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
gi

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

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

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

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

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
Twitter: @malakabed

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
Twitter: @smangul1

*These authors jointly supervised this work

Abstract

The recent explosion of omics data, made accessible by advances in high-throughput sequencing technologies and public repositories, enables unprecedented biological insights

through secondary analysis. Enhancing metadata availability in public repositories is critical to ensure data reusability and reproducibility and facilitate secondary analysis for novel biomedical discoveries. Yet, incomplete metadata accompanying public omics data limits the reproducibility, reusability, and diversity of millions of omics samples from public repositories. In this study, we performed a comprehensive assessment of metadata completeness in scientific publications and public repositories by analyzing over 253 studies encompassing over 164,000 samples. We observed that studies often omit sharing over a quarter of crucial phenotypes, with an average of only 74.8% of phenotypes being shared. We observed a greater metadata completeness for non-human studies compared to human studies, with 61% of non-human samples possessing available metadata, in contrast to 56% for human samples. Additionally, public repositories contained on average 3.5% more metadata information than original publications. The limited metadata availability reported by our study underscores the urgency to enhance metadata completeness and accessibility. Making metadata more accessible and reliable would greatly benefit the scientific community and beyond, enabling data-driven decision-making and policy development in fields 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 original publications or various public repositories[1]. The vast amount of omics data and its accompanying metadata enable unprecedented exploration of biological systems through secondary analysis, which also known as the re-analysis of omics data[2,3]. Secondary analysis drives 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 information, researchers can unravel the complex interplay of genes, proteins, cellular processes, and environmental factors in raw omics data. Metadata is essential for secondary analyses and the reusability and reproducibility of omics data. It provides essential details

about the fundamental attributes and basic phenotypes of raw data, such as 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].

Making metadata more accessible and accurate would greatly benefit the scientific community and society[1,4–8]. It would enable data-driven decision-making and policy development in fields such as healthcare, environmental sciences, and social sciences[1]. Metadata empowers researchers, stakeholders, regulatory authorities, and the public to make informed decisions based on reliable and relevant evidence. Metadata is essential for unlocking the full potential of omics data, enabling secondary analysis to accelerate scientific discovery and improve human health[6]. Metadata also plays a crucial role in verifying the authenticity and integrity of the data, facilitating the effective management of access and permissions[4]. Accurate and comprehensive metadata are critical for the efficient utilization, sharing, and subsequent re-analysis of omics data, which is also known as secondary analysis. To facilitate the reproducibility of omics studies and the effective secondary analysis of omics data, metadata completeness and accuracy are crucial[9,10]. 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].

Traditionally, metadata accompanying omics studies has been shared in two ways: in public repositories and in the text of publications. However, metadata shared only in original publications is insufficient to ensure that it is complete, accurate, and accessible. This is because metadata in publications can be scattered across multiple locations, such as the main text, supplementary materials, or figures, making it difficult for researchers to find and integrate. Additionally, metadata shared in publications is often summarized, shared at the study level, and not in a standardized format. This lack of sample-level metadata makes it difficult for researchers to conduct subsequent secondary analyses. Therefore, within the realm of metadata and data sharing, the primary role of public repositories is to make both omics data and metadata readily available to the biomedical community and to foster downstream

bioinformatics analyses[1,12]. If the metadata information accompanying raw data 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[13]. Mining and extracting metadata from the publications can be challenging due to the presence of misannotated, unstructured and absent metadata information. Even when metadata is shared in public repositories, its diverse formats and reporting practices can make researchers challenging to extract such information. The use of different ontologies and annotation systems across studies can create diversity in reported metadata, leading to additional challenges in detecting and resolving inconsistencies and errors[12]. Despite the challenges, public repositories play a crucial role in advancing biomedical research by facilitating the efficient and effective sharing and utilization of data accompanying metadata[1].

According to our previous studies, metadata accompanying the raw data is often incomplete[1,14,15]. The incompleteness of metadata impede the researchers from reproducing the results reported in the study and create additional challenges in the secondary analyses[6,16,17]. 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. Additionally, 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 repurposing and reproducing data than experiment-level metadata, because it provides detailed information about each sample rather than a summary of the entire study[18]. Researchers who want to replicate or expand on an experiment or perform secondary analyses on the raw data need sample-level metadata[19]. 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, such as the stratification of samples based on specific characteristics. When conducting analyses for new research questions, the specificity of sample-level metadata

enables researchers to derive more targeted insights and draw meaningful conclusions from the existing dataset.

In this study, we analyzed a total of 253 randomly selected scientific publications encompassing 164,909 samples across various disease conditions, phenotypes and organisms. We investigate the prevailing practice of metadata sharing for the six metadata information, including organism, sex, age, tissue types, strain type and race/ethicity/ancestry information, in the original publication and the corresponding public repositories. Our study found that many studies do not share all of the important information about their samples, with an average of only 74.8% of phenotypes being reported. Studies involving non-human samples were more likely to have complete metadata than studies involving human samples. Additionally, public repositories contained more metadata information than the original publications that reported the data. To generalize our results, we further examined 61,950 studies across 2,168,620 samples from Gene Expression Omnibus[20] (GEO) repository. The overall metadata availability in the GEO repository among the 61,950 studies is 63.2%. Additionally, non human studies has 16.1% more complete metadata availability than human studies in the GEO repository. Metadata availability was limited in studies published before 2010, but it notably improved in studies published after 2019. Our findings revealed the incomplete metadata reporting practices in both original publications and public repositories, providing valuable insights into the current state of metadata availability in the GEO repository. Furthermore, these findings reveal that the text of the publication does not contain much more metadata than the repository; instead many phenotypes are not shared in both sources. Our study underscores the importance of metadata sharing and standardization in promoting the reusability and repurposing 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, advance our understanding of complex diseases and their underlying mechanisms, and ultimately discover novel biomedical insights to improve human health[21].

Results

Assessment of completeness of public metadata accompanying omics studies

We have performed a comprehensive analysis of the completeness of the metadata reported in both the original scientific publications and the public repositories. We meticulously analyzed a total of 253 randomly selected scientific publications encompassing 164,909 samples across various disease conditions, phenotypes and organisms (D1 dataset). This dataset comprises 20,047 human samples from 153 studies and 144,862 non-human samples from 100 studies. 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). (Supplementary table 1). Out of the 253 studies, 153 involve human samples, while the remaining 100 involve non-human samples. Among the 153 studies, the majority (93.6%) were *Mus musculus* samples, followed by *Parus major* (2.2%) and *Glycine max* (1.7%), respectively (Figure S1). We examined per-sample metadata reported in human and non-human samples. We examined the four common phenotypes shared by both human and non-human samples: organism, sex, age, and tissue types. In addition to these four common phenotypes, we examined the human-specific phenotype of race/ethnicity/ancestry and the non-human-specific phenotype of strain information. To assess metadata availability, we manually reviewed the phenotypic information within the original publications, including both the main text and supplementary materials of such studies. To collect metadata from the public repositories for the 253 studies, we utilized our pre-existing Python scripts to retrieve metadata information from the GEO repository.

To increase the generalizability of our analysis, we randomly selecting 61,950 studies from the Gene Expression Omnibus[22] (GEO) repository, encompassing a total of 2,168,620 samples (D2 dataset). We conducted an examination of metadata accompanying raw omics data availability for the six phenotypes, tissue types, organism, sex, age, strain information (for non-human samples), and race/ethnicity/ancestry (for human samples) in a larger scale of analysis using a custom script. Among the studies, *Mus musculus* constitutes the majority at 86% of the samples, followed by *Rattus norvegicus* at 4.0%, and *Homo sapiens* at 2.3% **(Figure S4)**.

