The systematic assessment of completeness of public metadata accompanying omics studies

Yu-Ning Huang
Department of Clinical Pharmacy, Alfred E. Mann School of Pharmacy and Pharmaceutical Sciences, University of Southern California, Los Angeles, California, 90089, USA
yuninghu@usc.edu
ORCID: 0000-0003-1697-4267

Pooja Vinod Jaiswal
Department of Clinical Pharmacy, Alfred E. Mann School of Pharmacy, University of Southern California, Los Angeles, California, 90089, USA
pjaiswal@usc.edu
ORCID: 0009-0000-7957-3116

Anushka Rajesh
Department of Pharmacology and Pharmaceutical Sciences, Alfred E. Mann School of Pharmacy, University of Southern California, Los Angeles, CA, 90089, USA
anushkar@usc.edu

Anushka Yadav
Department of Pharmacology and Pharmaceutical Sciences, Alfred E. Mann School of Pharmacy, University of Southern California, Los Angeles, CA, USA
y.anushka248@gmail.com
ORCID: 0000-0002-4333-6028

Dottie Yu
Department of Quantitative and Computational Biology, Dana and David Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA, USA
dottieyu@usc.edu
ORCID: 0009-0004-5682-7362

Fangyun Liu
Department of Quantitative and Computational Biology, Dana and David Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA, USA
fangyunl@usc.edu

Grace Scheg
Department of Quantitative and Computational Biology, Dana and David Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA, USA
gscheg@usc.edu

Grigore Boldirev
Department of Computer Science, College of Arts and Sciences, Georgia State University, Atlanta, GA, 30303, USA
grigore.boldirev@gmail.com
ORCID: 0009-0008-8305-1650

Irina Nakashidze
Department of Natural Sciences and Health Care, Batumi Shota Rustaveli State University, 6010, Batumi, Georgia
irina.nakashidze@bsu.edu.ge
ORCID: 0000-0001-8934-6312

Aditya Sarkar
School of Computing and Electrical Engineering, Indian Institute of Technology Mandi, North Campus, Kamand, Mandi, HP 175005, India
asnov2k@gmail.com
ORCID: https://orcid.org/0000-0002-4496-8289

Jay Himanshu Mehta
Department of Pharmacology and Pharmaceutical Sciences, Alfred E. Mann School of Pharmacy, University of Southern California, Los Angeles, CA, USA
jayhiman@usc.edu

Ke Wang
Department of Translational Genomics, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
kwang558@usc.edu

Khooshbu Kantibhai Patel
Department of Clinical Pharmacy, Alfred E. Mann School of Pharmacy, University of Southern California, Los Angeles, California, 90089, USA
khooshbu@usc.edu
ORCID: 0000-0002-7575-4734

Mustafa Ali Baig Mirza
Department of Computer Science & Engineering, University of Bridgeport, Bridgeport, CT 06604, USA
mirzabaig.mustafaali@gmail.com

Kunali Chetan Hapani
Department of Clinical Pharmacy, Alfred E. Mann School of Pharmacy and Pharmaceutical Sciences, University of Southern California, Los Angeles, California, 90089, USA
khapani@usc.edu
ORCID: 0000-0002-7032-7952

Qiushi Peng

Department of Pharmacology and Pharmaceutical Sciences, Alfred E. Mann School of Pharmacy, University of Southern California, Los Angeles, CA, 90089, USA
qiushipe@usc.edu
ORCID: 0000-0001-5051-3435

Ram Ayyala
Department of Quantitative and Computational Biology, Dana and David Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA, USA
rayyala@usc.edu
ORCID: 0000-0001-7275-271X

Ruiwei Guo
Department of Pharmacology and Pharmaceutical Sciences, Alfred E. Mann School of Pharmacy, University of Southern California, Los Angeles, CA, 90089, USA
ruiweigu@usc.edu

Shaunak Kapur
Seven Lakes High School, 9251 South Fry Road, Katy, Texas, 77494, United States of America
shaunak1206@gmail.com
ORCID: 0009-0007-8053-8167

Tejasvene Ramesh
Center for Personalized Cancer Therapy, University of Massachusetts Boston, Boston, MA, 02125, USA
tejasvene.ramesh001@umb.edu
ORCID: 0000-0002-5021-1338

Malak S. Abedalthagafi
Department of Pathology & Laboratory Medicine, Emory University Hospital, Atlanta, GA, USA

King Salman Center for Disability Research, Riyadh, Saudi Arabia
malak.althgafi@emory.edu
ORCID: https://orcid.org/0000-0003-1786-3366

Serghei Mangul
Department of Clinical Pharmacy, Alfred E. Mann School of Pharmacy and Pharmaceutical Sciences, University of Southern California, Los Angeles, CA 90033, USA
Department of Quantitative and Computational Biology, USC Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA 90033, USA
mangul@usc.edu
ORCID: https://orcid.org/0000-0003-4770-3443

Abstract

Recent advances in high-throughput sequencing technologies have made it possible to collect and share a massive amount of omics data, along with its associated metadata. Enhancing metadata availability is critical to ensure data reusability and reproducibility and to facilitate novel biomedical discoveries through effective data reuse. Yet, incomplete metadata accompanying public omics data limits the reproducibility and reusability of millions of omics samples. In this study, we performed a comprehensive assessment of metadata completeness shared in both scientific publications and/or public repositories by analyzing over 253 studies encompassing over 164 thousands samples. We observed that studies often omit over a quarter of important phenotypes, with an average of only 74.8% of them shared either in the text of publication or the corresponding repository. Notably, public repositories alone contained 62% of the metadata, surpassing the textual content of publications by 3.5%. Only 11.5% of studies completely shared all phenotypes, while 37.9% shared less than 40% of the phenotypes. Studies involving non-human samples were more likely to share metadata than studies involving human samples. We observed similar results on the extended dataset spanning 2.1 million samples across over 61,000 studies from the Gene Expression Omnibus repository. The limited availability of metadata reported in our study emphasizes the necessity for improved metadata sharing

practices and standardized reporting. Finally, we discuss the numerous benefits of improving the availability and quality of metadata to the scientific community abd beyond, supporting data-driven decision-making and policy development in the field of biomedical research.

