# A COMPREHENSIVE ASSESSMENT OF SEQUENCE READ ARCHIVE METADATA COMPLETENESS

**Albert BAS*, Viorel MUNTEANU***

*Department of Software Engineering and Informatics, Technical University of Moldova,*
*Chisinau, Republic of Moldova*

* Corresponding authors: Albert Bas, albert.bas@iis.utm.md , Viorel Munteanu, viorel.munteanu@lt.utm.md

***Abstract.*** *Recent advances in high-throughput sequencing technologies have enabled the collection and sharing of a vast amount of omics data, along with its associated metadata. Enhancing the availability of this metadata is crucial to ensure the reusability and reproducibility of raw data, as well as for facilitating novel biomedical discoveries through efficient data reuse. In this study, we performed a comprehensive assessment of metadata completeness by analyzing over 26,000,000 experiments shared in the Sequence Read Archive (SRA) from 2008 to 2023. Our results show that the countries of Central Europe, the USA and China show dominance in generating sequencing data, corresponding to 45%, 16% and correspondingly 8% of total data in the SRA repository, the most frequently used platform is ILLUMINA (90%). Identified that some of the metadata contains inconsistencies in completeness: the absence of temporary identifiers (5.2%), the lack of assigned TaxonomyID (5%), and the absence of library strategy (8%). Our results highlight the urgent need for improved metadata sharing practices and the standardization of reporting.*

***Key words:*** *Metadata, data reusability, Sequence Read Archive, sequencing*

## Introduction

The quantity and diversity of genomic data continues to grow exponentially, it is essential that these data are discoverable, accessible, interoperable and reusable [1]. Metadata is information about data that describes its content, structure, and other characteristics. Metadata plays a crucial role in understanding accompanied data, as it provides essential information necessary to reproduce the data accurately [2,3,4]. Since the completion of the human genome and the creation of the second commercial next generation sequencing platform, DNA sequencing databases have been actively growing [5,6] and outpacing Moore's Law [7]. Recent estimates indicate that the global market for microbiome sequencing will continue growing. Unfortunately, most public databases rely on user input and do not have methods for identifying errors in the provided metadata, leading to the potential for error propagation and exacerbating issues of incompleteness or lack of standardization [8]. Poor metadata can significantly lower the value of sequencing experiments by limiting the reproducibility of the study and its reuse in meta analyses [9].

In order to delve into the nuances of data types and their sources, our objective is to procure the existing metadata from the SRA repository. This endeavor aims to evaluate both the comprehensiveness and quality of the data, thereby fostering a deeper comprehension of the research conducted. Such an assessment not only facilitates researchers in reusing findings and data in subsequent projects but also contributes to the broader scholarly discourse. Our study based on SRA data will elucidate the technological, geographical, temporal, and methodological intricacies of sequencing with indication of the completeness of the metadata.

**Materials and Methods**

We used the SRA Toolkit to extract information on 25.2 million experiments. Additional data on study descriptions, protocols, and design obtained by parsing information from the NCBI web resource with python scripts. Searching NCBI web resources also identified 1,467,839 new records, highlighting differences between the online version and the database accessible through the SRA Toolkit. This demonstrated a 94% correspondence between the sources. Information on the countries in which sequencing was performed was obtained from the domains of sites that Bing returned for queries of the organization's name, specified in metadata. The data were also subjected to country categorization using ChatGPT 3.5. The country was determined for 39916 (91%) of 43543 organizations covering 25909854 sequencing runs (92%) of 28124436. Manual evaluation on 100 samples showed that the accuracy of the country determination method is 90%.

**Results**

The metadata obtained from SRA by November 15, 2023, include describe about 28,124,436 sequencing runs conducted within 26,767,311 experiments on 23,692,425 unique samples from 184,134 organisms during the period from 2008 to 2023 had been obtained, utilizing 80 different sequencing models across 11 platforms. The experiments were conducted by 43540 research centers.

Between 2008 and 2012, fewer than 100,000 sequencing runs were recorded annually. Since 2013, there has been a significant increase in genomic sequencing, with sequencing runs reaching 1 million per year by 2016. By 2020, 3 million sequencing runs per year had been completed, with a record of over 6 million sequencing runs in 2022 (Fig. 1), driven by the COVID-19 pandemic, 47% of all sequencing runs for 2022 was for coronavirus.


