

Fig. S1

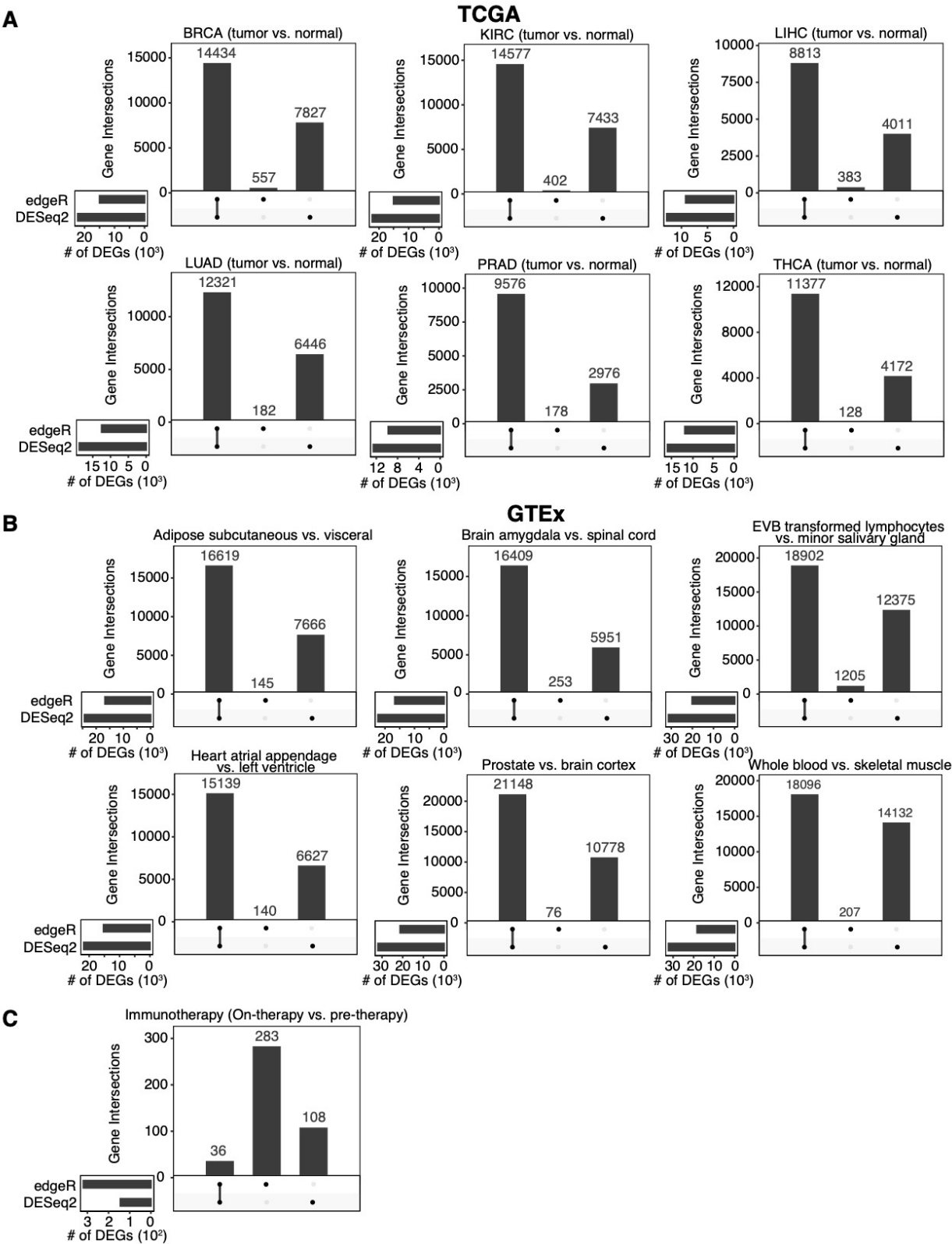


Fig. S1. The large discrepancies in the DEGs identified by DESeq2 and edgeR.

(A-C) Upset plots showing the intersections of the DEGs identified by DESeq2 and edgeR from TCGA **(A)**, GTEx **(B)**, and immunotherapy **(C)** RNA-seq datasets.

Fig. S2

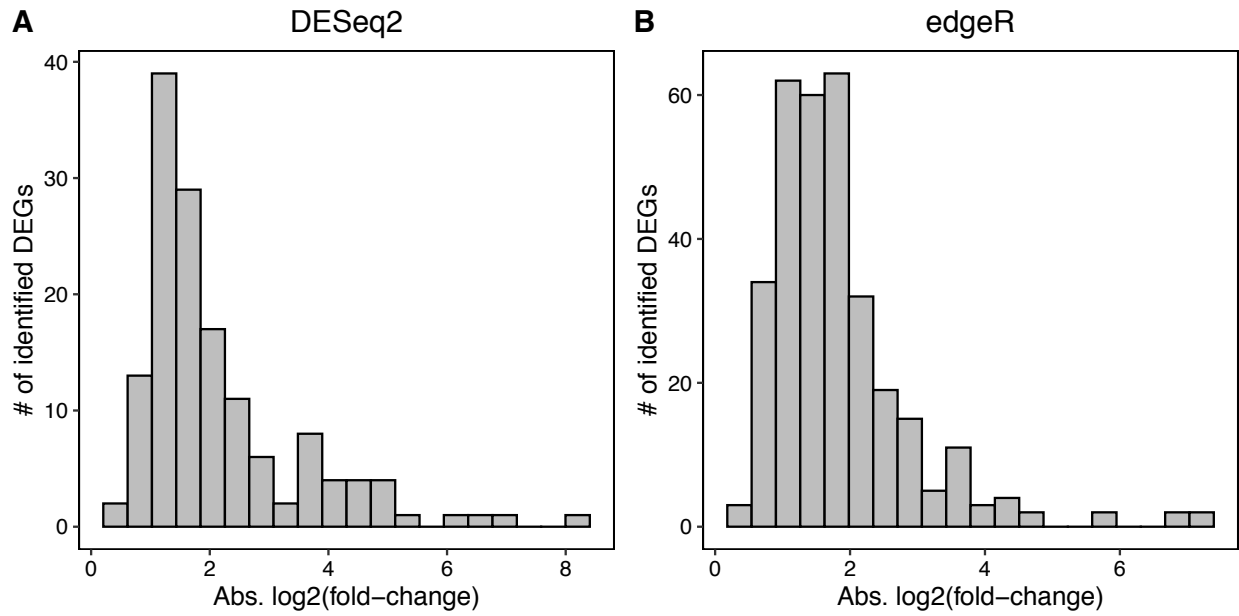


Fig. S2. The distribution of gene expression fold changes between pre-therapy and on-therapy samples in the original immunotherapy dataset calculated by DESeq2 (A) and edgeR (B).

Fig. S3

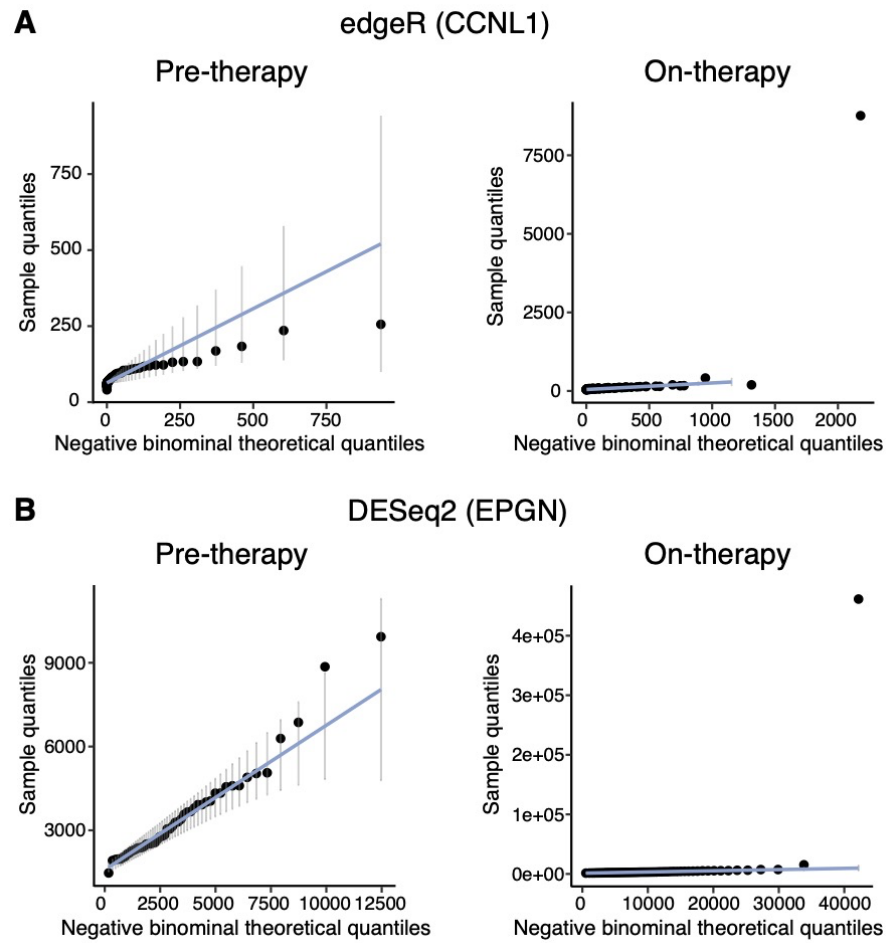


Fig. S3. Quantile-quantile (Q-Q) plots showing the discrepancy between observed read counts and the negative binominal theoretical quantiles estimated by edgeR and DESeq2.

A. The Q-Q plot for *CCNL1* with theoretical read counts in two conditions (pre-therapy and on-therapy) estimated by edgeR.

B. The Q-Q plot for *EPGN* with theoretical read counts in two conditions (pre-therapy and on-therapy) estimated by DESeq2.

Fig. S4

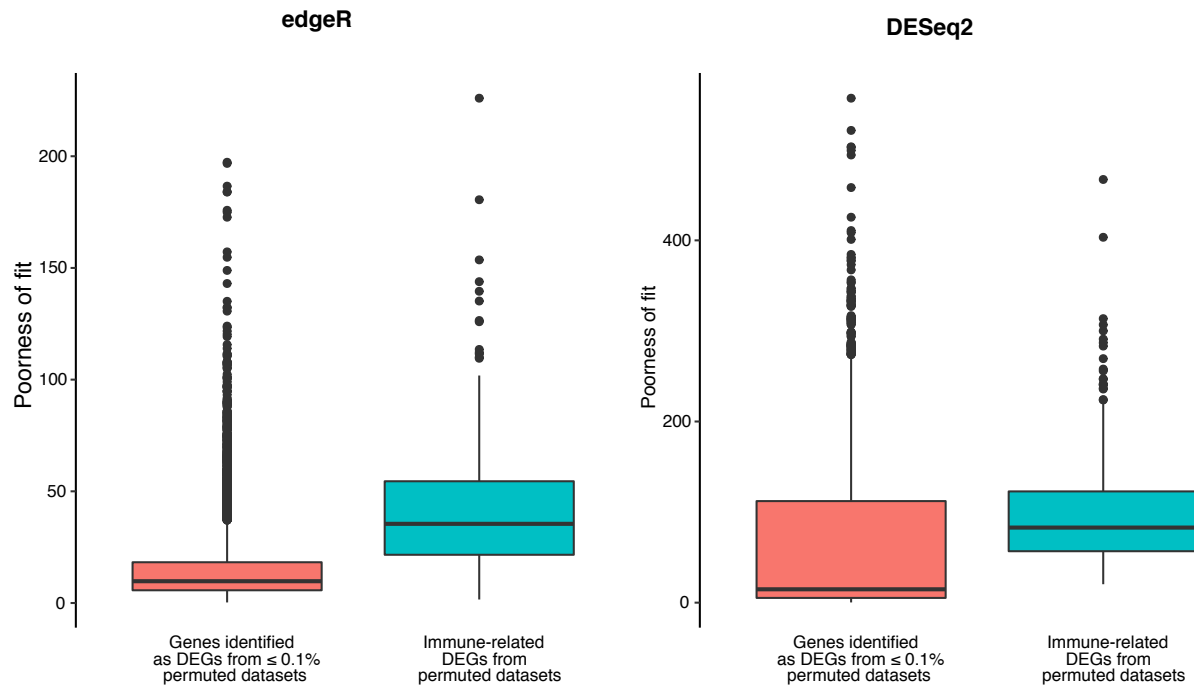


Fig. S4. Boxplots showing the poorness of fitting the negative binomial model to the immune-related spurious DEGs identified by DESeq2 or edgeR (“Immune-related DEGs from permuted datasets” in green) vs. the genes that were identified as DEGs from $\leq 0.1\%$ permuted datasets (“Non-DEGs from permuted datasets” in red). Immune-related spurious DEGs are the genes that were falsely identified as DEGs from at least 10% of the permuted datasets and were associated with the immune-related GO terms in Fig. 1E. The poorness of fit of each gene is defined as the negative $\log_{10}(\text{p-value})$ from the goodness-of-fit test for the negative binomial distribution based on estimated parameters by DESeq2 or edgeR. The p-value in each panel was calculated by the Wilcoxon rank-sum test to compare the two groups of genes' poorness-of-fit values.

Fig. S5

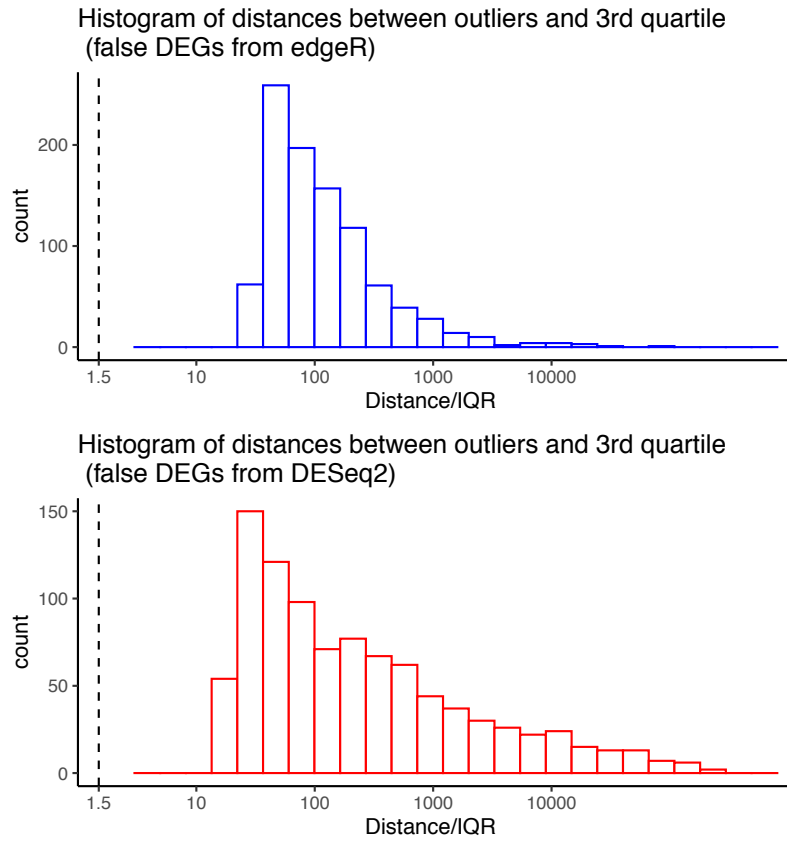


Fig. S5. Histograms showing the distances between the outliers and the third quartile of normalized gene expression values reported by edgeR (upper panel) or DESeq2 (lower panel). The results are for the genes that were falsely identified by each method as DEGs in at least 10 permuted datasets. The x-axis is (the distance between the largest outlier and the 3rd quartile of each gene) / (the inter-quartile range (IQR) of each gene).

