OPEN ACCESS

EDITORIAL

Background bias in functional enrichment analysis: Insights from *clusterProfiler*

Guangchuang Yu[1,*]

[1]Department of Bioinformatics, School of Basic Medical Sciences, Southern Medical University, Guangzhou 510515, China

*Correspondence: gcyu1@smu.edu.cn

Received: August 23, 2025; Accepted: October 13, 2025; Published Online: October 15, 2025; https://doi.org/10.59717/j.xinn-life.2026.100181

Citation: Yu G. (2026). Background bias in functional enrichment analysis: Insights from *clusterProfiler*. *The Innovation Life* **4**:100181.

Functional enrichment analysis has become one of the cornerstones of modern functional genomics. By identifying perturbed pathways and biological processes from large-scale omics datasets, enrichment methods enable researchers to move beyond lists of genes and toward mechanistic interpretation.[1] Among these methods, over-representation analysis (ORA) is by far the most widely used. Its appeal lies in its simplicity: a set of genes of interest, such as differentially expressed genes, is tested for statistical enrichment against curated biological pathways or functional categories.

Yet, despite its ubiquity, enrichment analysis is far from straightforward. Many in the community have already recognized one persistent challenge: the use of outdated knowledge bases. When annotations lag behind current biology, statistical results are distorted, newly discovered pathways are overlooked, and the resulting interpretations are misleading.[2] But while the problem of outdated databases has gained attention, another equally consequential source of error has remained largely neglected: background bias.

Most enrichment tools default to using the entire genome as the "background universe." This choice rests on an implicit assumption that every gene has an equal probability of being observed in the experiment. Such an assumption may hold in certain contexts—for instance, studies of germline variation, where any gene can, in principle, harbor a mutation. However, in functional genomics, it is rarely valid. Gene expression is highly tissue-specific. In any given sample, large fractions of the genome are silent and thus undetectable. Even more importantly, the scope of observable genes is constrained by technology. Single-cell RNA-seq, spatial transcriptomics, and proteomics all have inherent detection limits. Many genes are either not expressed or fall below detection thresholds. These genes can never appear among differentially expressed genes, and their inclusion in the background may artificially inflates the significance of enrichment results.

In practical terms, this means that many enrichment analyses risk telling us little more than the obvious. For example, when analyzing a brain transcriptome, one might "discover" that neural pathways are enriched. But such results are trivial if they merely reflect that neural genes are expressed in