**Figure 1. Increase in the number of sequencing runs over time in the SRA repository**

North American and European countries conduct 51.5% and 32.8% of sequencing runs, respectively (Fig. 2b). The U.S. leads in the number of studies conducted with 12,725,632 sequencing runs, accounting for 45% of the total data in SRA. The top 3 countries with the highest number of runs include England with 4,731,180 (16%) sequencing runs and China with 2,318,559 (8%) sequencing runs (Fig. 2c). The US has generated over 1 petabyte of information annually since 2018. As of 2018, China generates 250 to 500 terabytes of data annually, while England generates between 100 and 250 terabytes per year (Fig. 2a). Other countries such as Germany, Canada, Switzerland, Japan, Australia, and France also contributed to SRA data deposition, together submitting between 20,000 and 200,000 sequencing runs per year and generating between 10 and 50 terabytes of data per year (Fig. 2a).

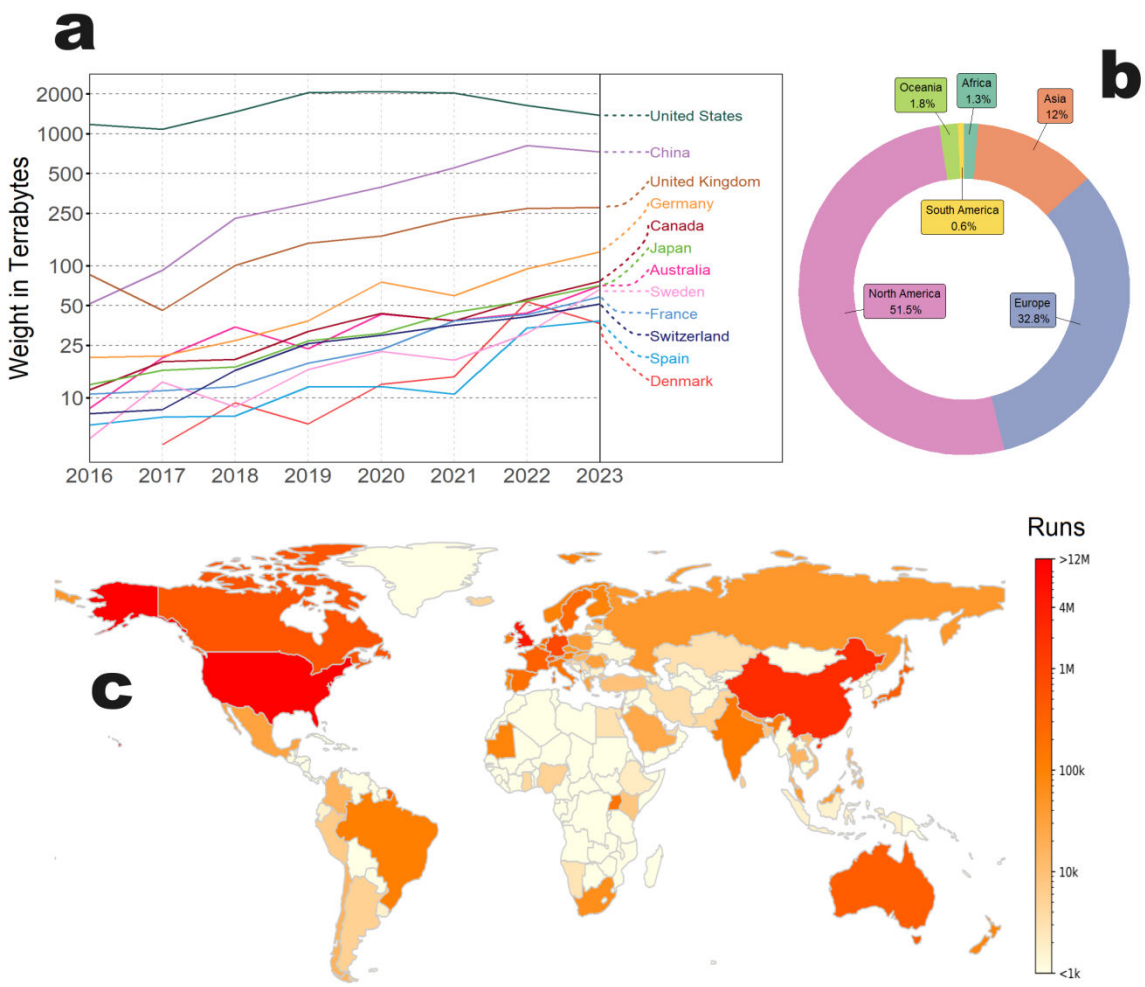


**Figure 2. Overall distribution of SRA runs. (a) Change in the amount of generated data for countries over time (sqrt scale for weight); (b) Share of regions by number of sequencing; (c) Heatmap for countries by number of sequences**

The Illumina platform accounts for 90% of sequencing cases and data volume, indicating the widespread use of short-read sequencing. The PacBio and Oxford Nanopore platforms are used for sequencing with shares of 2.6% and 2.3% respectively. The Ion Torrent share is 1.6%, while LS454 and Capillary each account for 1.3%. Other sequencing platforms collectively account for less than 0.5% of research. In terms of data generated, the new platforms BGIseq and DNBseq account for 2.1% and 1.5% respectively. PacBio represents 1.8%, while Oxford Nanopore accounts for 1.1%. The share of other platforms in the total data volume does not exceed 0.4%.

