SIN-CHN SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(4), 1465.
https://doi.org/10.62617/mcb1465

Article

Analysis of biological information detection technology based on the integration of non-parametric statistics and machine learning

Yang Hu*, Shuyuan Liu, Xingzhen Xu

College of Arts and Sciences, Northeast Agricultural University, Harbin 150006, China
*** Corresponding author:** Yang Hu, huyang985111@gmail.com

CITATION

Hu Y, Liu S, Xu X. Analysis of biological information detection technology based on the integration of non-parametric statistics and machine learning. Molecular & Cellular Biomechanics. 2025; 22(4): 1465. https://doi.org/10.62617/mcb1465

ARTICLE INFO

Received: 25 January 2025
Accepted: 4 March 2025
Available online: 21 March 2025

Abstract: This study is based on breast cancer data from the National Cancer Institute (NCI) database, focusing on triple-negative breast cancer (n = 200) and LumB subtype breast cancer (n = 400). A data generation and analysis process combining non-parametric statistics and machine learning was designed. In the initial stage, the wgain algorithm was developed by integrating Wasserstein Generative Adversarial Networks (WGAN) and Random Forest algorithms. The generated expanded dataset was consistent with the original data, with a Pearson correlation coefficient of approximately 0.9, and Principal Component Analysis (PCA) confirmed the high accuracy and consistency of the generated data. The optimal threshold for differential gene selection was determined using the High-Confidence (HC) high-order identification method, and significance analysis was performed using rank-sum tests, Kolmogorov-Smirnov (K-S) tests, and edgeR tests. The results indicated that the rank-sum test performed the best (False Discovery Rate (FDR) = 0.099). A comparison with GAN and Wasserstein GAN Gradient Penalty (WGAN-GP) algorithms showed that wgain had a significant advantage in data consistency and differential gene reproduction (accuracy 83%). This study demonstrates the advantages of combining non-parametric statistics with machine learning, providing a new method for biological data generation and precise analysis.

Keywords: rank-sum test; K-S test; HC high-order identification method; bioinformatics; wgain algorithm

1. Introduction

Generative Adversarial Networks (GANs) [1] have transformed gene expression analysis, particularly for small-sample datasets. While Marouf et al. employed WGAN-GP with gradient penalties for transcriptomic data, our wgain method extends this approach to tissue- and organ-specific bulk Ribonucleic Acid (RNA)-seq and microarray data. We further enhance biological relevance through word embeddings for categorical covariates. Traditional tools like SynTReN and GeneNetWeaver [2] focus on algorithm validation rather than small-sample analysis. Although DESeq2 and edgeR excel in low-replication scenarios, their efficacy diminishes with larger datasets. Recent deep learning approaches, such as GANs [3], address these limitations by capturing non-linear relationships. Building on this, wgain integrates GANs with Random Forests to improve differential gene expression (DEG) detection and pathway enrichment. Validated on public datasets, wgain replicates large-scale study patterns using limited samples [4]. Our enhanced Wasserstein Generative Adversarial Inference Network (WGAIN) method further refines DEG detection via Random Forests, balancing accuracy and stability. Challenges persist, including computational demands and training instability. This study introduces a hybrid framework combining multiple GAN variants with Random Forests and non-parametric methods (e.g., rank-

sum tests [5] and HC optimization [6]) to enhance DEG identification in small-sample contexts. Data preprocessing included normalization and batch correction, while GANs [7] generated synthetic samples to augment datasets [8]. Random Forests [9] improved DEG identification robustness, and non-parametric methods [10] optimized selection thresholds [11]. Validation via cross-validation and statistical tools confirmed the method's efficacy, overcoming small-sample limitations and advancing DEG analysis.

2. Materials and methods

2.1. Rank-sum test

The rank-sum test is a non-parametric method comparing two independent sample distributions without assuming normality. Key implementations include the Mann-Whitney U and Wilcoxon rank-sum tests, widely used for gene expression analysis in small or non-normal datasets.

1) Basic steps of the rank-sum test

The core principle of the rank-sum test is to assess whether the overall distributions of two independent sample groups are significantly different by calculating the sum of ranks for each group. The specific steps are as follows:

Combine the two sample groups: Merge the data from the two sample groups into one combined dataset, removing any group labels.

Ranking: Sort the combined dataset and assign a rank to each data point. If there are tied values in the data, assign the same rank to them, and the average of these ranks will be assigned as their rank.

Calculate the rank sums: Calculate the sum of ranks for each of the two groups of data.

Calculate the test statistic: Based on the rank sums, compute the test statistic and perform the hypothesis test.

2) Mann-Whitney U Test

The Mann-Whitney U Test is one of the most common Rank-Sum Test methods used to compare the differences between two independent samples and to test the null hypothesis that the two samples come from the same population distribution. The test is designed to determine whether there is a significant difference between the two groups. $X = \{x1, x2, x3, \ldots, xn\}$ and $y = \{y1, y2, y3, \ldots, yn\}$. Merge all the data from X and y, and sort the combined data in ascending order.

3) Calculate the rank sums R_X and R_Y

Calculate the rank sums for each sample group, denoted as R_X and R_Y, respectively:

$$U_x = R_x - \frac{n(n-1)}{2},$$

$$U_Y = R_Y - \frac{n(n-1)}{2},$$

where U_x and U_Y are the statistical measures for the X-group and Y-group, respectively [12].

4) Determine significance

The final test statistic is u which is the smaller of U_x and U_Y. If the value of U is small, the null hypothesis can be rejected, indicating that there is a significant difference between the two datasets. The distribution of this statistic approximates a standard normal distribution, allowing for a Z-test based on its value.

$$Z = \frac{U - \mu_U}{\sigma_U},$$

where μ_U and σ_U are the expected value and standard deviation of the statistic u:

$$\mu_U = \frac{mn}{2},$$

$$\sigma_U = \sqrt{\frac{mn(m+n+1)}{12}}.$$

Based on the Z value, we can look up the corresponding standard normal distribution to determine the p-value and thus make a decision on whether to reject the null hypothesis [13].

2.2. HC higher-order identification method

The HC higher-order identification method is typically combined with machine learning techniques for pattern recognition and classification in complex datasets. In certain specific fields, this method can be used to determine the threshold range for p-values. It is a data-driven approach that automatically extracts features from the data and makes inferences without relying on traditional hypothesis testing methods. The HC algorithm we use is based on the work by David Donoho and Jin, in which the e Influential Features PCA (IF-PCA) algorithm employs the KS test [14] to output p-values. These p-values are sorted, and the k-th HC score is computed, resulting in an improved HC identification algorithm. In our case, we modify this algorithm to integrate the rank-sum test. The steps for the rank-sum test within the HC identification process are as follows:

1) For each test, compute a test statistic and calculate the corresponding p-value based on the test statistic.

