A large-sample crisis? Exaggerated false positives by popular differential expression methods

Yumei Li[1,†], Xinzhou Ge[2,†], Fanglue Peng[3], Wei Li[1,*], Jingyi Jessica Li[2,4,5,6,7,*]

[1]Division of Computational Biomedicine, Department of Biological Chemistry, School of Medicine, University of California, Irvine, Irvine, CA 92697, USA

[2]Department of Statistics, University of California, Los Angeles, CA 90095

[3]Department of Molecular and Cellular Biology, Baylor College of Medicine, Houston, TX 77030, USA

[4]Interdepartmental Program in Bioinformatics, University of California, Los Angeles, CA 90095

[5]Department of Human Genetics, University of California, Los Angeles, CA 90095

[6]Department of Computational Medicine, University of California, Los Angeles, CA 90095

[7]Department of Biostatistics, University of California, Los Angeles, CA 90095

[†]These authors contributed equally to this work

*Corresponding authors. Email: wei.li@uci.edu and lijy03@g.ucla.edu

Abstract

We report a surprising phenomenon about identifying differentially expressed genes (DEGs) from population-level RNA-seq data: two popular bioinformatics methods, DESeq2 and edgeR, have unexpectedly high false discovery rates (FDRs). Via permutation analysis on an immunotherapy RNA-seq dataset, we observed that DESeq2 and edgeR identified even more DEGs after samples' condition labels were randomly permuted. Motivated by this, we evaluated five DEG identification methods (DESeq2, edgeR, limma-voom, NOISeq, and the Wilcoxon rank-sum test) on population-level RNA-seq datasets. We found that the three popular parametric methods---DESeq2, edgeR, and limma-voom---failed to control the FDR. In particular, the actual FDRs of DESeq2 and edgeR sometimes exceeded 20% when the target FDR threshold was only 5%. Although NOISeq, a non-parametric method used by GTEx, controlled the FDR better than the three parametric methods did, it had low power. Based on these results, for population-level RNA-seq studies, we recommend the classic non-parametric test: the Wilcoxon rank-sum test.

One-Sentence Summary

Popular differential expression methods have high false-positive risks on population-level RNA-seq data with large sample sizes.

Main text

RNA-seq is an approach to transcriptome profiling using deep-sequencing technologies (*1-3*). Since RNA-seq was developed more than one decade ago, it has become an indispensable tool for genome-wide transcriptomic studies. One primary research task in these studies is the

identification of DEGs between two conditions (e.g., tumor and normal samples) (*2*). A long-standing, core challenge in this task is the small sample size, which is typically two or three replicates per condition. Many statistical methods have been developed to address this issue by making parametric, restrictive distributional assumptions on RNA-seq data, and the two most popular methods of this type are DESeq2 (*4*) and edgeR (*5*). However, as sample sizes have become large (dozens to thousands) in population-level RNA-seq studies whose samples were collected from individuals (*6, 7*), a natural question to ask is whether DESeq2 and edgeR remain appropriate.

To evaluate the performance of DESeq2 and edgeR, we applied them to 13 population-level RNA-seq datasets with total sample sizes ranging from 100 to 1376 (**Supplementary Table 1**). We found that DESeq2 and edgeR had large discrepancies in the DEGs they identified on these datasets (**Supplementary Fig. 1**). In particular, 23.71%–75% of the DEGs identified by DESeq2 were missed by edgeR. The most surprising result is from an immunotherapy dataset (including 51 pre-nivolumab and 58 on-nivolumab anti-PD-1 therapy patients) (*8*): DESeq2 and edgeR had only an 8% overlap in the DEGs they identified (DESeq2 and edgeR identified 144 and 319 DEGs, respectively, with a union of 427 DEGs but only 36 DEGs in common). This phenomenon raises a critical question: did DESeq2 and edgeR reliably control their false discovery rates (FDRs) to the target 5% on this dataset?

