SUPPLEMENTARY MATERIAL SECTION A AND B

(A) MATHEMATICAL BACKGROUND AND TECHNICAL DETAILS

This supplementary material includes additional information to support the main manuscript. Here, we provide mathematical background for the receiver operating characteristic (ROC)-derived indices used for screening and technical details of the simulation study. Furthermore, we include additional information about the real-life RNA-Seq data, i.e., the cervical cancer dataset.

Receiver operating characteristic curve based screening indices

In the revised manuscript, differential expression (DE) inference is performed using DESeq2 as the primary hypothesis-testing framework, while ROC-derived quantities are used as complementary screening/ranking indices to prioritize candidate improper (non-monotonic) profiles. Here, we summarize the mathematical background of the ROC-derived indices used in this study: the classical area under the curve (cAUC), the generalized ROC curve (gROC) and its area (gAUC), and the length of the ROC curve (LROC). All ROC-derived indices were computed on preprocessed, variance-stabilized expression values, and they are interpreted as descriptive screening scores (not as standalone inferential DE tests).

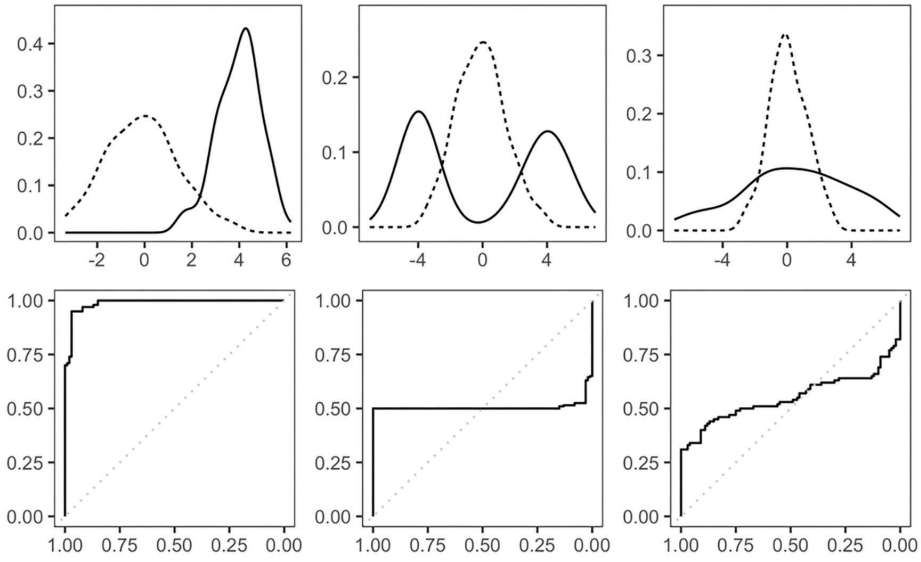

FIG. 1. (Revisited from main manuscript) Distribution of proper and improper gene expression profiles and corresponding ROC curves–(left) proper, (middle and right) improper; dashed lines correspond healthy group.

Classical ROC curve: The ROC curve depicts the trade-off between sensitivity and specificity across all thresholds of a biomarker. Let ξ and η denote gene expression values in diseased and healthy groups with cumulative distribution functions $F_{\xi}(\cdot)$ and $G_{\eta}(\cdot)$, respectively. The ROC curve is defined as:

$$R(t) = 1 - F_{\xi}\left(G_{\eta}^{-1}(1-t)\right), \qquad t \in [0,1] \text{ and } \xi > \eta \tag{1}$$

where t is the false positive rate (FPR). The area under the cAUC is calculated as:

$$cAUC = \int_0^1 R(t)dt \tag{2}$$

The cAUC is widely used as a summary index, with values near 1.0 indicating excellent discrimination and values around 0.5 suggesting no discrimination. However, when applied to improper distributions, the cAUC may result in misleading conclusions. In this study, empirical ROC curves were implemented by evaluating sensitivity and specificity across a grid of marker thresholds derived from the observed data (e.g., midpoints of consecutive ordered marker values).

Generalized ROC curve: The gROC extends the ROC framework by allowing classification rules based on flexible subsets of the marker distribution, such as $(-\infty, t_L] \cup [t_U, \infty)$ as shown Figure 2.

