The GSA Family in 2025: A Broadened Sharing Platform for Multi-Omics and Multimodal Data

Sisi Zhang[1,2,#], Xu Chen[1,2,#], Enhui Jin[1,2,3,#], Anke Wang[1,2,#], Tingting Chen[1,2,#], Xiaolong Zhang[1,2], Junwei Zhu[1,2], Lili Dong[1,2], Yanling Sun[1,2], Caixia Yu[1,2], Yubo Zhou[1,2], Zhuojing Fan[1,2], Huanxin Chen[1,2], Shuang Zhai[1,2], Yubin Sun[1,2], Qiancheng Chen[1,2], Jingfa Xiao[1,2,3], Shuhui Song[1,2,3], Zhang Zhang[1,2,3], Yiming Bao[1,2,3], Yanqing Wang[1,2*], Wenming Zhao[1,2,3*]

[1] *National Genomics Data Center, China National Center for Bioinformation, Beijing, China*

[2] *Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China*

[3] *University of Chinese Academy of Sciences, Beijing 100049, China*

[*] Corresponding author(s).

E-mail: zhaowm@big.ac.cn (Zhao W), wangyanqing@big.ac.cn (Wang Y).

Running title: *Zhang S et al / The GSA Family in 2024: A Broadened Sharing Platform for Multi-Omics and Multimodal Data*

Abstract

The Genome Sequence Archive Family (GSA family) provides a suite of database resources for archiving, retrieving, and sharing multi-omics data for the global academic and industrial communities. It currently comprises four distinct database members: the Genome Sequence Archive (GSA, https://ngdc.cncb.ac.cn/gsa/), the Genome Sequence Archive for Human (GSA-Human, https://ngdc.cncb.ac.cn/gsa-human/), the open Archive for Miscellaneous Data (OMIX, https://ngdc.cncb.ac.cn/omix/), and the Open Biomedical Imaging Archive (OBIA, https://ngdc.cncb.ac.cn/obia). Compared to the 2021 version, the GSA family has significantly expanded by introducing a new database member, the OBIA, and through comprehensive upgrades to the existing ones. Notable enhancements to the existing members include broadening the range of accepted data types, enhancing the quality control systems, fine-tuning the data retrieval system, and refining the data-sharing management mechanisms.

KEYWORDS: Genome Sequence Archive; GSA family; GSA; OBIA; OMIX

Introduction

Next-generation sequencing (NGS) technologies have revolutionized genomics, leading to an exponential increase in the amount of available raw sequence data and the expansion of researchers' understanding of genome structure, function, and complexity [1, 2]. The advancements in high-throughput omics technologies have empowered researchers to adopt multi-omics integration strategies to integrate and analyze data from different layers, including genomics, transcriptomics, epigenomics, proteomics, and metabolomics, thereby deepening our understanding of complex biological phenomena [3-5]. In recent years, advances in medical imaging technology (such as MRI, CT, and PET) and image analysis have enabled the extraction of imaging features and their association with multi-omics data, opening new directions for medical research [6-8]. These factors collectively have driven the transition from single-omics to multi-omics research and have advanced data management toward standardization, normalization, and platform-based sharing. The management and sharing of multi-omics and multimodal data present unique complexities, primarily in the following aspects: the large scale of data results in significant storage and processing challenges, requiring efficient solutions; the diverse sources and varying formats of data necessitate the implementation of flexible management strategies; and the imperative for data security and privacy protection demands the enforcement of stringent management measures. Therefore, a robust data management platform is needed to deliver comprehensive solutions for efficient storage, rapid retrieval, and secure sharing of the multi-omics and multimodal data in life sciences, to advance the progress of scientific research and clinical applications.

The Genome Sequence Archive [9] serves as a repository for raw sequence data in the National Genomics Data Center [10] at China National Center for Bioinformation (CNCB-NGDC)[10, 11], which has effectively addressed the longstanding challenges related to the unified submission and management of genomic data in China, promoting the sharing and reuse of data within the community. In 2021, the GSA family [12], comprising GSA, Genome Sequence Archive for Human (GSA-Human), and the Open

Archive for Miscellaneous Data (OMIX), was introduced, greatly enhancing the management and sharing capabilities for multi-omics data.

