantile of genes within each group, thereby promoting a more balanced and representative selection of biologically important genes across groups.

Despite these improvements, each of the three proposed variants has method-specific limitations that should be considered when applying them in practice. RFCell-DUBStepR relies on a stochastic negative-sample permutation procedure and Random Forest training, which introduces additional computational overhead relative to the original DUBStepR and may show some sensitivity to the number of trees and the permutation seed; its performance under different training configurations was not systematically explored here. Copula-DUBStepR requires fitting four candidate marginal distributions (Poisson, zero-inflated Poisson, negative binomial, and zero-inflated negative binomial) to every gene before the best-fitting model is selected, which increases runtime compared with the original method and may be sensitive to marginal misspecification for genes with extremely sparse or bimodal expression. Zqt-DUBStepR retains a fixed top-5% quantile threshold for gene selection; although this threshold performed well across the scenarios examined here, its optimal value may depend on dataset size, sparsity, or the number of underlying cell populations, and this sensitivity was not formally investigated.

The simulation study was also conducted under a single fixed setting of 1000 cells and 10,000 genes, so the robustness of these findings to smaller sample sizes has not yet been verified. We would expect RFCell-DUBStepR to be the most sensitive to a reduction in cell number—for example, to 500 cells—since fewer cells per group limit the number of negative-sample permutations available for Random Forest training and could increase the variance of the resulting gene importance scores; Copula-DUBStepR may likewise be affected, as reliably fitting marginal distributions to sparse genes becomes more difficult with fewer observations. In addition, our comparisons were restricted to the original DUBStepR

framework; future studies should also benchmark the proposed variants against other widely used feature selection strategies, such as Seurat's variance-stabilizing transformation for highly variable gene selection [25], the dropout-based M3Drop method [26], and scran's trend-based mean–variance modelling [27], to better position the relative advantages of the proposed approaches within the broader feature selection landscape. Finally, the real-data evaluation in this study was limited to two datasets spanning a mouse pancreatic cancer CTC model and human embryonic development; broader validation across additional tissue types and sequencing platforms would further strengthen the generalizability of these conclusions. These extensions are left as directions for future work.

Overall, the results provide compelling evidence that RFCell-DUBStepR, Copula-DUBStepR, and Zqt-DUBStepR each offer meaningful improvements over the original DUBStepR, with their relative advantages varying according to dataset complexity and cellular heterogeneity. These findings underscore the value of incorporating non-linear correlation modelling and expression-aware weighting into feature selection frameworks for single-cell transcriptomics and lay a strong foundation for future methodological development in this domain.

Author Contributions: Conceptualization, C.-A.T.; methodology, C.-A.T. and C.-H.C.; software, C.-A.T. and C.-H.C.; validation, C.-A.T. and C.-H.C.; formal analysis, C.-A.T. and C.-H.C.; investigation, C.-A.T. and C.-H.C.; resources, C.-A.T.; data curation, C.-A.T. and C.-H.C.; writing—original draft preparation, C.-A.T. and C.-H.C.; writing—review and editing, C.-A.T. and C.-H.C.; visualization, C.-A.T. and C.-H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by a grant from Taiwan's National Science and Technology Council (NSTC 113-2118-M-002-009-).

Data Availability Statement: The data used in the simulation study are generated using the GeneScape R package version 1.0. The raw read count dataset in Section 4 for the Petropoulos study is deposited in the ArrayExpress database under the accession number E-MTAB-3929 and the dataset for the Ting study is available on the National Institutes of Health under the accession number GSE51372.

Acknowledgments: During the preparation of this manuscript, the authors used Gemini (Google Gemini 2.0) to assist with text refinement and consistency checks. The authors reviewed and edited all outputs and take full responsibility for the content.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hwang, B.; Lee, J.H.; Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.* **2018**, *50*, 1–14. [CrossRef] [PubMed]
2. Adil, A.; Kumar, V.; Jan, A.T.; Asger, M. Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis. *Front. Neurosci.* **2021**, *15*, 591122. [CrossRef] [PubMed]
3. Su, K.; Yu, T.; Wu, H. Accurate feature selection improves single-cell RNA-seq cell clustering. *Brief. Bioinform.* **2021**, *22*, bbab034. [CrossRef] [PubMed]
4. Kiselev, V.Y.; Andrews, T.S.; Hemberg, M. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nat. Rev. Genet.* **2019**, *20*, 273–282. [CrossRef] [PubMed]
5. Jiang, P. Quality Control of Single-Cell RNA-seq. *Methods Mol. Biol.* **2019**, *1935*, 1–9. [CrossRef] [PubMed]
6. Vallejos, C.A.; Risso, D.; Scialdone, A.; Dudoit, S.; Marioni, J.C. Normalizing single-cell RNA sequencing data: Challenges and opportunities. *Nat. Method.* **2017**, *14*, 565. [CrossRef] [PubMed]
7. Slovin, S.; Carissimo, A.; Panariello, F.; Grimaldi, A.; Bouché, V.; Gambardella, G.; Cacchiarelli, D. Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview. *Methods Mol. Biol.* **2021**, *2284*, 343–365. [CrossRef] [PubMed]
8. Cheng, C.; Chen, W.; Jin, H.; Chen, X. A Review of Single-Cell RNA-Seq Annotation, Integration, and Cell-Cell Communication. *Cells* **2023**, *12*, 1970. [CrossRef] [PubMed]