Limited availability for essential phenotypes accompanying 253 studies

We examined the availability of essential phenotypes for the 253 studies in the D1 dataset, which includes 164,909 samples. The overall metadata availability for the 253 studies in the original publications or public repositories is 74.8% (Figure 2). Among the six phenotypes, organism is the mostly shared phenotype shared by all the studies (Figure 2). Approximately 80% of the studies have available tissue information, and similarly, 80% of the non-human samples have available strain information(Figure 2). Up to 60% of the studies have available sex and age information(Figure 2). The race/ethnicity/ancestry information is the least frequently reported phenotype with up to 20% of the human samples having available race/ethnicity/ancestry information(Figure 2). Next, we assessed trends in metadata availability in studies published between 2004 and 2022. Our analysis revealed a notable uptick in metadata reporting practices for publications released after 2018, with a 47.1% increase in metadata in 2018 alone. In contrast, publications from before 2017 exhibited considerably lower metadata availability, publications before 2017 had much lower metadata availability, with only 10.3% of publications reporting metadata (Figure 3).

Next, we examined whether metadata sharing was consistent within samples from the same study across the text of the publication and public repository. The degree of consistency was higher in the public repositories (62.6%) than in the original publications (51.7%) (Figure 7). Notably, 51.8% of the studies had less than 50% inconsistency in sharing metadata in the original publication, while 33.6% had less than 50% inconsistency in the public repositories. Of the six phenotypes, organism and disease information were the most consistently reported metadata, while race/ethnicity/ancestry information was the least consistently reported metadata. Age and sex information were inconsistently reported, with only half of the studies reporting this information in both sources (Figure 7). Strain information was more consistently reported in the public repositories than in the original publications (Figure 7).

Inconsistent practises of metadata sharing across original publications and corresponding omics repositories

We compared the six phenotypes for the 253 studies reported in both the original publications and the corresponding public repositories. The metadata availability in original publications (58.5%) was lower than the metadata availability in public repositories (62.0%) **(Figure 4)**. The overall metadata availability is higher than separate metadata availability in the original publication and public repositories because some missing metadata in the public repositories is complemented by the metadata presented in original publications for the same samples. Public repositories contain 3.5% more phenotypic information than original publications. In comparison to original publications, public repositories share more complete phenotypes for organism, sex, age, strain, and race/ethnicity/ancestry, while original publications share more complete tissue information (Figure 6).

Organism information is the most completely and consistently shared phenotype, which all such information is available across all the 164,909 samples (Figure 5). The order of consistency in sharing phenotypes is as follows: sex, race/ethnicity/ancestry, strain, and tissue information (Figure 6). In contrast, age is the least consistently shared phenotype between original publications and public repositories, with 17.9% more samples having age information in public repositories (Figure 6). Notably, organism information is fully available in the public repositories (Figure 5 and Figure 6). Tissue and strain information have similar availabilities, with 77.2% of samples having tissue information and 76.5% of non-human samples having strain information. Sex and age information exhibit a similar trend in such phenotype availability, both around 60%. Race/ethnicity/ancestry information is the least available phenotype, with only 22% of human samples reporting such information (Figure 5).

Non-human studies shared more complete metadata compared to human ones

The metadata availability in both original publications and public repositories for human samples is 56.1% and for non-human samples is 60.86% (Figure 8). Human and non-human samples had more complete metadata reported in public repositories (66.6% and 62.0%,

respectively) than in original publications (54.6% and 59.7%, respectively) (Figure 9). We analyzed metadata availability in original 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 original publications and public repositories. Specifically, they reported more comprehensive metadata in the original publications than in public repositories. (Figure S2)

We next separately investigate the availability of the five phenotypes, including organism, tissue, sex, age, and race/ethnicity/ancestry, across the 153 human studies and the availability of the five phenotypes, including organism, sex, age, tissue, and strain information, across 100 non-human studies. For the human dataset, organism and tissue information is widely available, with all the samples having available organism and tissue information in either public repositories or original publications **(Figure 10)**. However, less than half of the non-human samples contain available information for sex (49.4%) and age (48.9%), while only 22% of the samples include the race/ethnicity/ancestry information. **(Figure 10).** Organism is the only phenotype that is widely available for the non-human samples in the public repositories **(Figure 10)**. The remaining four phenotypes of the non-human samples show a consistent level of metadata availability in public repositories, each exhibiting over 50% metadata completeness **(Figure 10).** Strain information stands out as the second most comprehensively reported metadata, with 76.5% of non-human samples having available strain information. Tissue, age, and sex information all exhibit a consistent trend, with tissue information showing a metadata availability of 74%, while age and sex information have availability rates of 64.6% and 66.3%, respectively, across the non-human samples **(Figure 10).** When comparing the shared phenotypes between human and non-human samples, it's noteworthy that the “organism” metadata is completely shared among both groups in the public repositories. In contrast, human samples exhibit a higher level of completeness in sharing tissue information compared

to non-human samples. Conversely, age and sex information is more comprehensively shared among non-human samples than among human samples **(Figure 10).**

We observed that both human and non-human studies shared more complete metadata in public repositories than in original publications. Of the 153 human studies, 97 had consistent metadata sharing practices between original publications and public repositories, 36 had lower metadata availability in original publications, and 20 had higher metadata availability in original publications (Figure 11). Of the 100 non-human studies, 59 had more complete metadata in public repositories than in original publications, 28 had the same extent of metadata availability in both sources, and 13 had more complete metadata in original publications than in public repositories (Figure 11).

Sharing summerized metadata of the study is still a common practice

Experiment-level metadata availability is defined as only sharing the summary of the samples' phenotypes without sharing detailed per-sample phenotypic information. For human studies in original publications, age information was commonly reported in a summarized format (60% of samples), which we categorized as having only experiment-level metadata(Figure 12). Up to 20% of the human samples only shared experiment-level metadata for disease, sex, and race/ethnicity/ancestry information (Figure 12). Similarly, non-human studies in original publications frequently report sex information in a summarized format (40%) (Figure 12). 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 13).

Limited metadata availability for 153 human studies across seven disease condition

Next, we examined metadata sharing practices for seven disease conditions in human studies. Similar trends of availability of metadata reported can be seen across the studies in all seven therapeutics fields (Figure S3). All seven diseases have fully reported metadata for organism and tissue types (Figure S3). Notably, sepsis studies had 99.3% availability for disease

condition metadata, compared to 100% for other disease studies (Figure S3). Metadata availability varied across the seven diseases, particularly for sex, age, and race/ethnicity/ancestry. CVD studies had the highest sex metadata availability (75.9%), while MS studies had the lowest (10.8%). Sepsis studies had the highest age metadata availability (67.9%), while MS studies had the lowest (9.7%). Race/ethnicity/ancestry metadata availability was notably lower, with none of the samples in the IBD and MS studies reporting this information (Figure S3).

Limited availability of essential phenotypes shared at public repositories across 61,950 studies

To generalize the reported results we examined from the D1 dataset, we investigated the availability of the six phenotypes, organism, tissue types, sex, age, strain (for non-human samples) and race/ethnicity/ancestry information (for human samples), reported in the GEO repository among 61,950 studies across 2,168,620 samples (D2 dataset). The overall availability of metadata among the 61,950 studies is 63.2% (**Figure 14**). Notably, upon analyzing the cumulative metadata availability across various years, we discovered that studies preceding 2010 exhibited a lack of shared metadata within the GEO repository (**Figure 15**). We found that human studies had lower average metadata availability (47.4%) than non-human studies (63.5%) (**Figure 14**).