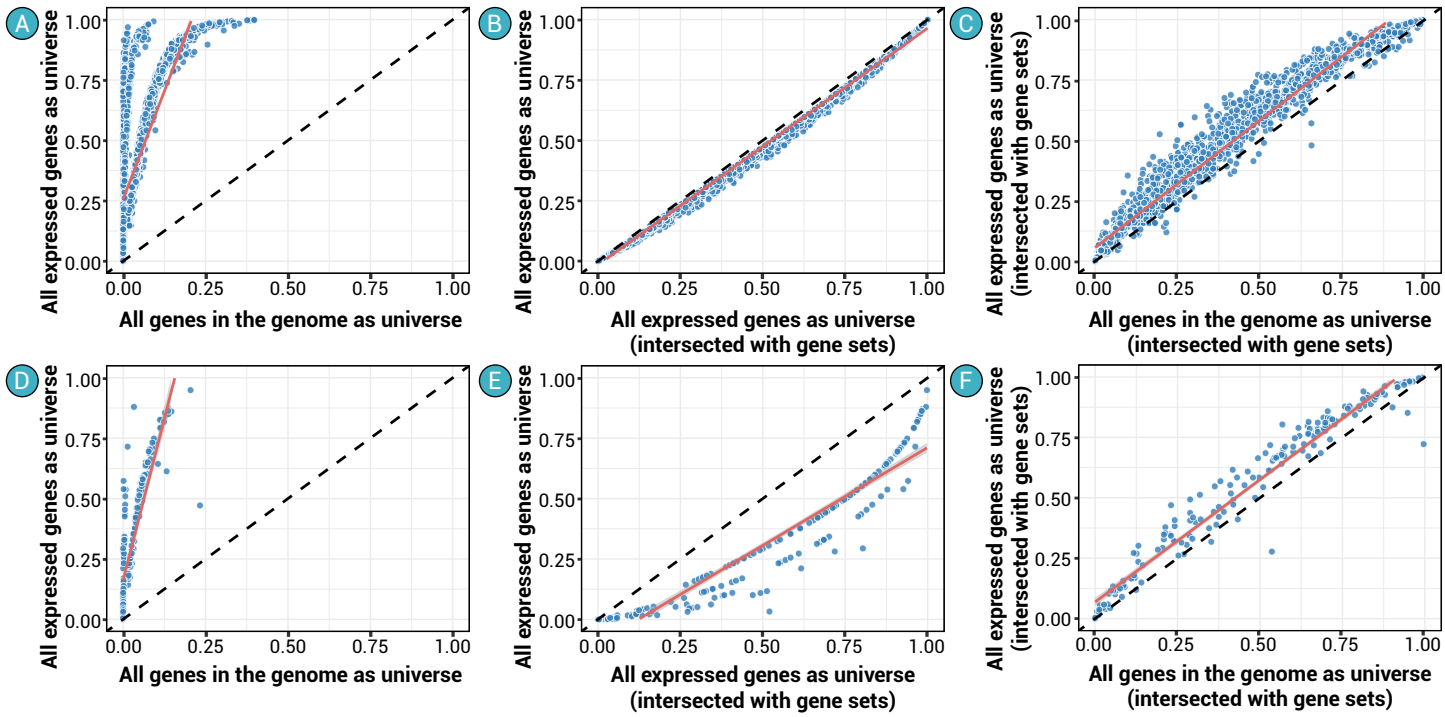

Figure 1. Comparison of enrichment *p*-values under different background universes and strategies (A, D) Using all expressed genes as the universe versus using all genomic genes as the universe. (B, E) Using all expressed genes as the universe versus using the intersection of expressed genes and the gene set as the universe. (C, F) Using the intersection of expressed genes and the gene set as the universe versus using the intersection of all genomic genes and the gene set as the universe. Panels A–C show results from Gene Ontology (biological process) enrichment analysis, whereas panels D–F show results from KEGG pathway analysis. Reproducible analysis scripts for Figure 1 are available at: https://github.com/YuLab-SMU/Background-bias-of-functional-enrichment-analysis.

brain tissue, rather than revealing any novel biology. A more appropriate background, therefore, is not the entire genome, but the subset of genes actually detectable in the experiment—those measured quantitatively or demonstrably within the scope of the assay.

Background bias does not end with detectability. Annotation coverage introduces a second layer of complexity. Consider the human genome: of roughly 20,000 protein-coding genes, only about 9,000 are annotated with KEGG pathways. If an experiment quantifies 15,000 genes, then approximately 6,000 fall outside pathway databases altogether. When such unannotated genes are included in the background, they are automatically classified as "not in the gene set." This misclassification introduces systematic error, leading to inflated or deflated p-values depending on the analysis. By contrast, restricting the background to the intersection of measurable genes and annotated genes produces more conservative—and ultimately more trustworthy—results. To emphasize this point, this article defines the concept of "annotation intersection," which refers to the overlap between the back-

ground genes used in enrichment analysis and the genes with functional annotations in databases such as GO and KEGG.

Our analyses illustrate this point vividly. When the entire genome is used as background, p-values are substantially smaller than when only detected genes are considered (Figures 1 A & D). Excluding unannotated genes shifts p-values upward again, further reducing false positives (Figures 1 B & E). Perhaps most importantly, intersecting the background with annotations serves as a corrective measure even when the genome-wide background is mistakenly applied. By discarding non-annotated genes, such as many non-coding genes, this strategy minimizes bias and brings results from different background definitions into closer agreement (Figures 1 C & F).

Customizing the background gene set by restricting it to genes detectable in the experiment was already noted as an advantageous practice more than twenty years ago in microarray analysis. However, even today, this approach is still not well supported and is often overlooked. Importantly, the current use of "the whole genome" as background is not the same as it was in the microarray era. With advances in technology, more and more non-protein-coding genes have been discovered, making annotation intersection increasingly important. When users neglect to specify a background gene set, using the whole genome typically results in the inclusion of a large number of non-coding genes, which are unlikely to be detected in transcriptomic assays and undetectable in proteomic assays. Notably, these non-coding genes usually lack functional annotations, and annotation intersection effectively removes them. This is why a simple adjustment in software can go a long way toward helping users minimize bias.