2) Rank the p-values, such that $\pi(1) < \pi(2) < \ldots < \pi(p)$;

3) Calculate the k-th HC score, corresponding to the second-order z-score.

$$HC_{p,k} = \sqrt{p}\left[\frac{\frac{k}{p} - \pi_k}{\sqrt{\pi_k - (1-\pi_k)}}\right].$$

4) Take the maximum value and calculate the corresponding $HCp* = \max 1 \leq k \leq p\ \alpha_0\{HC, k\}$, find the corresponding k, and consider the largest values, rejecting all $H(i)$, $i = 1, 2, \ldots, k$.

2.3. Combining random forest with generative adversarial networks—Wgain

Wgain (Weight Generative Adversarial Networks Integrated with Random Forest) is an innovative method that combines Generative Adversarial Networks (GANs) with a Random Forest classifier. The goal is to enhance the accuracy and

stability of gene expression pattern recognition by increasing data volume and diversity. The main implementation steps are as follows:

1) Training the Generative Adversarial Network (GAN) Model

In wgain, we use an ensemble model consisting of 10 GANs. Each GAN consists of two parts: the discriminator and the generator. The generator learns gene expression patterns specific to a particular phenotype from Z-normalized gene expression data and generates synthetic samples that resemble real data. The discriminator, on the other hand, accepts both real and generated data, with the task of distinguishing between the two. The generator and discriminator are trained alternately, and after 2000 training epochs, the generator can produce synthetic data that is almost indistinguishable from real gene expression data, greatly expanding the quantity and diversity of the training samples.

2) Restoring Gene Expression Scale

Since the synthetic data output by the GAN is Z-normalized, wgain further uses the mean and standard deviation from public datasets to perform a correction for the gene expression scale. The Random Forest classifier is used to classify each sample in the public dataset, determining its similarity to the target phenotype. Then, based on the mean and standard deviation of gene expression from the public samples, the absolute gene expression values of the synthetic data are restored, ensuring that the generated data is consistent in scale with the real data.

3. Research design

Data Preprocessing: The original dataset consists of 19,113 genes and includes 200 LumA samples and 400 LumB samples. For the data generation experiment, small sample sizes of 20 vs. 40 were selected. After removing low-expressed genes, 16,131 genes remained. The data were then transformed using the limma-voom method in the R package to ensure proper normalization and make the data suitable for the subsequent analysis.

To ensure the quality of the generated data, the values of each gene within the same class were standardized, resulting in a distribution with a mean of 0 and a standard deviation of 1. To enhance the accuracy and diversity of the generated data, multiple Generative Adversarial Network (GAN) architectures were used for data generation. A total of five GAN models were employed. WGAN-GP was used as the primary generator architecture for data generation in Class 1 and Class 2. Since WGAN-GP is an unsupervised learning model, the data for Class 1 and Class 2 were generated using different models.

The WGAN-GP architecture consists of a generator and a discriminator. The generator has three hidden layers with 256, 512, and 1024 nodes, while the discriminator has one hidden layer with 512 nodes. Each hidden layer uses a Leaky ReLU activation function with a slope of 0.5 to enhance the model's nonlinear expression ability. The input layer of the generator consists of 128 nodes, corresponding to a 128-dimensional Gaussian distribution vector. The number of nodes in the output layer of the generator and the input layer of the discriminator matches the total number of genes. The output layer of the discriminator has 256 nodes.

Additionally, to further enhance the diversity of the generated data, other types of GAN models were employed, including WGAN (Wasserstein GAN). Each GAN model was trained separately in different classes (such as Class 1 and Class 2) to generate more diverse sample data. The use of these models effectively avoided the homogeneity of the generated data and improved the quality of the final generated dataset.

During training, both the generator and the discriminator were optimized using the Adam optimizer. Specifically, the learning rate was set to 0.0002, and the gradient penalty coefficient was set to 10. The Adam optimizer combines momentum and adaptive learning rates, effectively handling sparse gradients and non-stationary objectives. In this experiment, the β1 was set to 0.5, and β2 was set to 0.999 to improve training stability. Each batch contained 2 samples. As pointed out by Gulrajani et al. [15], the training process included 2000 epochs, where the discriminator's parameters were updated every five iterations, and the generator's parameters were updated once per epoch. After training the WGAN-GP, the output data from the generator were restored to the log Counts Per Million (logCPM) range to ensure that the generated data were both usable and interpretable.

To enhance the sample size of the generated data, the sample size was set as an integer multiple of the real data, ranging from 1 to 20 times. For each multiplier, to minimize randomness in the generated data, five different datasets were generated after training WGAN-GP. For each multiplier, the generated data from Class 1 and Class 2 were merged to form five generated datasets, and the average Pearson correlation coefficient between these generated datasets and the real dataset was calculated.

Additionally, a random forest model was used for classification analysis. Random forests, as an ensemble learning method, were used to classify the generated and real data, further confirming the similarity of gene expression patterns between the generated and real data. The training process of the random forest model used samples generated from the augmented dataset, with the final classification accuracy serving as an important indicator of the quality of the generated data. After multiple training and cross-validation iterations, the random forest model showed stable performance in the classification task, further validating the reliability and application potential of the generated datasets. As illustrated in **Figure 1**, the experimental workflow diagram provides a clear depiction of the primary experimental steps.