Fig. S6

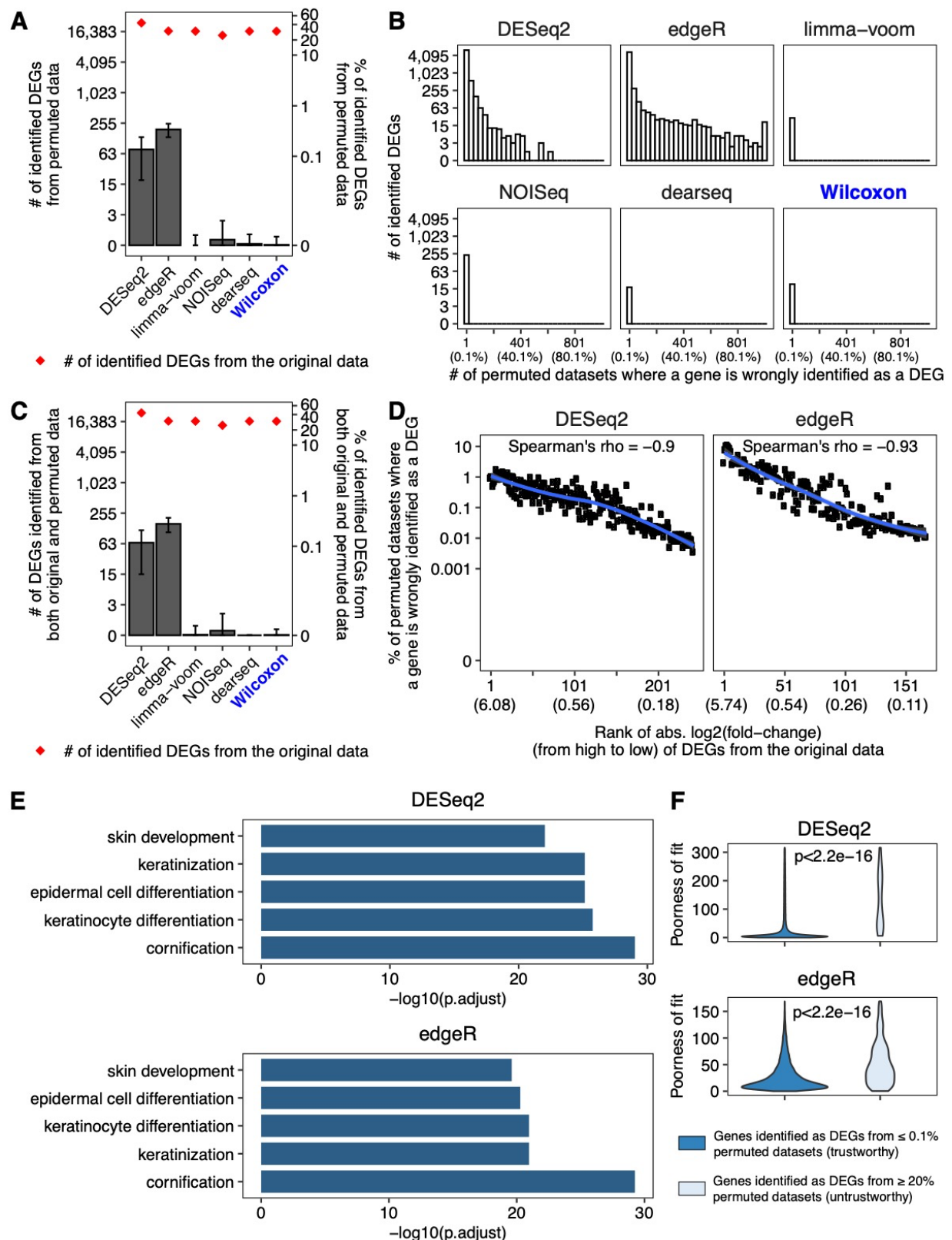


Fig. S6. Exaggerated false DEGs identified by DESeq2 and edgeR from adipose (subcutaneous vs. visceral) GTEx RNA-seq datasets.

A. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

B. The distributions of the number of permuted datasets where a gene was mistakenly identified as a DEG. The percentages corresponding to the numbers are listed in parentheses below the numbers.

C. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from both the original dataset and any of the 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

D. Percentage of permuted datasets where a DEG identified from the original dataset was also identified as a DEG. The genes are sorted by absolute $\log_2(\text{fold-change})$ in the original dataset in decreasing order and the average values of each 100 genes are shown. The absolute $\log_2(\text{fold-change})$ values corresponding to the ranks are listed in parentheses below the ranks. The line is fitted using the loess method, and the shaded areas represent 95% confidential intervals.

E. GO term enrichment for the DEGs identified from at least 10% permuted datasets. The top 5 enriched biological processes GO terms are shown. The analyses were performed using R package clusterProfiler. P.adjust represents the adjusted p-value using the Benjamini & Hochberg method.

F. Violin plots showing the poorness of fitting the negative binomial model to the genes identified by DESeq2 or edgeR as DEGs from $\geq 20\%$ vs. $\leq 0.1\%$ permuted datasets. The poorness of fit for each gene is defined as its negative $\log_{10}(\text{p-value})$ from the Pearson's chi-squared test for the negative binomial distribution. The p-value in each panel was calculated by the Wilcoxon rank-sum test to compare the two groups of genes' poorness-of-fit values.

Fig. S7

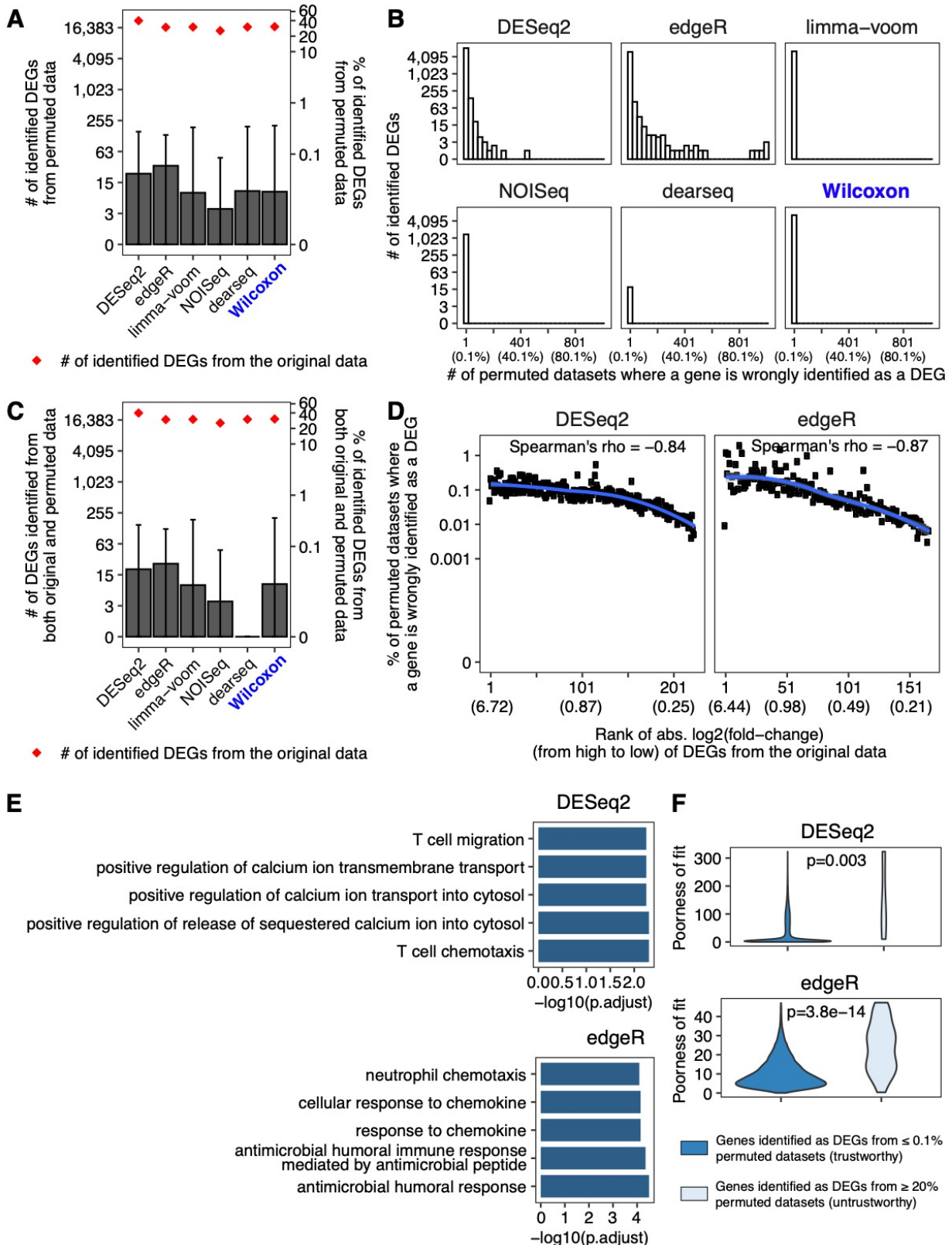


Fig. S7. Exaggerated false DEGs identified by DESeq2 and edgeR from brain (amygdala vs. spinal cord) GTEx RNA-seq datasets.

A. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of any genes (right y-axis) identified from 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

B. The distributions of the number of permuted datasets where a gene was mistakenly identified as a DEG. The percentages corresponding to the numbers are listed in parentheses below the numbers.

C. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of any genes (right y-axis) identified from both the original dataset and any of the 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

D. Percentage of permuted datasets where a DEG identified from the original dataset was also identified as a DEG. The genes are sorted by absolute $\log_2(\text{fold-change})$ in the original dataset in decreasing order and the average values of each 100 genes are shown. The absolute $\log_2(\text{fold-change})$ values corresponding to the ranks are listed in parentheses below the ranks. The line is fitted using the loess method, and the shaded areas represent 95% confidential intervals.

E. GO term enrichment for the DEGs identified from at least 10% permuted datasets. The top 5 enriched biological processes GO terms are shown. The analyses were performed using R package clusterProfiler. P.adjust represents the adjusted p-value using the Benjamini & Hochberg method.

F. Violin plots showing the poorness of fitting the negative binomial model to the genes identified by DESeq2 or edgeR as DEGs from $\geq 20\%$ vs. $\leq 0.1\%$ permuted datasets. The poorness of fit for each gene is defined as its negative $\log_{10}(\text{p-value})$ from the Pearson's chi-squared test for the negative binomial distribution. The p-value in each panel was calculated by the Wilcoxon rank-sum test to compare the two groups of genes' poorness-of-fit values.

Fig. S8

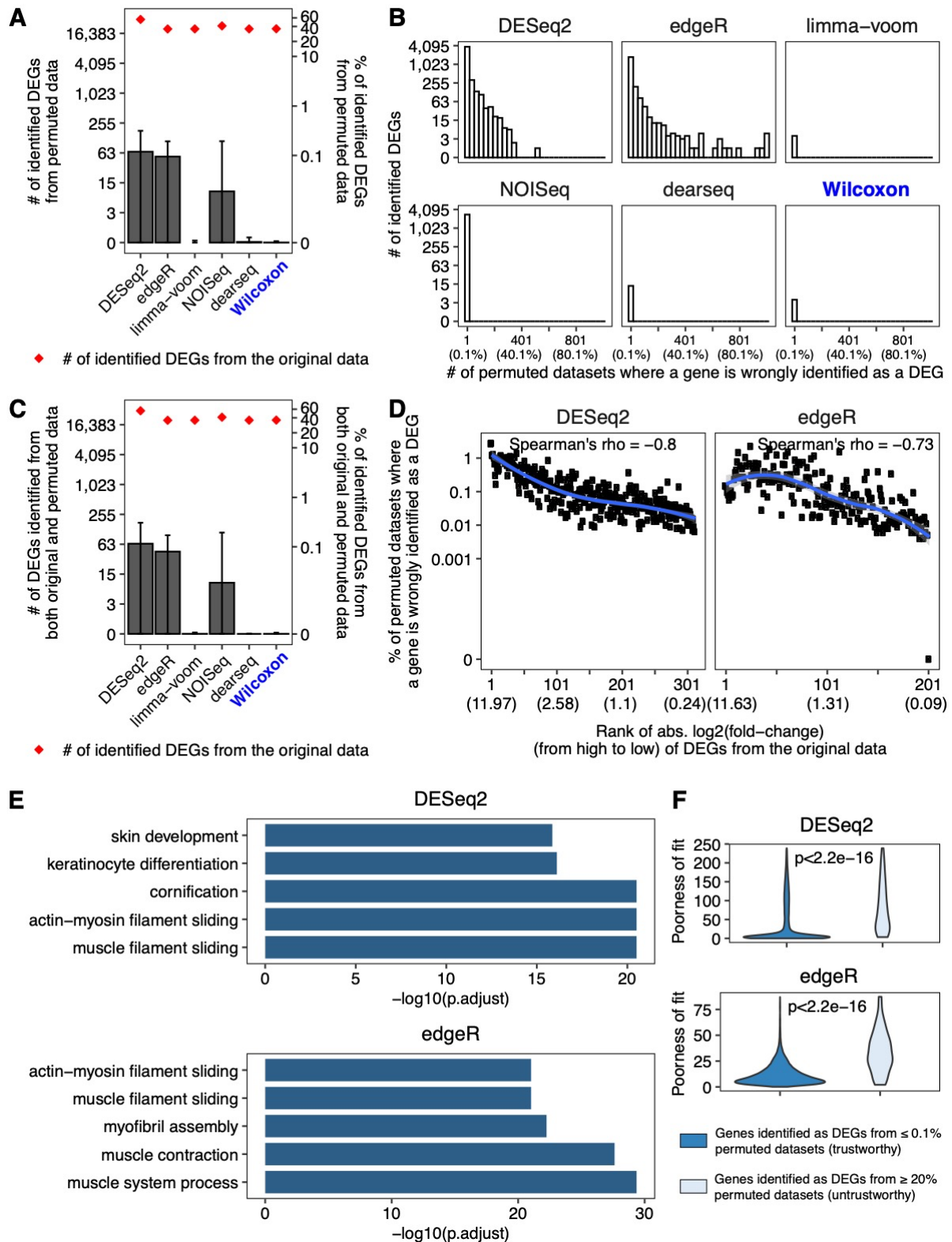


Fig. S8. Exaggerated false DEGs identified by DESeq2 and edgeR from EVB transformed lymphocytes vs. minor salivary gland GTEx RNA-seq datasets.

A. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

B. The distributions of the number of permuted datasets where a gene was mistakenly identified as a DEG. The percentages corresponding to the numbers are listed in parentheses below the numbers.

C. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from both the original dataset and any of the 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

D. Percentage of permuted datasets where a DEG identified from the original dataset was also identified as a DEG. The genes are sorted by absolute $\log_2(\text{fold-change})$ in the original dataset in decreasing order and the average values of each 100 genes are shown. The absolute $\log_2(\text{fold-change})$ values corresponding to the ranks are listed in parentheses below the ranks. The line is fitted using the loess method, and the shaded areas represent 95% confidential intervals.

E. GO term enrichment for the DEGs identified from at least 10% permuted datasets. The top 5 enriched biological processes GO terms are shown. The analyses were performed using R package clusterProfiler. P.adjust represents the adjusted p-value using the Benjamini & Hochberg method.

F. Violin plots showing the poorness of fitting the negative binomial model to the genes identified by DESeq2 or edgeR as DEGs from $\geq 20\%$ vs. $\leq 0.1\%$ permuted datasets. The poorness of fit for each gene is defined as its negative $\log_{10}(\text{p-value})$ from the Pearson's chi-squared test for the negative binomial distribution. The p-value in each panel was calculated by the Wilcoxon rank-sum test to compare the two groups of genes' poorness-of-fit values.

Fig. S9

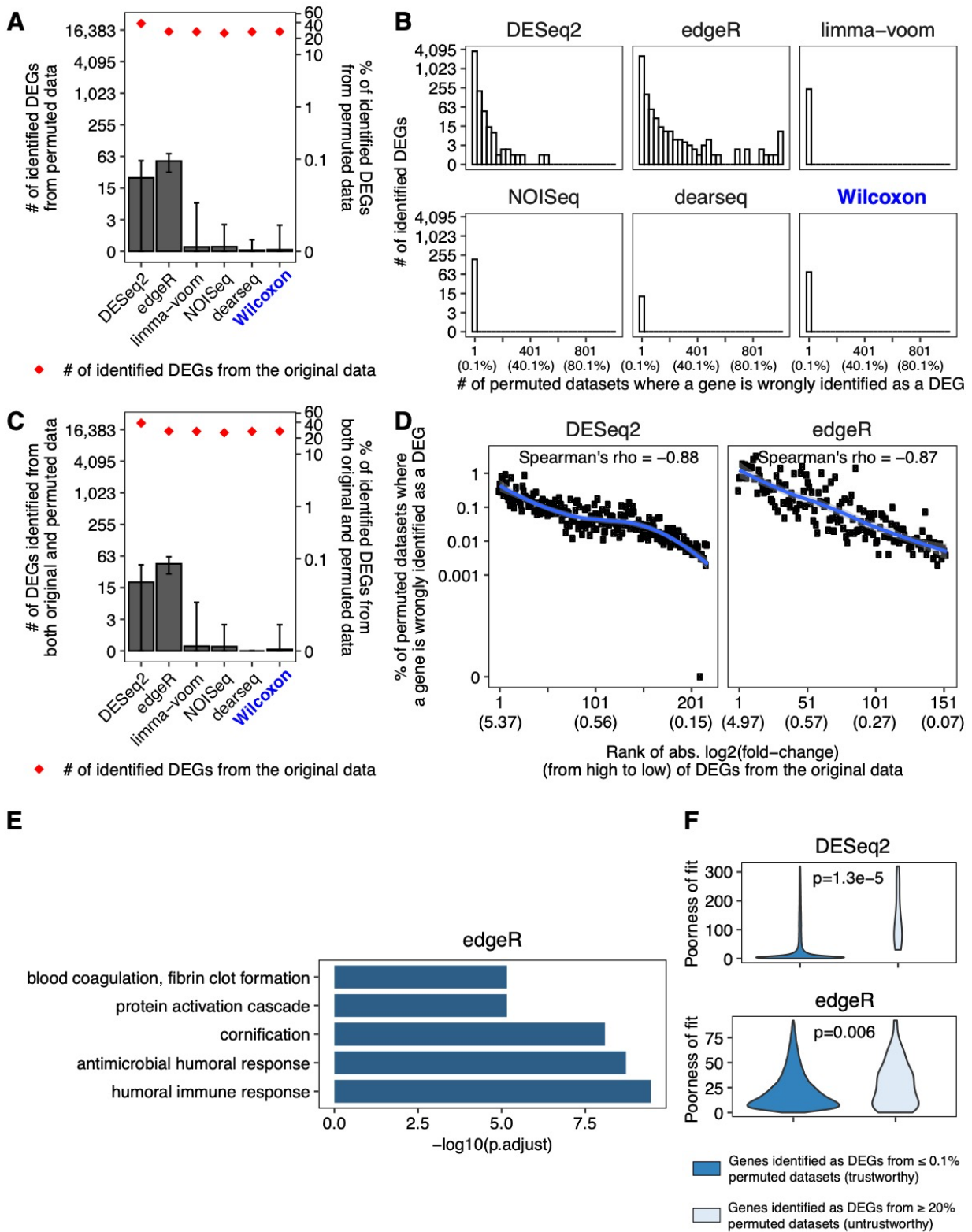


Fig. S9. Exaggerated false DEGs identified by DESeq2 and edgeR from heart (atrial appendage vs. left ventricle) GTEx RNA-seq datasets.

A. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

B. The distributions of the number of permuted datasets where a gene was mistakenly identified as a DEG. The percentages corresponding to the numbers are listed in parentheses below the numbers.

C. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from both the original dataset and any of the 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

D. Percentage of permuted datasets where a DEG identified from the original dataset was also identified as a DEG. The genes are sorted by absolute $\log_2(\text{fold-change})$ in the original dataset in decreasing order and the average values of each 100 genes are shown. The absolute $\log_2(\text{fold-change})$ values corresponding to the ranks are listed in parentheses below the ranks. The line is fitted using the loess method, and the shaded areas represent 95% confidential intervals.

E. GO term enrichment for the DEGs identified from at least 10% permuted datasets. The top 5 enriched biological processes GO terms are shown. The analyses were performed using R package clusterProfiler. P.adjust represents the adjusted p-value using the Benjamini & Hochberg method.

F. Violin plots showing the poorness of fitting the negative binomial model to the genes identified by DESeq2 or edgeR as DEGs from $\geq 20\%$ vs. $\leq 0.1\%$ permuted datasets. The poorness of fit for each gene is defined as its negative $\log_{10}(\text{p-value})$ from the Pearson's chi-squared test for the negative binomial distribution. The p-value in each panel was calculated by the Wilcoxon rank-sum test to compare the two groups of genes' poorness-of-fit values.

Fig. S10

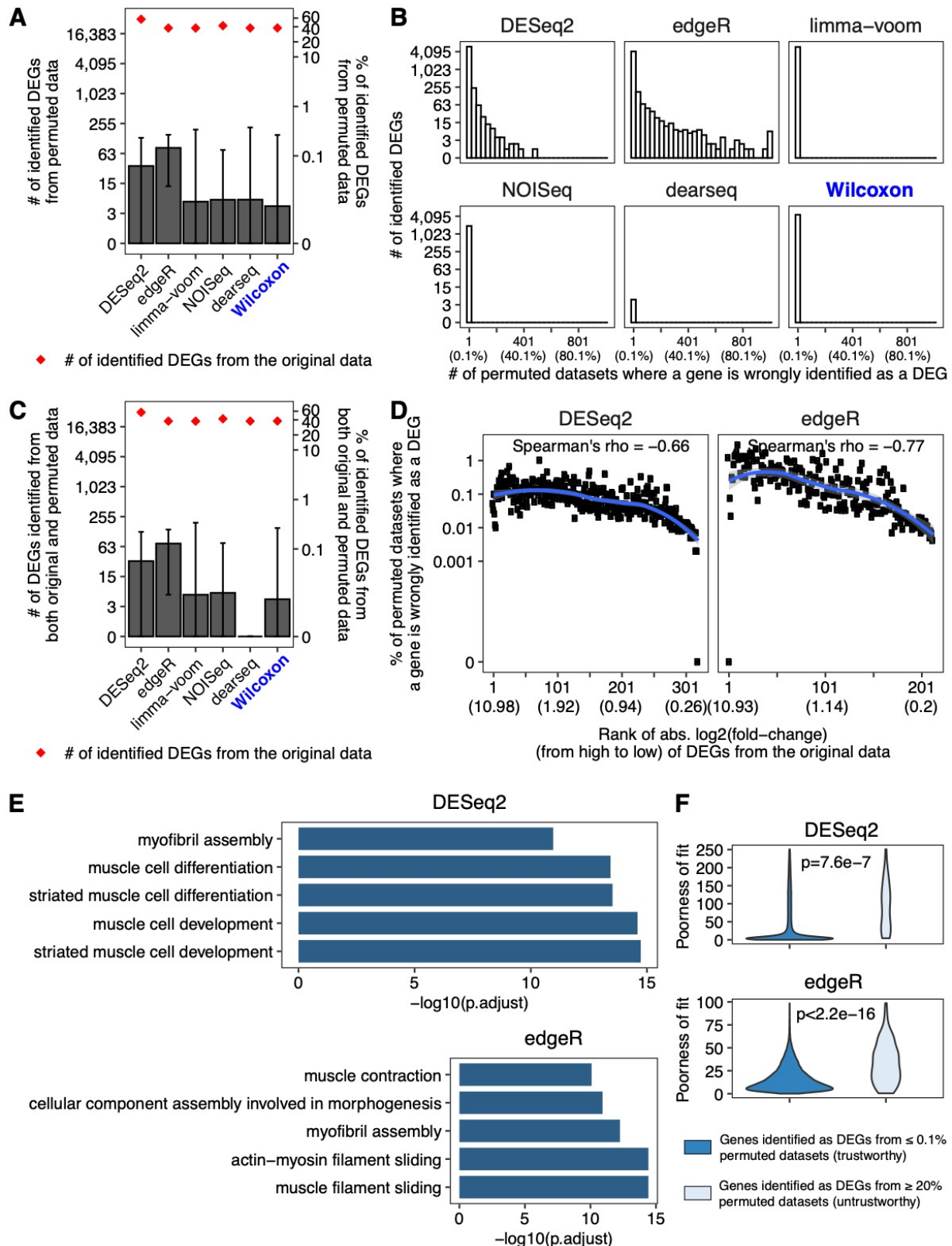


Fig. S10. Exaggerated false DEGs identified by DESeq2 and edgeR from prostate vs. brain cortex GTEx RNA-seq datasets.

A. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

B. The distributions of the number of permuted datasets where a gene was mistakenly identified as a DEG. The percentages corresponding to the numbers are listed in parentheses below the numbers.

C. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from both the original dataset and any of the 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

D. Percentage of permuted datasets where a DEG identified from the original dataset was also identified as a DEG. The genes are sorted by absolute $\log_2(\text{fold-change})$ in the original dataset in decreasing order and the average values of each 100 genes are shown. The absolute $\log_2(\text{fold-change})$ values corresponding to the ranks are listed in parentheses below the ranks. The line is fitted using the loess method, and the shaded areas represent 95% confidential intervals.

E. GO term enrichment for the DEGs identified from at least 10% permuted datasets. The top 5 enriched biological processes GO terms are shown. The analyses were performed using R package clusterProfiler. P.adjust represents the adjusted p-value using the Benjamini & Hochberg method.

F. Violin plots showing the poorness of fitting the negative binomial model to the genes identified by DESeq2 or edgeR as DEGs from $\geq 20\%$ vs. $\leq 0.1\%$ permuted datasets. The poorness of fit for each gene is defined as its negative $\log_{10}(\text{p-value})$ from the Pearson's chi-squared test for the negative binomial distribution. The p-value in each panel was calculated by the Wilcoxon rank-sum test to compare the two groups of genes' poorness-of-fit values.

Fig. S11

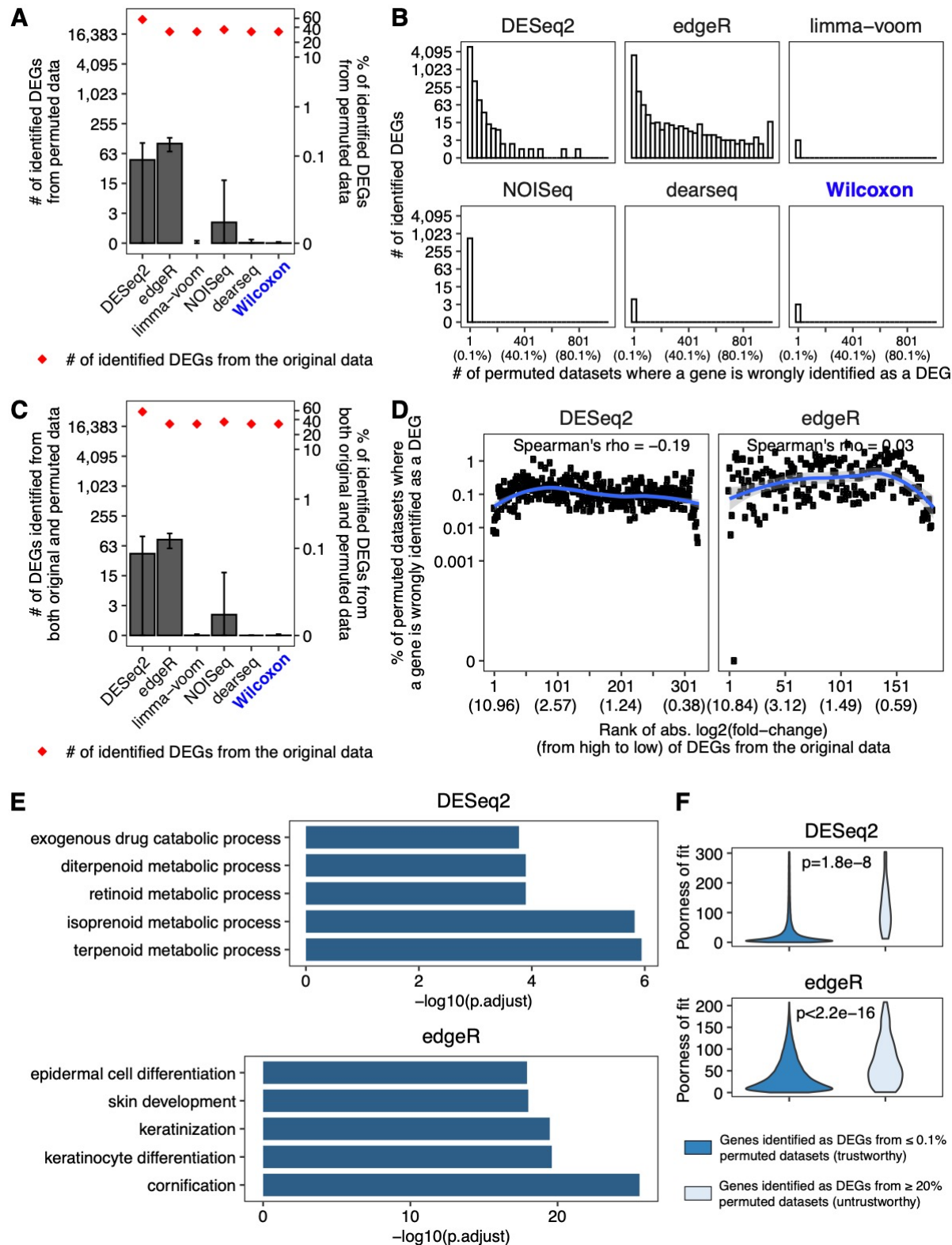


Fig. S11. Exaggerated false DEGs identified by DESeq2 and edgeR from whole blood vs. muscle GTEx RNA-seq datasets.

A. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

B. The distributions of the number of permuted datasets where a gene was mistakenly identified as a DEG. The percentages corresponding to the numbers are listed in parentheses below the numbers.

C. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from both the original dataset and any of the 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

D. Percentage of permuted datasets where a DEG identified from the original dataset was also identified as a DEG. The genes are sorted by absolute $\log_2(\text{fold-change})$ in the original dataset in decreasing order and the average values of each 100 genes are shown. The absolute $\log_2(\text{fold-change})$ values corresponding to the ranks are listed in parentheses below the ranks. The line is fitted using the loess method, and the shaded areas represent 95% confidential intervals.

E. GO term enrichment for the DEGs identified from at least 10% permuted datasets. The top 5 enriched biological processes GO terms are shown. The analyses were performed using R package clusterProfiler. P.adjust represents the adjusted p-value using the Benjamini & Hochberg method.

F. Violin plots showing the poorness of fitting the negative binomial model to the genes identified by DESeq2 or edgeR as DEGs from $\geq 20\%$ vs. $\leq 0.1\%$ permuted datasets. The poorness of fit for each gene is defined as its negative $\log_{10}(\text{p-value})$ from the Pearson's chi-squared test for the negative binomial distribution. The p-value in each panel was calculated by the Wilcoxon rank-sum test to compare the two groups of genes' poorness-of-fit values.

Fig. S12

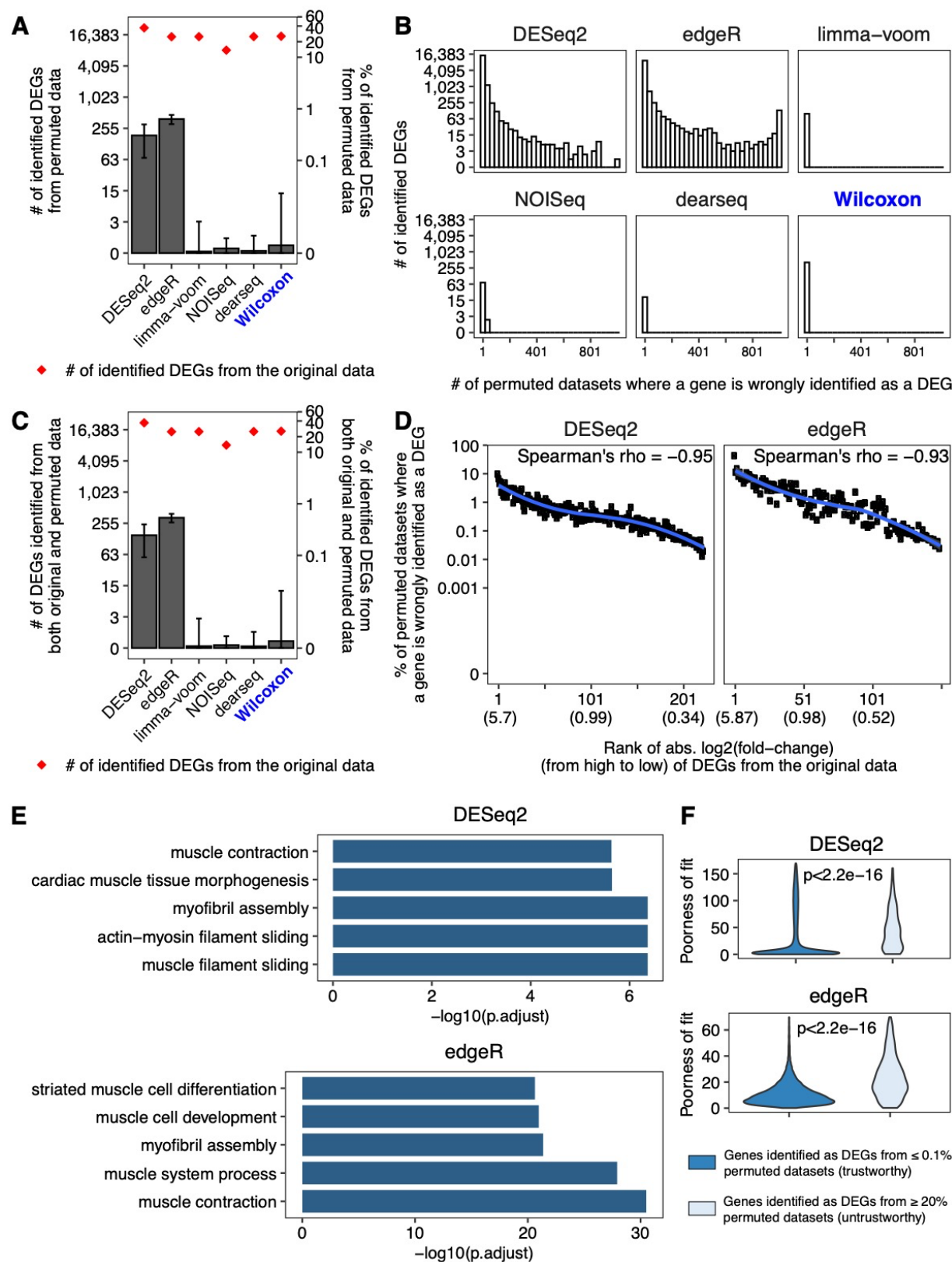


Fig. S12. Exaggerated false DEGs identified by DESeq2 and edgeR from BRCA (tumor vs. normal) TCGA RNA-seq datasets.

A. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

B. The distributions of the number of permuted datasets where a gene was mistakenly identified as a DEG. The percentages corresponding to the numbers are listed in parentheses below the numbers.

C. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from both the original dataset and any of the 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

D. Percentage of permuted datasets where a DEG identified from the original dataset was also identified as a DEG. The genes are sorted by absolute $\log_2(\text{fold-change})$ in the original dataset in decreasing order and the average values of each 100 genes are shown. The absolute $\log_2(\text{fold-change})$ values corresponding to the ranks are listed in parentheses below the ranks. The line is fitted using the loess method, and the shaded areas represent 95% confidential intervals.