Over the past several years, the GSA family has made significant efforts to address the challenges caused by the rapid accumulation of multi-omics and multimodal data. First, it has expanded the data types, such as biomedical imaging and clinical data, which can be archived in a new member database called the Open Biomedical Image Archive (OBIA) [13]. Second, the major functions of the existing databases, including the data quality control processes, the data search tools, the data-sharing mechanisms, and the data security measures, have been upgraded. Moreover, GSA has mirrored and integrated the raw sequence data from National Center for Biotechnology Information (NCBI) and provided unified retrieve and download services for the submitted and integrated data. These initiatives have significantly improved the GSA family's data management capabilities, enhanced the security and convenience of data access, and supported research and innovation in the life sciences.

Archival resources

Genome Sequence Archive

GSA, built based on the International Nucleotide Sequence Database Collaboration (INSDC) [14] data standards, is an open-access repository dedicated to archiving, retrieving, and sharing non-human raw sequence data. In 2023, GSA was selected as one of the Global Core Biodata Resources by the Global Biodata Coalition, demonstrating its pivotal role in the global landscape of biological data resources.

The metadata in GSA is systematically organized into four objects: Study, Sample, Experiment and Run (Figure 1). Each study in GSA is uniquely identified and represented by a BioProject (https://ngdc.cncb.ac.cn/bioproject), providing descriptive metadata about the research project. Analogously, each sample in GSA is assigned a BioSample identifier (https://ngdc.cncb.ac.cn/biosample), enabling the structured

storage and retrieval of biomaterial metadata. This standardized approach ensures that samples are accurately described, facilitating cross-study comparisons and enhancing the overall value of the data for scientific research and discovery.

To enhance the user experience and streamlined data accessibility, GSA has introduced a new data retrieval system tailored to meet the diverse needs of researchers. This advanced retrieval interface enables users to conduct sophisticated queries across multiple search fields, encompassing organism, file type, sequencing platform, sequencing strategy, and numerous other criteria. The system further empowers users by allowing them to construct intricate queries using logical operators, facilitating the precise targeting and swift retrieval of desired data. Once search results are obtained, users are presented with a range of filtering options to refine their search outcomes even further. This flexibility ensures that researchers can quickly zero in on the most relevant data sets. Moreover, GSA offers a convenient batch download tool, enabling the seamless acquisition of search results in various formats, significantly streamlining the process of data collection and analysis.

In addition, GSA has introduced a sequence taxonomic analysis tool, derived from the Sequence Taxonomic Analysis Tool (STAT) [15]. This tool offers a comprehensive platform for the taxonomic classification analysis of the submitted data, complemented by intuitive web-based visualizations of the analysis results. By leveraging this service, researchers can gain valuable insights into the quality and validity of the data from a species-centric perspective. This, in turn, empowers them to make informed decisions regarding data selection and utilization, ultimately facilitating the efficient acquisition of the required data for their studies.

Genome Sequence Archive for Human

GSA-Human is a data repository dedicated to archiving, managing, and sharing raw sequence data from human biomedical research endeavours. It employs a framework consisting of five distinct objects to organize and structure the metadata: Study, Individual, Sample, Experiment, and Run (Figure 1). Each research study in GSA-

Human is uniquely indicated by a BioProject, proving a clear and consistent reference point. While GSA-Human does not utilize the BioSample object for sample representation, it ensures that non-sensitive sample information is harmoniously synchronized with the BioSample database. This strategic integration enables the seamless sharing of sample metadata across various databases within the CNCB-NGDC ecosystem, such as the Genome Warehouse[16] and the Genome Variation Map [17], all facilitated through BioSample accessions. This collaborative approach ensures smooth data sharing across different databases while safeguarding data privacy and security, enhancing data interoperability, and supporting the discovery and utilization of human biomedical research data.

GSA-Human provides two types of data access modalities: open access and controlled access. Open access allows unrestricted data downloads, whereas controlled access requires prior authorization from the Data Access Committee (DAC), as designated by the data submitter. Furthermore, GSA-Human has formulated comprehensive policies governing the sharing of human genetic resources data (https://ngdc.cncb.ac.cn/gsa-human/policy), providing users with clear guidelines on data submission procedures, management practices, and sharing protocols. These measures ensure both the security of sensitive data and its efficient utilization by the scientific community.