Despite these clear implications, most enrichment tools have failed to address background bias adequately. Many do not allow users to define custom backgrounds at all, especially web-based platforms designed for ease of use. In such tools, users simply upload a differential gene list, select a species, and receive enrichment results without any control over background specification. The outcome is predictable: long lists of pathways with strikingly small p-values, many of which are irrelevant or trivial. Even among more flexible tools, virtually none implement annotation intersection by default. Thus, even careful users who correctly specify a background may still obtain biased results.

The clusterProfiler package stands out as a notable exception.[3] It supports user-defined backgrounds and, crucially, defaults to intersecting the background universe with annotated gene sets. This produces more conservative p-values and ensures robustness even if users provide a less-than-ideal background. As demonstrated in Figure 1, this design allows clusterProfiler to deliver more reliable enrichment outcomes regardless of background specification. Nevertheless, we encourage researchers to explicitly define biologically meaningful backgrounds whenever possible, especially in studies with limited detection universes such as single-cell or proteomic datasets. Of course, no single strategy fits all scenarios. The intersection approach assumes that gene sets are annotated at the genome-wide level. In specialized contexts—such as testing whether transcription factor targets are enriched within sex-specific gene lists—intersection with annotations is inappropriate. To accommodate such cases, clusterProfiler provides an option that disables intersection, offering flexibility while maintaining methodological rigor.

It is worth reiterating a fundamental point: the purpose of enrichment analysis is not to generate as many significant p-values as possible. Rather, it is to illuminate underlying biological processes and mechanisms. Inflated significance may be gratifying at first glance, but it risks obscuring true biology behind a haze of spurious associations. Unfortunately, background bias has long been overlooked. A survey of published enrichment studies found that only 4.1% explicitly described how the background was chosen.[4] Neither authors nor reviewers have consistently recognized its importance, and editorial policies have done little to encourage best practices. As a result, countless published analyses may have reported misleading conclusions—not due to negligence, but simply because the community has not treated background specification as a central methodological issue.

It is time for a cultural shift in how enrichment analysis is conducted and reported. Background bias is not a minor technical detail; it is a fundamental determinant of analytical validity. Addressing it will not only improve the rigor of individual studies but also enhance reproducibility and comparability across the field. Several steps are essential. Researchers must be educated about the consequences of background choice and encouraged to define biologically appropriate backgrounds for their data. Tool developers should implement defaults that minimize bias—such as annotation intersection—and provide clear documentation to guide users. Journals should require authors to report how backgrounds were defined and for reviewers to evaluate these choices critically.

Functional enrichment analysis has transformed our ability to interpret large-scale genomic data, but its power is undermined by background bias. By continuing to ignore this issue, we risk mistaking technical artifacts for biological insights. The good news is that the solution is within reach: through thoughtful background specification, improved software defaults, and greater editorial oversight, we can ensure that enrichment analysis fulfills its promise as a tool for discovery rather than distortion. It is time for the field to move beyond complacency. The genes we study deserve careful contextualization, and the pathways we report should reflect biology, not bias. Only then will enrichment analysis truly illuminate the mechanisms that underlie complex biological systems.

REFERENCES

1. Xu S., Hu E., Cai Y., et al. (2024). Using clusterProfiler to characterize multiomics data. *Nat. Protoc.* **19**:3292–320. DOI:10.1038/s41596-024-01020-z
2. Wadi L., Meyer M., Weiser J., et al. (2016). Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods* **13**:705–6. DOI:10.1038/nmeth.3963
3. Yu G. (2024). Thirteen years of clusterProfiler. *The Innovation* **5**:100722. DOI:10.1016/j.xinn.2024.100722
4. Wijesooriya K., Jadaan S. A., Perera K. L., et al. (2022). Urgent need for consistent standards in functional enrichment analysis. *PLOS Comput. Biol.* **18**:e1009935. DOI:10.1371/journal.pcbi.1009935

FUNDING AND ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China (no. 32270677). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

AUTHOR CONTRIBUTIONS

G. Y conceived the study, conducted the analyses, drafted the manuscript, and approved the final version.

DECLARATION OF INTERESTS

Guangchuang Yu is an editorial board member of The Innovation Life and was blinded from reviewing or making final decisions on the manuscript. Peer review was handled independently of this member and this research group. The author declares no conflicts of interest.

DATA AND CODE AVAILABILITY

The source code to reproduce Figure 1 is available at https://github.com/YuLab-SMU/Background-bias-of-functional-enrichment-analysis.

LEAD CONTACT WEBSITE

Guangchuang Yu:
https://yulab-smu.top