E. GO term enrichment for the DEGs identified from at least 10% permuted datasets. The top 5 enriched biological processes GO terms are shown. The analyses were performed using R package clusterProfiler. P.adjust represents the adjusted p-value using the Benjamini & Hochberg method.

F. Violin plots showing the poorness of fitting the negative binomial model to the genes identified by DESeq2 or edgeR as DEGs from $\geq 20\%$ vs. $\leq 0.1\%$ permuted datasets. The poorness of fit for each gene is defined as its negative $\log_{10}(\text{p-value})$ from the Pearson's chi-squared test for the negative binomial distribution. The p-value in each panel was calculated by the Wilcoxon rank-sum test to compare the two groups of genes' poorness-of-fit values.

Fig. S13

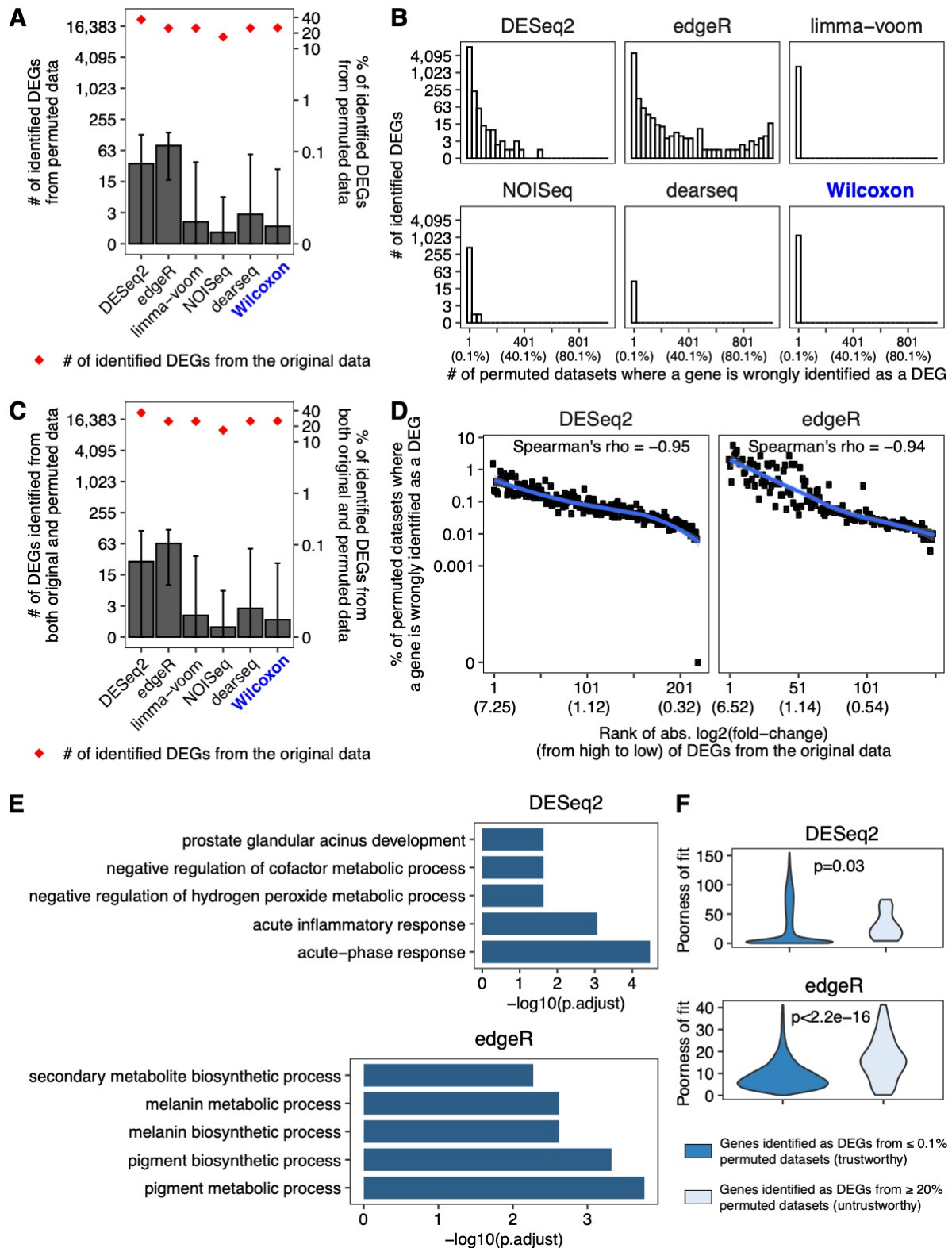


Fig. S13. Exaggerated false DEGs identified by DESeq2 and edgeR from KIRC (tumor vs. normal) TCGA RNA-seq datasets.

A. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

B. The distributions of the number of permuted datasets where a gene was mistakenly identified as a DEG. The percentages corresponding to the numbers are listed in parentheses below the numbers.

C. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from both the original dataset and any of the 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

D. Percentage of permuted datasets where a DEG identified from the original dataset was also identified as a DEG. The genes are sorted by absolute $\log_2(\text{fold-change})$ in the original dataset in decreasing order and the average values of each 100 genes are shown. The absolute $\log_2(\text{fold-change})$ values corresponding to the ranks are listed in parentheses below the ranks. The line is fitted using the loess method, and the shaded areas represent 95% confidential intervals.

E. GO term enrichment for the DEGs identified from at least 10% permuted datasets. The top 5 enriched biological processes GO terms are shown. The analyses were performed using R package clusterProfiler. P.adjust represents the adjusted p-value using the Benjamini & Hochberg method.

F. Violin plots showing the poorness of fitting the negative binomial model to the genes identified by DESeq2 or edgeR as DEGs from $\geq 20\%$ vs. $\leq 0.1\%$ permuted datasets. The poorness of fit for each gene is defined as its negative $\log_{10}(\text{p-value})$ from the Pearson's chi-squared test for the negative binomial distribution. The p-value in each panel was calculated by the Wilcoxon rank-sum test to compare the two groups of genes' poorness-of-fit values.

Fig. S14

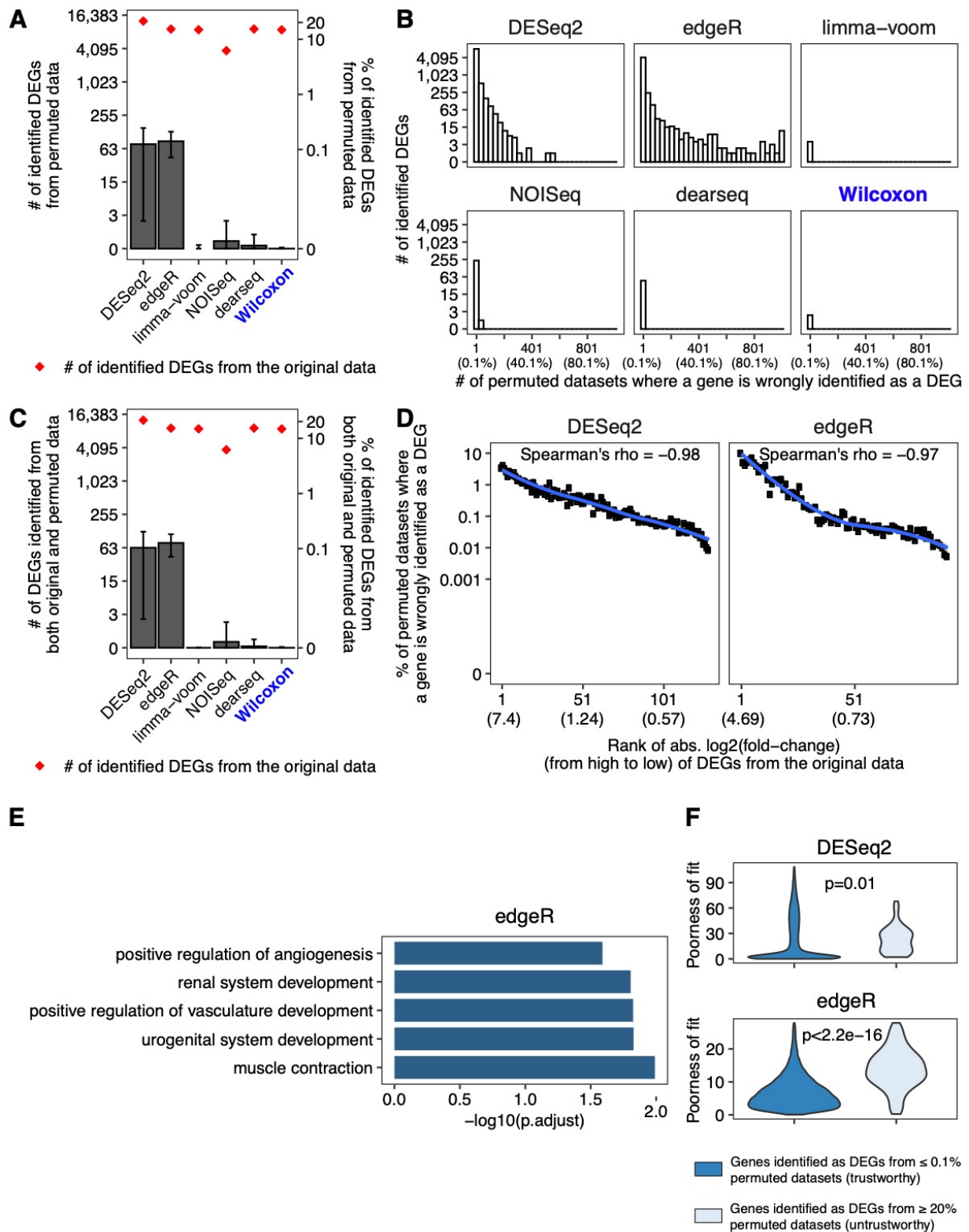


Fig. S14. Exaggerated false DEGs identified by DESeq2 and edgeR from LIHC (tumor vs. normal) TCGA RNA-seq datasets.

A. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

B. The distributions of the number of permuted datasets where a gene was mistakenly identified as a DEG. The percentages corresponding to the numbers are listed in parentheses below the numbers.

C. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from both the original dataset and any of the 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

D. Percentage of permuted datasets where a DEG identified from the original dataset was also identified as a DEG. The genes are sorted by absolute $\log_2(\text{fold-change})$ in the original dataset in decreasing order and the average values of each 100 genes are shown. The absolute $\log_2(\text{fold-change})$ values corresponding to the ranks are listed in parentheses below the ranks. The line is fitted using the loess method, and the shaded areas represent 95% confidential intervals.

E. GO term enrichment for the DEGs identified from at least 10% permuted datasets. The top 5 enriched biological processes GO terms are shown. The analyses were performed using R package clusterProfiler. P.adjust represents the adjusted p-value using the Benjamini & Hochberg method.

F. Violin plots showing the poorness of fitting the negative binomial model to the genes identified by DESeq2 or edgeR as DEGs from $\geq 20\%$ vs. $\leq 0.1\%$ permuted datasets. The poorness of fit for each gene is defined as its negative $\log_{10}(\text{p-value})$ from the Pearson's chi-squared test for the negative binomial distribution. The p-value in each panel was calculated by the Wilcoxon rank-sum test to compare the two groups of genes' poorness-of-fit values.

Fig. S15

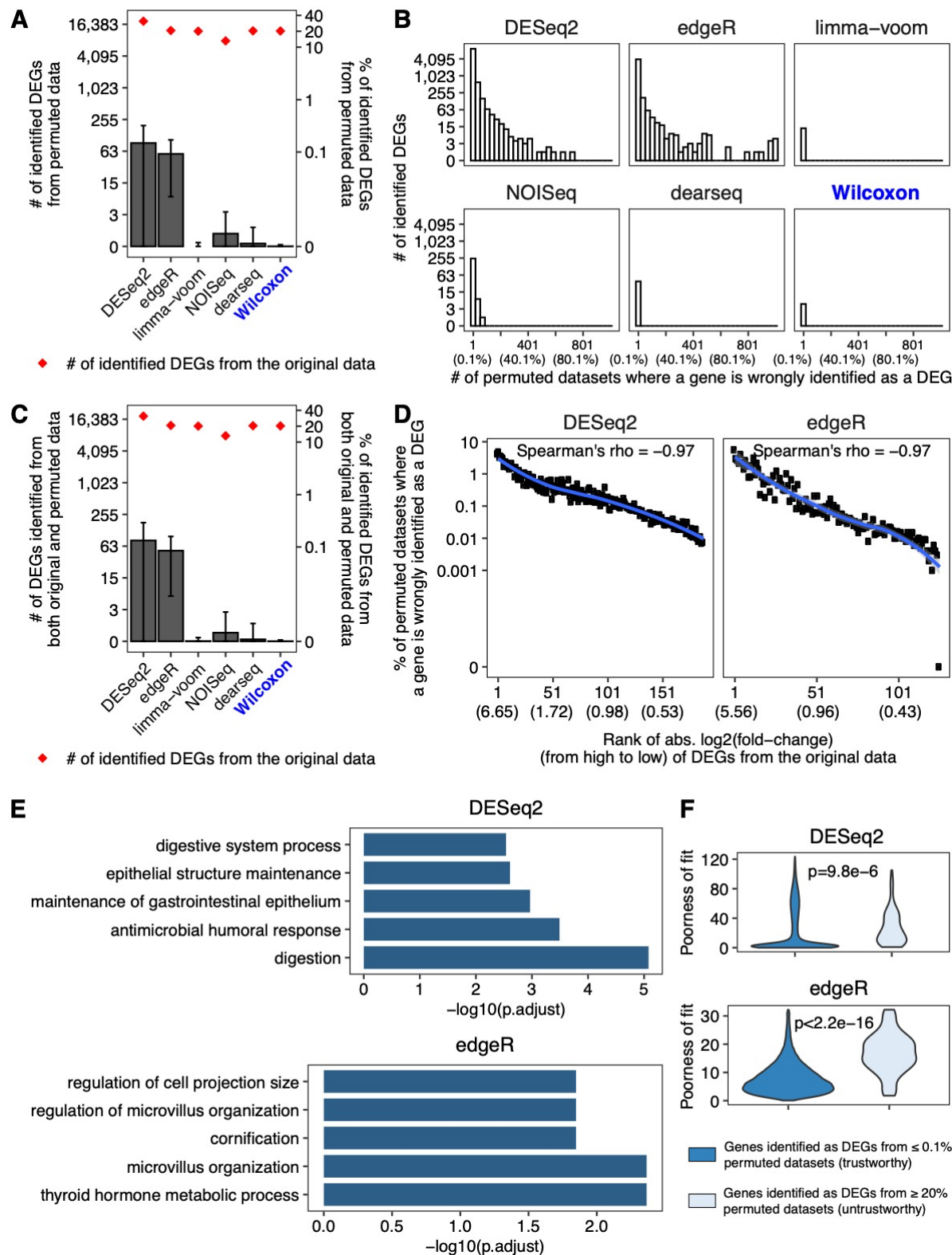


Fig. S15. Exaggerated false DEGs identified by DESeq2 and edgeR from LUAD (tumor vs. normal) TCGA RNA-seq datasets.

A. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

B. The distributions of the number of permuted datasets where a gene was mistakenly identified as a DEG. The percentages corresponding to the numbers are listed in parentheses below the numbers.

C. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from both the original dataset and any of the 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

D. Percentage of permuted datasets where a DEG identified from the original dataset was also identified as a DEG. The genes are sorted by absolute $\log_2(\text{fold-change})$ in the original dataset in decreasing order and the average values of each 100 genes are shown. The absolute $\log_2(\text{fold-change})$ values corresponding to the ranks are listed in parentheses below the ranks. The line is fitted using the loess method, and the shaded areas represent 95% confidential intervals.

E. GO term enrichment for the DEGs identified from at least 10% permuted datasets. The top 5 enriched biological processes GO terms are shown. The analyses were performed using R package clusterProfiler. P.adjust represents the adjusted p-value using the Benjamini & Hochberg method.

F. Violin plots showing the poorness of fitting the negative binomial model to the genes identified by DESeq2 or edgeR as DEGs from $\geq 20\%$ vs. $\leq 0.1\%$ permuted datasets. The poorness of fit for each gene is defined as its negative $\log_{10}(\text{p-value})$ from the Pearson's chi-squared test for the negative binomial distribution. The p-value in each panel was calculated by the Wilcoxon rank-sum test to compare the two groups of genes' poorness-of-fit values.