In recent years, the data volume in GSA-Human has experienced an exponential surge, with the current magnitude being a staggering sevenfold increase compared to 2021 levels. Notably, the most rapid expansion of 1 petabyte (PB) was achieved within a mere two days, posing significant challenges to data quality control procedures. To alleviate the pressure and improve data processing efficiency, we have upgraded the algorithm of the quality control process for FASTQ files by optimizing the data structure and minimizing memory consumption. The new algorithm reduces memory usage by 15 times, enabling the processing of terabyte-scale compressed files, and has been successfully integrated into GSA-Human and other relevant systems for processing raw sequence data.

In addition, GSA-Human has devised an intricate scoring system to evaluate the degree of sharing exhibited by the published datasets. Open access datasets automatically receive the full scores, while for controlled access datasets, a comprehensive evaluation incorporates various factors related to the DAC's (Data Access Committee) handling of data access requests. These factors include the percentages of processed and approved requests, and the average processing time for the requests, which together contribute to determining the final score. This scoring methodology offers an intuitive measure of dataset sharing levels, furnishing data requesters with clear and understandable insights. It is our hope that such a scoring system will foster a culture of data sharing and encourage the reuse of scientific data. The scoring details for the public datasets are available online (https://ngdc.cncb.ac.cn/gsa-human/browse/).

Open Archive for Miscellaneous Data

OMIX, built adhering to the FAIR principles[18], stands as a versatile data repository, dedicated to gathering, publishing, and sharing scientific data for biological researchers. OMIX employs a light-weighted data model in which all the metadata are decomposed into three objects: Study, Organism, and File (Figure 1). This data model endows OMIX with remarkable flexibility and scalability, enabling it to effortlessly expand its data type offerings to cater to the diverse data submission demands of researchers. Additionally, each OMIX study is linked to a BioProject, which connects OMIX with other resources, especially GSA and GSA-Human, thereby better-organizing multi-omics studies and facilitating user queries and usage. (*e.g.*, https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA006118). Currently, OMIX supports eight primary data types, including 31 sub-categories that span functional genomics, proteomics, metabolomics, and more (Table S1), underscoring its comprehensive capabilities and adaptability.

OMIX provides two distinct data access modalities: open access and controlled access, both available for human genetic resource data, while data from other species

is limited to open access. In managing and authorizing controlled access data, OMIX , OMIX uses a mechanism similar to GSA-Human but simplifies the "request and approval" process to better suit its own characteristics.. This simplified workflow ensures that data access requests are handled directly by the data submitter, fostering efficiency and expediency.

In addition, OMIX has designed intuitive interfaces that greatly enhance the experience for both data requesters and submitters when navigating the controlled data application process. The streamlined, one-step data request page enables data requesters to conveniently complete and simultaneously submit their access application to the data submitter. This allows the data submitter to approve requests with a single click, drastically reducing turnaround time. Moreover, OMIX incorporates real-time editing functions and feedback mechanisms, making the system more user-friendly and convenient, thus promoting data acquisition and utilization.

Open Biomedical Imaging Archive

OBIA, the new member of the GSA family, serves as an innovative data repository meticulously crafted to store and share multimodal medical imaging alongside its clinical data. It boasts a dual-pronged approach to data accessibility, encompassing both open and controlled access mechanisms, catering to varying user needs and security requirements. Within the intricate architecture of OBIA, five key objects interlock seamlessly: Collection, Individual, Study, Series, and Image (Figure 1). Notably, Individuals within this framework can be seamlessly interlinked to GSA-Human via unique accessions, streamlining the integration and correlation of imaging data with genomic information for comprehensive multi-omics research endeavours. This feature underscores OBIA's pivotal role in fostering collaborative research by bridging the gap between diverse data modalities and facilitating the synthesis of insights for advanced medical discoveries.

The primary challenge in constructing a medical image database is ensuring robust privacy protection. Images may contain protected health information (PHI) and require

appropriate processing before archived to minimize the risk of patient privacy breaches. OBIA addresses this by providing a unified de-identification and quality control mechanism based on the Digital Imaging and Communications in Medicine (DICOM) standard. DICOM tags that contain or could contain PHI are removed or blanked, and images with PHI in their pixel data are rigorously filtered to safeguard patient privacy.