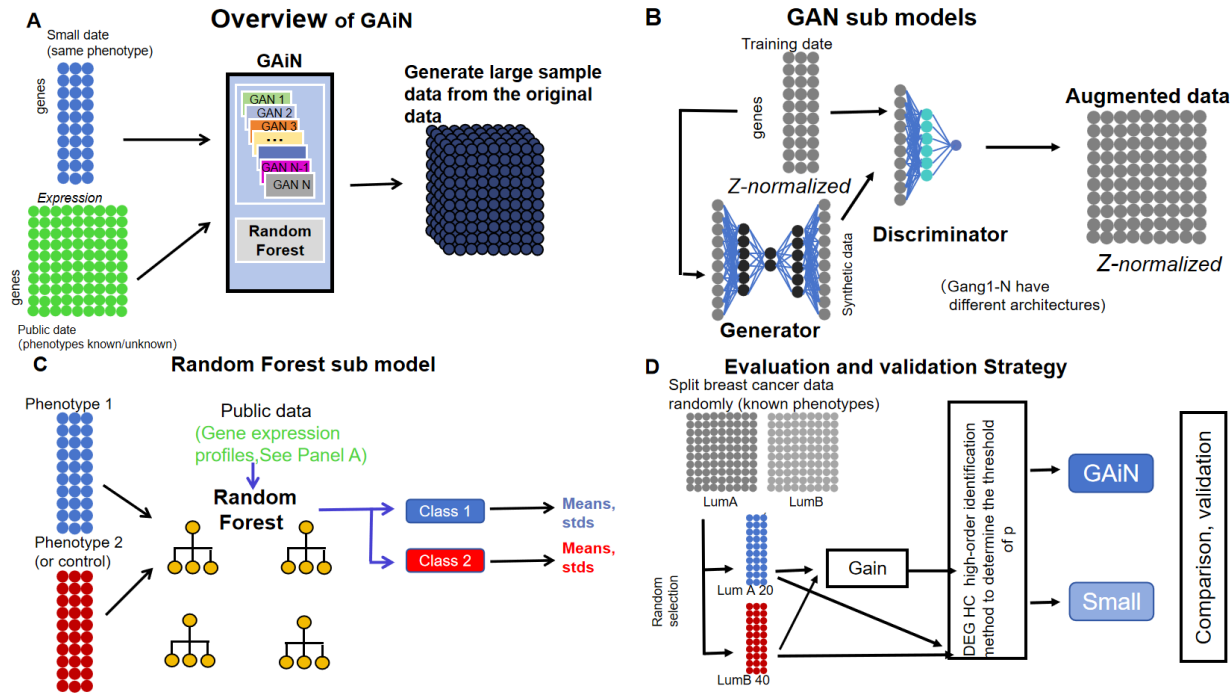

Figure 1. Experimental workflow diagram: **(a)** Overview of GAiN; **(b)** the principle of GAN; **(c)** the process of Random Forest; **(d)** the experimental process.

4. Evaluation of the quality of generated data

4.1. Pearson correlation coefficient

To validate the reliability of the data generation method, 20 and 40 samples were randomly selected from two sample groups to create small-scale datasets. These datasets were then used to generate expanded data through wgain. Subsequently, the Pearson correlation coefficients between the generated data and the original data were evaluated. As the multiple of generated data increased, the average Pearson correlation coefficient between the generated data and the original data showed a gradual upward trend, eventually stabilizing after reaching a certain multiple. This result indicates that the data generated by wgain effectively preserves the feature distribution of the original data and achieves high-quality extension of gene expression data under small sample conditions. This anviding support for subsequent gene differential expression analysis and model construction. As depicted in **Figure 2**, the average Pearson correlation coefficient plot provides a visual representation of the correlation analysis results.

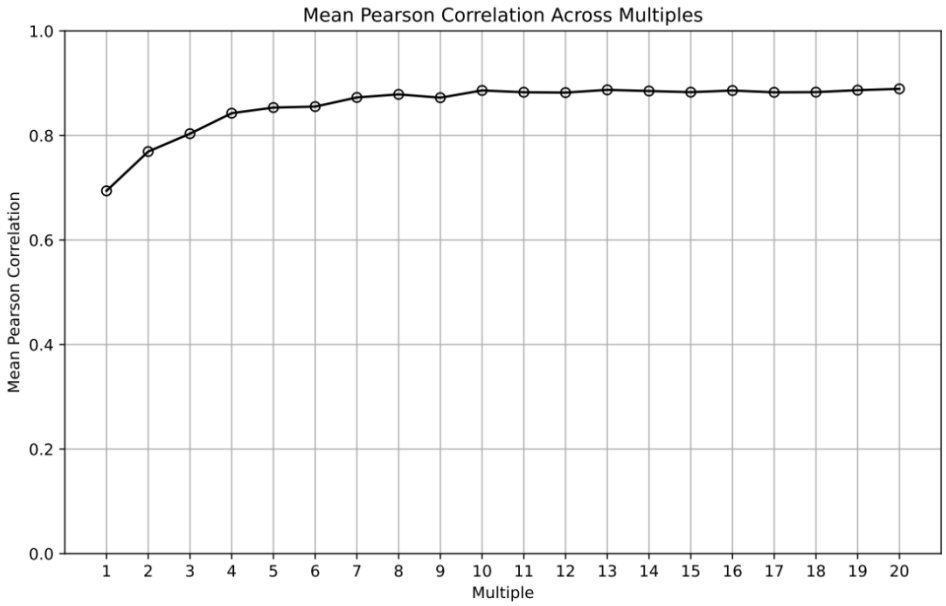

Figure 2. Average Pearson correlation coefficient plot.

4.2. Principal component analysis

In the two-dimensional visualization based on principal component analysis, the red points represent the original data, and the gray points represent the generated data. From the plot, it can be observed that the distribution pattern of the generated data

closely matches that of the original data in the principal component space, indicating that the generated data effectively simulates the main structure and features of the original data.

Regarding the distribution range of the first and second principal components, the generated data's distribution range covers the main area of the original data, with no significant deviations. This suggests that the generated data performs well in capturing the global distribution trends of the original data. Additionally, the density and point distribution characteristics of the generated data in the principal component space are highly similar to those of the original data, further confirming its effectiveness in preserving the local characteristics of the data.

In the overlapping region between the generated and original data, the distribution density of the gray points and red points shows high consistency, indicating that the generated data successfully replicates the statistical features of the original data. The distribution of the generated data in the boundary region also does not exhibit any abnormal or unreasonable clustering, demonstrating the stability of the generation model in handling the boundaries. **Figures 3** and **4** present the principal component plots of the original and generated data, respectively, providing a comparative visualization of their underlying data structures.

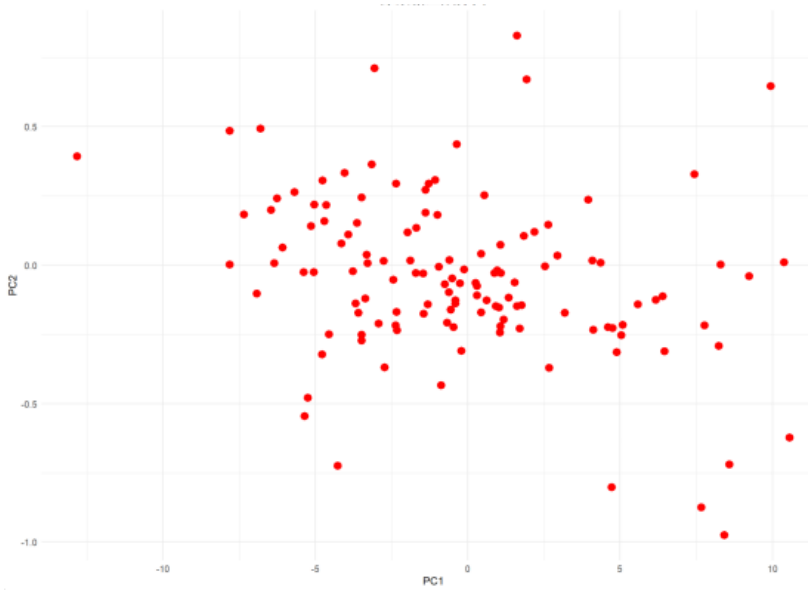

Figure 3. Principal component plot of original data.