Introduction

Advancements in the high throughput sequencing technologies over the last decade have made omics data readily available to the public, enabling researchers to access a vast array of data across various diseases and phenotypes from the textual content of publications or various public repositories[1]. The vast amount of omics data and its accompanying metadata enable unprecedented exploration of biological systems through the re-analysis of public omics data, which is also known as secondary analysis[2,3]. Secondary analysis is capable of driving transformative breakthroughs in biomedical research, fostering collaboration, accelerating scientific progress, and deepening the understanding of human biology and diseases. By leveraging this readily available wealth of omics information, researchers can unravel the complex interplay of genes, proteins, cellular processes, and environmental factors from raw omics data. To facilitate the reproducibility of omics studies and the effective secondary analysis of omics data, metadata completeness and accuracy are crucial[4,5]. Metadata enriches raw data by providing essential details about its fundamental attributes, including phenotype, age, sex, and disease condition, as well as comprehensive experimental and environmental information, such as the data generator, creation date, data format, and sequencing protocols[2]. As a result, accurate and comprehensive metadata are critical for the efficient utilization, sharing, and subsequent re-analysis of omics data[1,6–10]. Metadata also enable data-driven decision-making and policy development in fields such as biomedical sciences, clinical research, environmental sciences, and social sciences[1].

Although the biomedical community has made tremendous efforts in sharing omics data, little attention is allocated to ensure the completeness of metadata accompanying raw omics data[1,11]. We previously reported the limited availability of metadata accompanying sepsis-based transcriptomics studies[1] but the overall patterns of metadata sharing across other diseases and

organisms remains unknown[1,12,13]. Incomplete metadata hinders researchers' ability to utilize metadata information for subsequent downstream analysis[8,14,15]. Typically, metadata accompanying omics studies has been shared in two ways, in public repositories and in the text of publications. Metadata shared solely in the textual content of publications has many limitations and is insufficient to ensure that it is complete, accurate, accessible, and machine-actionable[16]. This is because metadata not in a standardized format in publications can be scattered across multiple locations, making it challenging and time consuming for researchers to locate and integrate metadata from multiple sources for further downstream analysis. Additionally, metadata shared in publications is often provided at the study level lacking per sample or individual information. The lack of sample-level metadata limits the reproducibility of conducted research, restricting the credibility and ultimately making secondary analysis impossible or extremely difficult to conduct. If the metadata is shared only in the publications, the researchers will require a time-consuming and laborious manual data-mining process to extract such metadata information, especially when looking at large scale studies[17]. Mining and extracting metadata from the publications can be challenging due to the presence of misannotated, unstructured and absent metadata information. In contrast to sharing metadata within the text of a publication, a highly effective method for disseminating metadata involves sharing it through publicly accessible repositories, making researchers easy to access this information and facilitate downstream secondary analyses[1,18]. As a result, public repositories play a crucial role in advancing biomedical research by facilitating the efficient and effective sharing and utilization of data accompanying metadata[1].

In this study, we analyzed a total of 253 randomly selected studies over 164 thousands samples across various disease conditions, phenotypes and organisms. We investigated the prevailing practice of metadata sharing in the textual content of publications and the corresponding public repositories. We observed that studies often omit over a quarter of crucial phenotypes, with an average of only 74.8% shared. Only 11.5% of studies completely shared all phenotypes, while 37.9% shared less than 40% of the phenotypes. Studies involving non-human samples were more likely to have complete metadata than studies involving human samples.

Additionally, public repositories contained more complete metadata compared to metadata shared in the textual content of publications. To generalize our results, we further examined over 61 thousand studies over 2.1 million samples from Gene Expression Omnibus[19] (GEO) repository. The overall metadata availability of 2.1 million GEO samples was 63.2%. Similar to the metadata availability of surveyed 253 studies, non-human studies have 16.1% more pheenotypes shared compared to human studies. Notably, the availability of metadata in published studies has increased substantially over time. In studies published before 2011, metadata availability was limited, with less than 1% of studies having metadata being shared. In contrast, in studies released after 2021, there has been a remarkable enhancement in metadata availability, with as many as 50% of studies now incorporating this valuable data.

Our findings revealed the incomplete metadata reporting practices in both the textual content of publications and public repositories. The limited metadata availability highlighted in our study emphasizes the pressing need to enhance metadata completeness and accessibility. Therefore, it is crucial to define best practices for metadata sharing and identify barriers to such practices. This will enable future researchers to ensure the availability, completeness, and accuracy of metadata. Increasing metadata accessibility and reliability would significantly benefit the scientific community and beyond, facilitating data-driven decisions and policy formulations in across diverse biomedical research domains. Surprisingly, these findings highlight that the textual content of the publication contains less metadata compared to the repository. Instead, it appears that numerous phenotypes are not shared between these two sources. Our study underscores the importance of metadata sharing in promoting the reusability of multi-omics data. By promoting the availability and quality of metadata accompanying raw omics data, we can enable more accurate and efficient secondary analyses, which in turn may advance our understanding of complex diseases and their underlying mechanisms, and ultimately discover novel biomedical insights to improve human health[20].