In addition to providing web interfaces for data query and browsing, OBIA offers innovative image retrieval capabilities[19]. The retrieval model is based solely on image features and uses the Hamming distance to return visually similar images. Users can upload images to search the database for the top 30 most similar images, significantly enhancing the user experience.

Data submission and mirroring

Data submission

The GSA family is equipped with the Single Sign-On (SSO) service, a user access control system that enables seamless authentication across multiple databases with a single, unified ID and password. To submit data to the GSA family, users are required to initially register an account within the SSO, and log into the specific database where they intend to initiate their submissions.

Each database within the GSA family provides a convenient web-based submission service that defines a series of metadata templates tailored to the characteristics of the accepted data, facilitating the collection of the required information. Take GSA-Human as an example, it offers specialized templates for human-related researches, encompassing diverse domains such as disease, cohort, cell line, clinical pathogen and human-associated metagenome studies. The templates cover essential attributes required for the data objects, spanning sample details, experimental conditions, and sequencing strategies, while also allowing users to incorporate custom attributes, greatly enhancing adaptability to personalized research requirements.

Moreover, each database within the GSA family has quality control and expert curation processes to ensure the accuracy and reliability of the submitted data. These

processes focus on verifying the integrity, validity, and consistency of the metadata and data files, as well as identifying and removing personal privacy information, thereby ensuring the preservation of high-quality data for future reference and analysis.

Data mirroring

The INSDC maintains the largest collection of raw sequence data that are not readily retrievable for many researchers in China, mainly due to the large volume and limited network bandwidth. In an effort to facilitate the seamless sharing and reuse of this global data, GSA has developed a system that mirrors the raw sequence data of INSDC sourced from NCBI. Metadata including BioProjects, BioSamples and SRAs are synchronously downloaded, parsed and stored locally, and the sequence files are also mirrored in accordance with the metadata. What's more, all the mirrored data are integrated into GSA and powered with the retrieval, browse, and download functions, offering users with convenient ways to access the data.

Data statistics

As of July 2024, the GSA family has archived 23,940 datasets, with a total data volume exceeding 51.52 petabytes (PB), encompassing a wide range of data types (Figure 2). The data accessions have been cited in 3,726 research articles and 629 scientific journals (https://ngdc.cncb.ac.cn/gsa/statistics?active=journals) (For detailed statistics on data submissions for each database, see Table 1). In terms of data sharing, the GSA family offers two main access modalities: open and controlled access. Through open access, the database provides researchers with unrestricted access to datasets. As of July 2024, the GSA family has published 11,447 open access datasets, amounting to 6.29 PB of data shared via FTP, HTTP, and Edge Turbo (exclusively for the GSA and GSA-Human database), with a total of 4,366,168 downloads. For controlled data sharing, the GSA family has published 3,369 controlled access datasets, receiving a total of 7,102 access requests from 3,844 requesters, with 2,411 requests approved, resulting in over 6.68 PB of data downloads (For detailed statistics on data sharing for each database,

see Table 2). Regarding international data mirroring, as of July 2024, GSA has mirrored 789,781 projects, 40,502,776 samples, 31,126,209 experiments, and 33,088,342 runs, as well as over 7.74 PB of sequence files from NCBI.

Future Direction

The GSA family has continuously enhanced its archival databases for raw sequence, multi-omics, and multimodal data to improve data submission and management capabilities and promote efficient data reuse. The China National Center for Bioinformation plans to strengthen its support for the GSA family through a comprehensive upgrade of the national big data infrastructure, including improvements in storage capacity, computing ability, and network bandwidth.

Future efforts will be concentrated on continuous optimizations of data models and management processes to meet evolving user demands. An advanced cloud-based data storage architecture, as long as a suit of innovative tools, will be established to improve the efficiency of storing and mining the omics big data. Efforts will also be focused on improving the data security management capability. Various measures such as data classification and management mechanisms, data backup and recovery systems, and data anonymization and encryption utilities will be implemented or enhanced to ensure the long-term secure maintenance and management of data.

Additionally, we actively advocate for global collaborations to collectively advance data standards, technologies, and methodologies, facilitating the widespread sharing of global biodiversity and health big data.

Data availability

GSA is open accessible at https://ngdc.cncb.ac.cn/gsa/.