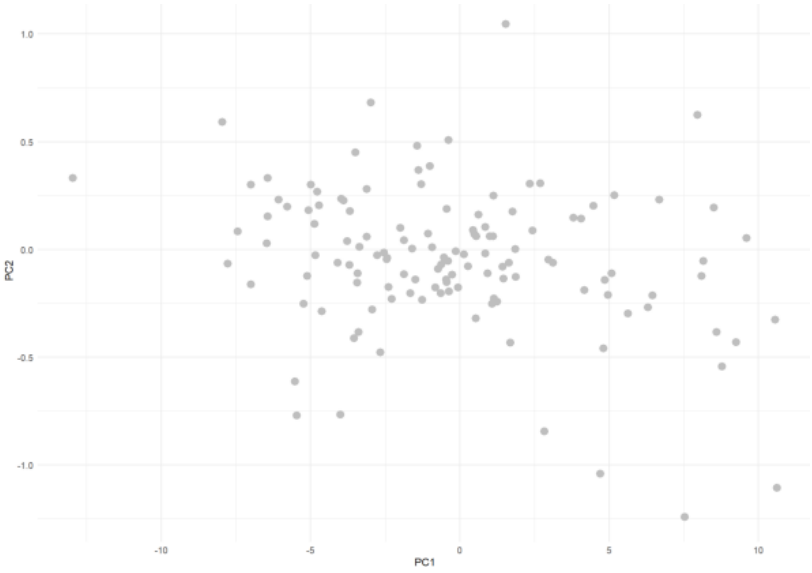

Figure 4. Principal component plot of generated data.

5. Generation data testing

5.1. Non-parametric statistics combined with HC higher-order identification method

For the generated data, we conducted significance analysis of gene expression data using Wilcoxon rank-sum test, K-S test, and the traditional biological method edgeR. To do this, we ranked the test statistics using the HC method and conducted a comprehensive analysis of their p-value distribution and HC corrected statistics. The experimental results showed that the null hypothesis was rejected at $p = 0.00049$, $p = 0.00049$, $p = 0.00049$, $p = 0.00042$, $p = 0.00042$, $p = 0.00042$, and $p = 0.005$, $p = 0.005$, $p = 0.005$, indicating significant statistical differences for the corresponding gene features. Further analysis of the trend in HC corrected statistics revealed that the HC value peaked when the number of features reached 3876, indicating the optimal signal

strength for this feature set. In contrast, the K-S test results indicated that the corresponding number of features was 6123, while the edgeR method indicated 6318 features. The results are shown in the figure below. The rank-sum test achieved optimal FDR control (**Figures 5–8**).

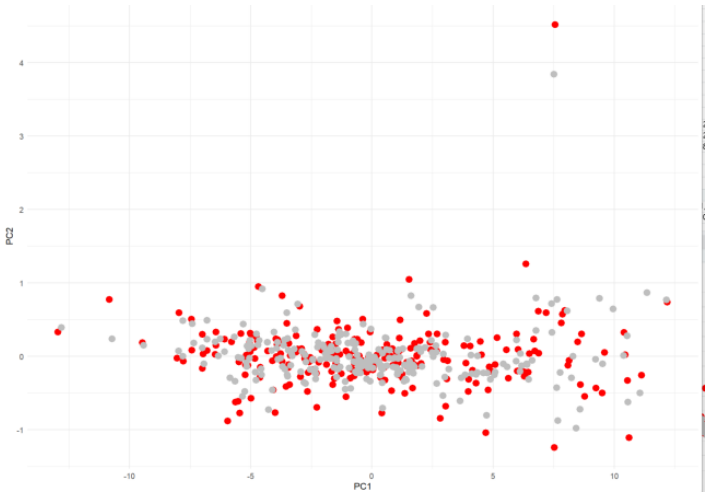

Figure 5. Comparison of principal components.