The NovaSeq 6000 and MiSeq models lead in the number of sequencing runs generated, producing 7 million and 6.2 million sequencing runs, respectively (Fig. 3d). Other models range from 1.3 million to 3.4 million sequencing runs. The expensive X Ten model conducted 1 million sequencing runs (Fig. 3d) but leads in the total data volume - 7000 terabytes, 1800 terabytes more than the NovaSeq 6000 (Fig. 3a). The HiSeq 2000, 2500, 4000 generated 3500, 3200, and 1800 terabytes of data, respectively. Together, all other models including Miseq produced less than 2600 terabytes of data (Fig. 3a).

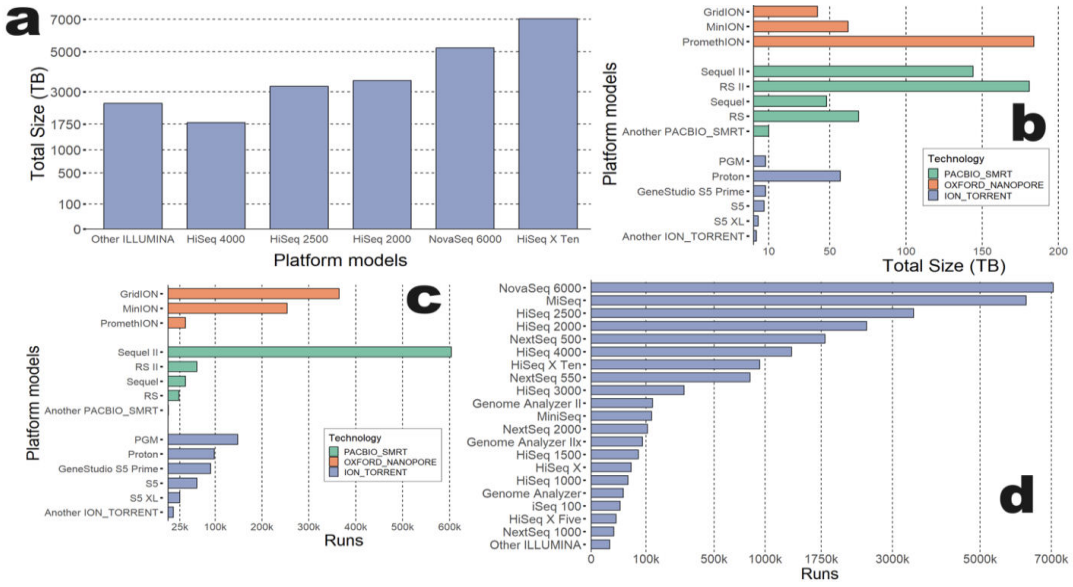


**Figure 3. Distribution of the number of sequencing runs and the amount of data generated by the sequencing model. (a) Number of data generated by ILLUMINA models; (b) Number of data generated by Pacbio, Oxford Nanopore, Ion Torrent models. (c) Number of sequencing by Pacbio, Oxford Nanopore, Ion Torrent models; (d) Number of sequencing runs by ILLUMINA models.**

Models such as Sequel II from Pacific Biosciences, Personal Genome Machine (PGM) from Ion Torrent, and GridION from Oxford Nanopore Platforms stand out among less common technologies, demonstrating volumes of 150,000 to 600,000 sequencing runs (Fig. 3c). In terms of the volumes of generated data, models like PromethION, Sequel II, and RS II generate between 145 and 180 terabytes of data (Fig. 3b).

Analyzing methodologies we noticed that researchers primarily focus on whole-genome sequencing studies, achieving 9 million sequencing runs, leading with the human genome. With the onset of the coronavirus pandemic, 6.1 million viral RNA sequencing runs were conducted, of which more than 95% were related to the SARS-CoV-2. Transcriptomic and metagenomic studies show similar volumes, with about 4.6-4.7 million sequencing runs each. Other research directions attract significantly less interest, with a total volume of less than 800,000 sequencing runs.Amplicon sequencing, which targets specific genomic regions, is the most common library strategy with 12 million cases. RNA-seq and Whole Genome Sequencing (WGS) were used in 4.8 million and 4.4 million sequencing runs, respectively. Less common methods find application in a range from 30,000 to 43,000 sequencing runs.