GSA-Human is freely accessible at https://ngdc.cncb.ac.cn/gsa-human/.

OMIX is freely accessible at https://ngdc.cncb.ac.cn/omix/.

OBIA is freely accessible at https://ngdc.cncb.ac.cn/obia.

Authors' contributions

Sisi Zhang: Investigation, Methodology, Data curation, Writing - original draft. Xu Chen: Software, Writing - original draft. Enhui Jin: Software, Investigation, Data curation, Writing - original draft. Anke Wang: Software, Writing - original draft. Tingting Chen: Investigation, Methodology, Data curation, Writing - original draft. Xiaolong Zhang Software, Writing - original draft. Junwei Zhu: Software. Lili Dong: Data curation. Yanling Sun: Data curation. Caixia Yu: Data curation. Yubo Zhou: Software. Zhuojing Fan: Resources. Huanxin Chen: Resources. Shuang Zhai: Resources. Yubin Sun: Resources. Qiancheng Chen: Resources. Jingfa Xiao: Conceptualization. Shuhui Song: Conceptualization, Funding acquisition. Zhang Zhang: Conceptualization, Funding acquisition. Yiming Bao: Conceptualization, Funding acquisition. Yanqing Wang: Conceptualization, Investigation, Methodology, Software, Writing - review & editing, Project administration, Funding acquisition. Wenming Zhao: Conceptualization, Methodology, Writing - review & editing, Supervision, Funding acquisition. All authors have read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by National Key R&D Program of China (Grant Nos. 2023YFC2605702 to WZ, 2023YFC2604402 to YW, 2021YFF0703701 to YB); National Natural Science Foundation of China (Grant Nos. 32100511 to YW, 32170678 to WZ; 32030021 to ZZ; 92374201 to SS); Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDB38050300 to WZ).

References

[1] Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. Human Immunology 2021;82:801-11.

[2] Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, et al. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology 2023;12:997.

[3] Chen C, Wang J, Pan D, Wang X, Xu Y, Yan J, et al. Applications of multi-omics analysis in human diseases. MedComm 2023;4:e315.

[4] Picard M, Scott-Boyer M-P, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Computational and Structural Biotechnology Journal 2021;19:3735-46.

[5] Jude G, Rifaï K, Daures M, Dubois L, Bignon Y-J, Penault-Llorca F, et al. High-throughput 'Omics' technologies: New tools for the study of triple-negative breast cancer. Cancer Letters 2016;38:77-85.

[6] Zanfardino M, Franzese M, Pane K, Cavaliere C, Monti S, Esposito G, et al. Bringing radiomics into a multi-omics framework for a comprehensive genotype-phenotype characterization of oncological diseases. Journal of Translational Medicine 2019;17:337.

[7] Incoronato M, Aiello M, Infante T, Cavaliere C, Grimaldi AM, Mirabelli P, et al. Radiogenomic Analysis of Oncological Data A Technical Survey. International Journal of Molecular Sciences 2017;18:805.

[8] Leithner D, Horvat JV, Ochoa-Albiztegui RE, Thakur S, Wengert G, Morris EA, et al. Imaging and the completion of the omics paradigm in breast cancer. Radiologe 2018;58:7-13.

[9] Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, et al. GSA: Genome Sequence Archive. Genomics, Proteomics & Bioinformatics 2017;15:14-8.

[10] Partners C-NMa. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. Nucleic Acids Research 2024;52:D18-D32.

[11] Bao Y, Xue Y. From BIG Data Center to China National Center for Bioinformation. Genomics, Proteomics & Bioinformatics 2023;21:900-3.

[12] Chen T, Chen X, Zhang S, Zhu J, Tang B, Wang A, et al. The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types. Genomics, Proteomics & Bioinformatics 2021;19:578-83.

[13] Jin E, Zhao D, Wu G, Zhu J, Wang Z, Wei Z, et al. OBIA: An Open Biomedical Imaging Archive. Genomics, Proteomics & Bioinformatics 2023;21:1059-65.

[14] Cochrane G, Karsch-Mizrachi I, Takagi T, International Nucleotide Sequence Database C. The International Nucleotide Sequence Database Collaboration. Nucleic Acids Research 2016;44:D48-50.

[15] Katz KS, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. STAT: a fast, scalable, MinHash-based k-mer tool to assess Sequence Read Archive next-generation sequence submissions. Genome Biology 2021;22:270.