Results

Assessing the completeness of public metadata accompanying omics studies

We have performed a comprehensive analysis of the availability of the metadata reported in the textual content of the original scientific publications and the corresponding public repositories. We meticulously analyzed a total of 253 randomly selected scientific publications encompassing over 164 thousand samples across various disease conditions, phenotypes and organisms (D1 dataset). The human studies encompassed various disease conditions, namely Alzheimer's disease (AD), acute myeloid leukemia (AML), cardiovascular disease (CVD), inflammatory bowel disease (IBD), multiple sclerosis (MS), sepsis, and tuberculosis (TB) (Table S1) (Figure S1). Among the 253 studies, 153 pertained to human studies, while 100 focused on non-human studies. To assess metadata availability, we manually reviewed the phenotypic information within the textual content of publications, including both the main text and supplementary materials. To retrieve metadata from the public repositories, we developed the custom Python scripts (Methods Section). To increase the generalizability of our analysis, we randomly selected over 61 thousands studies from the Gene Expression Omnibus[21] (GEO) repository, encompassing a total of over 2.1 million samples (D2 dataset). For both D1 and D2 datasets, we examined the four common phenotypes shared by both human and non-human studies, which included organism, sex, age, and tissue types. In addition to the common phenotypes, we investigated the human-specific phenotype of race/ethnicity/ancestry and the non-human-specific phenotype of strain information.

Limited availability for essential phenotypes accompanying 253 studies

Our analysis has unveiled that more than a quarter of phenotypes are not shared in either the textual context of the publication or public repositories (Figure S2). While all studies shared at least one phenotype, on average, only 11.5% of studies managed to encompass the entire set of phenotypes. In contrast, over a quarter of studies shared a significant 80% of the phenotypes, and more than a third of studies shared 40% of the phenotypes. Additionally, 4.7% of the studies shared less than 20% of the phenotypes (Figure 1a). Among the six phenotypes, organism information was the most commonly shared phenotype with all the studies sharing such information (Figure 1b). Meanwhile, up to 80% of the studies reported tissue information, and similarly, 80% of the non-human samples reported strain information (Figure 1b). About 60% of

the studies have available sex and age information (Figure 1b), with, on average, half of the samples within the studies reporting this information in both sources. Importantly, the race/ethnicity/ancestry information was the least frequently reported phenotype with only up to 20% of the human studies reported this essential phenotype information (Figure 1b), with only about 10% of the samples, on average, sharing such information within the studies in both sources.

We noticed that metadata availability has improved substantially over the years (Figure 1c). Our analysis revealed a substantial improvement in metadata reporting practices, particularly in studies released after 2018, where we observed a remarkable 47.1% increase in metadata reporting in that year alone. In contrast, studies predating 2017 had substantially lower metadata availability with only 10.3% metadata available (Figure 1c). Most of the phenotypes experienced a drastic increase in availability after 2018, including age, organism, sex, tissue, and strain information. In contrast, studies published before 2010 show a significant limitation in the availability of the six phenotypes (Figure S3).

Inconsistent practices of metadata sharing across the textual content of publications and corresponding public omics repositories

We compared metadata availability across the 253 studies reported in the textual content of publications versus in the corresponding public repositories. The metadata availability in original publications was more than 3% lower than the metadata availability in public repositories, which was 62.0% (Figure 1d). In comparison to the textual context of the publications, public repositories share more complete phenotypes for organism, sex, age, strain, and race/ethnicity/ancestry, while the textual content of publications share more complete tissue information (Figure 1e). Organism information was the most completely and consistently shared phenotype in both public repositories and the textual content of the publications with all such information available across all the 164,909 samples (Figure 1e). Notably, organism information was fully available in the public repositories in contrast to 99.2% availability in the textual content of the publication (Figure 1e). In contrast, age was the least consistently shared

phenotype between the textual content of publications and public repositories, with 17.9% more samples having age information in public repositories (Figure 1e). Tissue information was reported in both platforms for over 60% of the samples, 20% of samples only report such information in the textual content of the publications, while only about 10% shared it only in public repositories. About 50% of the samples shared sex and race/ethnicity/ancestry information between both sources, and 30% of the samples only shared such information in public repositories (Figure 1f).

Non-human studies demonstrated an enhanced commitment to sharing metadata

We compared the metadata reporting practices for the common phenotypes across 153 non-human and 100 human studies. The metadata availability for human samples was 56.1% and for non-human samples was 60.86% (Figure 1g). Human and non-human samples had more complete metadata reported in public repositories (66.6% and 62.0%, respectively) than in the textual content of publications (54.6% and 59.7%, respectively) (Figure S4). Non-human studies reported more complete metadata for age and sex, while human studies reported more complete metadata for tissue information (Figure 1h). In the human and non-human studies, organism information is widely available (Figure 1h). Tissue information is also widely available in human studies; conversely, only up to 60% of the non-human samples had available tissue information. While organism and tissue information was commonly reported in human studies, sex data was provided by only 49.4% of human studies, age data by 48.9%, and race/ethnicity/ancestry details by a mere 22% of human studies (Figure 1h). For non-human studies, other than the organism data, strain information was the most reported, available for 76.5% of samples. Age, and sex information has over 60% availability in non-human studies (Figure 1h). We observed that both human and non-human studies shared more complete metadata in public repositories than in the textual content of publications. Of the 153 human studies, 97 had consistent metadata sharing practices between original publications and public repositories, 36 had lower metadata availability in the textual content of publications, and 20 had higher metadata availability in the textual content of publications (Figure 1i). Of the 100 non-human studies, 59 had more complete metadata in public repositories than in the textual content

of publications, 28 had the same extent of metadata availability in both sources, and 13 had more complete metadata in the textual content of publications than in public repositories (Figure 1i).