In the case of human samples, the organism information is consistently shared across all samples, achieving a 100% availability rate. Additionally, tissue information is almost universally provided, with 99.9% of samples containing available tissue or source details (**Figure 16**). However, the availability of sex and age information is relatively limited, with only 19.0% of human samples disclosing age and 13.8% indicating sex(**Figure 16**). Furthermore, race information is the least reported among human samples, with only 4.2% of the samples including ancestral data; In the context of non-human samples, both organism and tissue information exhibit complete availability, with all the non-human samples shared available organism and tissue information (**Figure 16**). Notably, strain information is reported for 56.7%

of non-human samples, while age information is shared by 37.4% of the samples (**Figure 16**). Among these phenotypic characteristics, sex information stands out as the least consistently reported metadata for non-human samples, with only 24.15% of samples providing such details (**Figure 16**). In summary, the availability of metadata in non-human samples surpasses that in human samples (**Figure 14 and Figure 15**). Tissue, sex, and age metadata are more comprehensively reported in non-human samples compared to their human counterparts (**Figure 16**).

Discussion

Our study is the first to systematically analyze metadata completeness across multiple organisms and phenotypes, revealing limited availability of metadata both in original publications and public repositories. We found substantial disparities in metadata completeness across phenotypes, with organism and tissue types almost universally reported, but age, sex, strain, and race/ethnicity/ancestry metadata often lacking. Similar inconsistencies in reporting diverse phenotypes across various diseases suggest that authors had varying levels of consideration and interest towards the inclusion of such information in the research (Figure S3). In the original publications, a majority of metadata is only available at the experiment-level. Additionally, our study is the first to investigate metadata availability at both the experiment and sample levels. We found that half of the human samples with available metadata for sex, age, and race/ethnicity/ancestry have only experiment-level metadata, which limits reproducibility in original publications and may not be as valuable for secondary analysis as sample-level metadata. Note that all studies have shared some part of information about the samples' metadata, indicating that the authors have some awareness toward sharing the samples' metadata information. In general, there is insufficient metadata availability for the sex, age, strain and race/ethnicity/ancestry information in the GEO repositories in a larger scale.

identify that non-human samples contains more complete metadata than human samples. 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. Non-human studies tend to have a more standardized and comprehensive data collection process due to their specific research objectives. In contrast, human studies often encompass a broader range of research questions, resulting in more variable data collection practices. Ethical considerations also come into play. Human studies are typically subject to stricter ethical guidelines and consent requirements[23–25]. These regulations may limit the extent of metadata that can be collected, whereas non-human studies may not encounter the same ethical constraints[23,25–27]. Moreover, data accessibility varies between these two categories. Non-human studies may enjoy easier access to metadata due to less stringent privacy and consent regulations, facilitating more extensive data collection efforts[28–30]. Data sharing norms 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[31].

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 original publications or from the request to the authors since the both processes can be time-consuming and error-prone[11,32–34]. 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[10,35,36]. Improving the availability of metadata in public repositories could provide valuable and accurate information for downstream analyses, and further enhance the usefulness of the GEO database. Additionally, the standardization of metadata can ensure that the data can be accurately compared and integrated with other datasets, increasing rigor and robustness of data driven biomedical research for new insights and discoveries.

Challenges in metadata sharing practices include ensuring metadata quality and completeness[19], a lack of uniform reporting standards[9], potential privacy concerns limiting data disclosure[28], and study design limitations impacting metadata collection. For instance, researchers face issues with inconsistent reporting of phenotypic information, such as tissue types, in public repositories. Legal and ethical concerns, for example the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union, may deter researchers from sharing certain phenotypic data, affecting data trustworthiness and scientific progress[30,37,38]. Additionally, study designs may not include plans for collecting complete metadata, further affecting metadata availability. Addressing these challenges is crucial for enhancing metadata quality and accessibility in scientific research. To address the issue of incomplete metadata, the scientific community should develop standardized metadata reporting guidelines and prioritize responsible data sharing. Establishing proper metadata sharing policies, similar to those for raw omics data, will standardize reported metadata for easier secondary analysis. For example, the FAIR (Findable, Accessible, Interoperable, and Reusable) guidelines emphasize the importance of making data and its accompanying metadata easily accessible and complete to the scientific community[39]. This requires comprehensive reporting of metadata alongside raw data to ensure that the data is easily findable and accessible ensuring the FAIRification of the raw data and metadata[40–42]. Additionally, increasing biomedical researchers' awareness of the value of rich phenotypic data and including metadata plans in Institutional Review Board (IRB) applications can enhance metadata availability and reduce regulatory barriers. Well-designed metadata

sharing policies from repositories and journals can raise authors' awareness of collecting complete sample metadata, improving metadata quality. Complete and standardized metadata shared on public repositories are crucial for data availability, reproducibility, and reliability.

There are several limitations in our study. First, we were able to only extract commonly reported metadata 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, accuracy, or consistency of metadata reported.

Methods

Examine the metadata availability for the D1 dataset

We assessed the per-sample metadata availability within the D1 dataset, spanning 253 multi-omics studies and comprising a total of 164,909 samples. First, we identified a total of 153 human studies pertaining to seven different diseases, including Alzheimer's disease (AD), acute myeloid leukemia (AML), cardiovascular disease (CVD), inflammatory bowel disease (IBD), multiple sclerosis (MS), sepsis, and tuberculosis (TB), in PubMed and randomly selected 100 non-human studies in the Gene Expression Omnibus (GEO) repository, and identified the corresponding publications through the available GEO study accession number. .

Extraction of metadata for the from the text of the original publications

For the human samples, we scrutinized the availability of six per-sample metadata categories, which include disease condition, organism, sex, age, tissue types, and race/ethnicity/ancestry information. Meanwhile, for the non-human samples, we assessed the availability of five per-sample metadata categories, including disease condition, organism, sex, age, tissue types, and strain information. We assessed metadata availability in the original publications through manual examination of the phenotypic information in the text of the publication or in the supplementary materials in such studies. Based on the origin of the metadata from either the text of publications or in the public repositories, we categorized the metadata availability into two categories, sample-level metadata available and experiment-level metadata available. 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 existing Python scripts. We have validated the script and can provide the validation results obtained from running it on the 253 studies. 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.

Examine the metadata availability in the original publications and the public repositories of the D1-dataset

Following the collection of metadata availability data from both the original publications and public repositories within the D1-dataset, we harnessed Python libraries, namely pandas, seaborn, and matplotlib, for data manipulation and visualization purposes. Utilizing these tools, we computed the metadata availability across both the original publications and public repositories. Subsequently, we conducted a thorough comparison between the metadata reported in publications and the corresponding metadata available in the repositories, aiming to evaluate the consistency of metadata reporting practices between original publications and public repositories. Our analysis then extended to the examination of the availability of metadata in the original publications and public repositories for human and non-human samples.

Large scale extraction of phenotypes from the GEO repositories

By utilizing and expanding upon the pre-existing Python script, we expanded the analysis to a larger scale of studies, including 61,950 studies across 2,168,620, in the GEO repository. The outcomes of this comprehensive analysis will be referred to as D2-dataset. We conducted the examination of metadata available across six metadata, including organism, tissue types, sex, age, strain information (for non-human samples) and race/ethnicity/ancestry information (for human samples). We also excluded the disease condition phenotype, as our code-based solutions could not determine if the samples collected for the study were from individuals with a particular disease condition. We validated our code-based approach by comparing it to the gold standard data, which was obtained through manual curation of the data. We leveraged Python libraries, namely pandas, seaborn, and matplotlib, for data manipulation and visualization purposes and computed the above six metadata availability in the GEO repository.

Contributions

SM supervised the work. YH analyzes the collected data. 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. All co-authors contributed to the revision of the manuscript.