[16] Chen M, Ma Y, Wu S, Zheng X, Kang H, Sang J, et al. Genome Warehouse: A Public Repository Housing Genome-Scale Data. Genomics, Proteomics & Bioinformatics 2021;19:584-9.

[17] Li C, Tian D, Tang B, Liu i, Teng X, Zhao W, et al. Genome Variation Map: a worldwide collection of genome variations across multiple species. Nucleic Acids Research 2021;49:D1186-D91.

[18] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016;3:160018.

[19] Wu G, Jin E, Sun Y, Tang B, Zhao W. Deep Attention Fusion Hashing (DAFH) Model for Medical Image Retrieval. Bioengineering 2024;11:673.

Figures legends

Figure 1 GSA family data model

BioProject and BioSample are two independent meta-information databases, which offers structured storage for research purposes and biomaterials, respectively. All members of the GSA family are linked to BioProject, thereby creating a comprehensive framework for multi-omics data. GSA-Human established a data synchronization process with BioSample. Individuals in OBIA can be linked to GSA-Human by accessions. OMIX is for various types of data (that are unsuitable for another member).

Figure 2 Data statistics of the GSA family

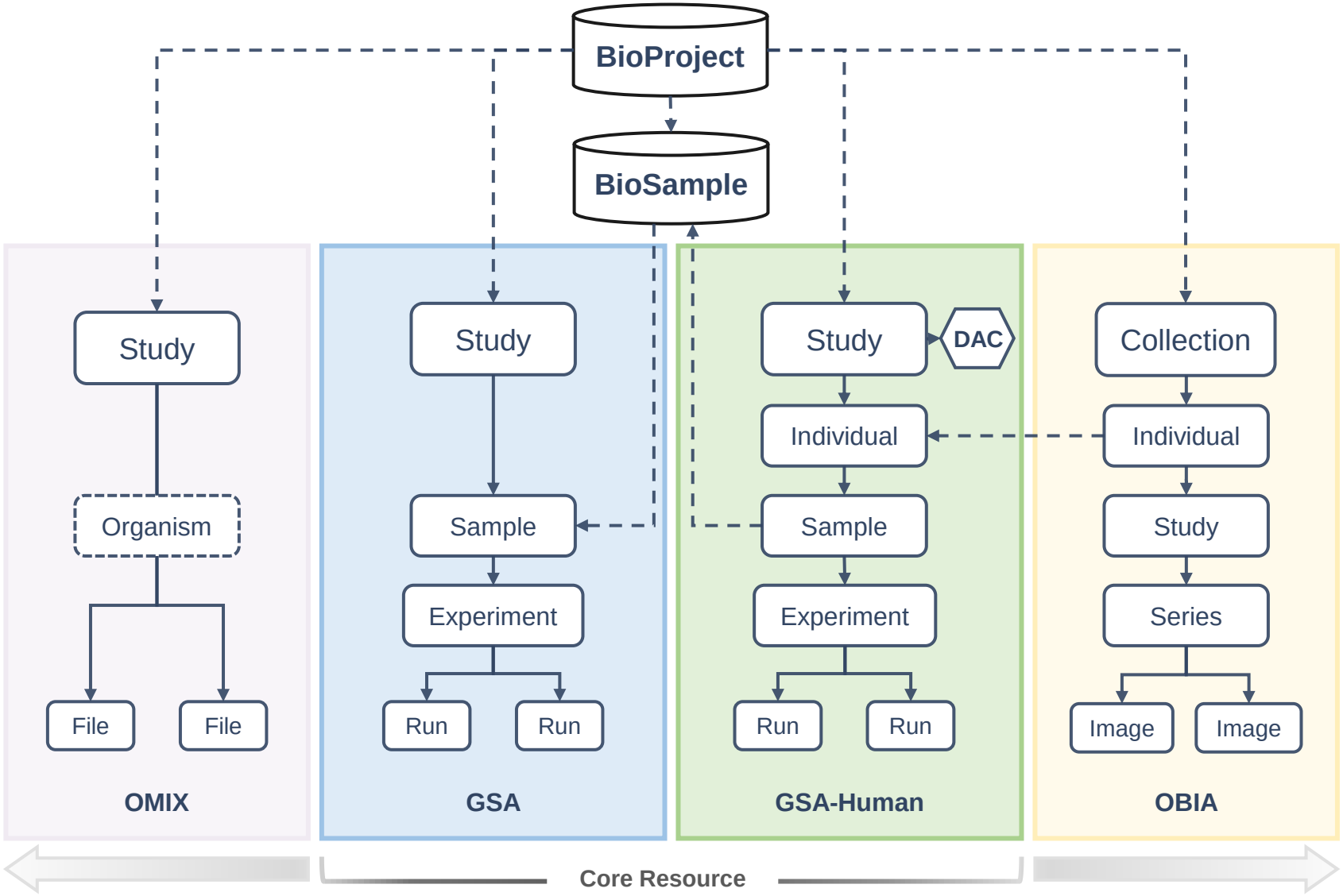
A. Number of datasets by year accumulated from 2015 to 2024 from GSA family, with thirteen major data types indicated. B. Increase in the volume of submitted raw sequence data over time, with the time required to accumulate each PB of data indicated. All statistics were derived from GSA and GSA-Human as of July 2024. PB, petabyte; d, days.