We observed significant discrepancies in metadata sharing among non-human species and aim to assess the overall completeness of metadata across these species. We analyzed metadata availability in the textual content of publications and public repositories for the top five non-human organisms in the 153 non-human studies, including *Mus musculus, Parus major, Glycine max, Danio rerio,* and *Gallus gallus*. Mus musculus and Parus major had similar metadata availability levels (up to 60%) in both sources. In contrast, *Glycine max, Danio rerio*, and *Gallus gallus* demonstrate a notable difference in metadata availability between the textual content of publications and public repositories. Specifically, they reported more comprehensive metadata in the textual content of publications than in public repositories (Figure S5).

Sharing experiment-level metadata remains a common practice

The concept of experiment-level metadata availability was characterized by the sharing of only summarized information regarding the phenotypes of the samples, omitting detailed per-sample phenotypic information. It remains a common practice to offer an overall description of the study or experiment's participants, while refraining from providing detailed descriptions of each individual's phenotypes. Among human studies, 14.8% of metadata was exclusively shared at the experiment-level, while non-human studies exhibited a comparable pattern, with 6.8% of metadata exclusively shared at the experiment-level. In human studies, approximately 60% of the samples reported available age information only at the experiment-level (Figure S6). Up to 20% of the human samples only shared experiment-level metadata for disease, sex, and race/ethnicity/ancestry information (Figure S6 and Figure S7). In non-human studies, 20% of the non-human samples only reported tissue information at the experiment-level. Sex information was often presented in a summarized format, with approximately 40% of this information available at the experiment level. In contrast, less than 10% of the sex information was available at the sample level (Figure S6 and Figure S7). Notably, the experiment-level metadata for sex,

age, and race/ethnicity/ancestry information accounts for approximately half of the human samples' phenotypic information availability (Figure S7).

Discrepancies in metadata sharing practices were observed across seven disease conditions

Next, we examined metadata sharing practices for seven disease conditions, including Alzheimer's disease (AD), acute myeloid leukemia (AML), cardiovascular disease (CVD), inflammatory bowel disease (IBD), multiple sclerosis (MS), sepsis, and tuberculosis (TB), in human studies. The metadata completeness for the six phenotypes varies across the seven disease condition studies (Figure 2). All seven diseases have fully reported metadata for organism and tissue types (Figure 2). The reporting practices of three phenotypes (sex, age, and race/ethnicity/ancestry information) exhibited variability across the seven diseases. For the sex information, CVD studies had reported the highest availability for such information (75.9%), followed by AD studies, sepsis studies, AML studies, IBD studies, and TB studies , while MS studies had the lowest (10.8%). For the age information, sepsis studies had reported the highest availability of such information (67.9%), followed by AD studies, AML studies, IBD studies, CVD studies, and TB studies (32.4%), while MS studies had the lowest (9.7%). For the race/ethnicity/ancestry information, most studies rarely reported such information, and notably, none of the samples in the IBD and MS studies reported race/ethnicity/ancestry information (Figure 2).

Analysis of 61,000 studies encompassing 2.1 million samples across public repositories confirms limited metadata availability

We investigated the availability of the six phenotypes reported for 2,168,620 samples from 61,950 studies available at the Gene Expression Omnibus (GEO) (D2 dataset). The overall availability of metadata among the 61,950 studies was 63.2% (Figure S8). Notably, we discovered that studies preceding 2010 exhibited a lack of shared metadata (Figure S9). We discovered that human studies exhibited a lower average metadata completeness (47.4%) compared to non-human studies (63.5%) (Figure S8). As a result, the availability of metadata in non-human samples surpasses that in human samples (Figure S8 and Figure S9). Human samples invariably share

organism and tissue information (100% and 99.9%, respectively), but less frequently report age (19%) and sex (13.8%), with race being the least disclosed at 4.2% (Figure S10); Conversely, all non-human samples report organism and tissue data, with strain and age information present in 56.7% and 37.4% of samples, respectively, while only 24.15% include sex information (Figure S10). Tissue, sex, and age metadata are more comprehensively reported in non-human samples compared to their human counterparts (Figure S10).

Discussion

Our study is the first to systematically analyze metadata completeness across multiple organisms and phenotypes, revealing limited availability of metadata both in textual content of publications and public repositories. We found substantial disparities in the completeness of metadata across different phenotypes, with organism and tissue types consistently reported, in contrast to age, sex, strain, and race/ethnicity/ancestry information often not shared in either the text of publications or public repositories (Figure S2). Notably, non-human studies demonstrated an enhanced commitment to sharing metadata than human studies. Several factors may contribute to the observed differences in metadata availability between human and non-human studies. Firstly, the nature of the research focus plays a crucial role. Human studies are typically subject to stricter ethical guidelines and consent requirements, and data sharing regulations[22–24]. These regulations may limit the extent of metadata that can be collected, whereas non-human studies may not encounter the same constraints[22,24–26]. As a result, non-human studies may enjoy easier access to metadata due to less stringent privacy and consent regulations, facilitating more extensive data collection efforts[27–29]. Data sharing norms may further influence metadata availability. Non-human studies often have a stronger tradition of data sharing, motivating researchers to provide more complete metadata to promote transparency and collaboration[30].