Fig. S16

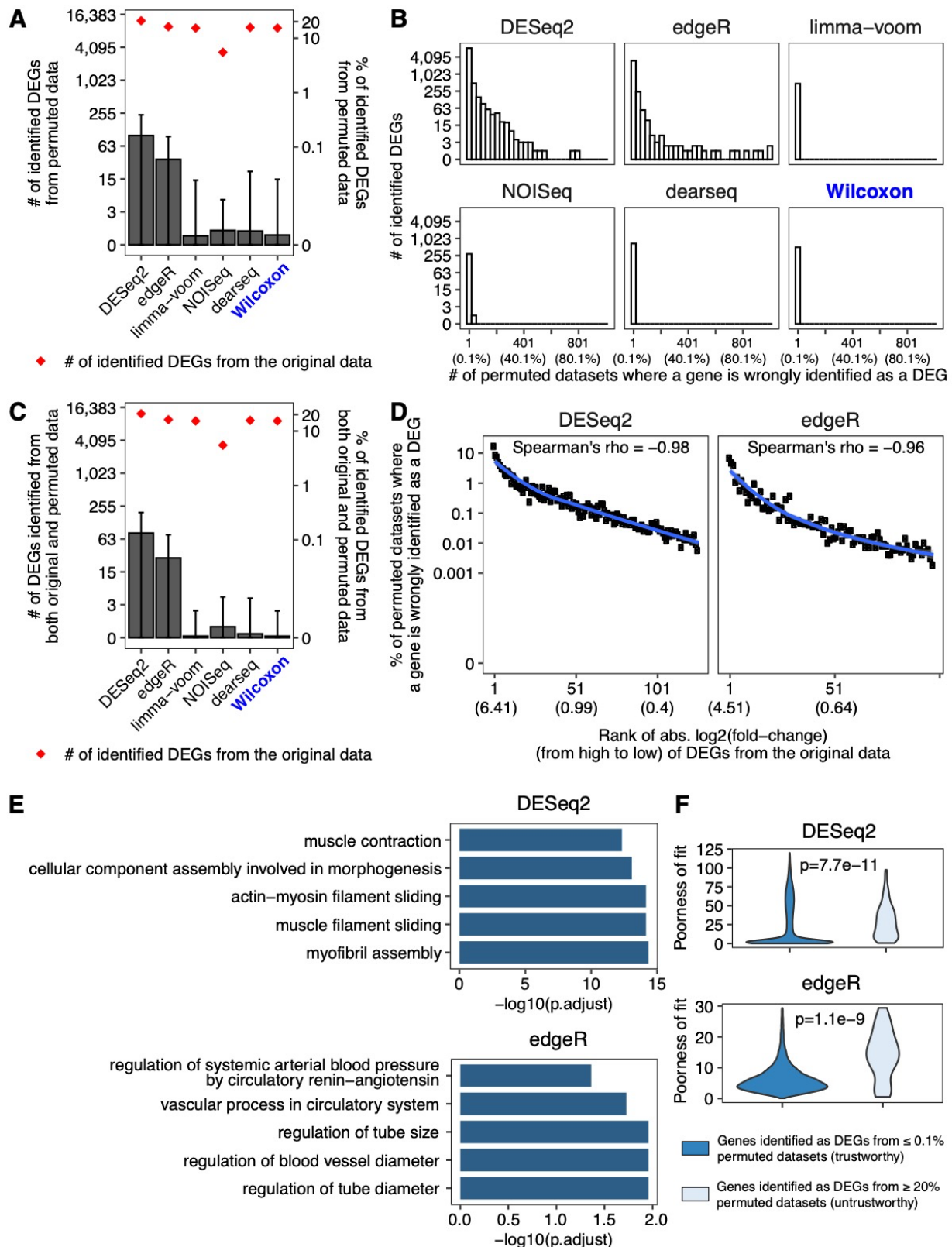


Fig. S16. Exaggerated false DEGs identified by DESeq2 and edgeR from PRAD (tumor vs. normal) TCGA RNA-seq datasets.

A. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

B. The distributions of the number of permuted datasets where a gene was mistakenly identified as a DEG. The percentages corresponding to the numbers are listed in parentheses below the numbers.

C. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from both the original dataset and any of the 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

D. Percentage of permuted datasets where a DEG identified from the original dataset was also identified as a DEG. The genes are sorted by absolute $\log_2(\text{fold-change})$ in the original dataset in decreasing order and the average values of each 100 genes are shown. The absolute $\log_2(\text{fold-change})$ values corresponding to the ranks are listed in parentheses below the ranks. The line is fitted using the loess method, and the shaded areas represent 95% confidential intervals.

E. GO term enrichment for the DEGs identified from at least 10% permuted datasets. The top 5 enriched biological processes GO terms are shown. The analyses were performed using R package clusterProfiler. P.adjust represents the adjusted p-value using the Benjamini & Hochberg method.

F. Violin plots showing the poorness of fitting the negative binomial model to the genes identified by DESeq2 or edgeR as DEGs from $\geq 20\%$ vs. $\leq 0.1\%$ permuted datasets. The poorness of fit for each gene is defined as its negative $\log_{10}(\text{p-value})$ from the Pearson's chi-squared test for the negative binomial distribution. The p-value in each panel was calculated by the Wilcoxon rank-sum test to compare the two groups of genes' poorness-of-fit values.

Fig. S17

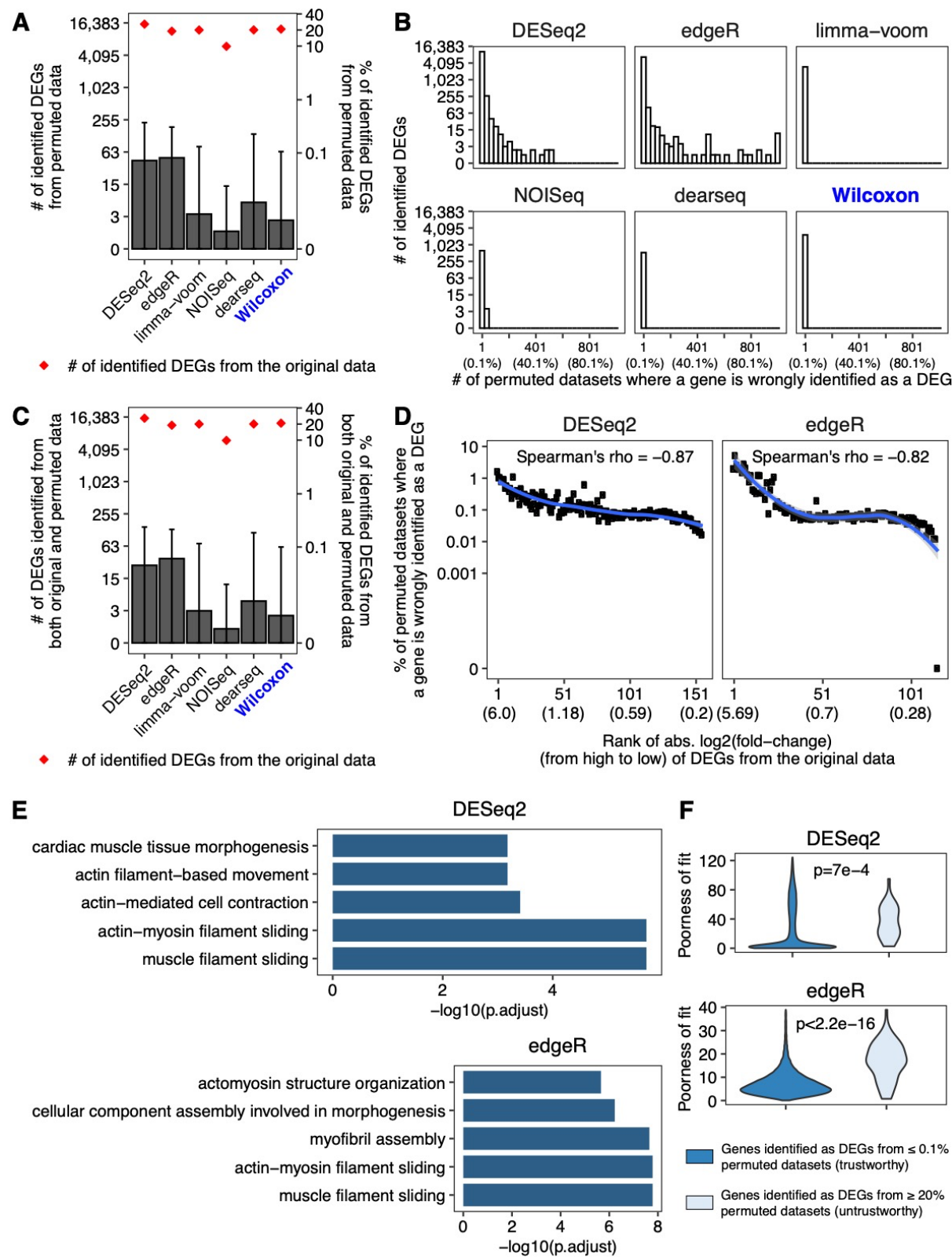


Fig. S17. Exaggerated false DEGs identified by DESeq2 and edgeR from THCA (tumor vs. normal) TCGA RNA-seq datasets.

A. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

B. The distributions of the number of permuted datasets where a gene was mistakenly identified as a DEG. The percentages corresponding to the numbers are listed in parentheses below the numbers.

C. Barplot showing the average numbers of DEGs (left y-axis) and the proportion of DEGs out of all genes (right y-axis) identified from both the original dataset and any of the 1000 permuted datasets. The error bars represent the standard deviations of 1000 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

D. Percentage of permuted datasets where a DEG identified from the original dataset was also identified as a DEG. The genes are sorted by absolute $\log_2(\text{fold-change})$ in the original dataset in decreasing order and the average values of each 100 genes are shown. The absolute $\log_2(\text{fold-change})$ values corresponding to the ranks are listed in parentheses below the ranks. The line is fitted using the loess method, and the shaded areas represent 95% confidential intervals.

E. GO term enrichment for the DEGs identified from at least 10% permuted datasets. The top 5 enriched biological processes GO terms are shown. The analyses were performed using R package clusterProfiler. P.adjust represents the adjusted p-value using the Benjamini & Hochberg method.

F. Violin plots showing the poorness of fitting the negative binomial model to the genes identified by DESeq2 or edgeR as DEGs from $\geq 20\%$ vs. $\leq 0.1\%$ permuted datasets. The poorness of fit for each gene is defined as its negative $\log_{10}(\text{p-value})$ from the Pearson's chi-squared test for the negative binomial distribution. The p-value in each panel was calculated by the Wilcoxon rank-sum test to compare the two groups of genes' poorness-of-fit values.

Fig. S18

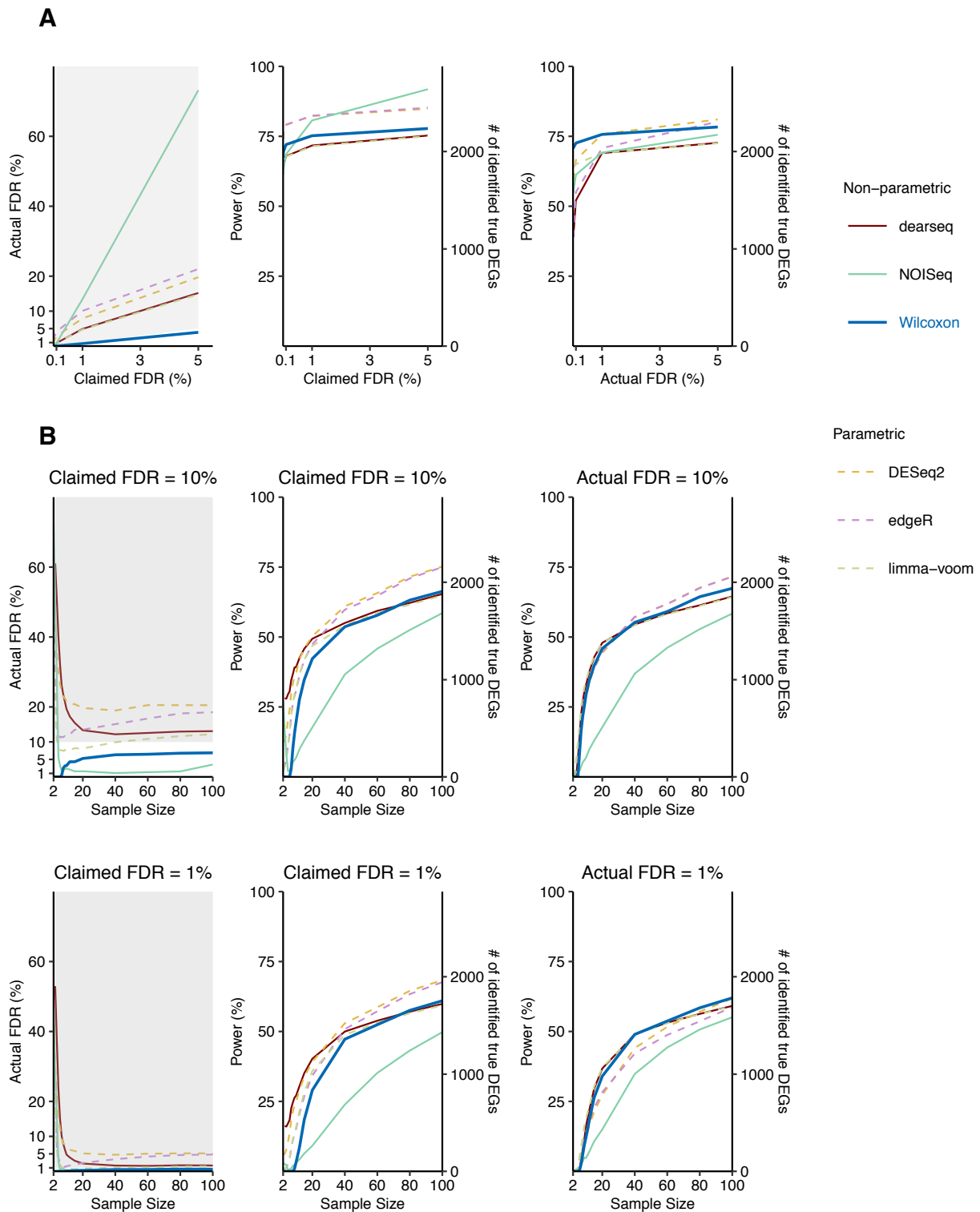


Fig. S18. The Wilcoxon rank-sum test has the best FDR control and power on heart left ventricle vs. atrial appendage GTEx datasets with semi-synthetic ground truths.

A. The FDR control (left panel), power (middle panel) given the claimed FDRs, and power given the actual FDRs (right panel) under a range of FDR thresholds from 0.001% to 5%.

B. The FDR control (left), power given the claimed FDRs (middle), and power given the actual FDRs (right) for a range of per-condition sample sizes from 2 to 100, under FDR thresholds 10% (top panels) and 1% (bottom panels). The claimed FDRs, actual FDRs, and power were all calculated as the averages of 50 randomly down-sampled datasets.

Fig. S19

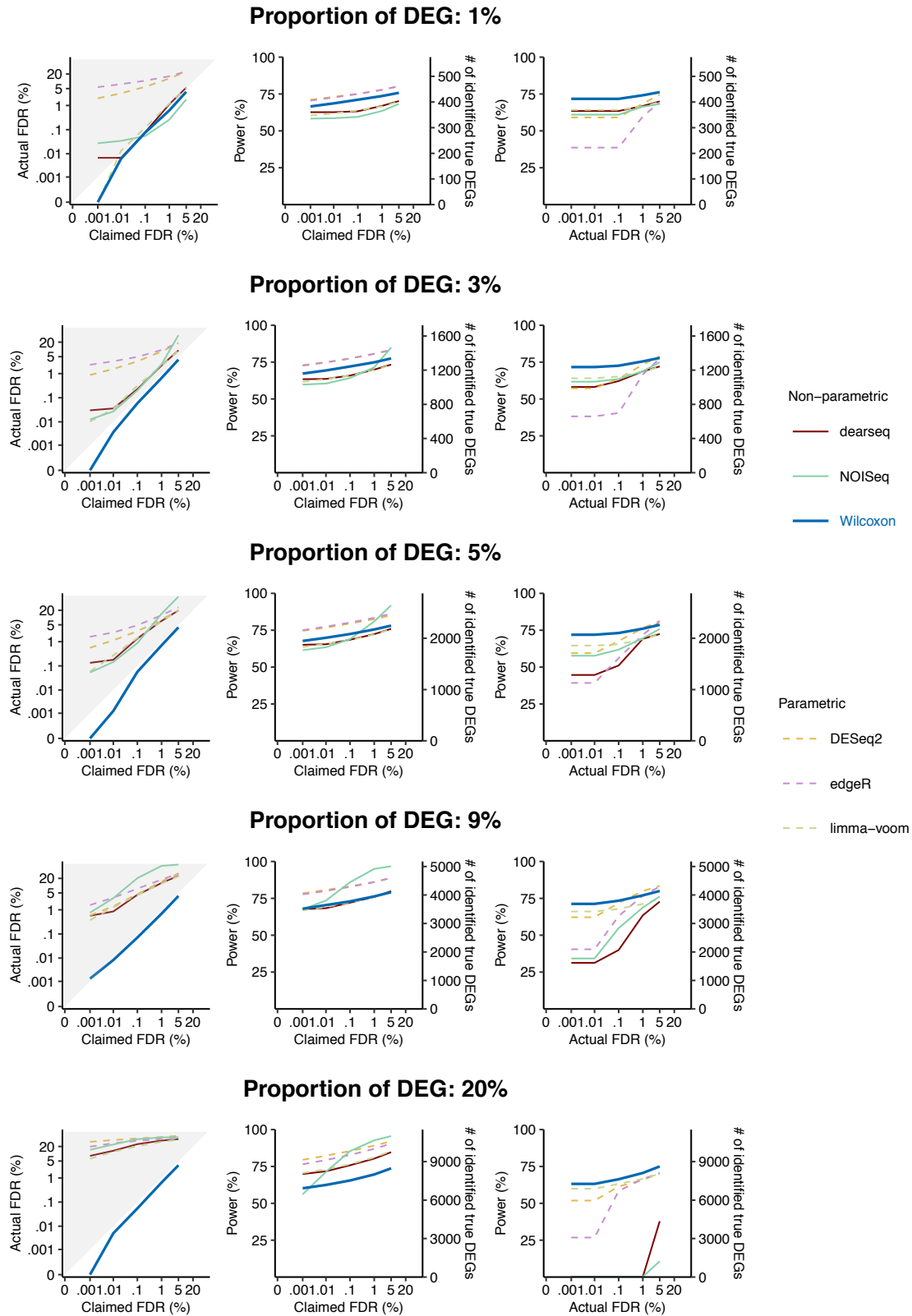


Fig. S19. The Wilcoxon rank-sum test has the best FDR control and power on heart left ventricle vs. atrial appendage GTEx datasets with semi-synthetic ground truths of varying proportions of DEGs.