Acknowledgements

We acknowledge all authors sharing the information with us. SM and YH 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.

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. Qin, J. & D'ignazio, J. The Central Role of Metadata in a Science Data Literacy Course. *J. Libr. Metadata* **10**, 188–204 (2010).
5. 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).
6. Musen, M. A. Without appropriate metadata, data-sharing mandates are pointless. *Nature* **609**, 222–222 (2022).
7. Brito, J. J. *et al.* Recommendations to enhance rigor and reproducibility in biomedical research. *GigaScience* **9**, giaa056 (2020).
8. 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).
9. Schriml, L. M. *et al.* COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci. Data* **7**, 188 (2020).
10. Kim, Y.-M., Poline, J.-B. & Dumas, G. Experimenting with reproducibility: a case study of

robustness in bioinformatics. *GigaScience* **7**, giy077 (2018).

11. Gozashti, L. & Corbett-Detig, R. Shortcomings of SARS-CoV-2 genomic metadata. *BMC Res. Notes* **14**, 189 (2021).

12. 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).

13. 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).

14. Toczydlowski, R. H. *et al.* Poor data stewardship will hinder global genetic diversity surveillance. *Proc. Natl. Acad. Sci.* **118**, (2021).

15. 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).

16. Rung, J. & Brazma, A. Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.* **14**, 89–99 (2013).

17. Musen, M. A. *et al.* The center for expanded data annotation and retrieval. *J. Am. Med. Inform. Assoc. JAMIA* **22**, 1148–1152 (2015).

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

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

20. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* **41**, D991-995 (2013).

21. 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).

22. 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.

23. 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).

24. 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).

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

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

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

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

29. 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).

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

31. 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).

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

33. Stodden, V., Seiler, J. & Ma, Z. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl. Acad. Sci.* **115**, 2584–2589 (2018).

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

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

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

37. Heys, M. & Smyth, R. L. General data protection regulation: What does this mean for research? *Arch. Dis. Child. Educ. Pract. Ed.* **105**, 296–297 (2020).

38. Vlahou, A. *et al.* Data Sharing Under the General Data Protection Regulation: Time to Harmonize Law and Research Ethics? *Hypertens. Dallas Tex 1979* **77**, 1029–1035 (2021).

39. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

40. Scheffler, M. *et al.* FAIR data enabling new horizons for materials research. *Nature* **604**, 635–642 (2022).

41. Alharbi, E. *et al.* Selection of data sets for FAIRification in drug discovery and development: Which, why, and how? *Drug Discov. Today* **27**, 2080–2085 (2022).

42. Harrow, I., Balakrishnan, R., Küçük McGinty, H., Plasterer, T. & Romacker, M. Maximizing data value for biopharma through FAIR and quality implementation: FAIR plus Q. *Drug Discov. Today* **27**, 1441–1447 (2022).

Figures

Figure 1. The metadata availability reported in the D1-dataset from the original publications and public repositories. (Red line: the mean metadata availability among the samples in the D1-dataset.)

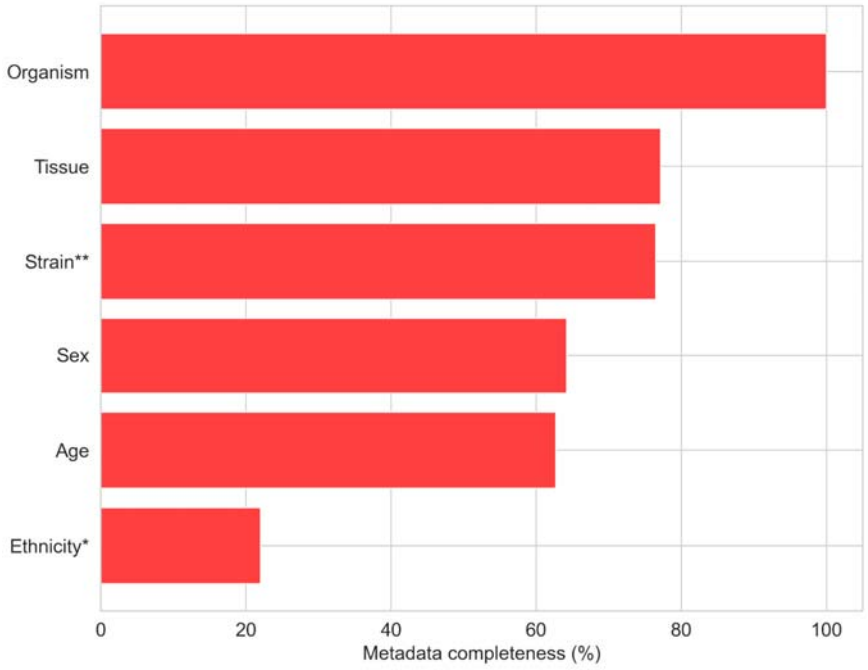

Figure 2. The availability of the six phenotypes information in the original publications and public repositories in the D1-dataset. (The ethnicity(*) (race/ethnicity/ancestry) information is

analyzed over the human samples only; The strain(**) information is analyzed over the non-human samples only.)

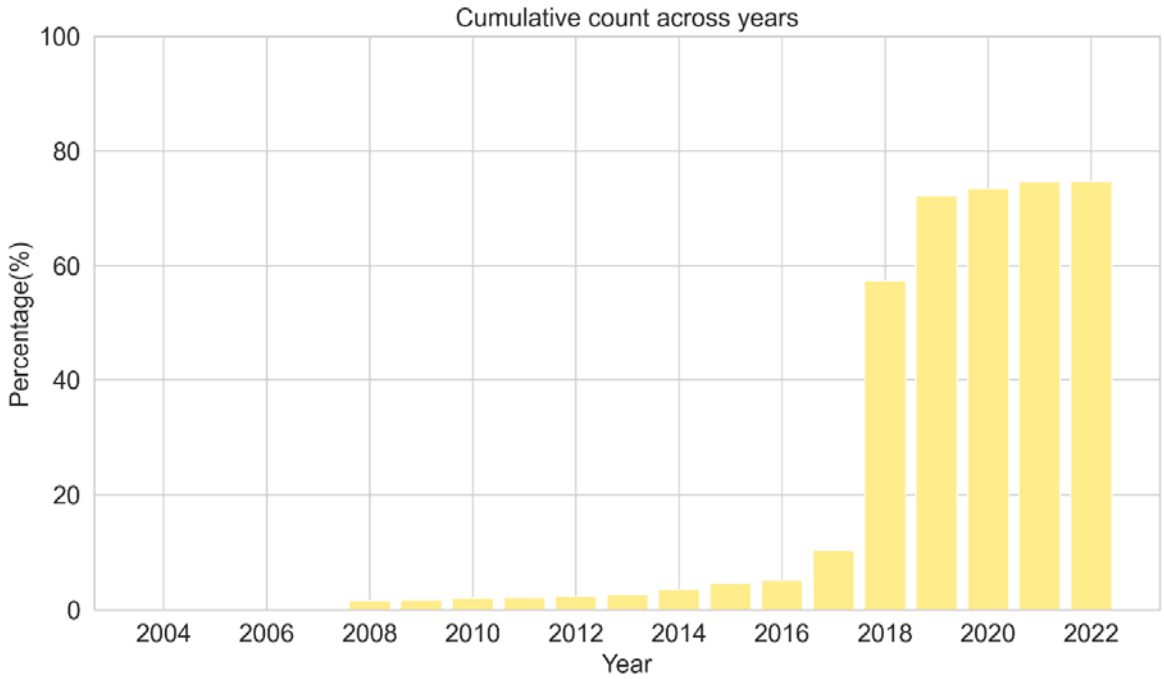

Figure 3. The cumulative metadata availability of the D1-dataset over the years.