Our study is also the first to investigate metadata availability at both the experiment and sample levels. Our previous analysis did not distinguish among the various methods of metadata sharing, whether it involved sharing metadata information for each sample individually or just

providing summary statistics for the entire study[1]. Typically, sample-level metadata is more valuable for ensuring the transparency and reproducibility of reported results and for secondary analysis than experiment-level metadata, because it provides detailed information about each sample rather than a summary of the entire study[31]. Researchers who want to replicate or expand on an experiment or perform secondary analyses on the raw data need sample-level metadata[32]. In contrast, experiment-level metadata provides a comprehensive description of the study, but it may not be enough for many types of analyses that require a granular level of detail. When conducting analyses for new research questions, the specificity of sample-level metadata enables researchers to reuse the existing dataset to answer novel research questions and draw meaningful conclusions. We observed that metadata is often reported at the experiment-level, not the sample-level. In most textual content of publications, metadata is only available at the experiment level. We found that half of the human samples with available metadata for sex, age, and race/ethnicity/ancestry have only experiment-level metadata. Note that all studies have shared a portion of information pertaining to the metadata of the samples, implying a conscientious effort by the authors to provide relevant details about the samples' metadata; however, such effort remains incomplete. However, there are several limitations in our study. First, we were able to only extract commonly reported phenotypes and were unable to assess the availability of study-specific phenotypes. Additionally, our analysis focused solely on the availability of metadata reported and did not address the quality or accuracy of metadata reported.

Making metadata widely available can enhance the transparency and reproducibility of research, facilitate cross-study comparisons, and contribute to the advancement of scientific knowledge by providing comprehensive contextual information about datasets. To ensure the quality and usefulness of clinical data, metadata reporting in all categories must be improved and shared across domains on public repositories, in addition to the text of publications[1]. Additionally, sharing metadata solely within the text of a publication is not an optimal approach for metadata sharing for several reasons. Firstly, when metadata is embedded within textual content, it becomes a challenge to locate and extract, especially for those researchers who may

be sourcing multiple papers for comprehensive data. Additionally, the metadata presented within a publication might not always be complete; key details might be overlooked or omitted due to space constraints or editorial decisions. This potential for incompleteness can lead to inaccuracies when researchers rely on this data for their work. Moreover, when metadata is constrained within a publication, it's not easily shareable. Digital repositories or databases, on the other hand, allow for streamlined sharing and collaboration. Thus, relegating metadata solely to the text of a publication creates barriers to efficient research and collaboration. Sharing metadata on public repositories can reduce the efforts in metadata mining directly from the textual content of publications or from the request to the authors since the both processes can be time-consuming and error-prone[11,33–35]. Leveraging previously published data for novel biological discoveries could be facilitated when the metadata accompanying its raw omics data is reported, present in a standardized format, and made available in online repositories[5,36,37]. Improving the availability of metadata in public repositories could provide valuable and accurate information for downstream analyses, and further enhance the usefulness of the public repositories.

Our overall findings highlight the need for improved metadata reporting and sharing practices in biomedical research. The limited availability of metadata in both textual content of publications and public repositories impedes data reuse and reproducibility, and the disparities in completeness across different phenotypes make it difficult to conduct secondary analyses. We encourage authors to routinely report all relevant metadata, including organism, tissue type, age, sex, strain, and race/ethnicity/ancestry information, in both their publications and public repositories. Improving the accessibility and reliability of metadata would significantly benefit the broader scientific community by facilitating data-driven research and fostering secondary analysis within the biomedical research field.

Methods

Datasets

We examine the per-sample metadata completeness within the D1 dataset, spanning 253 multi-omics studies and comprising a total of 164,909 samples. This dataset included 20,047 human samples from 153 studies and 144,862 non-human samples from 100 studies. The human studies covered a variety of diseases, including Alzheimer's disease, acute myeloid leukemia, cardiovascular disease, inflammatory bowel disease, multiple sclerosis, sepsis, and tuberculosis. The remaining 100 studies involved non-human samples. Additionally, for the D2 dataset, we randomly selected 61,950 studies from the GEO repository, encompassing a total of 2,168,620 samples. Among the studies in the D2-dataset, 2,119,749 samples were non-human samples, while 48,871 samples were human samples.

Examined phenotypes

We examined the metadata for the six phenotypes among the human and non-human studies in the D1 and D2 dataset. For human samples, we examined the human-specific phenotype of race/ethnicity/ancestry information, along with four common phenotypes, including organism, age, sex, and tissue types. Conversely, for non-human samples, we investigated the additional non-human-specific phenotype of strain information, along with the four common phenotypes, including organism, age, sex, and tissue types. In the D2 dataset, we examined the metadata for the same six phenotypes as in the D1 dataset, including tissue types, organism, sex, age, strain information (non-human samples), and race/ethnicity/ancestry (human samples).

