LETTER

Thirteen years of clusterProfiler

Guangchuang Yu[1,*]

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

*Correspondence: gcyu1@smu.edu.cn

Received: October 10, 2024; Accepted: October 18, 2024; Published Online: October 21, 2024; https://doi.org/10.1016/j.xinn.2024.100722

Citation: Yu G. (2024). Thirteen years of clusterProfiler. The Innovation **5(6)**, 100722.

Dear Editor,

When the human genome was fully sequenced in 2003, research focus shifted to functional genomics, particularly the spatiotemporal expression of genes, which is crucial for understanding organism development, functional regulation, and disease mechanisms. A key step in this process is uncovering the biological pathways involved. The first bioinformatics tool for analyzing biological pathways using Gene Ontology (GO) was GO::TermFinder, a Perl module published in 2004 that implemented the over-representation analysis method.[1] Shortly thereafter, in 2005, the gene set enrichment analysis (GSEA) method was introduced.[2] Various information content-based methods for measuring semantic similarity were adapted for use with GO, and in 2007, Wang proposed a graph-based approach to measure GO semantic similarity. In 2008, I developed GOSemSim, which implemented multiple GO semantic similarity measures, including information content and graph structure algorithms.[3] These tools, which mine biological knowledge, rely heavily on gene functional information accumulated during the Human Genome Project era.

However, these tools were primarily designed for model organisms. One of the motivations behind developing clusterProfiler was my desire to extend pathway analysis to non-model organisms. Additionally, all tools at that time were created for case-control experimental designs. I wanted to apply pathway analysis to more complex biological experiments with multiple conditions, which is the inspiration for the software's name—it profiles biological themes across different gene clusters (Figure 1). This comparative approach to biological themes is an innovation of clusterProfiler. In the v.4.0 paper, we applied it to compare the pathways perturbed by different drugs over time.[4] In a protocol published in 2024, we demonstrated its use in comparing microbiome and metabolome data across