The FDR control (left panel), power (middle panel) given the claimed FDRs, and power given the actual FDRs (right panel) under a range of FDR thresholds from 0.001% to 5% for semi-synthetic datasets with five proportions of DEGs (1%, 3%, 5%, 9%, and 20%).

Fig. S20

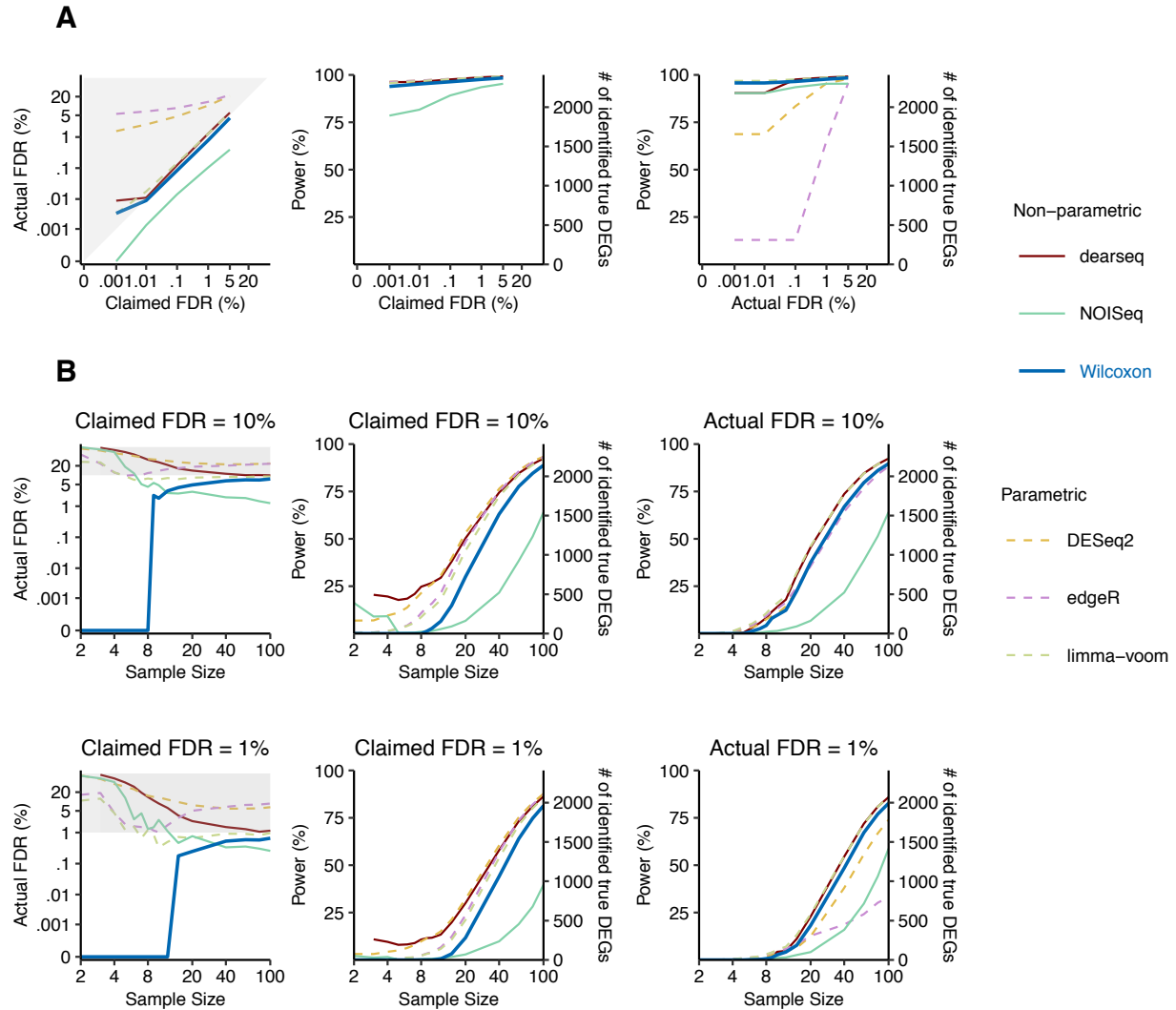


Fig. S20. The Wilcoxon rank-sum test has the best FDR control and power on adipose (subcutaneous vs. visceral) GTEx datasets with semi-synthetic ground truths.

A. The FDR control (left panel), power (middle panel) given the claimed FDRs, and power given the actual FDRs (right panel) under a range of FDR thresholds from 0.001% to 5%.

B. The FDR control (left), power given the claimed FDRs (middle), and power given the actual FDRs (right) for a range of per-condition sample sizes from 2 to 100, under FDR thresholds 10% (top panels) and 1% (bottom panels). The claimed FDRs, actual FDRs, and power were all calculated as the averages of 50 randomly down-sampled datasets.

Fig. S21

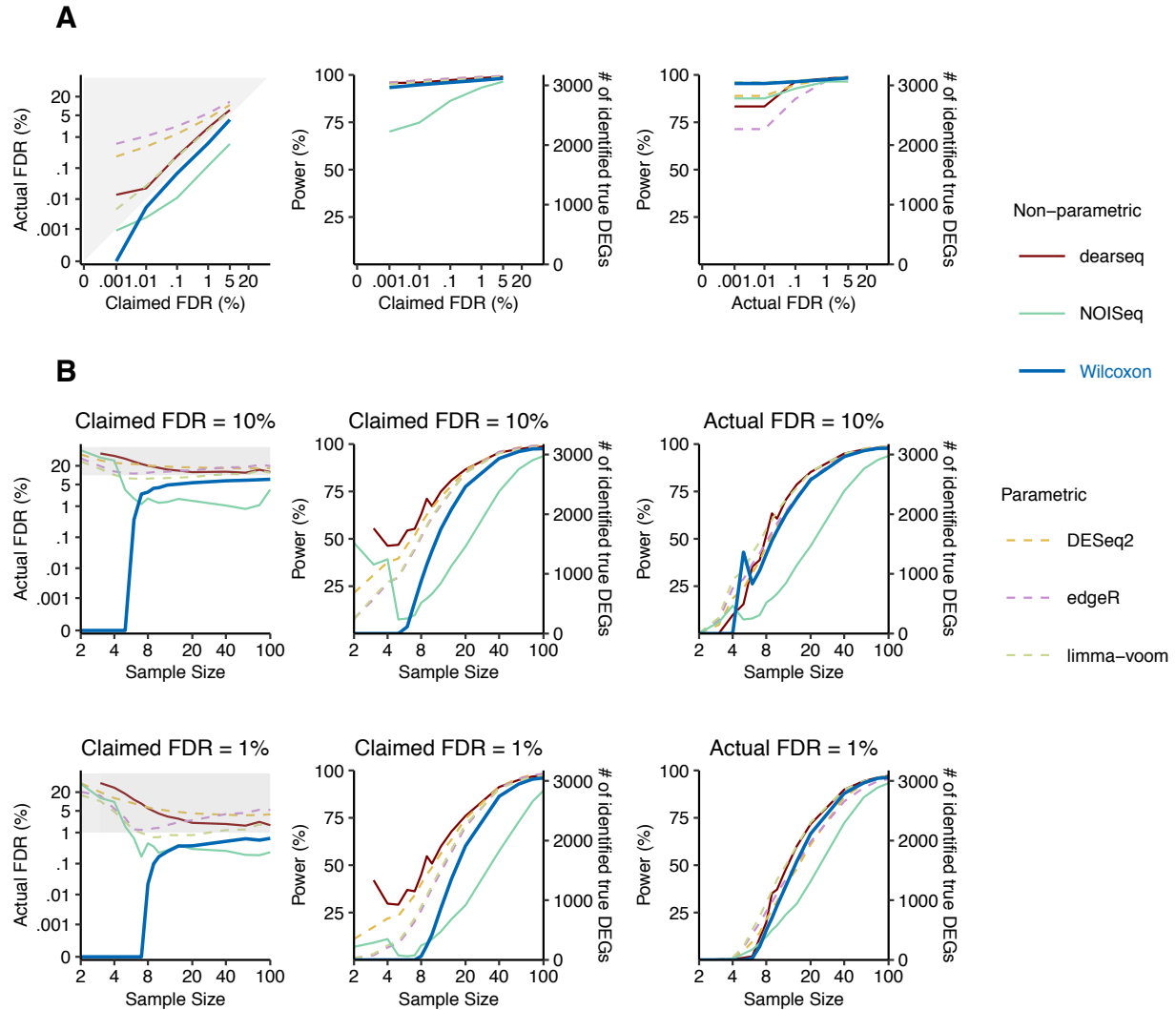


Fig. S21. The Wilcoxon rank-sum test has the best FDR control and power on brain (amygdala vs. spinal cord) GTEx datasets with semi-synthetic ground truths.

A. The FDR control (left panel), power (middle panel) given the claimed FDRs, and power given the actual FDRs (right panel) under a range of FDR thresholds from 0.001% to 5%.

B. The FDR control (left), power given the claimed FDRs (middle), and power given the actual FDRs (right) for a range of per-condition sample sizes from 2 to 100, under FDR thresholds 10% (top panels) and 1% (bottom panels). The claimed FDRs, actual FDRs, and power were all calculated as the averages of 50 randomly down-sampled datasets.

Fig. S22

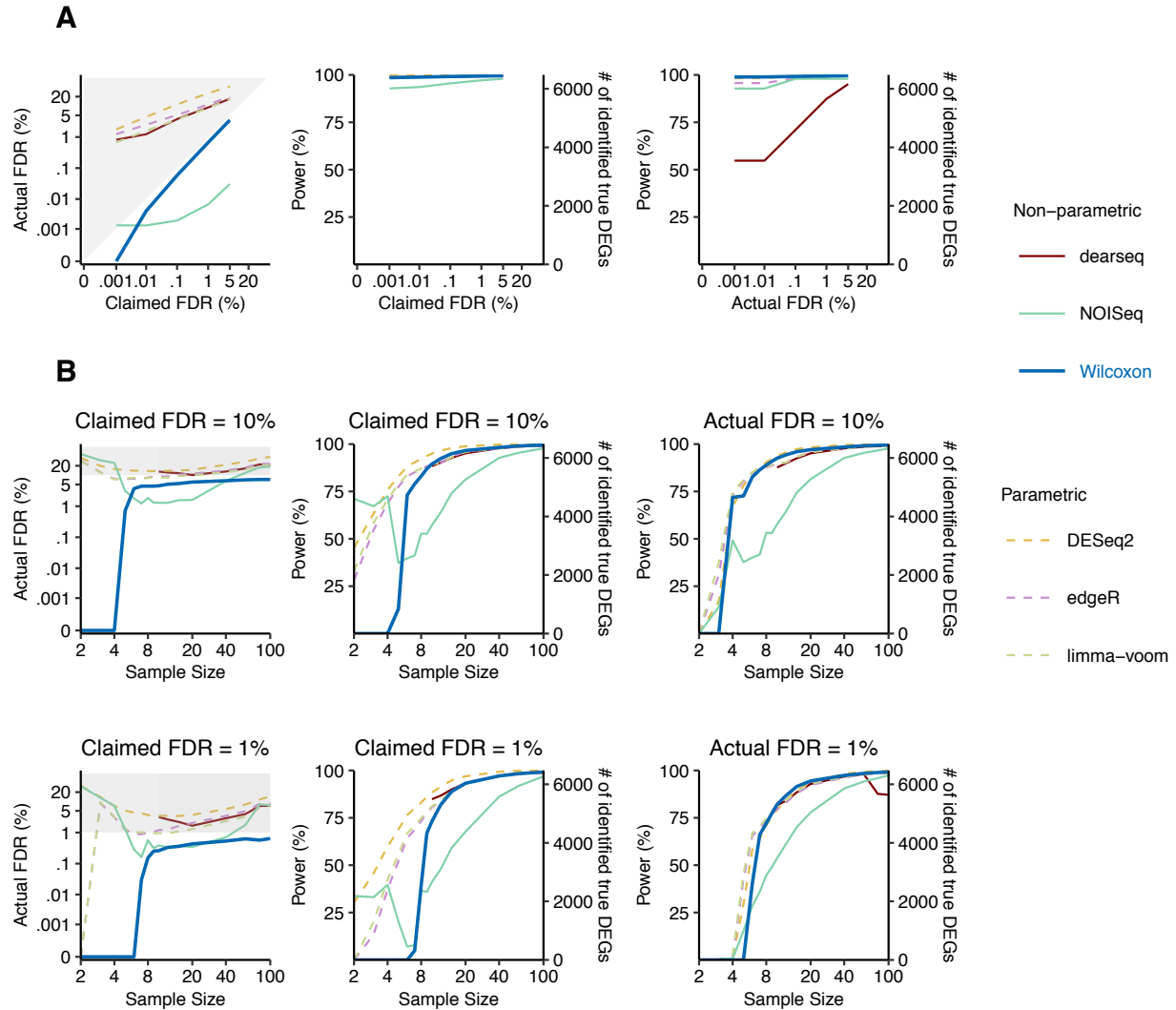


Fig. S22. The Wilcoxon rank-sum test has the best FDR control and power on EVB transformed lymphocytes vs. minor salivary gland GTEx datasets with semi-synthetic ground truths.

A. The FDR control (left panel), power (middle panel) given the claimed FDRs, and power given the actual FDRs (right panel) under a range of FDR thresholds from 0.001% to 5%.

B. The FDR control (left), power given the claimed FDRs (middle), and power given the actual FDRs (right) for a range of per-condition sample sizes from 2 to 100, under FDR thresholds 10%

(top panels) and 1% (bottom panels). The claimed FDRs, actual FDRs, and power were all calculated as the averages of 50 randomly down-sampled datasets.

Fig. S23

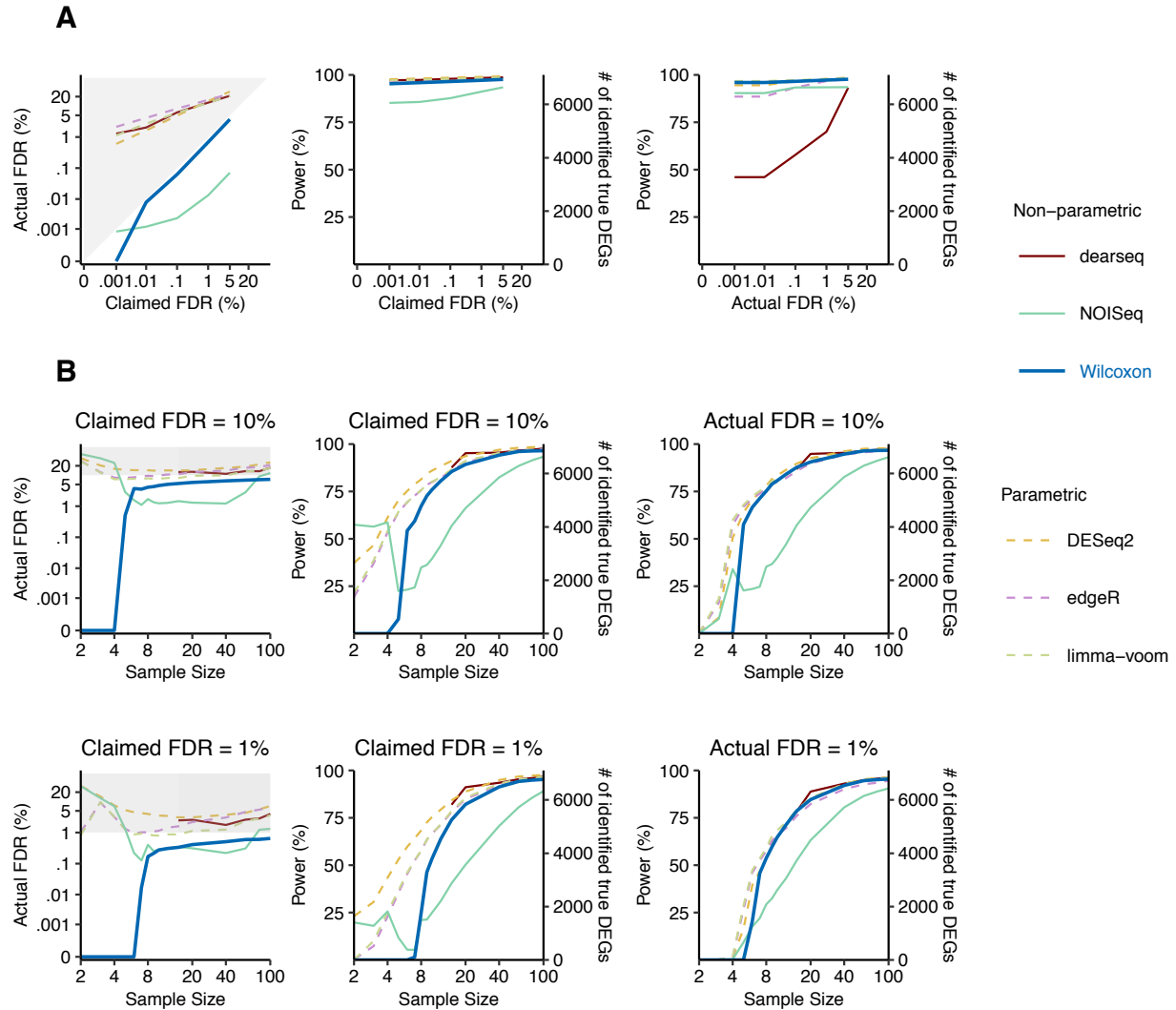


Fig. S23. The Wilcoxon rank-sum test has the best FDR control and power on prostate vs. brain cortex GTEx datasets with semi-synthetic ground truths.