Extraction of metadata from the textual content of publications

We manually extracted metadata from the text of the publication or in the supplementary materials. Based on the information provided from the textual content of publications or supplementary materials, we categorized the metadata availability into two categories, sample-level metadata available and experiment-level metadata available. Additionally, we inferred sample-level metadata from the textual content of publications when we were able to infer the metadata for all samples, assuming that all samples shared the same phenotypes. We identified the sample-level metadata is available only if the study shared individual samples' metadata

information. If the studies that only shared the summary information of the metadata with not available sample-level metadata information in the text of the publications, we categorized such studies as providing experiment-level metadata while not having available sample-level metadata.

Extraction of metadata from the public genomic repositories

To extract metadata from the public genomic repositories, we downloaded the metadata information from the GEO repository using our custom Python scripts. The script was specifically designed to extract the metadata that was reported in the GEO repository. We first scraped GEO datasets' metadata by logging into NCBI FTP using ftp python module and iterated over all the GEO accession numbers listed in a text file, and downloaded it into an Extensible Markup Language **(**XML) file. XML files are difficult to interpret and work with, so we parsed all the files into readable Comma Separated Values (CSV) files. The CSV sheet was structured in such a way that it contains submission date, release date, last update date, title, accession, type, source, organism, sex, age, BMI, braak stage, brain bank, replicate, molecule, extract protocol and description.

Contributions

SM supervised the work. YH analyzes the collected data. PJ validated the collected data from human studies. AR collected the data from sepsis studies. RG and EL collected the data from tuberculosis studies. IN collected the data from cardiovascular disease studies. RA and MYW collected data from acute myeloid leukemia. JH collected data from inflammatory bowel disease studies. AS and AN collected data from Alzheimer's disease studies. SK collected data from non-human studies. GB and MM developed customized Python script to extract metadata from GEO. All co-authors contributed to the revision of the manuscript.

Acknowledgements

SM and Y-NH is supported by the National Science Foundation (NSF) grants 2041984, 2135954 and 2316223 and National Institutes of Health (NIH) grant R01AI173172.

Code Availability

All code for the study can be found at https://github.com/Mangul-Lab-USC/metadata-completeness.

Data Availability

All data for the study can be found at https://github.com/Mangul-Lab-USC/metadata-completeness.

Conflict of interest

None declared.

References

1. Rajesh, A. *et al.* Improving the completeness of public metadata accompanying omics studies. *Genome Biol.* **22**, 106 (2021).
2. Johnston, M. P. Secondary Data Analysis: A Method of which the Time Has Come. *Qual. Quant. Methods Libr.* **3**, 619–626 (2017).
3. Ruggiano, N. & Perry, T. E. Conducting secondary analysis of qualitative data: Should we, can we, and how? *Qual. Soc. Work* **18**, 81–97 (2019).
4. Schriml, L. M. *et al.* COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci. Data* **7**, 188 (2020).
5. Kim, Y.-M., Poline, J.-B. & Dumas, G. Experimenting with reproducibility: a case study of robustness in bioinformatics. *GigaScience* **7**, giy077 (2018).
6. Qin, J. & D'ignazio, J. The Central Role of Metadata in a Science Data Literacy Course. *J. Libr. Metadata* **10**, 188–204 (2010).
7. Ghiringhelli, L. M. *et al.* Towards efficient data exchange and sharing for big-data driven materials science: metadata and data formats. *Npj Comput. Mater.* **3**, 1–9 (2017).
8. Musen, M. A. Without appropriate metadata, data-sharing mandates are pointless. *Nature* **609**, 222–222 (2022).

9. Brito, J. J. *et al.* Recommendations to enhance rigor and reproducibility in biomedical research. *GigaScience* **9**, giaa056 (2020).
10. Clarke, R. I. & Schoonmaker, S. Metadata for diversity: Identification and implications of potential access points for diverse library resources. *J. Doc.* **76**, 173–196 (2019).
11. Gozashti, L. & Corbett-Detig, R. Shortcomings of SARS-CoV-2 genomic metadata. *BMC Res. Notes* **14**, 189 (2021).
12. Toczydlowski, R. H. *et al.* Poor data stewardship will hinder global genetic diversity surveillance. *Proc. Natl. Acad. Sci.* **118**, (2021).
13. Ryan, M. J. *et al.* Towards a unified data infrastructure to support European and global microbiome research: a call to action. *Environ. Microbiol.* **23**, 372–375 (2021).
14. Rung, J. & Brazma, A. Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.* **14**, 89–99 (2013).
15. Musen, M. A. *et al.* The center for expanded data annotation and retrieval. *J. Am. Med. Inform. Assoc. JAMIA* **22**, 1148–1152 (2015).
16. Batista, D., Gonzalez-Beltran, A., Sansone, S.-A. & Rocca-Serra, P. Machine actionable metadata models. *Sci. Data* **9**, 592 (2022).
17. P, Y. *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **29**, (2011).
18. Corrêa, F. B., Saraiva, J. P., Stadler, P. F. & da Rocha, U. N. TerrestrialMetagenomeDB: a public repository of curated and standardized metadata for terrestrial metagenomes. *Nucleic Acids Res.* **48**, D626–D632 (2020).
19. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* **41**, D991-995 (2013).
20. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).
21. Clough, E. & Barrett, T. The Gene Expression Omnibus Database. in *Statistical Genomics: Methods and Protocols* (eds. Mathé, E. & Davis, S.) 93–110 (Springer, 2016). doi:10.1007/978-1-4939-3578-9_5.

22. Committee on Population, Division of Behavioral and Social Sciences and Education, & Engineering The National Academies of Sciences. *Exploring the Ethical Imperative for Data Sharing*. *Sharing Research Data to Improve Public Health in Africa: A Workshop Summary* (National Academies Press (US), 2015).