To answer this question, we generated 668 negative-control datasets by randomly permuting the two-condition labels (pre-nivolumab and on-nivolumab) of the 109 RNA-seq samples in this immunotherapy dataset (**Methods**). Since any DEGs identified from these permuted datasets are

known as false discoveries, we used these permuted datasets to evaluate the FDRs of DESeq2 and edgeR. Surprisingly, DESeq2 and edgeR identified more DEGs from 84.88% and 78.89% of these permuted datasets than from the original dataset (**Fig. 1A**). In particular, DESeq2 and edgeR mistakenly identified 16,791 and 13,448 genes as DEGs, respectively, from at least one permuted dataset (**Fig. 1B**). Even more, as many as 56.25% and 89.34% of the DEGs, which DEseq2 and edgeR identified from the original dataset, were also identified as DEGs from at least one permuted dataset, suggesting that these DEGs were spurious (**Fig. 1C**). These results raise the caution about exaggerated false positives by DESeq2 and edgeR on the original dataset.

What's more counter-intuitive, the genes with larger fold changes (between the two conditions) in the original dataset were more likely identified as DEGs from the permuted datasets (**Fig. 1D and Supplementary Fig. 2**). As biologists tend to believe that these genes are more likely true DEGs (which is not necessarily true because a dataset may contain no true DEGs at all), the fact that these genes are false positives would waste experimental validation efforts.

Out of curiosity and as a means of verification, we investigated the biological functions of the spurious DEGs identified by DESeq2 or edgeR from the permuted datasets. Unexpectedly, these spurious DEGs were enriched in immune-related gene ontology (GO) terms (**Fig. 1E**). Hence, if these spurious DEGs were not removed by FDR control, they would mislead researchers to believe that there was an immune response difference between pre-nivolumab and on-nivolumab patients, a surely undesirable consequence that DEG analysis must avoid as much as possible.

Then a question follows: why did DESeq2 and edgeR make so many false discoveries from this immunotherapy dataset? Our immediate hypothetical reason was the violation of the negative binomial model assumed by both DESeq2 and edgeR (*9*). To check this hypothesis, we separated all genes into two groups: (1) the genes identified as DEGs from any permuted datasets and (2) all the other genes; then we evaluated how well the negative binomial model fit to the genes in each group. In line with our hypothesis, the model fitting was worse for the genes in the first group, consistent with the fact these genes were spurious DEGs (**Fig. 1F and Supplementary Fig. 3**).

Motivated by these findings, we benchmarked another two popular DEG identification methods—limma-voom (*10, 11*) and NOISeq (*12*)—and the classic Wilcoxon rank-sum test (*13*) on this immunotherapy dataset and the other 12 population-level RNA-seq datasets from the Genotype-Tissue Expression (GTEx) project (*7*) and the Cancer Genome Atlas (TCGA) (*6*) (**Supplementary Table 1**). Note that DESeq2, edgeR, and limma-voom are parametric methods that assume parametric models for data distribution, while NOISeq and the Wilcoxon rank-sum test are non-parametric methods that are less restrictive but require large sample sizes to have good power. (The GTEx project used DESeq2 and NOISeq for DEG identification.) Using permutation analysis on these datasets, we found that DESeq2 and edgeR consistently showed exaggerated false positives (reflected by underestimated FDRs) compared to the other three methods (**Supplementary Fig. 4-15**).

While permutation analysis creates true negatives (non-DEGs) to allow FDR evaluation, it cannot not be used to evaluate the power, which requires true positives (DEGs) to be known.

Hence, we generated 50 (identically and independently distributed) semi-synthetic datasets with known true DEGs and non-DEGs from each original dataset (from the GTEx and TCGA datasets) to compare the FDRs and power of the five DEG identification methods (**Methods**). In the comparison between 386 heart left ventricle samples with the 372 atrial appendage samples from GTEx datasets, only the Wilcoxon rank-sum test consistently controlled the FDR under a range of thresholds (**Fig. 2A**). In contrast, other methods, especially DESeq2 and edgeR, failed to control the FDR consistently. Moreover, we compared the power of the five methods conditional on their actual FDRs (**Methods**). (Due to the tradeoff between FDR and power, power comparison is only valid when FDR is equal.) As shown in **Fig. 2A**, the Wilcoxon rank-sum test outperformed the other four methods in terms of power.