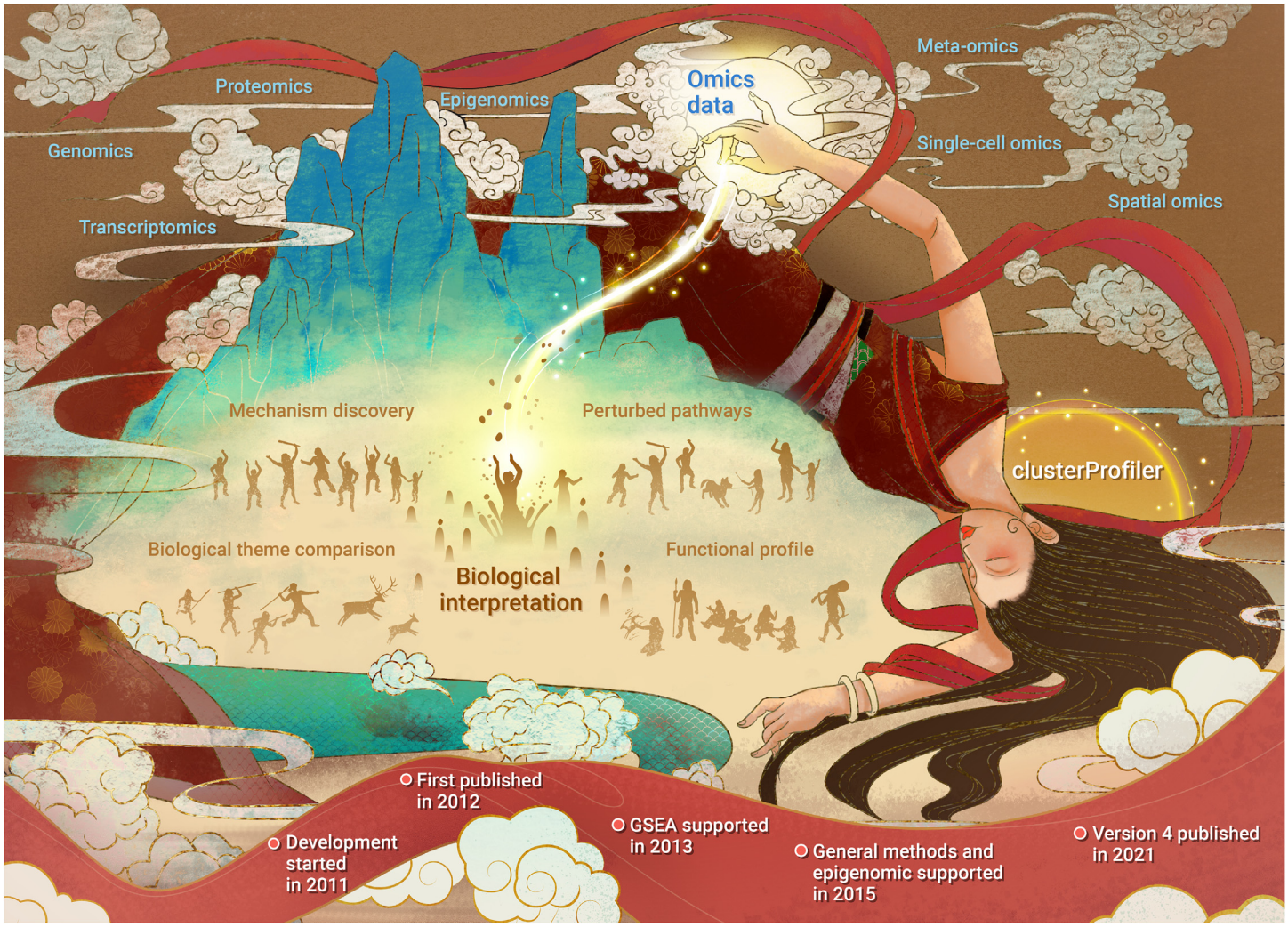

Figure 1. clusterProfiler: Elucidating genomic insights with a Nuwa allegory Through the lens of the Chinese myth of Nuwa creating humans, depicted in the style of Dunhuang murals, we can illustrate the function and value of clusterProfiler. In this analogy, clusterProfiler is the architect, like Nuwa, carefully analyzing coding and non-coding multi-omics data across species. Its work is grounded in updated gene annotation references, represented by the mountain in the mural. Just as Nuwa shaped humanity, clusterProfiler provides a flexible platform for systematically exploring biological mechanisms and states, deepening our understanding of complex phenomena. The results it generates can be compared to the miniature humans Nuwa created. Like her hands, the clean interface of clusterProfiler allows researchers to easily access, manage, and visualize enrichment results. Moreover, as Nuwa created tribes with unique traits, clusterProfiler compares data from multiple treatments and time points in a single run, simplifying the process of identifying functional similarities and differences across conditions.

disease subtypes, characterizing transcription factors and their functions activated under stress at different time points and analyzing cell type enrichment in single-cell clusters.[5]

The development of clusterProfiler began in 2011, with the first version published in 2012.[6] We initially applied it to study biological pathways regulated by cobalt- and nickel-binding proteins in *Streptococcus pneumoniae* and to compare host pathways regulated by human virus-encoded microRNAs.[7,8] We have continuously maintained and updated clusterProfiler, with over 12,000 commits to the codebase. In 2013, we added the GSEA method, and in 2016, we adopted the fgsea algorithm to accelerate the computation.[9] In early versions, KEGG.db was used as the data source of KEGG pathway analysis, but with changes to KEGG's licensing, KEGG.db stopped updating. In 2015, clusterProfiler began supporting KEGG analysis by fetching the latest data online via HTTP, allowing analysis for all species available on the KEGG website. clusterProfiler also supports WikiPathways and PathwayCommons, and we developed tools for analyzing disease ontology, Reactome pathways, and Medical Subject Headings.[10]

Developing methods for specific biological knowledge databases cannot always keep up with newly emerging resources or support custom user-defined databases. To address this, in 2015, clusterProfiler began supporting general pathway enrichment analysis methods, enabling users to analyze new or custom databases, expanding the scope beyond just biological pathways. My other R packages have also supported clusterProfiler's capabilities. These include GOSemSim for calculating semantic similarity, which can be used to remove redundant pathways for enrichment results[3]; ChIPseeker for annotating genomic locations, applicable to functional enrichment analysis of epigenomic data[11]; and ggtree for displaying the hierarchical relationships in enrichment results. Additionally, in the enrichplot package, we have continually developed new visualization methods to help users better interpret and present enrichment analysis results.

Each month, clusterProfiler is downloaded over 18,000 times via Bioconductor and has been integrated into more than 40 bioinformatics software tools, making it one of the foundational tools in bioinformatics analysis. Over 13 years of development, we have seen clusterProfiler applied to explore individual development, molecular mechanisms of diseases, and drug mechanisms of action. We have also witnessed its use in analyzing data from new technologies, including single-cell transcriptomics and spatial transcriptomics. In the future, I will continue to maintain, update, and add new features to meet the needs of new applications.

REFERENCES

1. Boyle, E.I., Weng, S., Gollub, J., et al. (2004). GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics ***20***: 3710–3715. https://doi.org/10.1093/bioinformatics/bth456.
2. Subramanian, A., Tamayo, P., Mootha, V.K., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA ***102***: 15545–15550. https://doi.org/10.1073/pnas.0506580102.
3. Yu, G., Li, F., Qin, Y., et al. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics ***26***: 976–978. https://doi.org/10.1093/bioinformatics/btq064.
4. Wu, T., Hu, E., Xu, S., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation ***2***: 100141. https://doi.org/10.1016/j.xinn.2021.100141.
5. Xu, S., Hu, E., Cai, Y., et al. (2024). Using clusterProfiler to characterize multiomics data. Nat. Protoc. ***19***: 3292–3320. https://doi.org/10.1038/s41596-024-01020-z.
6. Yu, G., Wang, L.G., Han, Y., et al. (2012). clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS J. Integr. Biol. ***16***: 284–287. https://doi.org/10.1089/omi.2011.0118.
7. Sun, X., Yu, G., Xu, Q., et al. (2013). Putative cobalt- and nickel-binding proteins and motifs in Streptococcus pneumoniae. Metallomics ***5***: 928–935. https://doi.org/10.1039/C3MT00126A.
8. Yu, G., and He, Q.Y. (2011). Functional similarity analysis of human virus-encoded miRNAs. J. Clin. Bioinf. ***1***: 15. https://doi.org/10.1186/2043-9113-1-15.
9. Korotkevich, G., Sukhov, V., Budin, N., et al. (2021). Fast gene set enrichment analysis. Preprint at bioRxiv. https://doi.org/10.1101/060012.
10. Yu, G. (2018). Using meshes for MeSH term enrichment and semantic analyses. Bioinformatics ***34***: 3766–3767. https://doi.org/10.1093/bioinformatics/bty410.
11. Wang, Q., Li, M., Wu, T., et al. (2022). Exploring Epigenomic Datasets by ChIPseeker. Curr. Protoc. ***2***: e585. https://doi.org/10.1002/cpz1.585.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Natural Science Foundation of China (32270677).

DECLARATION OF INTERESTS

The authors declared no competing interest.