Analysis of the SRA database revealed 183,102 unique organisms classified into vertebrates, invertebrates, metagenomes, bacteria, viruses, fungi, protists, and archaea taxonomic groups. A total of 9,100,000 vertebrate sequencing runs were conducted (Fig. 7), with a combined volume of 18,000 terabytes, primarily focusing on human (*Homo sapiens*) genomes (5.1 million) and mice (*Mus musculus*) (2.6 million). A significant number of studies have also been conducted on domestic animals, in particular cattle (*Bos taurus*), pigs (*Sus scrofa*), chickens (*Gallus Gallus*) and dogs (*Canis lupus familiaris*), ranging from 37,000 to 87,000 sequencing runs. Viruses underwent 6.7 million sequencing runs (Fig. 7) with a total data volume of 555 terabytes, predominantly targeting the SARS-CoV-2 (95%). Cultivated plants such as rice (*Oryza sativa*), wheat (*Triticum Aestivum*), barley (*Hordeum vulgare*), as well as model plants like *Arabidopsis thaliana*, underwent sequencing runs ranging from 50,000 to 180,000. Invertebrates of particular interest include mosquitoes (*Anopheles gambiae*) and fruit flies (*Drosophila melanogaster*) with 65,000 and 165,000 sequencing runs respectively, as well as organisms similar to *Hydra* with 3,000 sequencing. Bacterial research focused on pathogenic microbes such as *Salmonella* (609k), *E. coli* (390k), *Streptococcus* (208k), *Mycobacterium tuberculosis* (202k), and *Staphylococcus*

(150k). Among fungi, yeast (*Saccharomyces cerevisiae*) the most interest with 200,000 sequencing runs. Among protists, *Plasmodium falciparum*, which causes malaria in humans, stood out with 247000 sequencing runs.

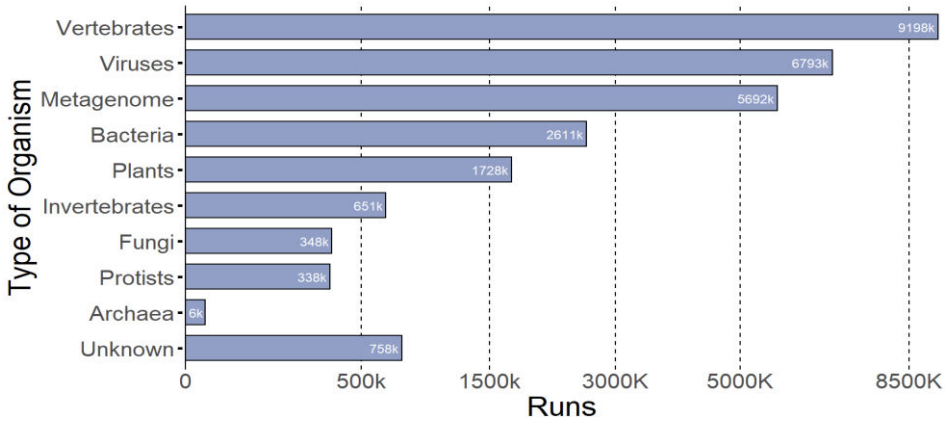


**Figure 4. Distribution of sequencing runs by organism classes in SRA repository**

Metagenomic analysis comprised 5,690,000 sequencing runs (Fig. 4), representing 350 unique metagenome samples in the SRA database. Researchers are particularly interested in human metagenomes, for instance, the study of gut metagenomes led to 1 million sequencing runs. Metagenomes from the oral cavity and skin also attracted attention, with 120,000 and 72,000 sequencing runs respectively. Among environmental metagenomes, soil research stands out, encompassing 778,000 sequencing runs. Various aquatic metagenomes also pique interest, ranging from 15,000 to 176,000 sequencing runs. Mouse metagenomes are frequently encountered in animal studies, with 253,000 sequencing runs conducted, along with metagenomes from pigs, cattle, and insects ranging from 63,000 to 66,000 sequencing runs. In artificial environments, researchers conduct runs on human waste and wastewater, totaling 90 terabytes in data volume.