Finally, to investigate how sample sizes influence the performance of the five methods, we down-sampled each semi-synthetic dataset to obtain per-condition sample sizes ranging from 2 to 100. Again, only the Wilcoxon rank-sum test consistently controlled the FDR at all sample sizes, while the other four methods achieved the best FDR control when the per-condition sample size was around 10 (**Fig. 2B**). Although the Wilcoxon rank-sum test had small power when the per-condition sample size was smaller than 8, an expected phenomenon for its nonparametric nature, it achieved comparable or even better power compared with the three parametric methods (DESeq2, edgeR, and limma-voom) when the per-condition sample size exceeded 8 (**Fig. 2B**). The other non-parametric method, NOIseq, however, had underestimated FDRs (thus exaggerated false positives) when the per-condition sample size was under 5 and was consistently outpowered by the Wilcoxon rank-sum test. These observations were consistent across all 600 semi-synthetic datasets **(Supplementary Fig. 16-26**). In summary, when the per-

condition sample size is less than 10, parametric methods may be used because their power advantage may outweigh their possibly underestimated FDRs; however, for large-sample-size data, the Wilcoxon rank-sum test is the best choice for its solid FDR control and good power.

The three parametric methods for DEGs identification—DESeq2, edgeR, and limma—have long been the dominant methods for transcriptomic studies. For example, the GTEx project, a consortium effort studying gene expression and regulation in normal human tissues, used DESeq2 coupled with NOISeq to find DEGs between tissues (*14*); several studies applied edgeR or limma to TCGA RNA-seq data to find DEGs between tumor and normal samples (*15-17*); with increasing attention on the immunotherapy, researchers used DESeq2 to detect DEGs between responders and non-responders of the immunotherapy (*8, 18*). However, while the three parametric methods were initially designed to address the small-sample-size issue, these population-level studies had much larger sample sizes (at least dozens) and thus no longer needed restrictive parametric assumptions. Moreover, violation of the parametric distributions assumed by the three popular methods may cause ill-behaved P-values and thus lead to failed FDR control (*19*), an issue independent of the sample size.

In this study, we showed the superiority of the Wilcoxon rank-sum test, a classic non-parametric test also known as the Mann-Whitney test developed in the 1940s (*13, 20*), for analyzing large-sample-size RNA-seq datasets. Our results also echo the importance of verifying FDR control by permutation analysis.

Beyond RNA-seq data analysis, our study suggests that, for population-level studies with large sample sizes, classic non-parametric statistical methods should be used as the baseline methods for data analysis and new method benchmarking.

Acknowledgements

We thank Jason Sheng Li of Wei Li lab for suggestions on the title. We also thank other members of Wei Li lab and Jingyi Jessica Li lab for helpful discussions.

Funding

This work was supported by the following grants: The U.S. National Institutes of Health R01CA193466 and R01CA228140 (to W.L.); NIH/NIGMS R01GM120507 and R35GM140888, NSF DBI-1846216 and DMS-2113754, Johnson & Johnson WiSTEM2D Award, Sloan Research Fellowship, and UCLA David Geffen School of Medicine W.M. Keck Foundation Junior Faculty Award (to J.J.L.).

Author contributions

Y.L., W.L. and J.J.L. conceived and supervised this project. Y.L. and X.Z. performed the data analysis. Y.L., X.Z., F.P., J.J.L and W.L. interpreted the data and wrote the manuscript.

Competing interests

The authors declare no competing financial interests.

Data availability

All the permuted and semi-synthetic datasets used to generated results can be found at Zenodo via https://doi.org/10.5281/zenodo.5241320.

Code availability

All the codes used to generate results can be found at GitHub via URL

https://github.com/xihuimeijing/DEGs_Analysis_FDR.