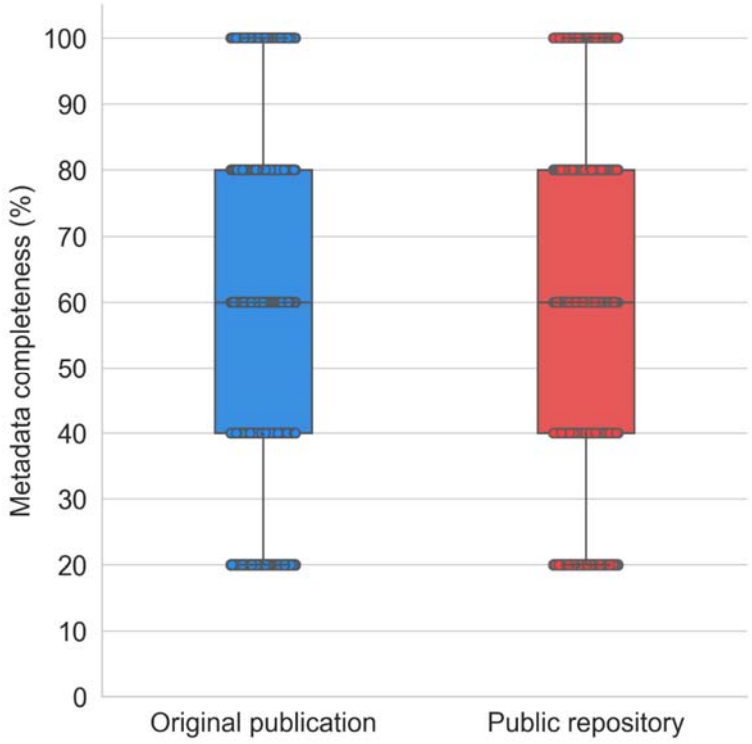

Figure 4. The overall metadata availability in the original publications and public repositories in the D1-dataset.

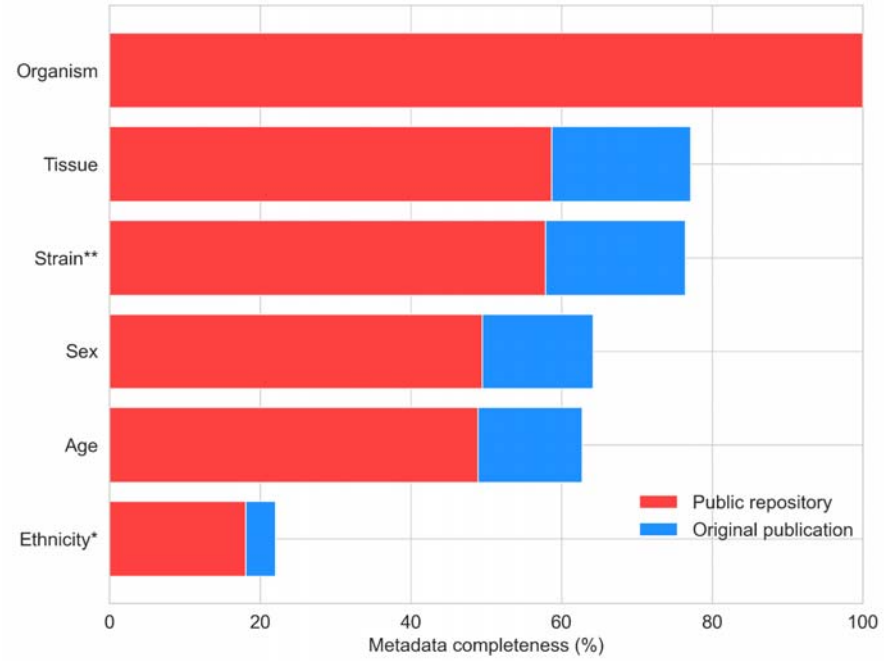

Figure 5. The overall composition of availability of the metadata accompanying raw data across seven phenotypes in the human and non-human studies in the D1-dataset. (The ethnicity(*) (race/ethnicity/ancestry) information is analyzed over the human samples only; The strain(**) information is analyzed over the non-human samples only.) (Public repository: The sample-level metadata reported in the public repository; Original publication: The original publications provided additional sample-level metadata that was not reported in the public repositories where the samples were deposited.)

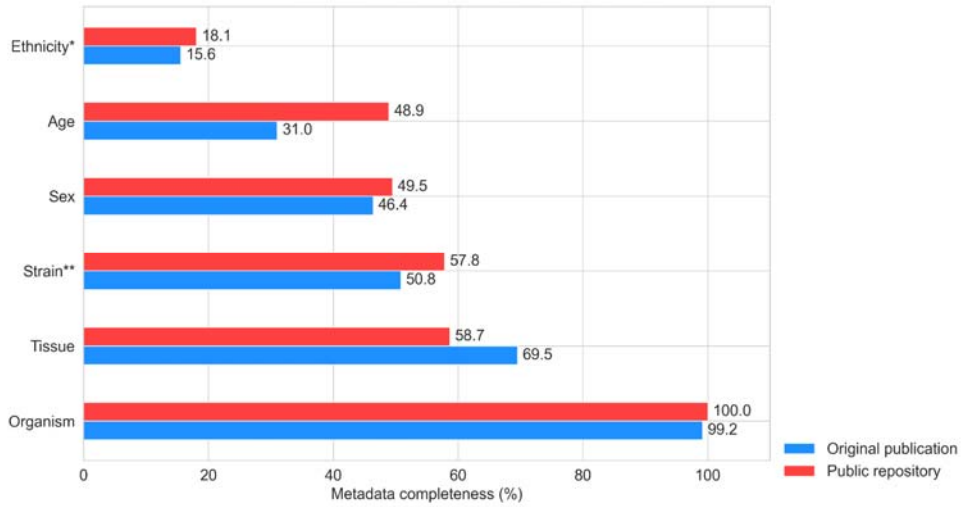

Figure 6. The sample-level metadata across seven phenotypes among all the samples in the D1-dataset reported in the original publications and public repositories. (The disease(*) and ethnicity(*) (race/ethnicity/ancestry) information is analyzed over the human samples only; The strain(**) information is analyzed over the non-human samples only.)

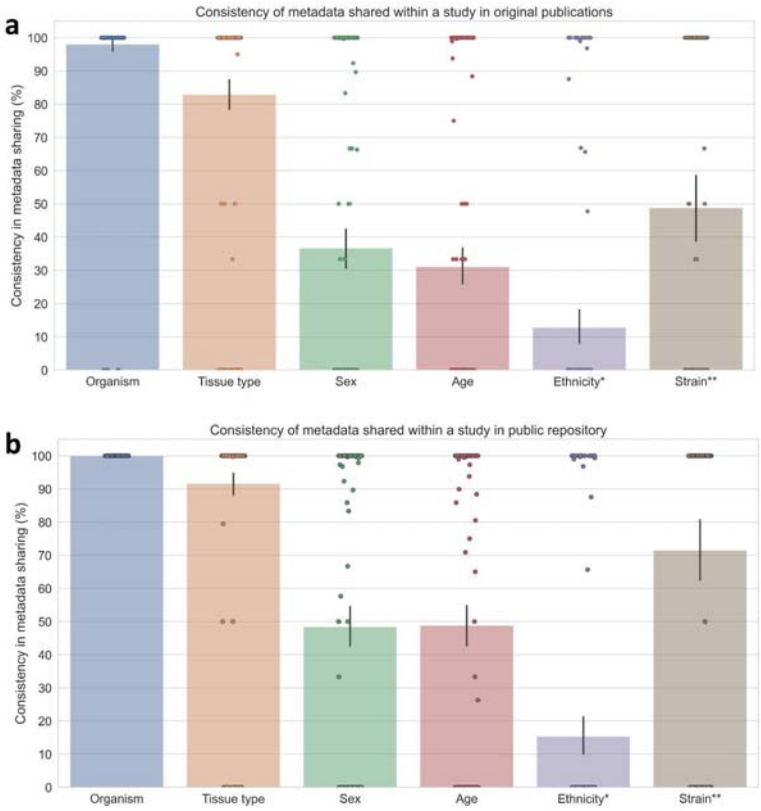

Figure 7. The degree of consistency with which the six phenotypes were shared across the 253 studies in the D1-dataset in both the original publications and public repositories.

