The reusability of public omics data across 5 million research publications

Viorel Munteanu[1], Nicolae Drabcinski[1] , Dumitru Ciorba[1], Serghei Mangul[2], Viorel Bostan[1]

[1] Technical University of Moldova, viorel.munteanu@lt.utm.md, nicolae.drabcinski1@lt.utm.md, dumitru.ciorba@fcim.utm.md, viorel.bostan@adm.utm.md, ORCID: 0000-0002-4133-5945, 0009-0008-4381-836X, 0000-0002-3157-5072, 0000-0002-2422-3538

[2]University of Southern California, serghei.mangul@gmail.com, ORCID: 0000-0003-4770-3443

Keywords: Reproducibility, public omics data, data reuse, secondary analysis

Abstract. Publicly accessible omics data are a vital resource for the scientific community, enabling re-analysis, experiments, and meta-analyses that promote reproducibility and fuel new discoveries [1]. Despite their importance, the patterns and extent of secondary data reuse are not well understood. In this comprehensive study, we analyzed over five million open-access publications from 2001 to 2024, identifying 400,000 papers focused on omics data [2]. Among these, 58% of the publications reused publicly available datasets. Notably, from 2016 to 2024, there was a significant 30% increase in publications utilizing reused gene expression data [3], surpassing the number of studies using newly generated data. For the study, we collected 5,547,235 open-access publications from PubMed Central (PMC), spanning the years 2001 to 2024. We identified 276,642 publications that mentioned omics datasets, such as those from the Sequence Read Archive (SRA) and Gene Expression Omnibus (GEO), using text mining and regular expressions. The publications were classified as either primary or secondary analyses based on the dataset release date. Our validation process, based on a curated dataset, achieved high accuracy: 97.6% for primary analyses and 97.4% for secondary analyses, with misclassifications primarily occurring in incomplete texts or pre-print journals [4]. Our findings show that datasets requiring minimal

computational resources or more advanced analytical methods had higher rates of reuse. We introduced the normalized reusability index (NRI), which revealed that over 16% of omics datasets are reused at least ten times annually, while at least 56% of datasets are reused at least once. This analysis provides critical insights into trends in omics data reuse and highlights methodological inconsistencies in the field.

Acknowledgments. This work was partially supported by the State Program LIFETECH No. 020404 at the Technical University of Moldova.

References

1. Rajesh, A. *et al.* Improving the completeness of public metadata accompanying omics studies. *Genome Biol.* 22, 106 (2021).
2. The Replication Crisis: How Can Open Science Improve the Scale of Reproducibility? *Pubs - Bio-IT World* https://www.bio-itworld.com/news/2024/05/10/the-replication-crisis-how-can-open-science-improve-the-scale-of-reproducibility.
3. Oza, V. H. *et al.* Ten simple rules for using public biological data for your research. *PLOS Comput. Biol.* 19, e1010749 (2023).
4. Bioinformatics-Lab-TUM/Reusability_omics_data: This repository contains the analysis and tools developed for a comprehensive study into the reuse of public omics datasets across over 5 million research publications. https://github.com/Bioinformatics-Lab-TUM/Reusability_omics_data/tree/main.