Supplementary Materials

Materials and Methods

Supplementary Figs. 1 to 26

Supplementary Table 1

References

1. U. Nagalakshmi *et al.*, The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349 (2008).
2. R. Stark, M. Grzelak, J. Hadfield, RNA sequencing: the teenage years. *Nat Rev Genet* **20**, 631-656 (2019).
3. Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).
4. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
5. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).
6. N. Cancer Genome Atlas Research *et al.*, The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113-1120 (2013).
7. G. T. Consortium, The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).
8. N. Riaz *et al.*, Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* **171**, 934-949 e916 (2017).
9. S. Hawinkel, J. C. W. Rayner, L. Bijnens, O. Thas, Sequence count data are poorly fit by the negative binomial distribution. *PLoS One* **15**, e0224909 (2020).
10. M. E. Ritchie *et al.*, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
11. C. W. Law, Y. Chen, W. Shi, G. K. Smyth, voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29 (2014).
12. S. Tarazona *et al.*, Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res* **43**, e140 (2015).
13. F. Wilcoxon, Individual comparisons of grouped data by ranking methods. *J Econ Entomol* **39**, 269 (1946).
14. M. Mele *et al.*, The human transcriptome across tissues and individuals. *Science* **348**, 660-665 (2015).
15. L. Peng *et al.*, Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types. *Sci Rep* **5**, 13413 (2015).
16. S. R. Rosario *et al.*, Pan-cancer analysis of transcriptional metabolic dysregulation using The Cancer Genome Atlas. *Nat Commun* **9**, (2018).
17. Z. Tang *et al.*, GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res* **45**, W98-W102 (2017).
18. T. N. Gide *et al.*, Distinct Immune Cell Populations Define Response to Anti-PD-1 Monotherapy and Anti-PD-1/Anti-CTLA-4 Combined Therapy. *Cancer Cell* **35**, 238-255 e236 (2019).

19. Y. Benjamini, Y. Hochberg, Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J R Stat Soc B* **57**, 289-300 (1995).
20. H. B. Mann, D. R. Whitney, On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics* **18**, 50-60 (1947).