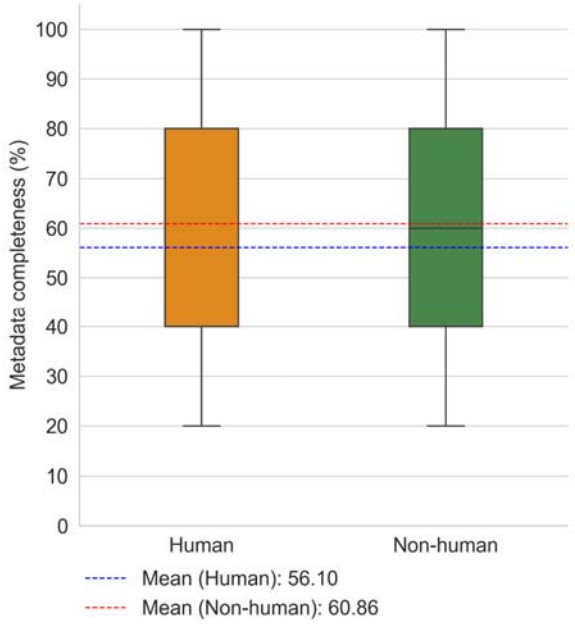

Figure 8. The overall metadata availability for human and non-human samples in the D1-dataset. (Red line displayed the mean for non-human samples; blue line displayed the mean for human samples.)

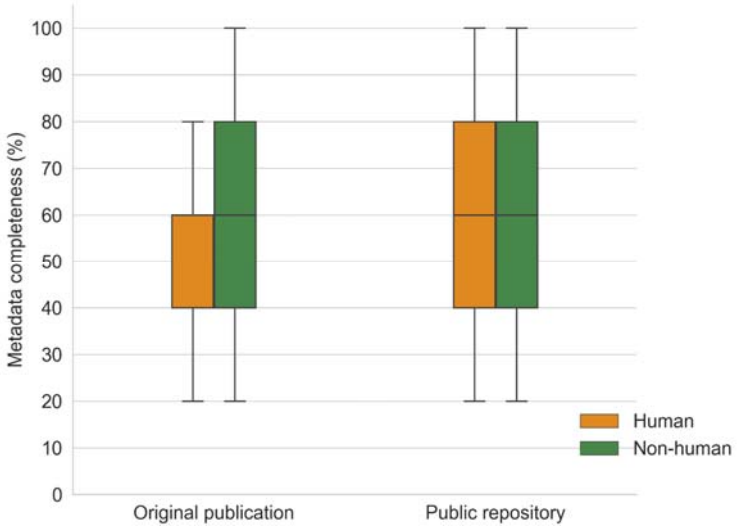

Figure 9. The human and non-human metadata availability in the original publications and public repositories for the D1-dataset.

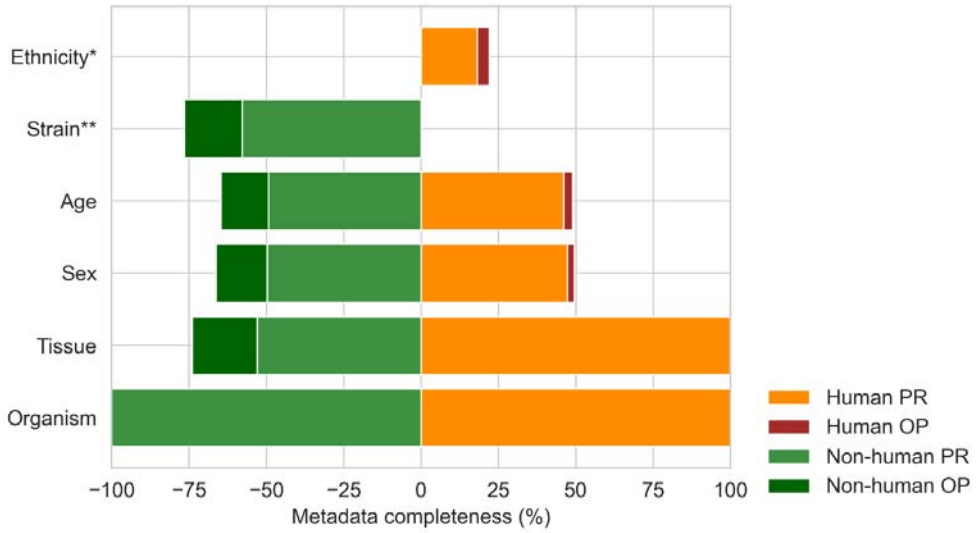

Figure 10. Right: The overall composition of sample-level metadata availability across six phenotypes in the original publications and public repositories across the human samples in the D1-dataset; Left: The overall composition of sample-level metadata availability across five

phenotypes in the original publications and public repositories across the non-human samples in the D1-dataset.

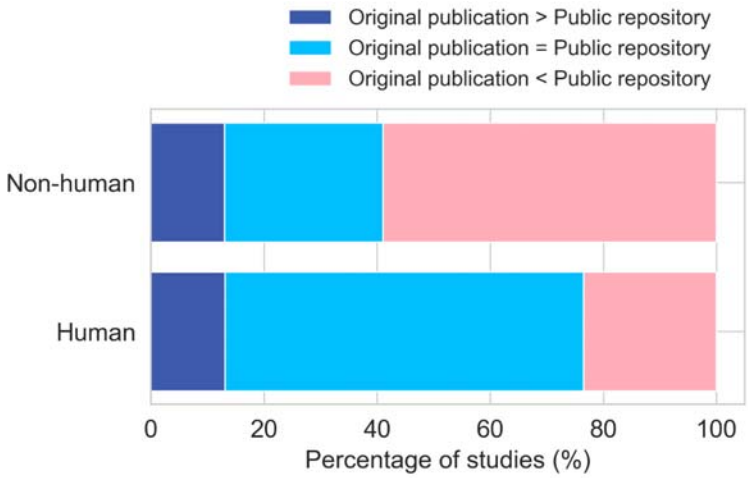

Figure 11. The percentage of non-human and human studies in the D1-dataset that reported the sample-level metadata in the original publications the same as, exceeds, or falls below the sample-level metadata reported in the public repository.