23. Garett, R. & Young, S. D. Ethical Views on Sharing Digital Data for Public Health Surveillance: Analysis of Survey Data Among Patients. *Front. Big Data* **5**, (2022).

24. Meyer, M. N. Practical Tips for Ethical Data Sharing. *Adv. Methods Pract. Psychol. Sci.* **1**, 131–144 (2018).

25. Carroll, M. W. Sharing Research Data and Intellectual Property Law: A Primer. *PLOS Biol.* **13**, e1002235 (2015).

26. Corrie, B. D. *et al.* Data Sharing and Reuse: A Method by the AIRR Community. *Methods Mol. Biol. Clifton Nj* **2453**, 447–476 (2022).

27. Bonomi, L., Huang, Y. & Ohno-Machado, L. Privacy challenges and research opportunities for genomic data sharing. *Nat. Genet.* **52**, 646–654 (2020).

28. Nass, S. J., Levit, L. A., Gostin, L. O. & Rule, I. of M. (US) C. on H. R. and the P. of H. I. T. H. P. The Value and Importance of Health Information Privacy. in *Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research* (National Academies Press (US), 2009).

29. Kels, C. G. HIPAA in the Era of Data Sharing. *JAMA* **323**, 476–477 (2020).

30. Thoegersen, J. L. & Borlund, P. Researcher attitudes toward data sharing in public data repositories: a meta-evaluation of studies on researcher data sharing. *J. Doc.* **78**, 1–17 (2021).

31. Sansone, S.-A. *et al.* Toward interoperable bioscience data. *Nat. Genet.* **44**, 121–126 (2012).

32. Gonçalves, R. S. & Musen, M. A. The variable quality of metadata about biological samples used in biomedical experiments. *Sci. Data* **6**, 190021 (2019).

33. Huang, Y.-N. *et al.* Ancestral diversity is limited in published T cell receptor sequencing studies. *Immunity* **54**, 2177–2179 (2021).

34. Stodden, V., Seiler, J. & Ma, Z. An empirical analysis of journal policy effectiveness for

computational reproducibility. *Proc. Natl. Acad. Sci.* **115**, 2584–2589 (2018).

35. Sarkans, U. *et al.* REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology. *Nat. Methods* **18**, 1418–1422 (2021).

36. Halbert, M. D. Advancing Reproducibility at the NSF. *Computer* **55**, 31–39 (2022).

37. Ryan, J. *et al.* MethodsJ2: a software tool to capture metadata and generate comprehensive microscopy methods text. *Nat. Methods* **18**, 1414–1416 (2021).