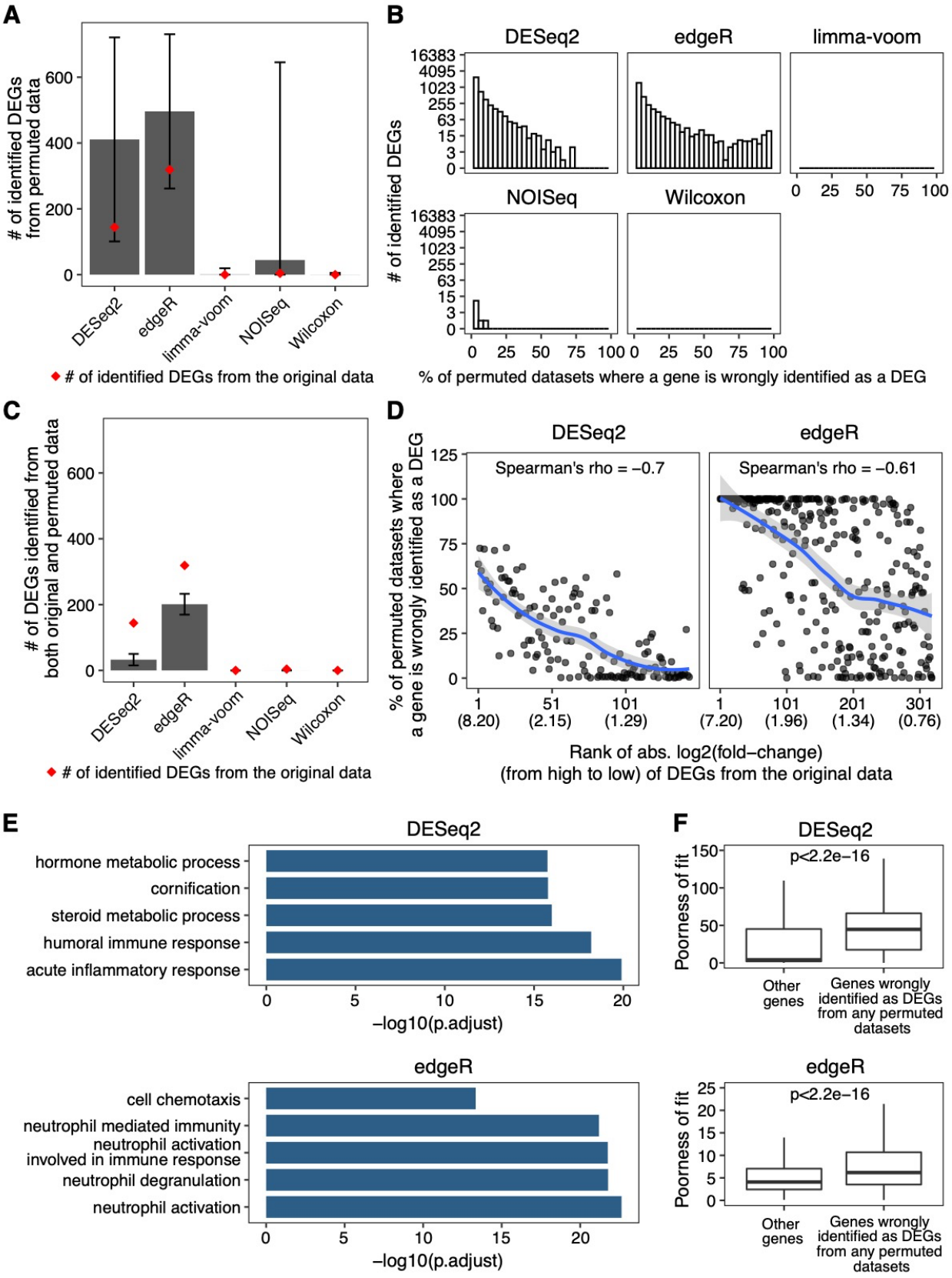
A
# of identified DEGs from permuted data
600
400
200
0
DESeq2
edgeR
limma-voom
NOISeq
Wilcoxon
# of identified DEGs from the original data
B
DESeq2
edgeR
limma-voom
NOISeq
Wilcoxon
# of identified DEGs
16383
4095
1023
255
63
15
3
0
0 25 50 75 100
% of permuted datasets where a gene is wrongly identified as a DEG
C
# of DEGs identified from both original and permuted data
600
400
200
0
DESeq2
edgeR
limma-voom
NOISeq
Wilcoxon
# of identified DEGs from the original data
D
DESeq2
edgeR
Spearman's rho = −0.7
Spearman's rho = −0.61
% of permuted datasets where a gene is wrongly identified as a DEG
125
100
75
50
25
0
1 (8.20)
51 (2.15)
101 (1.29)
1 (7.20)
101 (1.96)
201 (1.34)
301 (0.76)
Rank of abs. log2(fold−change)
(from high to low) of DEGs from the original data
E
DESeq2
hormone metabolic process
cornification
steroid metabolic process
humoral immune response
acute inflammatory response
0 5 10 15 20
−log10(p.adjust)
edgeR
cell chemotaxis
neutrophil mediated immunity
neutrophil activation involved in immune response
neutrophil degranulation
neutrophil activation
0 5 10 15 20
−log10(p.adjust)
F
DESeq2
p<2.2e−16
Poorness of fit
150
100
50
0
Other genes
Genes wrongly identified as DEGs from any permuted datasets
edgeR
p<2.2e−16
Poorness of fit
25
20
15
10
5
0
Other genes
Genes wrongly identified as DEGs from any permuted datasets

Fig. 1. Exaggerated false DEGs identified by DESeq2 and edgeR from anti-PD-1 therapy RNA-seq datasets.

A. Barplot showing the average numbers of DEGs identified from 668 permuted datasets. The error bars represent the standard deviations of 668 permutations. The red dots indicate the numbers of DEGs identified from the original dataset.

B. The distributions of the percentage of permuted datasets where a gene was mistakenly identified as a DEG.

C. Barplot showing the average numbers of DEGs identified from both the original dataset and the permuted datasets. The error bars represent the standard deviations of 668 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 log2(fold-change) in the original dataset in decreasing order. The absolute log2(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. Boxplots showing the poorness of fitting the negative binomial model to the genes identified by DESeq2 or edgeR as DEGs from any permuted datasets vs. all the other genes. The poorness of fit for each gene is defined as its negative $\log_{10}$(P-value) from the goodness-of-fit test for the negative binomial distributions estimated 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.