A. The FDR control (left panel), power (middle panel) given the claimed FDRs, and power given the actual FDRs (right panel) under a range of FDR thresholds from 0.001% to 5%.

B. The FDR control (left), power given the claimed FDRs (middle), and power given the actual FDRs (right) for a range of per-condition sample sizes from 2 to 100, under FDR thresholds 10% (top panels) and 1% (bottom panels). The claimed FDRs, actual FDRs, and power were all calculated as the averages of 50 randomly down-sampled datasets.

Fig. S24

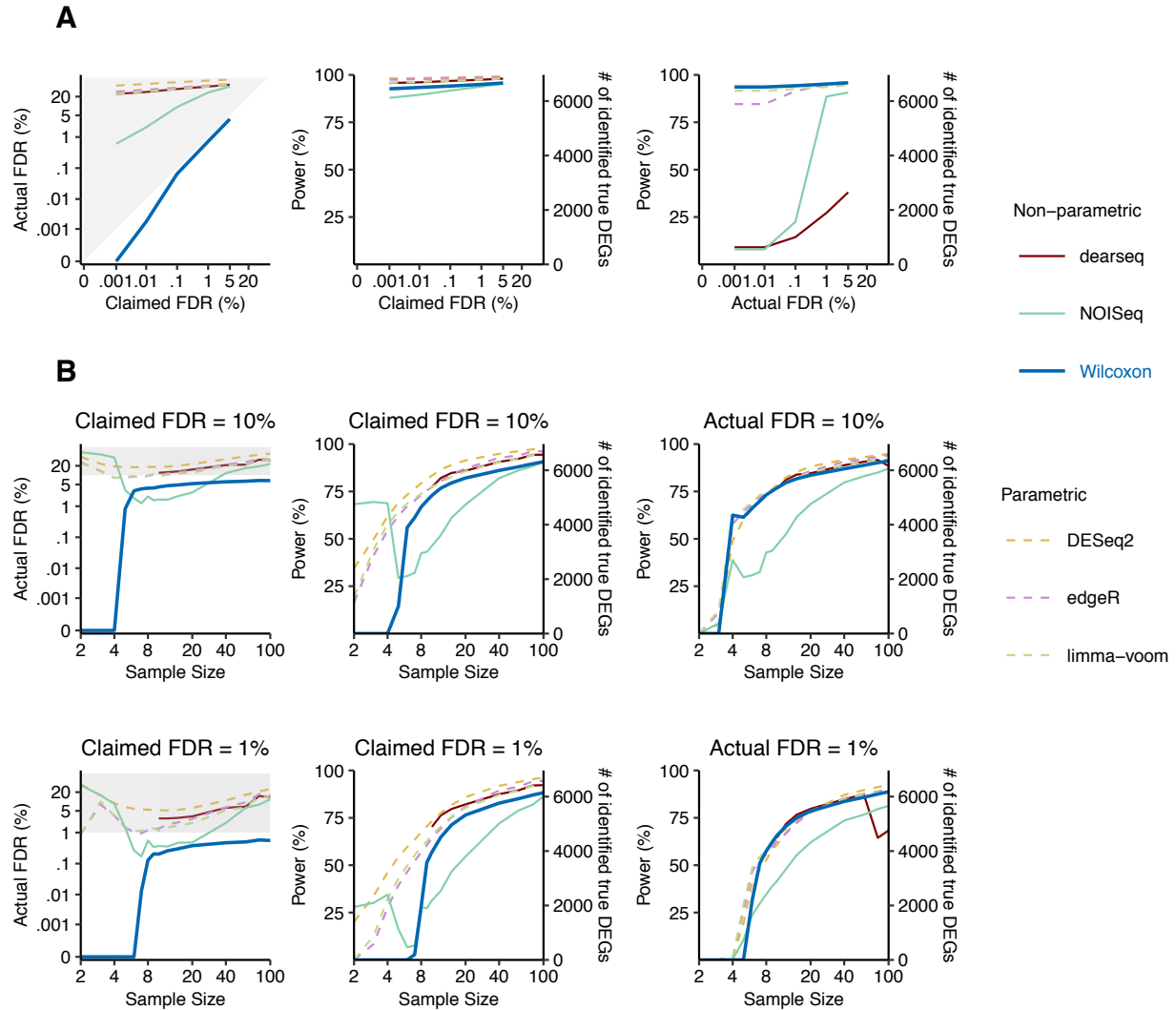


Fig. S24. The Wilcoxon rank-sum test has the best FDR control and power on whole blood vs. muscle GTEx datasets with semi-synthetic ground truths.

A. The FDR control (left panel), power (middle panel) given the claimed FDRs, and power given the actual FDRs (right panel) under a range of FDR thresholds from 0.001% to 5%.

B. The FDR control (left), power given the claimed FDRs (middle), and power given the actual FDRs (right) for a range of per-condition sample sizes from 2 to 100, under FDR thresholds 10% (top panels) and 1% (bottom panels). The claimed FDRs, actual FDRs, and power were all calculated as the averages of 50 randomly down-sampled datasets.

Fig. S25

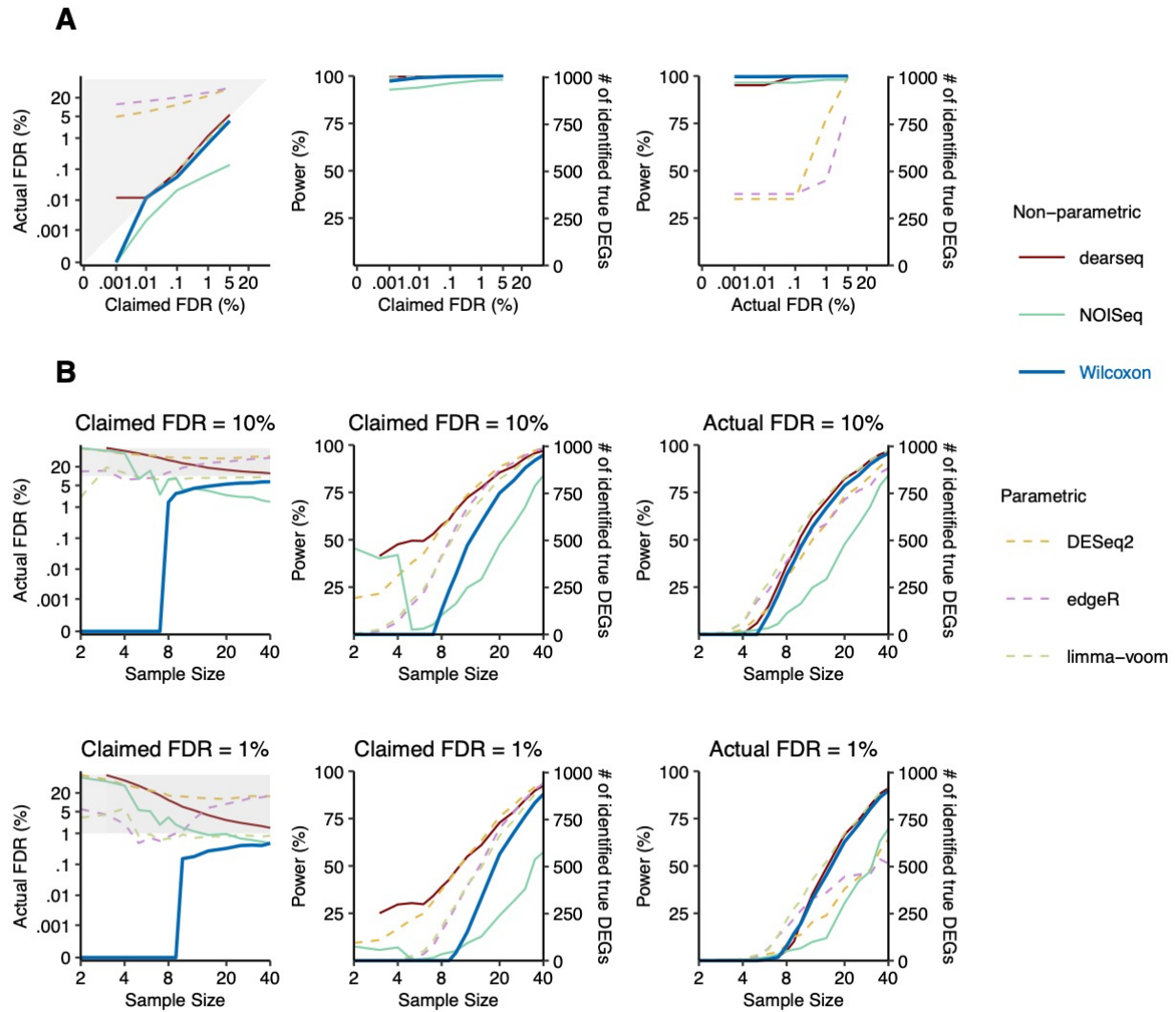


Fig. S25. The Wilcoxon rank-sum test has the best FDR control and power on BRCA TCGA datasets with semi-synthetic ground truths.

A. The FDR control (left panel), power (middle panel) given the claimed FDRs, and power given the actual FDRs (right panel) under a range of FDR thresholds from 0.001% to 5%.

B. The FDR control (left), power given the claimed FDRs (middle), and power given the actual FDRs (right) for a range of per-condition sample sizes from 2 to 100, under FDR thresholds 10% (top panels) and 1% (bottom panels). The claimed FDRs, actual FDRs, and power were all calculated as the averages of 50 randomly down-sampled datasets.

Fig. S26

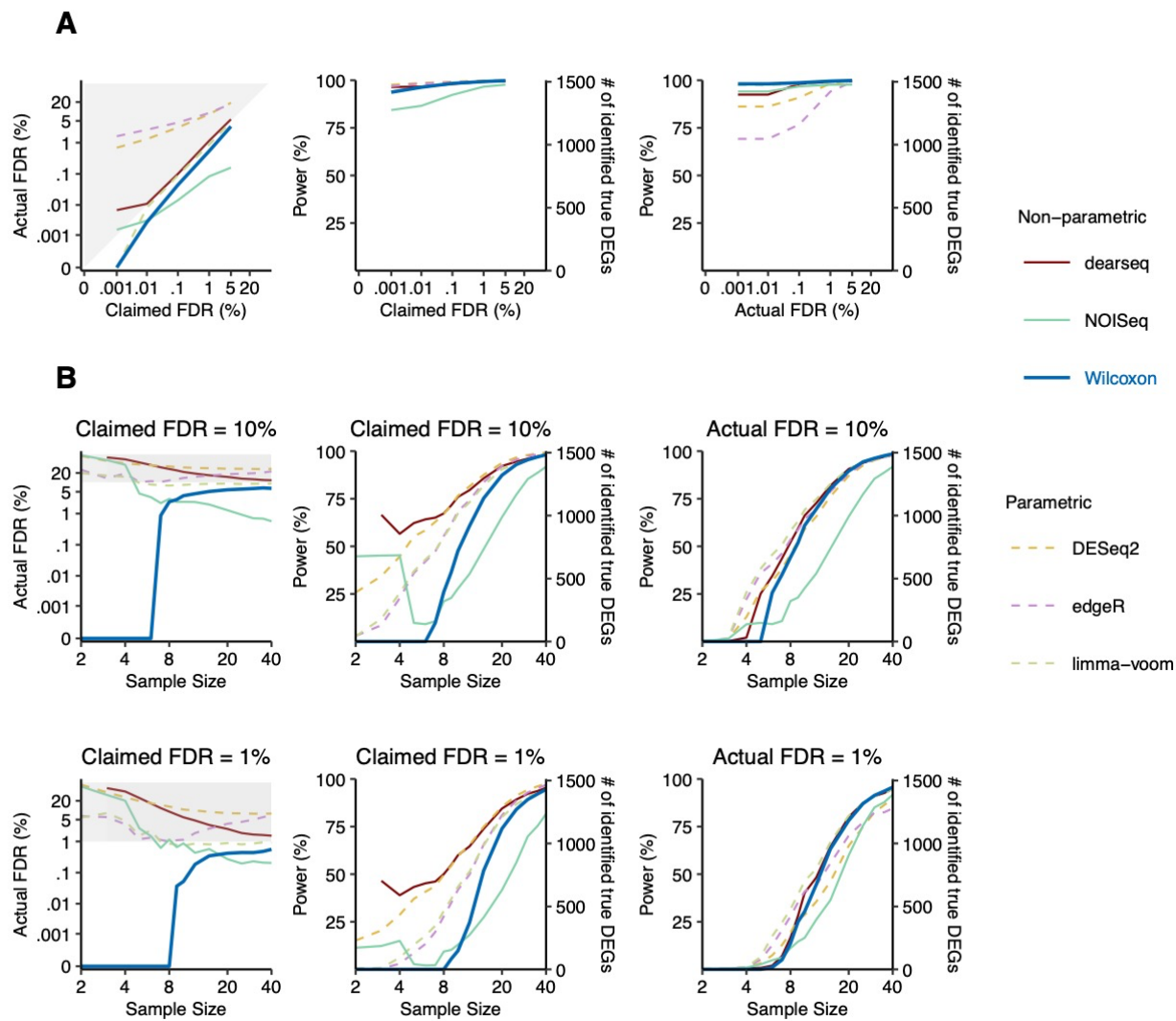


Fig. S26. The Wilcoxon rank-sum test has the best FDR control and power on KIRC

TCGA datasets with semi-synthetic ground truths.

A. The FDR control (left panel), power (middle panel) given the claimed FDRs, and power given the actual FDRs (right panel) under a range of FDR thresholds from 0.001% to 5%.

B. The FDR control (left), power given the claimed FDRs (middle), and power given the actual FDRs (right) for a range of per-condition sample sizes from 2 to 100, under FDR thresholds 10% (top panels) and 1% (bottom panels). The claimed FDRs, actual FDRs, and power were all calculated as the averages of 50 randomly down-sampled datasets.

Fig. S27

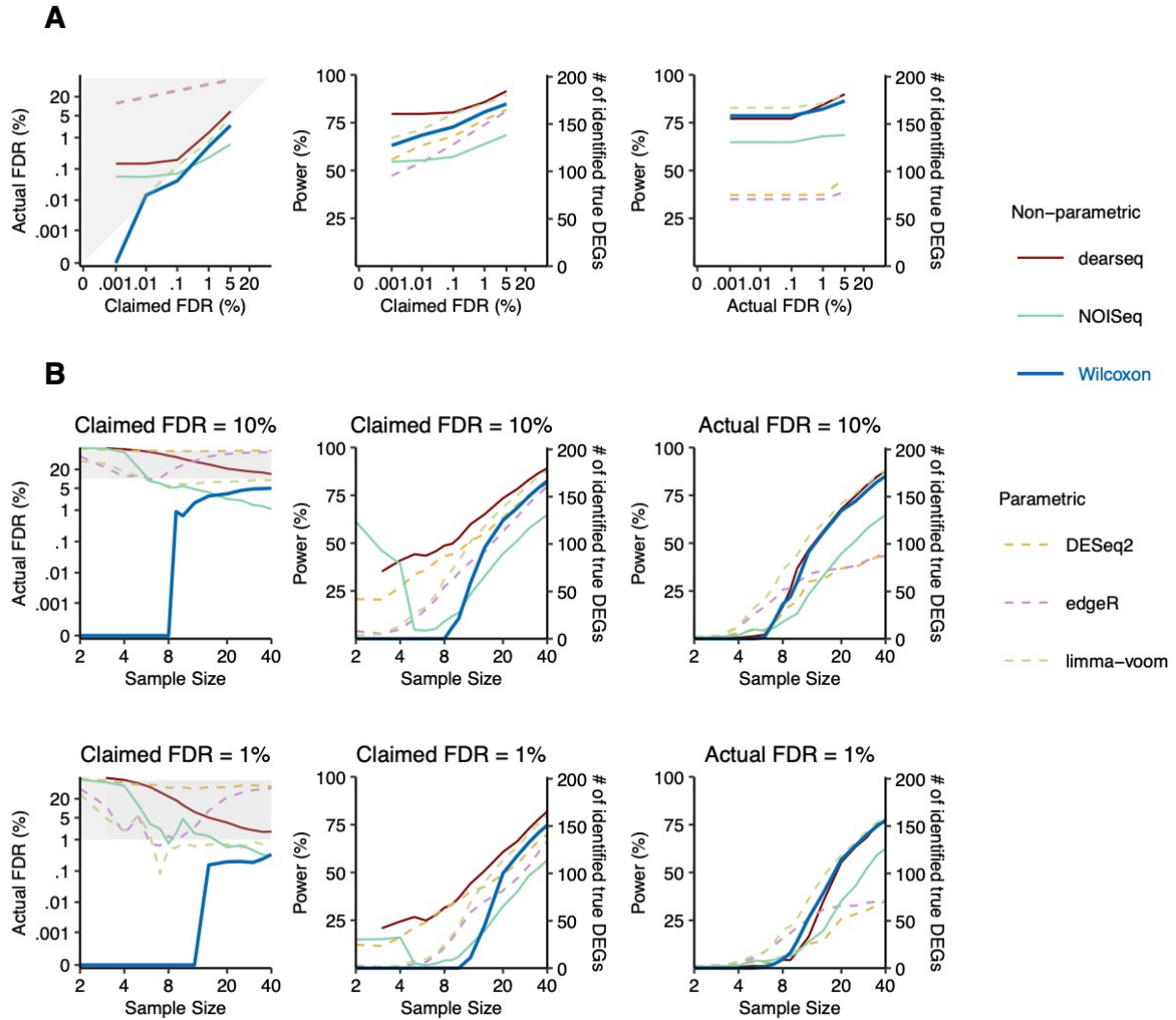


Fig. S27. The Wilcoxon rank-sum test has the best FDR control and power on LIHC TCGA datasets with semi-synthetic ground truths.