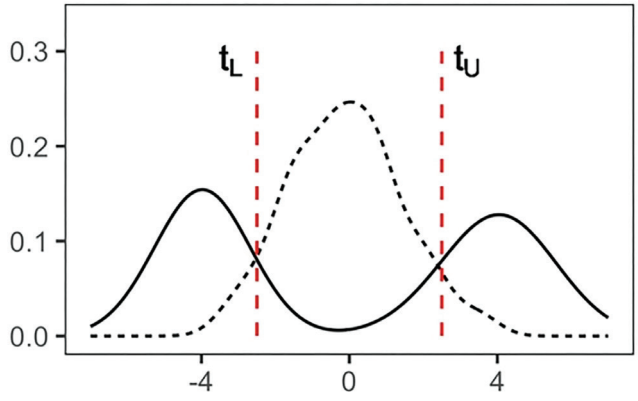

FIG. 2. Generalized ROC curve (gROC) and the corresponding gAUC screening index.

For a given FPR (i.e., *t*), the gROC is defined as

$$R_g(t) = \sup_{[l,u]\in\mathcal{I}_t} \left\{1 - F_\xi(u) + F_\xi^-(l)\right\}, \quad t \in [0,1] \tag{3}$$

where $\mathcal{I}_t = \{[l,u]: G_\eta(u) - G_\eta(l) = 1 - t\}$, which fixes the *specificity* at $1 - t$. Finally, the generalized area under the curve (gAUC) is calculated as

$$gAUC = \int_0^1 R_g(t)dt. \tag{4}$$

In the main manuscript, we use gAUC as a descriptive screening score to rank features for exploratory prioritization.

Length of ROC curve (LROC): The LROC, proposed by Franco-Pereira et al.,[1] measures the arc LROC trajectory in the unit square as an index for evaluating the ability of a biomarker to discriminate between groups. The LROC, under parametric binormal ROC assumption, is defined as

$$LROC = \int_0^1 \sqrt{1 + \left(\frac{dR(t)}{dt}\right)^2}\, dt \tag{5}$$

where R(t) denotes the ROC curve as a function of *t* calculated via equation 1. This measure captures the degree of curvature of the ROC function. Longer curves correspond to stronger separation between groups. For better interpretability, Franco-Pereira et al.[1] applied a min-max normalization on LROC such as

$$LROC^* = \frac{LROC - LROC_{min}}{LROC_{max} - LROC_{min}} \tag{6}$$

where $LROC_{min} = \sqrt{2} \approx 1.414$ and $LROC_{max} = 2$. This transformation scales $LROC^*$ between 0 and 1 as *no* discrimination and perfect discrimination, respectively. In this study, we report and apply the unnormalized LROC (on its natural scale between $\sqrt{2}$ and 2) as a screening index; therefore, the heuristic cervical cut-off reported in the main manuscript (e.g., LROC > 1.48) refers to the raw LROC scale, not to $LROC^*$. Unlike cAUC, LROC emphasizes the shape and curvature of the ROC function, potentially capturing signals that cAUC may underestimate.

Simulation study

To evaluate how an inferential DE method (DESeq2) and ROC-derived screening indices (cAUC, gAUC, and LROC) prioritize improper genes (IGs) under realistic RNA-Seq conditions, we designed a comprehensive simulation study. DESeq2 was included as a reference benchmark for DE inference, and ROC-derived indices were treated as complementary descriptive scores computed on variance-stabilized values. The following parameter combinations defined simulation scenarios (18 scenarios in total). We generated gene expression data from a negative binomial (NB) distribution, constructing improper profiles according to the high-low mixture scenario in Figure 1 (middle panel). Mathematical background of data generation procedure and the parametrization of overdispersion and offset parameters can be found in the related papers.[2]

- Sample size (n): 100, 300, and 500
- Number of features (p): 3000
- Proportion of differentially expressed genes (π_{DEGs}): 0.05 (low) and 0.30 (high)
- Proportion of DEGs with improper profiles (π_{IR}): 0.25
- Overdispersion (ϕ): 0.01 (low), 0.10 (moderate), and 1 (high)
- Offset parameter (std. deviation of log-FoldChange values): 1.5

For each scenario, 1,000 datasets were generated in the final runs reported in the revised manuscript (i.e., setting the number of replicates accordingly in the simulation scripts). The analysis pipeline followed in the simulation study is presented in Figure 3.