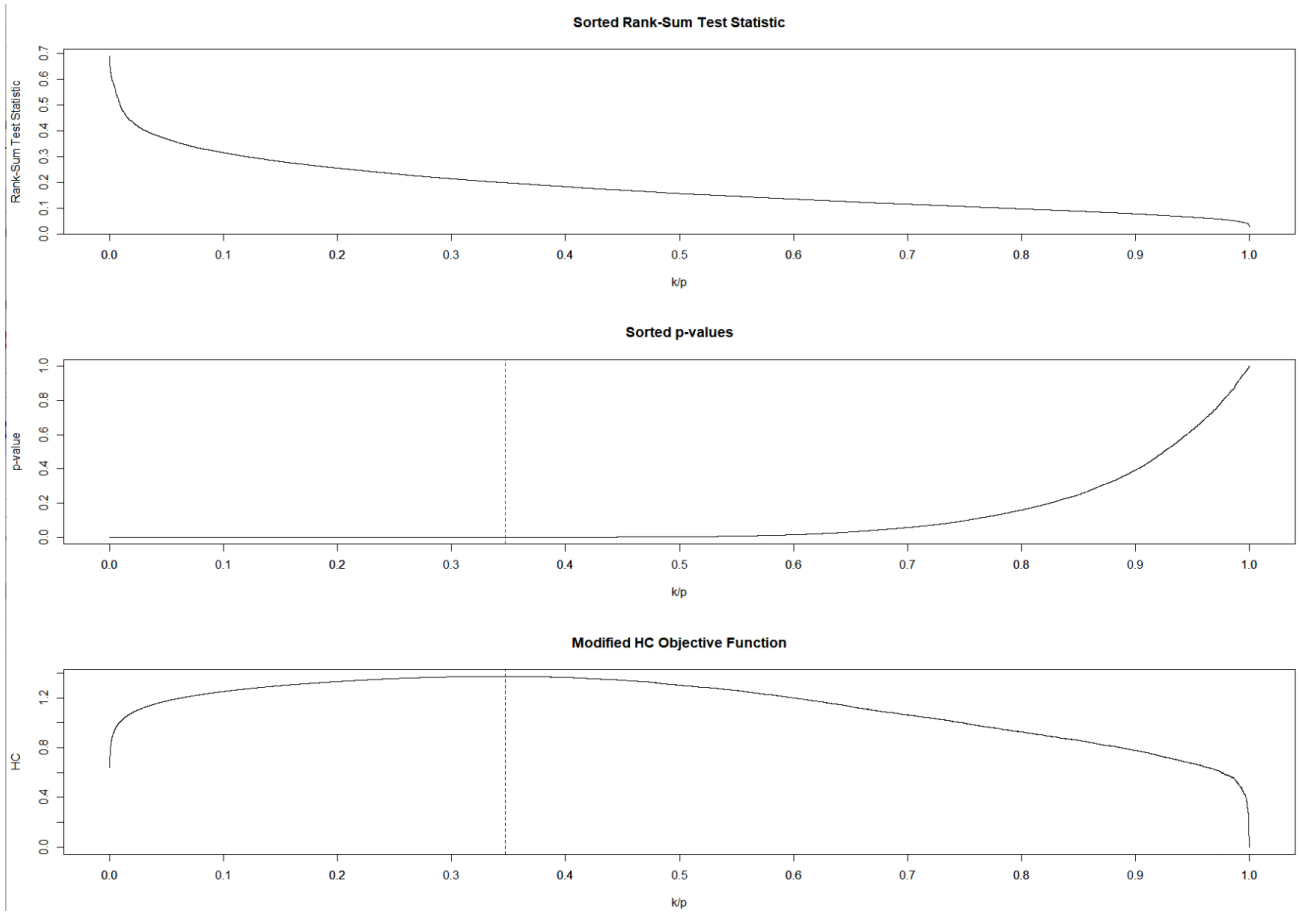

Figure 6. Rank-sum test HC results plot.

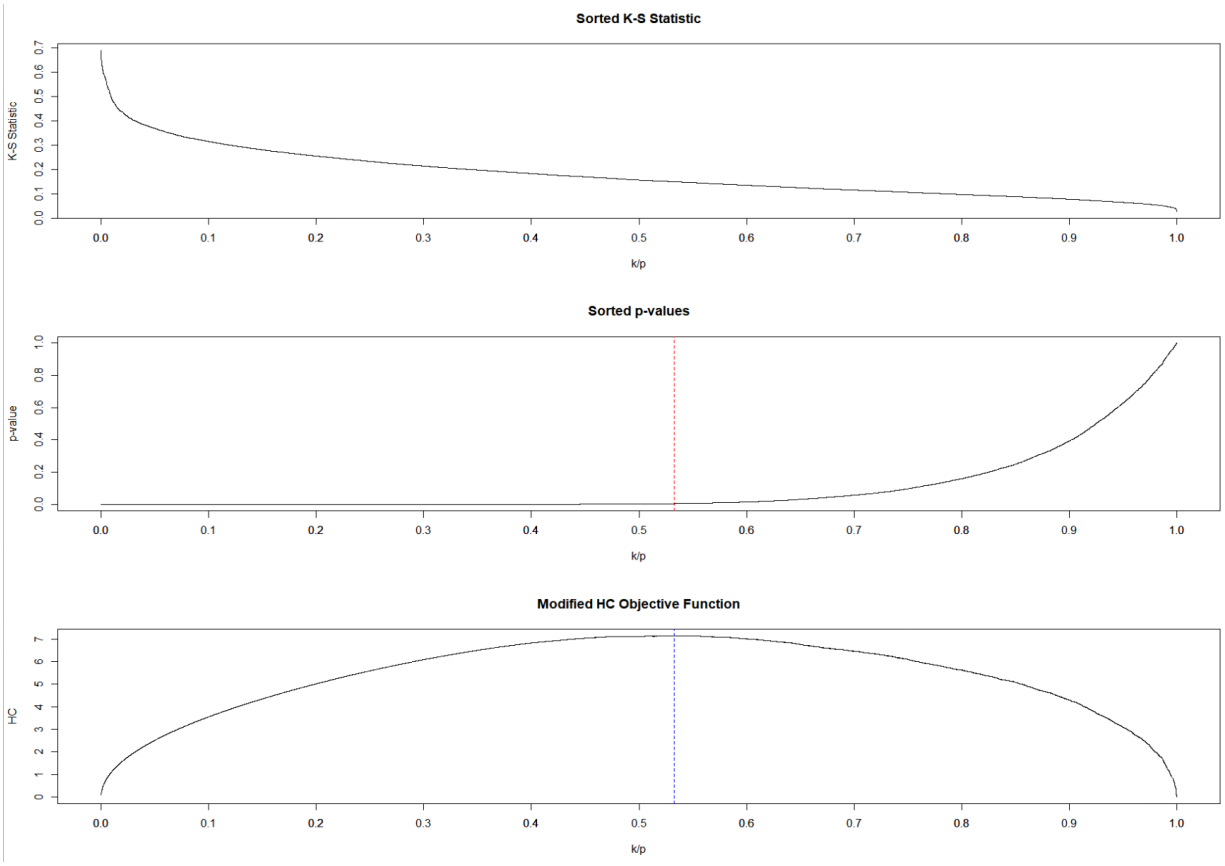

Figure 7. K-S test HC results plot.