The SRA database contains detailed descriptions of the data, methods, technologies and organisms studied. However, in some cases there are gaps or incomplete metadata. Although the completeness of the source data is high, even 0.85% of missing data represents 200 thousand mislabeled data, which is significant for certain domains and data reusability. We found that 5.2% for XXX samples lack submission date information. Additionally, XXX samples that represent 9% lack submitter information. About 9.4% of the metagenomic data are incompletely described, lacking an indication of where a particular metagenome was derived from, for 8% of the data the library strategies used are unknown. Also, 5% of the metadata do not have a specified taxid and 41% of the data do not have information on the attribute that specifies the design of the conducted experiment. For 2% of the data there is no description, it should be noted that this is not always critical. A small amount of metadata includes information about the presence of diseases, present in less than 1% of the human metadata.

**Discussion**

Our study based on the SRA database demonstrates that many countries are involved in DNA research. ILLUMINA-based models are the most commonly used, but long reads-based models are gradually increasing in popularity. A large amount of research is focused on the human genome, pathogens, and organisms that are pets or used as food. Among metagenomes, human body parts, soil, and wastewater account for a large proportion. The absence of standardized metadata schema compounds issues the veracity of experiments design and downstream analysis, hampering efforts to harness datasets for advanced computational analyses and integrative bioinformatics approaches. Metadata accompanying the raw data in public repositories allows us to better understand global trends in DNA research and reuse it for new-hypothesis testing analyses and more advanced studies, however, it has been found that some of the data contains critical lacks

in the metadata, making it difficult to reuse. The incompleteness of the metadata identified by the submission date makes it difficult to compare results and assess the consistency between different genetic changes and different phenotypic or environmental conditions occurring in the world. The absence of a taxonomyID that describes the object of study makes the data themselves unusable. Variations in the quality and organization of metadata yield considerable analytical uncertainties and can trigger a cascade of interpretative inaccuracies. Such inconsistencies could be major obstacles to the reproducibility of research findings and their subsequent extension into new scientific inquiries. It is necessary to ensure the data meets minimum requirements specified in the metadata standards. We also highlight the significant benefits that the improved availability and quality of metadata can offer, facilitating broader reuse within the scientific community. Improvement over data partitioning will improve the quality and completeness index of databases, this in turn will have a positive impact on storage efficiency, speed of research and information retrieval and data reuse.

**Bibliography**

[1] Yilmaz P (2017) Metadata Standards for Genomic Sequence Data: Past and Future of MIxS Standards Family. Proceedings of TDWG 1: e20423. https://doi.org/10.3897/tdwgproceedings.1.20423

[2] Matsuda T. Importance of experimental information (metadata) for archived sequence data: case of specific gene bias due to lag time between sample harvest and RNA protection in RNA sequencing. PeerJ. 2021 Aug 25;9:e11875. doi: 10.7717/peerj.11875. PMID: 34527435; PMCID: PMC8401820.

[3] "Metadata, FAIR principles, and their importance in genomics." [Online]. Avalaible: https://genestack.com/assets/pdfs/The%20importance%20of%20metadata%20in%20genomics%20and%20the%20FAIR%20principles%20ebook.pdf

[4] Moresis, A., Restivo, L., Bromilow, S. et al. A minimal metadata set (MNMS) to repurpose nonclinical in vivo data for biomedical research. Lab Anim 53, 67–79 (2024). https://doi.org/10.1038/s41684-024-01335-0

[5] Kodama Y, Shumway M, Leinonen R; International Nucleotide Sequence Database Collaboration. The Sequence Read Archive: explosive growth of sequencing data. Nucleic Acids Res. 2012 Jan;40(Database issue):D54-6. doi: 10.1093/nar/gkr854. Epub 2011 Oct 18. PMID: 22009675; PMCID: PMC3245110.

[6] Satam, H.; Joshi, K.; Mangrolia, U.; Waghoo, S.; Zaidi, G.; Rawool, S.; Thakare, R.P.; Banday, S.; Mishra, A.K.; Das, G.; et al. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology 2023, 12, 997. https://doi.org/10.3390/biology12070997

[7] Edgar, R.C., Taylor, B., Lin, V. et al. Petabase-scale sequence alignment catalyses viral discovery. Nature 602, 142–147 (2022). https://doi.org/10.1038/s41586-021-04332-2

[8] Bagheri H, Severin AJ, Rajan H. Detecting and correcting misclassified sequences in the large-scale public databases. Bioinformatics. 2020 Sep 15;36(18):4699-4705. doi: 10.1093/bioinformatics/btaa586. PMID: 32579213; PMCID: PMC7821992.

[9] Stevens I, Mukarram AK, Hörtenhuber M, Meehan TF, Rung J, Daub CO. Ten simple rules for annotating sequencing experiments. PLoS Comput Biol. 2020 Oct 5;16(10):e1008260. doi: 10.1371/journal.pcbi.1008260. PMID: 33017400; PMCID: PMC7535046.