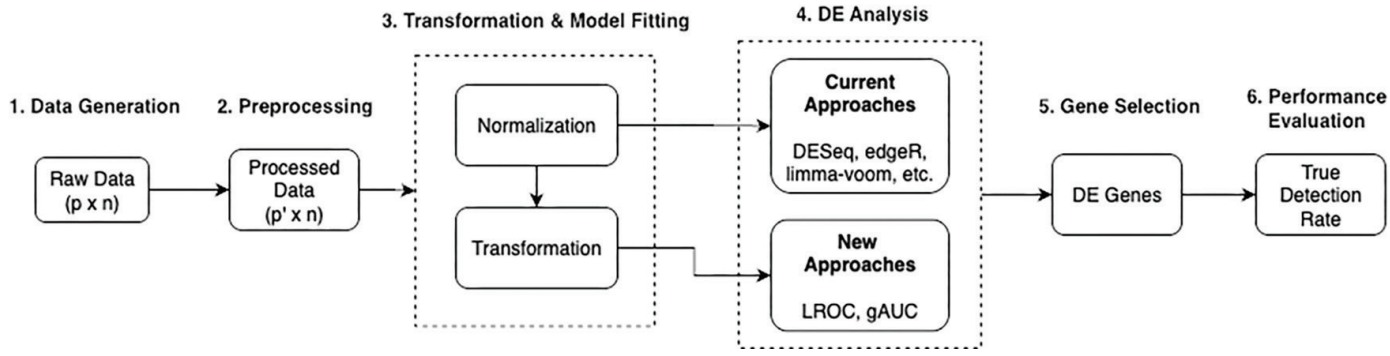

FIG. 3. (Revisited from main manuscript) simulation workflow. LROC, length of the receiver operating characteristic; gAUC, generalized area under the curve; DE, differential expression.

In the simulation workflow, the gene expression distributions of improper profiles were defined according to the high-low mixture construction in Figure 1 (middle panel). The proportion of differentially expressed genes was set at two levels, 0.05 (low) and 0.30 (high). Among these differentially expressed genes, the proportion of genes with improper profiles was fixed at 0.25. For downstream analysis, DE inference was performed using DESeq2 on raw counts, while ROC-derived indices (cAUC/gAUC/LROC) were computed on preprocessed, variance-stabilized expression values and used as screening/ranking scores. Method performance was summarized using sensitivity-analysis metrics reported in the main manuscript (e.g., TPR/PPV across selected-set sizes), both overall and within the subset of IGs. All simulations were implemented in the R programming language (https://cran.r-project.org), using custom scripts and primarily the DESeq2,[3] nsROC,[4] and pROC[5] packages.

Cervical cancer dataset

In addition to the simulation study, we conducted analyses on a real RNA-Seq dataset originally published by Witten et al.[6] This dataset comprises small RNA sequencing measurements obtained from 58 human cervical tissue samples, including 29 tumor and 29 matched non-tumor controls. The sequencing was performed using the Illumina (Solexa) platform, and the resulting expression profiles represent counts for 714 distinct microRNAs (miRNAs). The data are publicly available through the Gene Expression Omnibus under accession number GSE20592.

The original study aimed to identify differentially expressed miRNAs associated with cervical cancer and to explore novel small RNA species. In our simulations, we modeled counts as if they originated from protein-coding gene expression (mRNAs). By contrast, the real dataset analyzed here consists of sequencing counts from microRNAs (miRNAs), where identifiers such as miR-34b, miR-135a, and

let-7d denote mature miRNA species quantified directly from sequencing. Following the original study by Witten et al.,[6] we therefore refer to these abundance profiles as "miRNA expression" in the real data analyses, while retaining "gene expression" terminology for the simulated datasets. This distinction avoids ambiguity and maintains consistency with the source study.

For our analysis, we used the preprocessed, variance-stabilized miRNA expression values to compute ROC-derived screening indices and to illustrate how they can complement DESeq2 by prioritizing candidate non-monotonic (improper) miRNA profiles for follow-up. The cervical cancer dataset was selected because it exhibits moderate to high dispersion and contains biologically heterogeneous tumor subtypes, providing a suitable real-world context to assess the robustness of screening indices under heterogeneous expression patterns.

Differential expression analysis using DESeq2

In this study, DESeq2 was used as the primary inferential DE method for count-based RNA-seq/miRNA-seq data. Below, we summarize the model and the analysis workflow used in our scripts (simulation_codes.R, real_data_codes.R, and helper_functions.R) to support reproducibility. For full technical details, we refer readers to the DESeq2 package documentation and vignette on Bioconductor (https://bioconductor.org/packages/DESeq2).