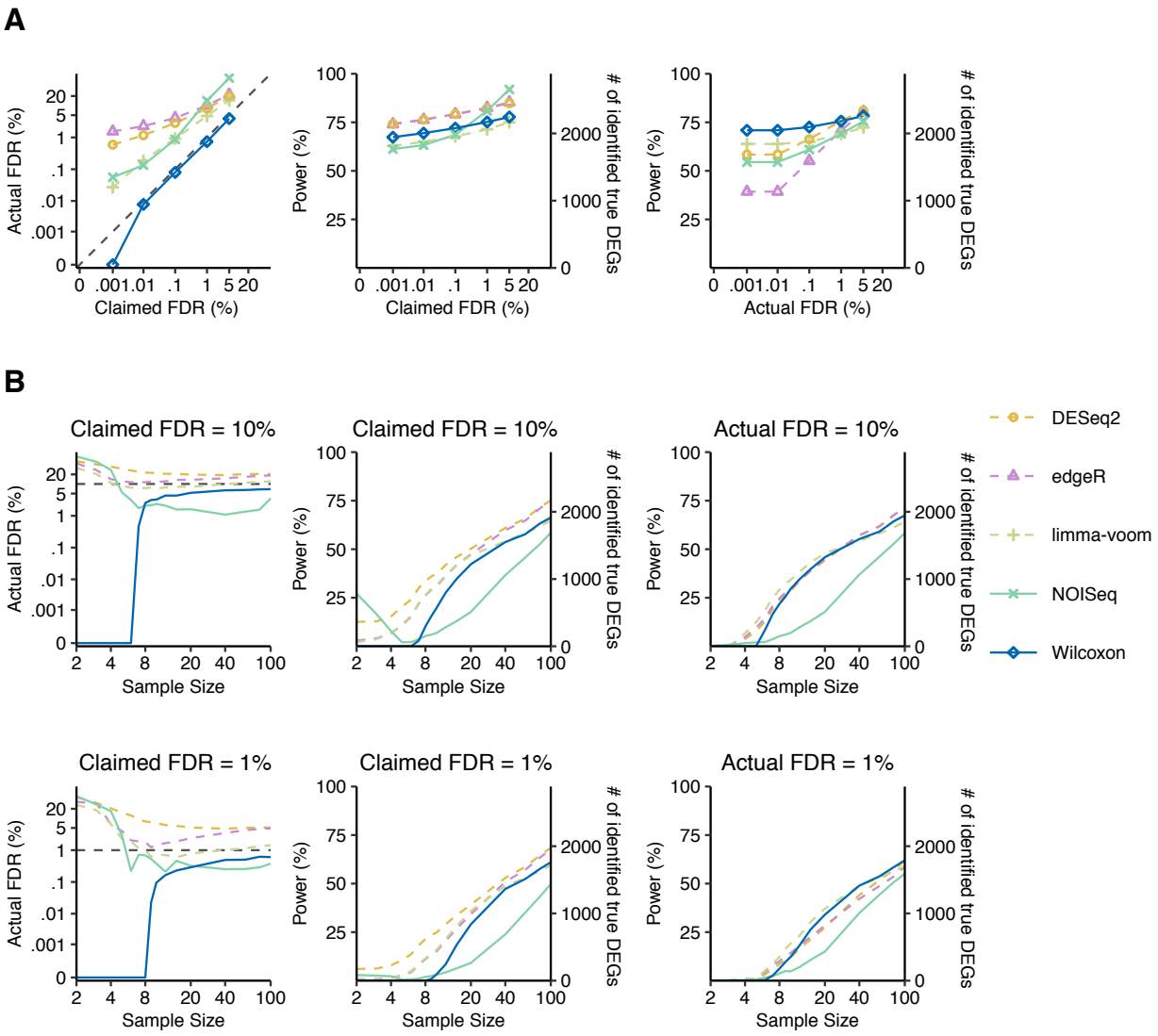

Fig. 2. Wilcoxon test has the best FDR control and power on heart left ventricle vs. atrial appendage GTEx datasets with synthetic ground truths.

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.