A. The FDR control (left panel), power (middle panel) given the claimed FDRs, and power given the actual FDRs (right panel) under a range of FDR thresholds from 0.001% to 5%.

B. The FDR control (left), power given the claimed FDRs (middle), and power given the actual FDRs (right) for a range of per-condition sample sizes from 2 to 100, under FDR thresholds 10% (top panels) and 1% (bottom panels). The claimed FDRs, actual FDRs, and power were all calculated as the averages of 50 randomly down-sampled datasets.

Fig. S28

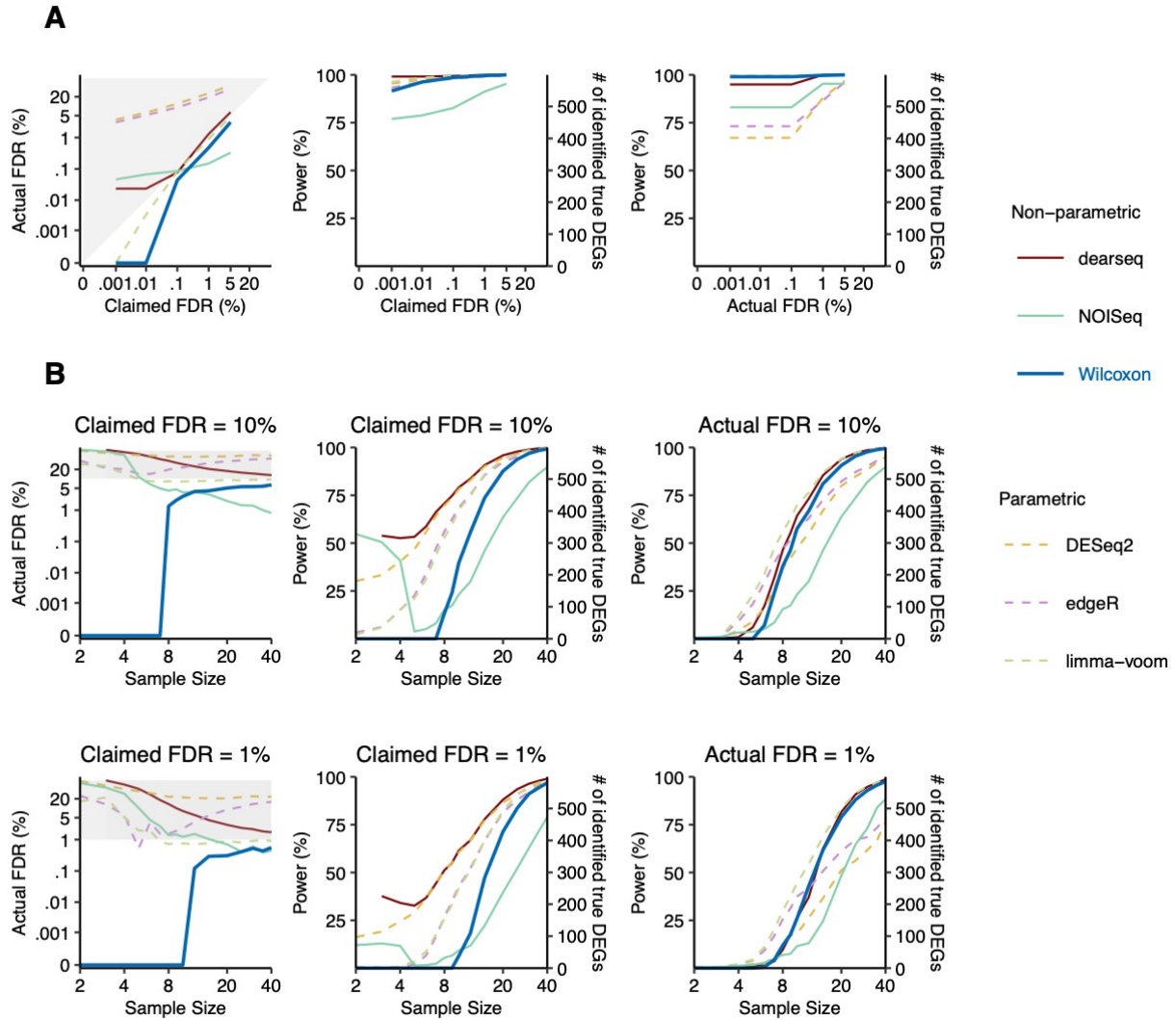


Fig. S28. The Wilcoxon rank-sum test has the best FDR control and power on LUAD

TCGA datasets with semi-synthetic ground truths.

A. The FDR control (left panel), power (middle panel) given the claimed FDRs, and power given the actual FDRs (right panel) under a range of FDR thresholds from 0.001% to 5%.

B. The FDR control (left), power given the claimed FDRs (middle), and power given the actual FDRs (right) for a range of per-condition sample sizes from 2 to 100, under FDR thresholds 10% (top panels) and 1% (bottom panels). The claimed FDRs, actual FDRs, and power were all calculated as the averages of 50 randomly down-sampled datasets.

Fig. S29

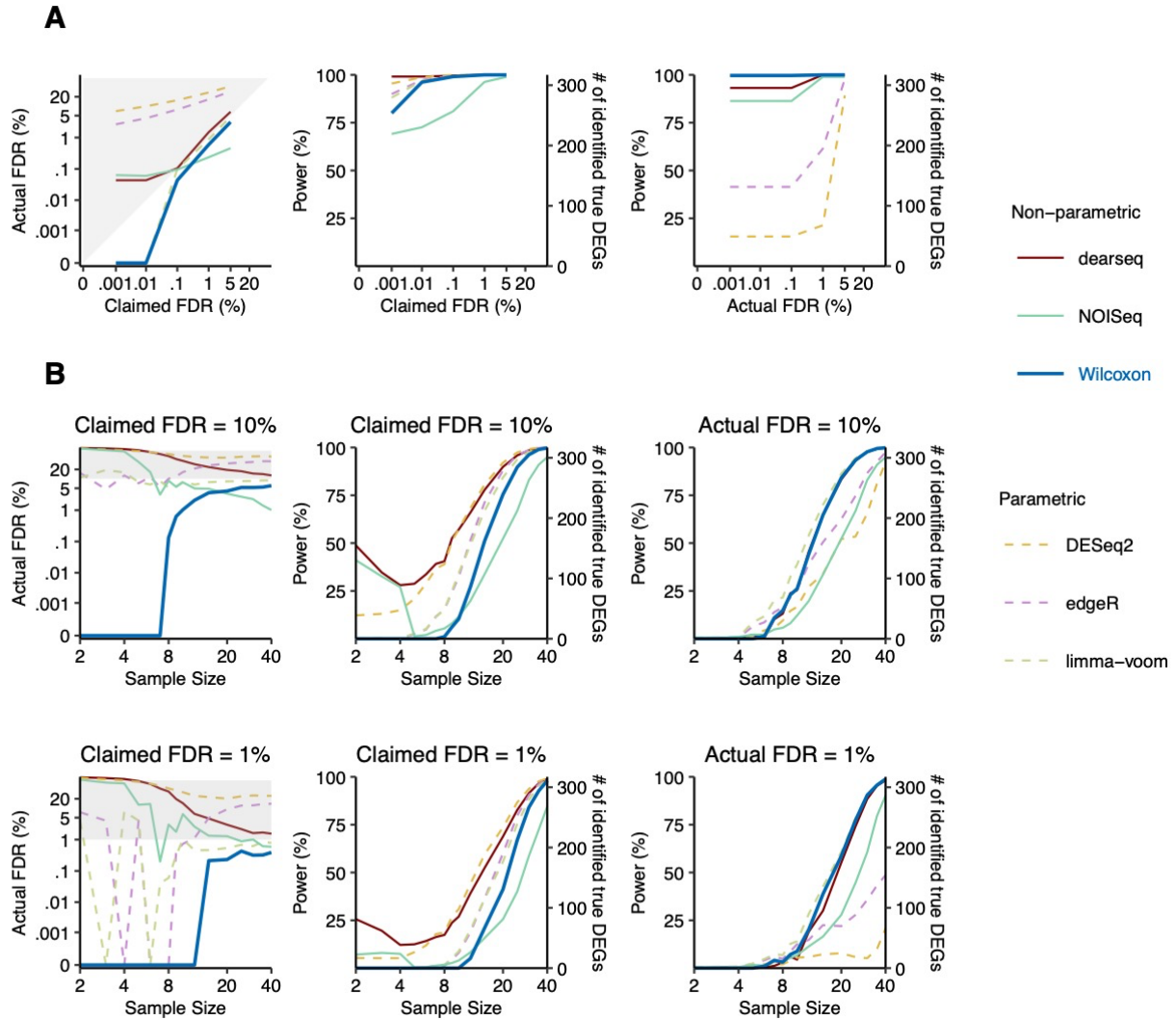


Fig. S29. The Wilcoxon rank-sum test has the best FDR control and power on PRAD

TCGA datasets with semi-synthetic ground truths.

A. The FDR control (left panel), power (middle panel) given the claimed FDRs, and power given the actual FDRs (right panel) under a range of FDR thresholds from 0.001% to 5%.

B. The FDR control (left), power given the claimed FDRs (middle), and power given the actual FDRs (right) for a range of per-condition sample sizes from 2 to 100, under FDR thresholds 10% (top panels) and 1% (bottom panels). The claimed FDRs, actual FDRs, and power were all calculated as the averages of 50 randomly down-sampled datasets.

Fig. S30

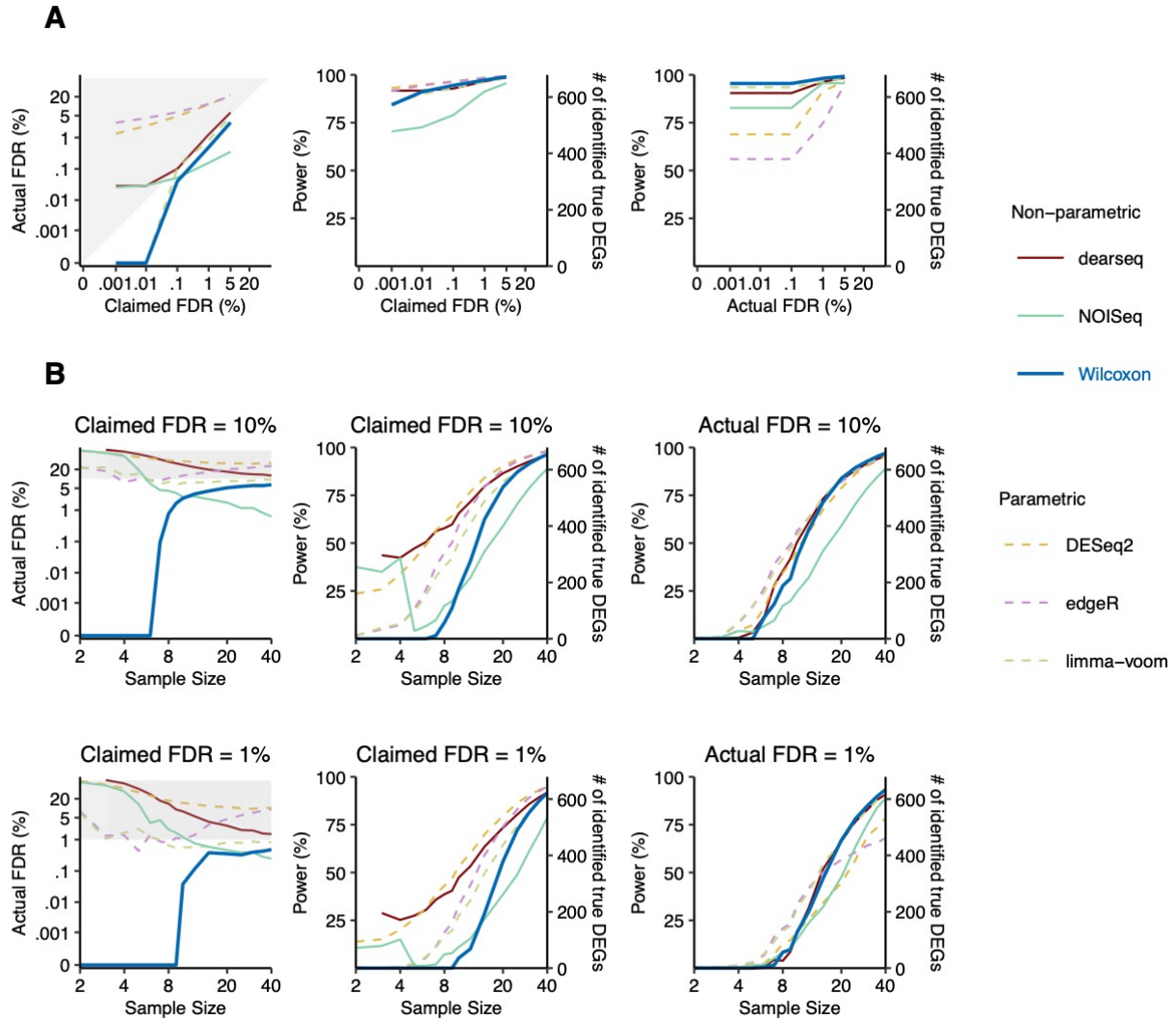


Fig. S30. The Wilcoxon rank-sum test has the best FDR control and power on THCA

TCGA datasets with semi-synthetic ground truths.

A. The FDR control (left panel), power (middle panel) given the claimed FDRs, and power given the actual FDRs (right panel) under a range of FDR thresholds from 0.001% to 5%.

B. The FDR control (left), power given the claimed FDRs (middle), and power given the actual FDRs (right) for a range of per-condition sample sizes from 2 to 100, under FDR thresholds 10% (top panels) and 1% (bottom panels). The claimed FDRs, actual FDRs, and power were all calculated as the averages of 50 randomly down-sampled datasets.

Table S1. Summary of information for samples used in this study

Data source	Condition 1	Condition 2	Sample size (Condition 1 vs. Condition 2)	Accession
Immunotherapy study	Pre-therapy	On-therapy	51 vs. 58	GSE91061
TCGA	BRCA normal tissue ¹	BRCA tumor tissue	112 vs. 112	GDC Xena Hub
	KIRC normal tissue ²	KIRC tumor tissue	72 vs. 72	
	THCA normal tissue ³	THCA tumor tissue	58 vs. 58	
	LUAD normal tissue ⁴	LUAD tumor tissue	57 vs. 57	
	PRAD normal tissue ⁵	PRAD tumor tissue	52 vs. 52	
	LIHC normal tissue ⁶	LIHC tumor tissue	50 vs. 50	
GTEX	Whole blood	Muscle - skeletal	670 vs. 706	GTEX Portal
	Adipose - subcutaneous	Adipose - visceral	581 vs. 469	
	Heart - atrial appendage	Heart - Left ventricle	372 vs. 386	
	Prostate	Brain - cortex	221 vs. 205	
	Cells - EVB transformed lymphocytes	Minor salivary gland	147 vs. 144	
	Brain - amygdala	Brain - spinal cord	129 vs. 126	

¹BRCA: Breast invasive carcinoma

²KIRC: Kidney renal clear cell carcinoma

³THCA: Thyroid carcinoma

⁴LUAD: Lung adenocarcinoma

⁵PRAD: Prostate adenocarcinoma

⁶LIHC: Liver hepatocellular carcinoma