Model and hypothesis testing (brief)

Let X_{ij} denote the raw read count for feature (gene/miRNA) *i* in sample *j*. DESeq2 models counts with a NB generalized linear model,

$$X_{ij} \sim \mathrm{NB}\left(\mu_{ij}, \alpha_i\right) \tag{7}$$

where μ_{ij} is the mean and α_i is the feature-specific dispersion parameter. Under the standard NB parametrization,

$$\mathrm{Var}\left(X_{ij}\right) = \mu_{ij} + \alpha_i \mu_{ij}^2 \tag{8}$$

Mean counts are decomposed into a sample-specific size factor s_j and a normalized mean q_{ij},

$$\mu_{ij} = s_j q_{ij} \tag{9}$$

and the normalized mean is linked to covariates through the design matrix $\boldsymbol{D}$,

$$\log\left(q_{ij}\right) = d_j^{\mathsf{T}} \beta_i \tag{10}$$

where d_j is the *j*-th row of $\boldsymbol{D}$ and β_i is the vector of regression coefficients for feature *i*. For the group effect ("condition"), the null hypothesis is $H_0: \beta_{i,\mathrm{condition}} = 0$, which is tested using the default Wald test in DESeq2; resulting p-values are adjusted for multiple testing using the Benjamini–Hochberg (BH) false discovery rate (FDR) procedure, reported as padj.

Implementation details used in this study

Design formula and inputs: We constructed a DESeqDataSet via DESeqDataSetFromMatrix function using raw integer count matrices and a two-level condition factor with design ~condition. In simulations, samples were labeled C1/C2 (see createDDSobject in helper_functions.R); in the cervical dataset, condition was set to Normal / Tumor (see real_data_codes.R).

Prefiltering (low counts and near-zero variance): Before DE inference, we applied two deterministic filters (function filterCounts in helper_functions.R): (i) near-zero variance filtering using caret::nearZeroVar on the sample-wise count profiles, and (ii) a *low-count* prefilter keeping features with at least 10 reads in at least 3 samples, i.e.,

$$\sum_j \mathbb{I}\left(X_{ij} \geq 10\right) \geq 3 \tag{11}$$

These filters were applied identically in the simulation and real-data workflows.

DESeq2 fitting and results extraction: Differential expression was performed using the default DESeq pipeline (function diffExp in helper_functions.R), i.e., DESeq(dds) followed by results(dds) with default arguments. We used the padj values returned by DESeq2 (BH-adjusted p-values) and did not apply any additional multiple-testing adjustment outside the package.

Handling zeros (pseudocount option used in scripts): In our scripts, DESeq2 inference was run with an optional safeguard nonzero=TRUE, which adds a constant 1 to all counts if any zero is present (counts(dds) <- counts(dds) + 1L). This was used to avoid undefined geometric means in size-factor estimation under the default median-ratio method when zeros are present.

Ranking for screening comparisons and DE list definition: For simulation summaries that compare ranking behavior across methods, features were ranked by increasing padj from DESeq2 ($\log_2$ fold changes were not used for ranking), and the top-*K* features were selected at each *K* value (see selectDEfeatures_DESeq in helper_functions.R). For the cervical cancer dataset, we additionally defined a DESeq2 "significant" miRNA set using a fixed decision rule (absolute $\log_2$ fold-change > 0.6 and padj < 0.05) for reporting and for downstream comparisons (e.g., the volcano plot and Venn diagram in the main Results); this threshold-based definition is distinct from the top-*K* ranking used in the simulation sensitivity analyses.

Dispersion summaries (real data): To summarize dispersion levels in the cervical dataset, we additionally computed gene-wise dispersion estimates using estimateSizeFactors and estimateDispersions, and then extracted dispersions via dispersions(dds) (see real_data_codes.R).

Variance-stabilizing transformation (used outside DESeq2 inference): For ROC-derived screening indices (reported elsewhere), we computed variance-stabilized expression values using DESeq2's varianceStabilizingTransformation after size-factor estimation (function preProcessCounts in helper_functions.R). These transformed values were used for descriptive screening and were not treated as inferential DESeq2 test statistics.

Data and code availability

All analysis scripts, simulation codes, and data used in this study are available for reproducibility and further exploration. The full repository, including code and processed datasets, can be accessed at https://github.com/dncR/ImproperGeneProfiles, and key materials are also provided as supplementary files.