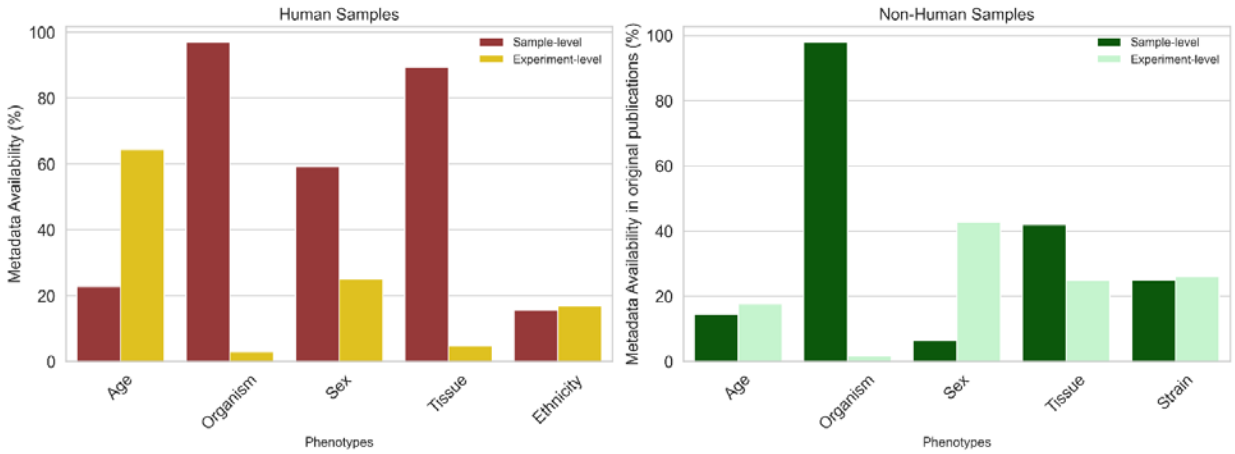

Figure 12. The comparison of the availability of sample-level and experiment-level metadata between human samples (six phenotypes) and non-human samples (five phenotypes) in the D1-dataset.

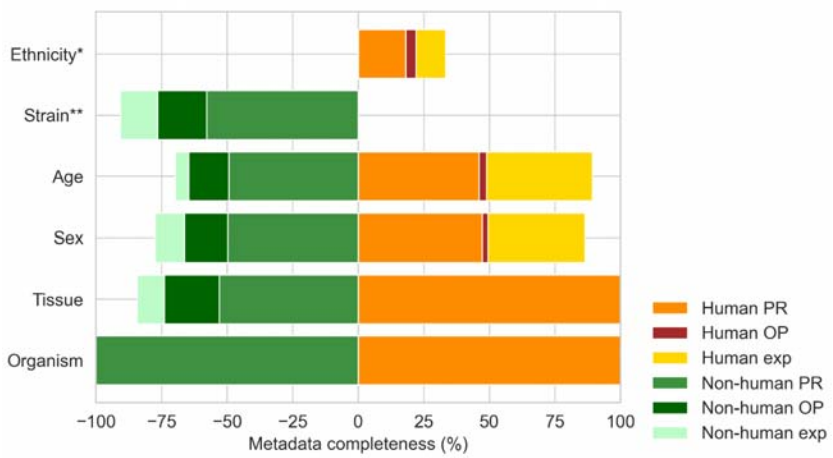

Figure 13. Right: The overall composition of sample-level and experiment-level metadata availability across six phenotypes in the original publications and public repositories across the human samples in the D1-dataset; Left: The overall composition of sample-level and experiment-level metadata availability across five phenotypes in the original publications and public repositories across the non-human samples in the D1-dataset.

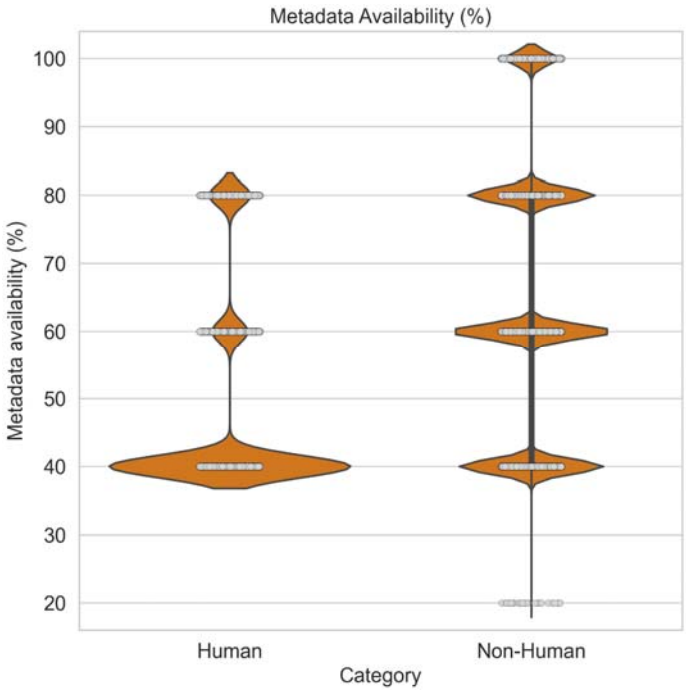

Figure 14. The comparison of the human and non-human metadata availability in the D2-dataset.

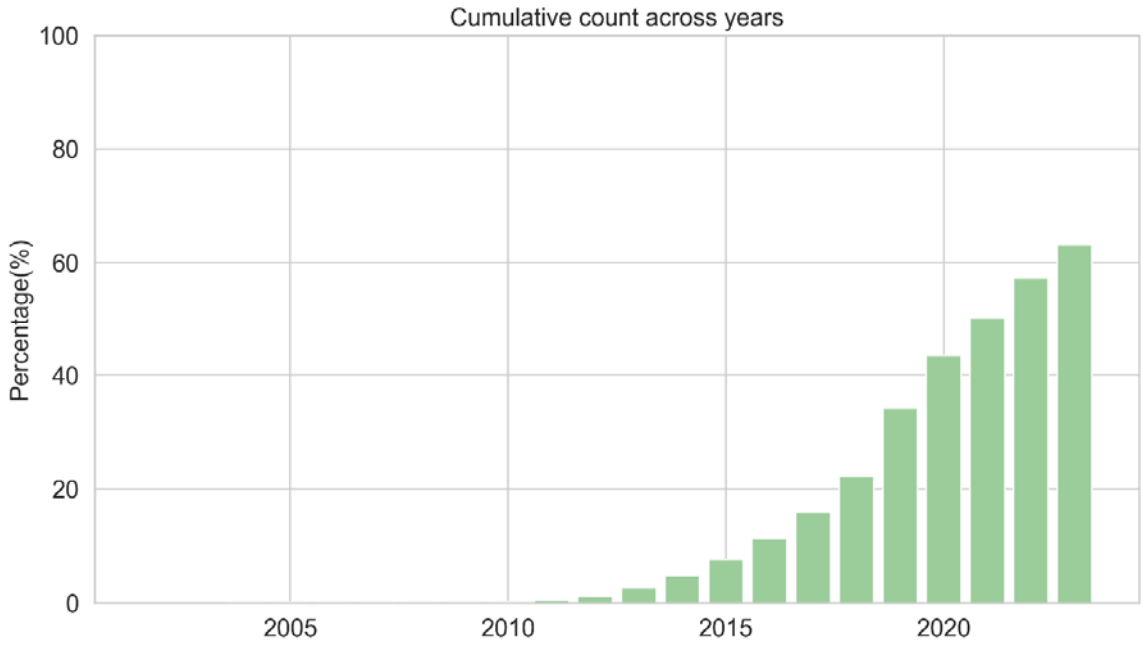

Figure 15. The cumulative metadata availability of the D2-dataset over the years.