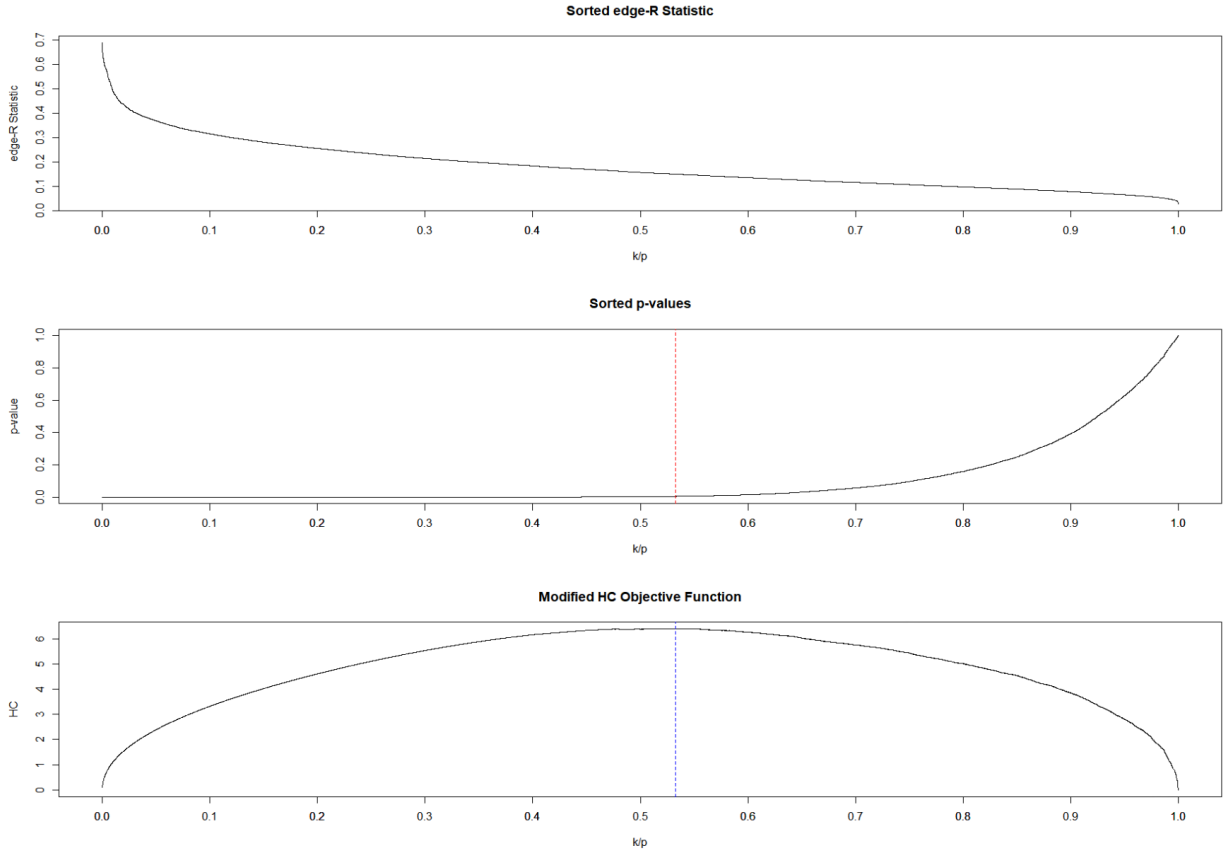

Figure 8. EdgeR test HC results plot.

As shown in the table below, this study systematically identified and analyzed differentially expressed genes (DEGs) using the edgeR method, Wilcoxon rank-sum test, and K-S test, combined with the HC higher-order identification method. The aim was to evaluate the applicability and effectiveness of each method in large-scale gene data analysis. **Table 1** presents the non-parametric statistical results, providing detailed insights into the statistical analysis conducted in this study.

Table 1. Non-parametric statistical results.

	p	**Number of Differentially. Expressed Genes**	**Number of False Discoveries**	**False Discovery Rate**
Rank-Sum Test	0.0005	3876	384	0.099071207
K-S Test	0.0049	6123	738	0.120529152
edgeR Test	0.005	6318	907	0.143558088

From the perspective of data analysis, the Wilcoxon rank-sum test demonstrates high accuracy, with its false discovery rate significantly lower than the other two methods. In balancing detection sensitivity and false discovery control, the Wilcoxon rank-sum test achieves a superior equilibrium.

As a non-parametric statistical method, the Wilcoxon rank-sum test is particularly suitable for scenarios with large sample sizes, unknown or biased data distributions. It has lower assumptions regarding distribution, making it more stable and robust. This allows it to provide accurate and reliable results without adding extra complexity. In contrast, while the edgeR method is widely used in gene expression analysis, it relies on generalized linear models with stricter assumptions and has a stronger dependence on data distribution and false discovery control. The K-S test, although more flexible, is limited in sensitivity under specific distributions and does not offer the same level of false discovery control as the rank-sum test.

Based on the above results and analysis, it can be concluded that under traditional large-sample conditions, the Wilcoxon rank-sum test is recommended as the primary method for gene expression analysis in non-parametric statistics. Its stable performance in large-scale gene data and its effective control of false discoveries provide a solid foundation for ensuring the reliability and scientific validity of the analysis results. This conclusion serves as an important reference for differential gene expression research and subsequent analyses.

5.2. Comparison of wgain with other models

By comparing the differential gene (DEGs) detection results of wgain with the mainstream algorithms GAN and WGAN-GP, the performance of each method was evaluated in terms of the accuracy of the generated data and the control of false discovery rate (FDR). A total of 3876 differential genes were detected in the original data using wgain. The following are the specific analysis results for the generated data, wgain outperformed in accuracy and FDR (**Figures 9** and **10**). **Table 2** provides a comprehensive comparison of different methods, highlighting their respective strengths and limitations.

Table 2. Method comparison.

	The number of differential genes	overlap count
wgain	3629	3245
gan	3920	3089
wgan-gp	4130	3200

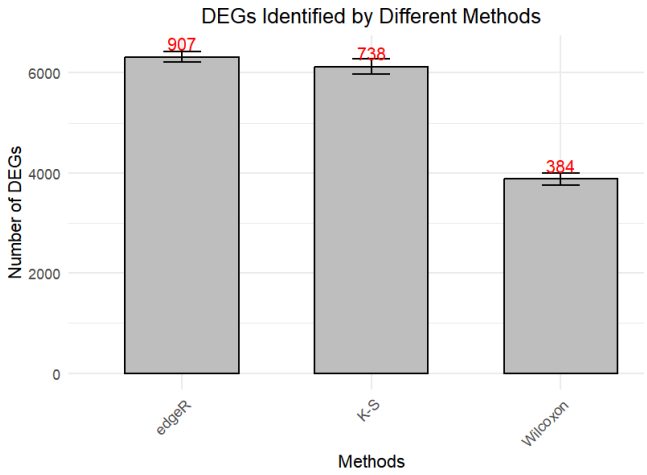

Figure 9. Comparison of statistical results plot.

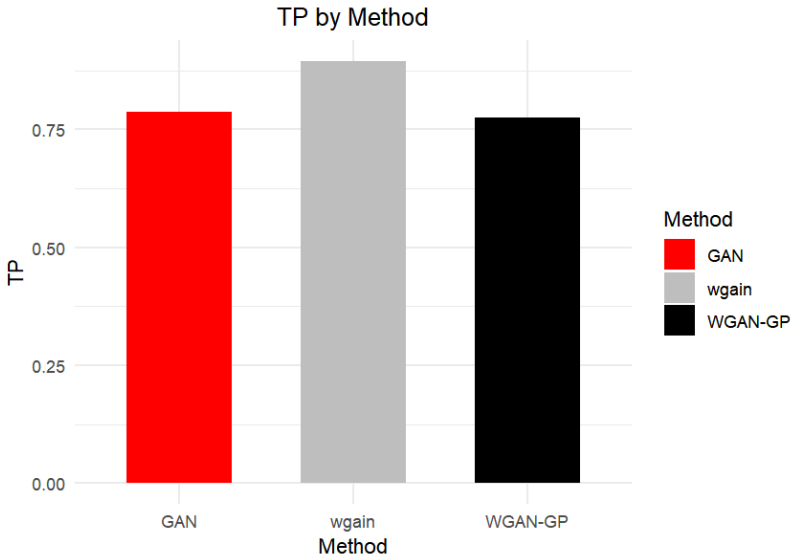

Figure 10. Accuracy comparison plot.