REFERENCES

1. Franco-Pereira AM, Nakas CT, Pardo MC. Biomarker assessment in ROC curve analysis using the length of the curve as an index of diagnostic accuracy: the binormal model framework. *AStA Advances in Statistical Analysis*. 2020;104:625-647. [CrossRef]
2. Göksülük D, Karaağaoğlu AE. Classification of RNA-sequencing data via poisson and negative binomial linear discriminant analyses: a methodological study. *Türkiye Klinikleri Journal of Biostatistics*. 2023;15:150-160. [CrossRef]
3. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550. [CrossRef]
4. Fernandez SP. nsROC: non-standard ROC curve analysis; 2018, R package version 1.1. [CrossRef]
5. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12:77. [CrossRef]
6. Witten D, Tibshirani R, Gu SG, Fire A, Lui WO. Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biol*. 2010;8:58. [CrossRef]

(B) SIMULATION BENCHMARKS

Sensitivity analysis for low-DE scenario: false positive burden of the methods

To provide an empirical benchmark for "null behavior" (specificity) in our simulation study, we report false positive rates (FPR) across the sensitivity analysis grid of selected-set sizes (top-*K* features). Under the low- differential expression (DE) scenario (very low DE prevalence), the estimated FPR values are typically small and between-method differences are visually subtle, yielding largely overlapping curves in Figures 4-5. We therefore emphasize PPV in the main manuscript because it more directly summarizes the false-positive burden within the selected set and is easier to compare across methods; however, because PPV depends on prevalence, we provide FPR here as a complementary benchmark focused on specificity.

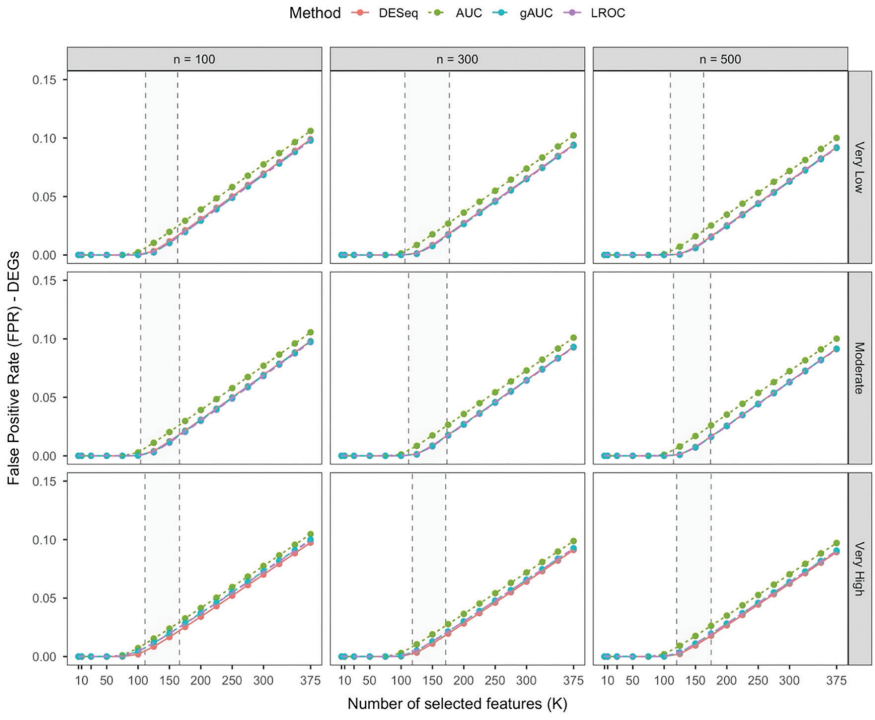

FIG. 4. Sensitivity analysis of the methods across simulations under the low-DE scenario – false positive burden of the methods (DEGs). LROC, length of the receiver operating characteristic; gAUC, generalized area under the curve; AUC, area under the curve; DEGs, differentially expressed genes; DE, differential expression.