9. Wang, T.; Li, B.; Nelson, C.E.; Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinform.* **2019**, *20*, 40. [CrossRef] [PubMed]
10. Townes, F.W.; Hicks, S.C.; Aryee, M.J.; Irizarry, R.A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **2019**, *20*, 295. [CrossRef] [PubMed]
11. Zhao, Y.; Fang, Z.Y.; Lin, C.X.; Deng, C.; Xu, Y.P.; Li, H.D. RFCell: A Gene Selection Approach for scRNA-seq Clustering Based on Permutation and Random Forest. *Front. Genet.* **2021**, *12*, 665843. [CrossRef] [PubMed]
12. Ranjan, B.; Sun, W.; Park, J.; Mishra, K.; Schmidt, F.; Xie, R.; Alipour, F.; Singhal, V.; Joanito, I.; Honardoost, M.A.; et al. DUBStepR is a scalable correlation-based feature selection method for accurately clustering single-cell data. *Nat. Commun.* **2021**, *12*, 5849. [CrossRef] [PubMed]
13. Schober, P.; Boer, C.; Schwarte, L.A. Correlation Coefficients: Appropriate Use and Interpretation. *Anesth. Analg.* **2018**, *126*, 1763–1768. [CrossRef] [PubMed]
14. Hou, J.; Ye, X.; Feng, W.; Zhang, Q.; Han, Y.; Liu, Y.; Li, Y.; Wei, Y. Distance correlation application to gene co-expression network analysis. *BMC Bioinform.* **2022**, *23*, 81. [CrossRef] [PubMed]
15. Sun, T.; Song, D.; Li, W.V.; Li, J.J. scDesign2: A transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured. *Genome Biol.* **2021**, *22*, 163. [CrossRef] [PubMed]
16. Kiselev, V.Y.; Kirschner, K.; Schaub, M.T.; Andrews, T.; Yiu, A.; Chandra, T.; Natarajan, K.N.; Reik, W.; Barahona, M.; Green, A.R.; et al. SC3: Consensus clustering of single-cell RNA-seq data. *Nat. Methods* **2017**, *14*, 483–486. [CrossRef] [PubMed]
17. Avramidis, A.N.; Channouf, N.; L'Ecuyer, P. Efficient correlation matching for fitting discrete multivariate distributions with arbitrary marginals and normal-copula dependence. *INFORMS J. Comput.* **2009**, *21*, 88–106. [CrossRef]
18. Lebrun, R.; Dutfoy, A. An innovating analysis of the Nataf transformation from the copula viewpoint. *Probabilistic Eng. Mech.* **2009**, *24*, 312–320. [CrossRef]
19. Wilkerson, M.D.; Hayes, D.N. ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics* **2010**, *26*, 1572–1573. [CrossRef] [PubMed]
20. Gao, Q. *GeneScape: Simulation of Single Cell RNA-Seq Data with Complex Structure*, R package version 1.0; Comprehensive R Archive Network (CRAN): Vienna, Austria, 2023. Available online: https://cran.r-project.org/package=GeneScape (accessed on 28 July 2026).
21. Ting, D.T.; Wittner, B.S.; Ligorio, M.; Jordan, N.V.; Shah, A.M.; Miyamoto, D.T.; Aceto, N.; Bersani, F.; Brannigan, B.W.; Xega, K.; et al. Single-cell RNA sequencing identifies extracellular matrix gene expression by pancreatic circulating tumor cells. *Cell Rep.* **2014**, *8*, 1905–1918. [CrossRef] [PubMed]
22. Petropoulos, S.; Edsgärd, D.; Reinius, B.; Deng, Q.; Panula, S.P.; Codeluppi, S.; Plaza, R.A.; Linnarsson, S.; Sandberg, R.; Lanner, F. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **2016**, *165*, 1012–1026. [CrossRef] [PubMed]
23. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
24. Linderman, G.C.; Steinerberger, S. Clustering with T-SNE, provably. *SIAM J. Math. Data Sci.* **2019**, *1*, 313–332. [CrossRef] [PubMed]
25. Hafemeister, C.; Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **2019**, *20*, 296. [CrossRef] [PubMed]
26. Andrews, T.S.; Hemberg, M. M3Drop: Dropout-based feature selection for scRNASeq. *Bioinformatics* **2019**, *35*, 2865–2867. [CrossRef] [PubMed]
27. Lun, A.T.L.; McCarthy, D.J.; Marioni, J.C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research* **2016**, *5*, 2122. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.