The performance differences of three generative models—wgain, GAN, and WGAN-GP—in differential gene detection were compared using two key metrics: FDP and TP [16]. FDP measures the proportion of false discoveries, reflecting the accuracy of the generated data, while TP evaluates the model's ability to identify true differential genes in the generated data, indicating how well it retains the characteristics of the real data.

From the FDP results, wgain demonstrated a significant advantage, showing the lowest false discovery rate, indicating that the data generated by wgain are more reliable in terms of consistency with real data. In contrast, GAN had a significantly higher FDP than wgain, suggesting that a relatively larger proportion of genes in the

generated data were falsely detected. WGAN-GP exhibited an even higher FDP than the other two methods, indicating a notable deficiency in accurately identifying differential genes in the generated data. A lower FDP is an important indicator of generative model quality, and wgain's excellent performance in this aspect suggests it is more suitable for bioinformatics analyses that require high accuracy. **Figure 11** presents an error rate comparison plot, visually illustrating the differences in error rates among various methods or conditions.

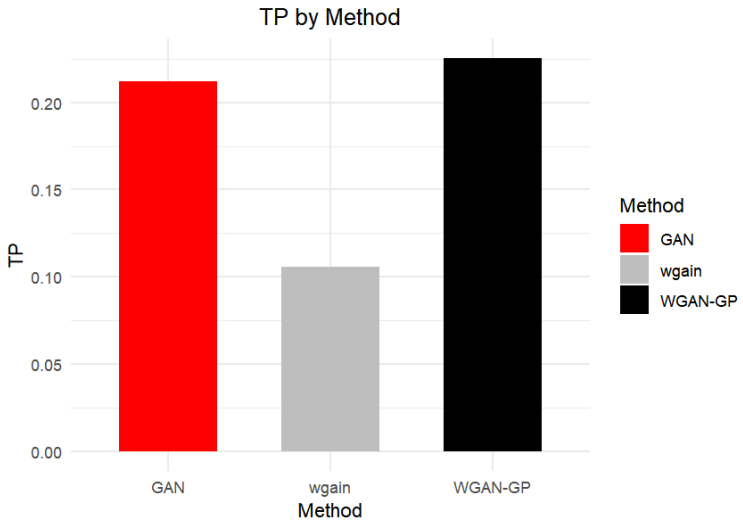

Figure 11. Error rate comparison plot.

Regarding the TP results, wgain also performed excellently, with the highest true positive rate and the greatest success in replicating the true differential genes from the original data. This indicates that wgain is effective at retaining the key information from the original data, thus enhancing the scientific value of the generated data. The TP for GAN was slightly lower than wgain but still at a relatively high level, suggesting that it can somewhat replicate the characteristics of the real data. WGAN-GP showed a slightly lower TP than both wgain and GAN, reflecting its somewhat reduced capability to capture the true differential gene features.

In summary, based on the analysis of both FDP and TP results, wgain outperforms GAN and WGAN-GP in both the accuracy of the generated data and its ability to preserve key features. wgain not only significantly reduces the false discovery rate but also effectively increases the detection of true differential genes. This indicates that wgain has significant advantages in differential gene analysis and biological data generation, and when combined with the Wilcoxon rank-sum test, it can serve as a major method for enhancing data in current biostatistical practices.

5.3. Conclusion

In this study, we proposed and implemented a data augmentation approach—termed the WGAIN algorithm—that integrates WGAN-GP with a random forest classifier to address the limitations of differential gene expression analysis using small sample datasets. Our experimental results demonstrated that the WGAIN algorithm outperforms traditional methods in terms of data generation stability, data quality, and the accuracy of differential gene identification. Traditional GAN-based methods (e.g.,

the GAIN method proposed by Waters et al. [4]) are prone to mode collapse during training, which can compromise the stability of the generated data. In contrast, we employ WGAN-GP, which introduces a gradient penalty term to improve the training balance between the generator and the discriminator, thereby significantly enhancing the robustness and quality of the generated data.

Additionally, existing research (e.g., Li et al. [2]) has indicated that the rank-sum test exhibits superior statistical performance in large-sample differential expression analyses. However, such methods depend on a preset p-value threshold (typically 0.05), and this subjective thresholding may exaggerate statistical significance when applied to small sample datasets. To address this issue, we developed a novel threshold determination strategy based on the Higher Criticism (HC) approach. This data-driven method uses a mathematical formulation to automatically determine the optimal *p*-value threshold, thus overcoming the limitations of subjective thresholding inherent in traditional methods. This strategy not only ensures the objectivity and precision of statistical assessments under small-sample conditions but also demonstrates remarkable stability across various tests.

Furthermore, we validated the similarity between the generated and real data using a random forest model, further confirming that the WGAIN algorithm effectively captures and preserves key gene expression features during data augmentation, thereby providing a robust foundation for subsequent differential gene screening. Overall, the WGAIN algorithm significantly enhances the performance of small-sample data analysis, offering a novel and efficient technical pathway to address data scarcity issues, with considerable potential for broader application and further development.

5.4. Potential weaknesses

Despite the strong performance of wgain in data generation, there are several potential weaknesses that may impact its effectiveness in practical applications:

5.4.1. Training instability

Although WGAN-GP mitigates some of the training instability issues of traditional GANs through gradient penalties, GANs still face challenges such as mode collapse, particularly when dealing with complex or high-dimensional data. This can lead to instability during training and affect the quality and consistency of the generated data.

5.4.2. Overfitting

Due to the model's complexity and long training periods, overfitting may occur, especially when the training data is small or the data distribution is imbalanced. This can cause the model to overly fit specific training samples, limiting its ability to generate diverse, representative data.

5.4.3. High computational cost

The use of multiple GAN architectures increases the computational load, particularly when each model needs to be trained independently. This significantly increases training time and hardware resource consumption, posing a challenge for researchers with limited computational resources.

5.4.4. Lack of data diversity

While the use of multiple GAN architectures aims to enhance data diversity, the generated data may still be concentrated around certain patterns and fail to capture all the variations present in the real data. This limits the comprehensiveness and practical applicability of the generated data.