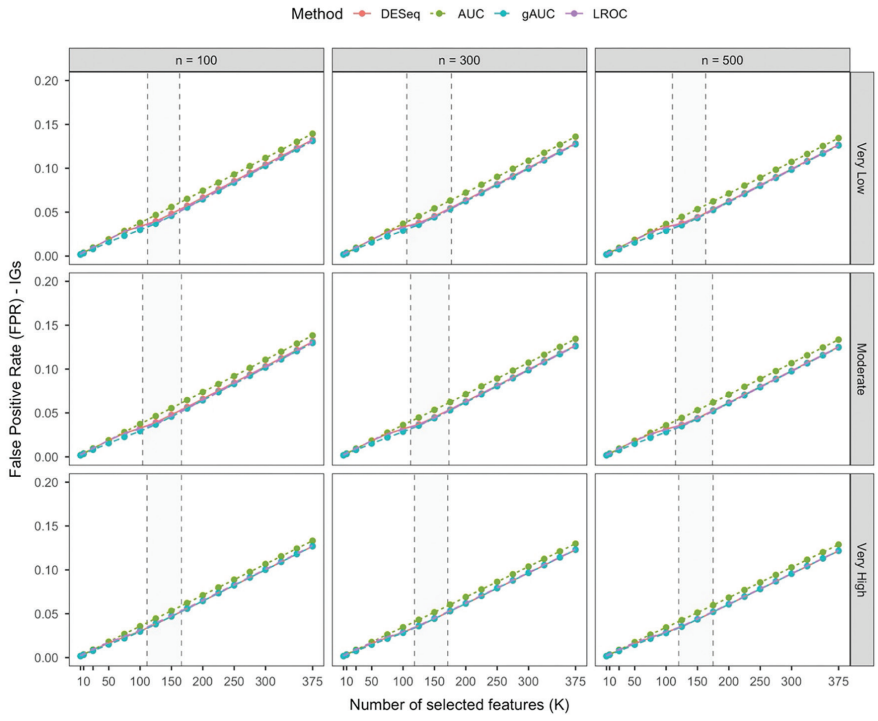

FIG. 5. Sensitivity analysis of the methods across simulations under the low-DE scenario–false positive burden of the methods (IGs). LROC, length of the receiver operating characteristic; gAUC, generalized area under the curve; AUC, area under the curve; IGs, improper genes.

Simulation results for high-DE scenario

We provide the results of the simulation study for the high-DE scenario in the following figure.

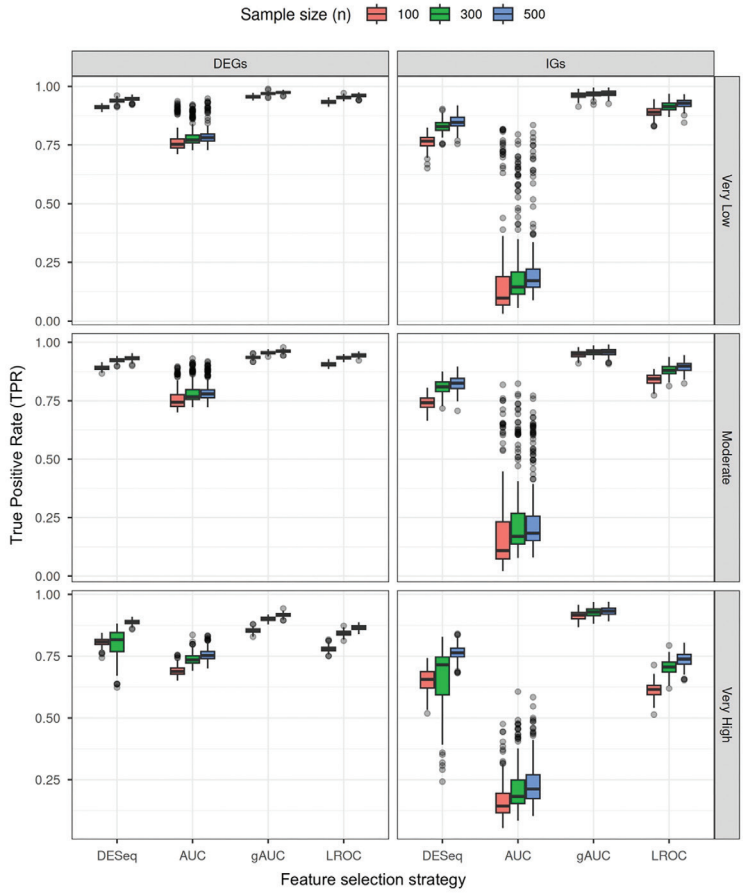

FIG. 6. Simulation results for high-DE scenario at the one-shot benchmark. LROC, length of the receiver operating characteristic; gAUC, generalized area under the curve; AUC, area under the curve; DE, differential expression.