Figures

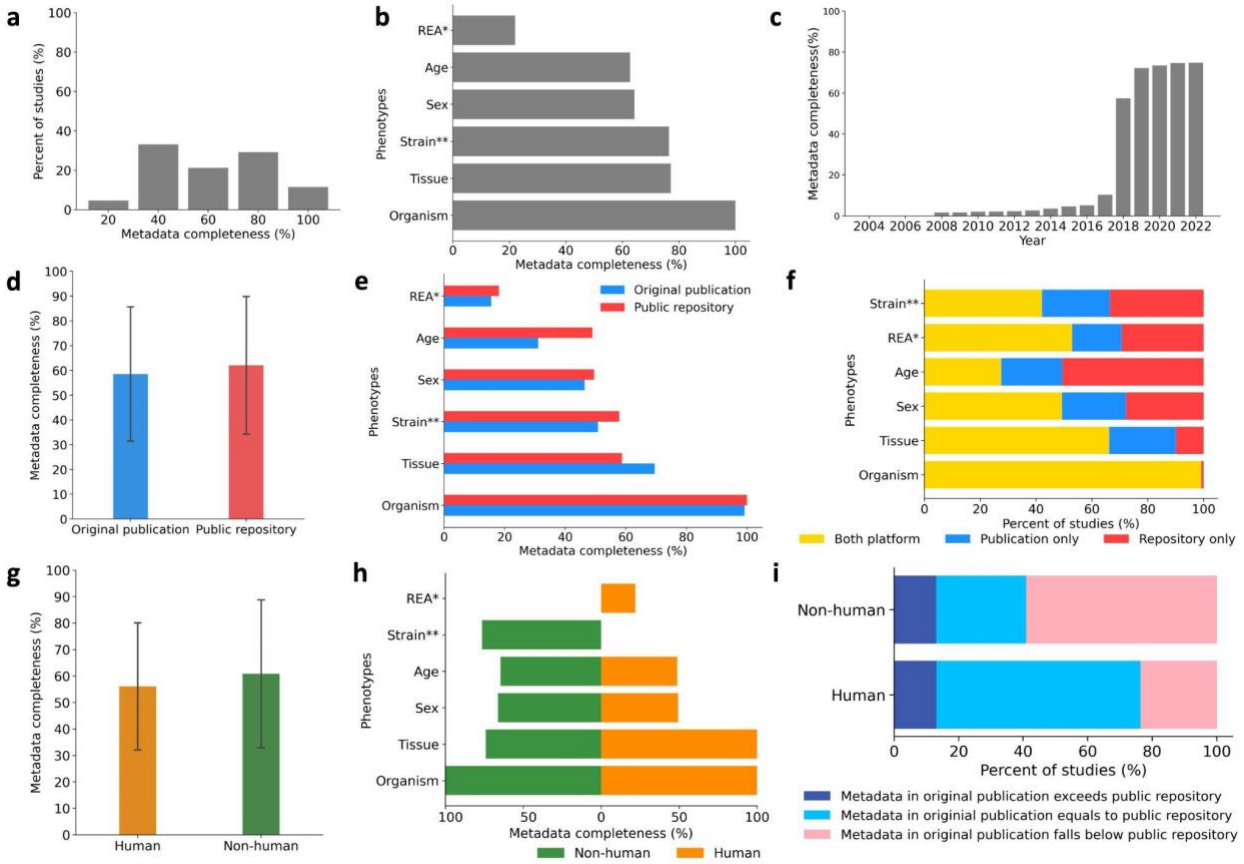

Figure 1. Completeness of public metadata accompanying omics studies (a) The distribution of studies that share 100% of metadata (all phenotypes), 80% of metadata (four phenotypes), 60% of metadata (three phenotypes), 40% of metadata (two phenotypes), and 20% of metadata (one phenotype) in the D1 dataset. (b) The availability of the six phenotypes information in the textual content of publications and/or public repositories in the D1-dataset. (c) The cumulative metadata availability of the D1-dataset over the years. (d) The overall metadata availability in the textual

content of publications and public repositories in the D1-dataset. (e) The sample-level metadata across seven phenotypes among all the samples in the D1-dataset reported in the textual content of publications and public repositories. (f) The proportion between the studies shared the six phenotypes in both sources, including textual content of the publication and public repositories, the studies shared the six phenotypes only in the textual content of the publication, and the studies shared the six phenotypes only in the public repositories. (g) The overall metadata availability for human and non-human samples in the D1-dataset. (h) Right: The overall composition of sample-level metadata availability across six phenotypes in the textual content of publications and public repositories across the human samples in the D1-dataset; Left: The overall composition of sample-level metadata availability across five phenotypes in the textual content of publications and public repositories across the non-human samples in the D1-dataset. (i) The percentage of non-human and human studies in the D1-dataset that reported the sample-level metadata in the textual content of publications the same as, exceeds, or falls below the sample-level metadata reported in the public repository. (The REA* (race/ethnicity/ancestry) information is analyzed over the human samples only; The strain** information is analyzed over the non-human samples only.)

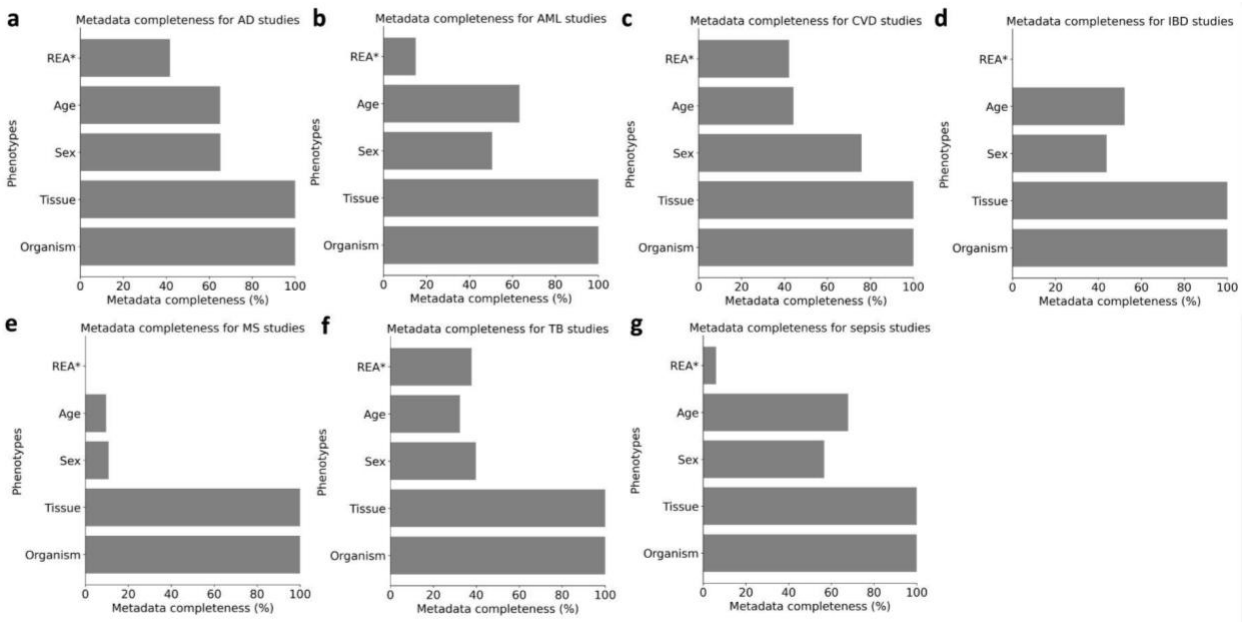

Figure 2. Metadata completeness across human studies. (a). Metadata completeness for Alzheimer's disease (AD) studies. (b) Metadata completeness for acute myeloid leukemia (AML) studies. (c) Metadata completeness for cardiovascular disease (CVD) studies. (d) Metadata completeness for inflammatory bowel disease (IBD) studies. (e) Metadata completeness for multiple sclerosis (MS) studies. (f) Metadata completeness for tuberculosis (TB) studies. (g) Metadata completeness for sepsis studies. (The REA* (race/ethnicity/ancestry) information is analyzed over the human samples only; The strain** information is analyzed over the non-human samples only.)