6. Conclusion and future outlook

This study demonstrates an innovative method with significant advantages in differential gene analysis and biological data generation by developing and validating the wgain algorithm (combined with the Wilcoxon Rank-Sum Test). By effectively addressing the issue of limited sample size through data augmentation techniques, our approach shows remarkable potential for applications in the medical field, particularly in areas such as gene expression analysis and tumor classification. In many medical studies, especially those targeting rare diseases or early diagnosis, the number of available samples is often limited, posing serious challenges for data analysis and model training. Traditional methods frequently fail to fully extract the latent information contained in small sample datasets. In contrast, our wgain algorithm leverages WGAN-GP to generate high-quality, extended data that closely conforms to the distribution of the original data, and incorporates Random Forest classification for performance validation, thereby significantly enhancing data utilization efficiency and the accuracy of differential gene identification. Our results indicate that wgain can efficiently generate extended data that highly match the original data characteristics, and that incorporating non-parametric statistical methods such as the Wilcoxon Rank-Sum Test further improves the accuracy and reliability of the data analysis. This provides important technical support for the analysis and modeling of high-dimensional biological data.

In future research, the combined application of the wgain algorithm and non-parametric statistical methods holds vast potential for exploration. The importance of this method in the medical field is reflected in several key aspects. First, by expanding the sample size through data augmentation, the statistical significance of analyses and the stability of model training are enhanced, offering a new solution for the utilization of limited clinical sample data. This is particularly critical for early cancer screening, disease prognosis evaluation, and precision medicine, where data scarcity and imbalanced samples are common challenges. Second, the use of non-parametric statistical methods (such as the Wilcoxon Rank-Sum Test) makes the data analysis more objective and precise, reducing the false positive rate and providing more reliable technical support for biomarker discovery. Effective identification of differential genes not only aids in a deeper understanding of the molecular mechanisms underlying diseases but also provides key evidence for the development of personalized treatment strategies.[17]

Furthermore, this method can be extended to the analysis of genome data with larger sample sizes and various cancer subtypes, in order to validate its generalizability and adaptability across diverse datasets. In addition, the wgain algorithm possesses strong adaptability and scalability. It is not only applicable to gene expression data, but can also be extended to other types of medical data, such as imaging data, clinical

test data, and multi-omics data. By applying it across different domains, its universality and robustness can be further verified, providing novel technical means for multimodal data integration and comprehensive diagnosis. Particularly in the fields of precision medicine and tumor feature discovery, the large-scale, high-quality data generated by wgain can reveal subtle differences among various tumor subtypes, offering precise data support for the development of personalized treatment plans.

In summary, the wgain algorithm, through its innovative data augmentation approach, not only effectively mitigates the challenges posed by small sample sizes in medical research but also provides a high-quality data foundation for bioinformatics analysis. It shows great potential in enhancing data analysis accuracy, optimizing model training, and supporting personalized medical decision-making. Looking ahead, by integrating with other advanced machine learning models, wgain is expected to further improve its data generation and analysis capabilities, offering a more comprehensive and reliable technical guarantee for early diagnosis, disease monitoring, and precision treatment in the medical field [18,19].

Author contributions: Conceptualization, YH; methodology, YH; software, YH; validation, YH and SL; formal analysis, SL; investigation, YH and SL; resources, SL; data curation, XX; writing—original draft preparation, YH; writing—review and editing, SL; visualization, YH and XX; supervision, SL; project administration, XX; funding acquisition, YH, SL and XX. All authors have read and agreed to the published version of the manuscript.

Ethical approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. Torné RV, Bryson K. Adversarial generation of gene expression data [Master's thesis]. University College London; 2018.
2. Li Y, Ge X, Peng F, et al. Exaggerated false positives by popular differential expression methods when analyzing human population samples. Genome Biology. 2022; 23(1). doi: 10.1186/s13059-022-02648-4
3. Viñas R, Andrés-Terré H, Liò P, et al. Adversarial generation of gene expression data. Bioinformatics. 2021; 38(3): 730-737. doi: 10.1093/bioinformatics/btab035
4. Waters MR, Inkman M, Jayachandran K, et al. GAiN: An integrative tool utilizing generative adversarial neural networks for augmented gene expression analysis. Patterns. 2024; 5(2): 100910. doi: 10.1016/j.patter.2023.100910
5. Yang W. Application of non-parametric statistical analysis in multi-sample research—Example of the biological effect of normal liver RNA on cancer cells. Today Wealth Magazine. 2016.
6. Liu M, Wang B, Ta L, et al. Stereological analysis of the ultrastructure of human breast cancer cells and the rank-sum test. Progress in Biomedical Engineering. 2011; 32(02): 74-76.
7. Jiao CN, Gao YL, Yu N, et al. Hyper-Graph Regularized Constrained NMF for Selecting Differentially Expressed Genes and Tumor Classification. IEEE Journal of Biomedical and Health Informatics. 2020; 24(10): 3002-3011. doi: 10.1109/jbhi.2020.2975199
8. Zhang S. Research on GAN data augmentation methods for brain print recognition. Information Engineering University of Strategic Support Forces; 2023.
9. Zou H. Financial time series forecasting based on deep forest generative adversarial networks. Dalian Maritime University; 2021.
10. Zhou F. General Non-Parametric Tests for Differential Gene Expression Analysis [PhD thesis]. University of California, Berkeley; 2023.

11. Stupniko A, McInerney CE, Savage KI, et al. Robustness of differential gene expression analysis of RNA-seq. Computational and structural biotechnology journal. 2021; 19: 3470-3481.
12. Tang Y. Empirical analysis of non-parametric test statistics in survival analysis [PhD thesis]. Dalian University of Technology; 2018.
13. Yang Y, Zhao P. Non-parametric tests for two independent samples in teaching of non-parametric statistics. Science and Education Journal (Upper Volume). 2013; (04): 45-46.
14. Zhang X. Development and application of non-parametric KS test software. Yangzhou University; 2024.
15. Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of Wasserstein GANs. arXiv preprint arXiv:1704.00028. 2017. doi: 10.48550/arXiv.1704.00028
16. Fang J, Liu L. Overview of various biological statistical tests and their conditions of use. Ecology Journal. 1995; (03): 67-70.
17. Hu Z. Research on clinical treatment protocols for diabetes combined with coronary heart disease. Chengdu University of Traditional Chinese Medicine; 2015.
18. Cheng N. Summary of the New Drug Biostatistics Seminar. Chinese Journal of Clinical Pharmacology and Therapeutics. 1996; (02): 142-145.
19. Gordon GJ, Jensen RV, Hsiao LL, et al. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer research. 2002; 62(17): 4963-4967.