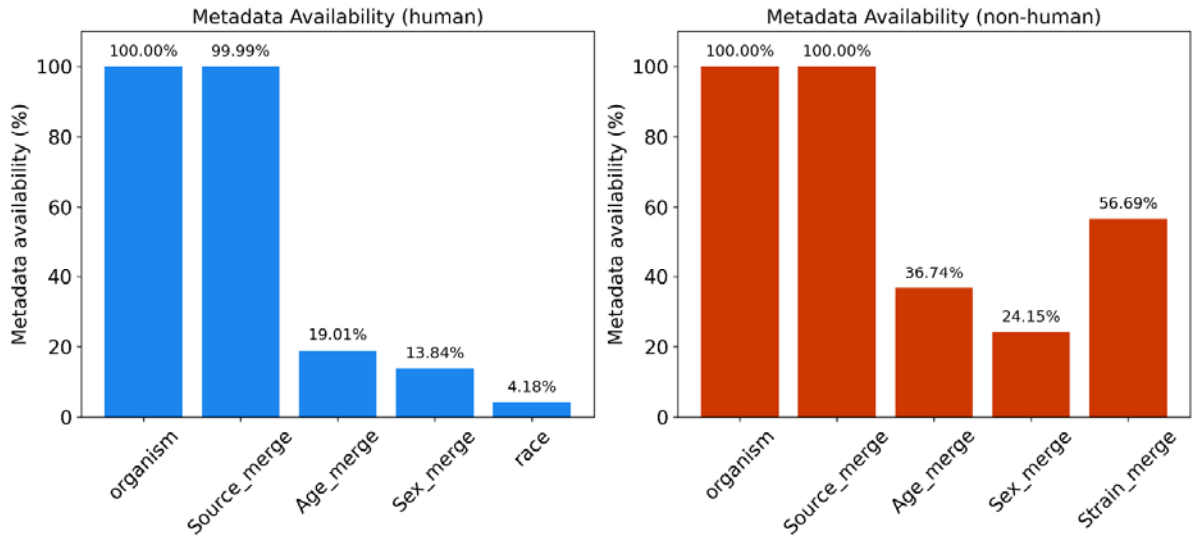

Figure 16. The separate metadata availability of human and non-human samples in the D2-dataset.

Supplementary Materials

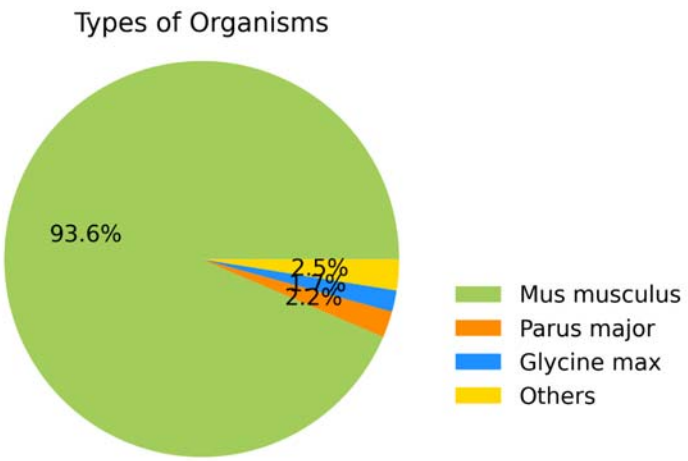

Figure S1. The proportion of non-human samples across 100 studies encompassing 144,862 samples in the D1 dataset.

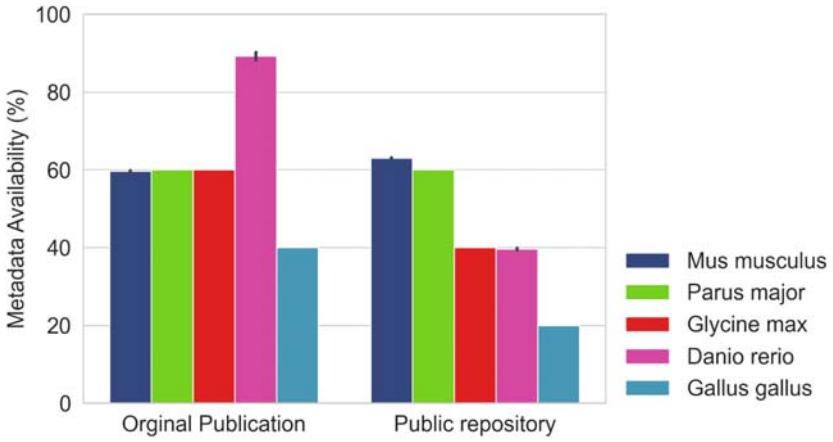

Figure S2. The metadata availability between the original publications and public repositories among the top five non-human organisms represented in the D1 dataset

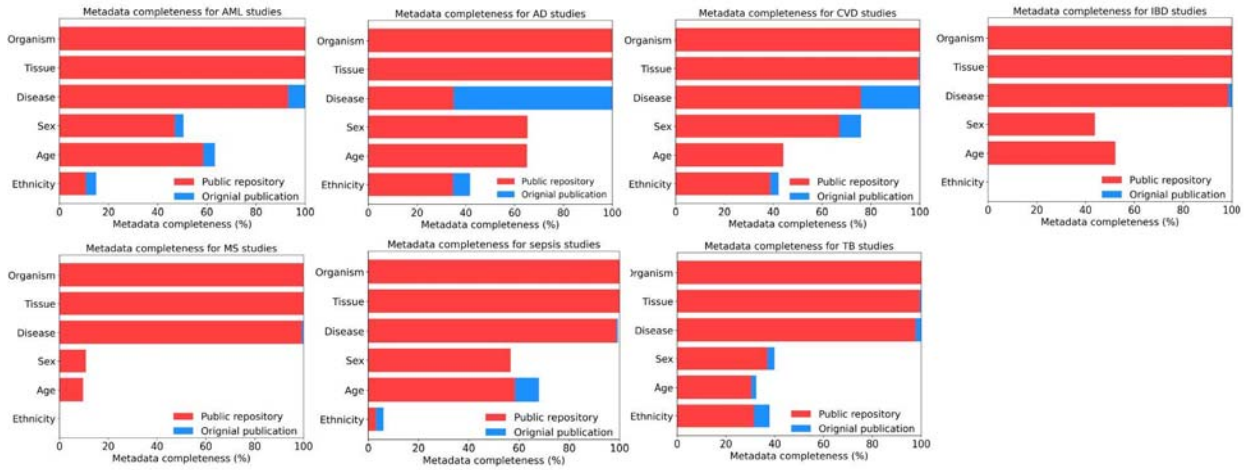

Figure S3. The overall composition of sample-level metadata availability across six phenotypes for the human studies encompassing seven diseases, including Alzheimer's disease (AD), acute myeloid leukemia (AML), cardiovascular disease (CVD), inflammatory bowel disease (IBD), multiple sclerosis (MS), sepsis, and tuberculosis (TB), in the D1-dataset.

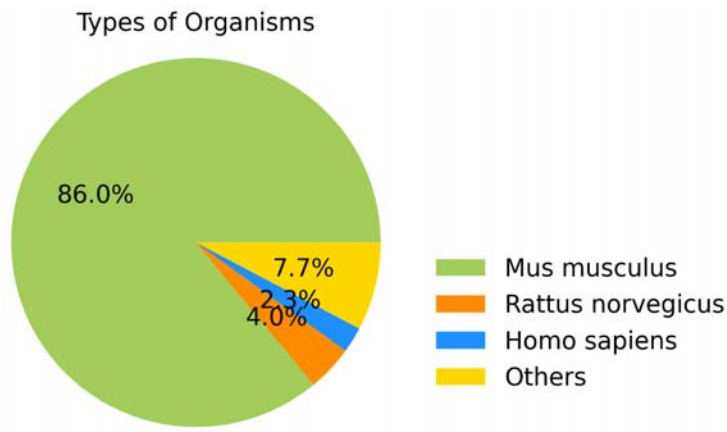

Figure S4. The proportion of organism among the 2,168,620 samples of the 61,950 studies from the D2 dataset.