Table 1 Data submissions within the GSA Family

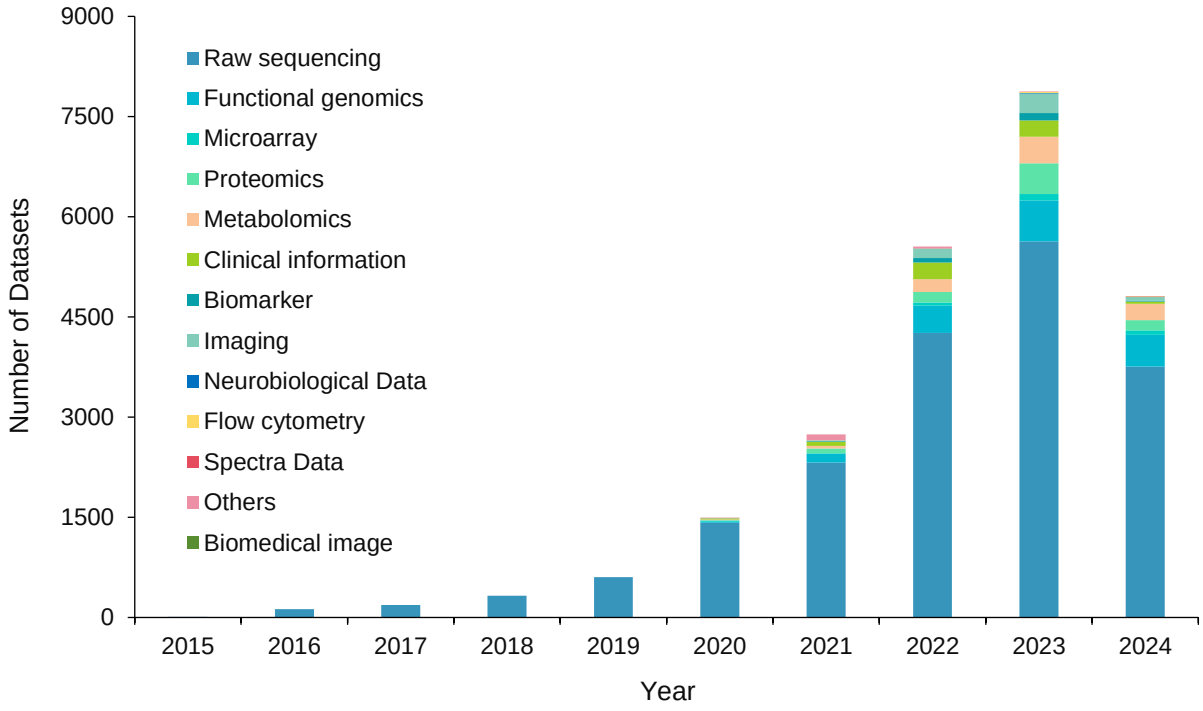
Table 2 Data sharing within the GSA family

Supplementary materials

Table S1 Data type of OMIX database


BioProject
BioSample
Study
Organism
File
File
OMIX
Study
Sample
Experiment
Run
Run
GSA
Study
DAC
Individual
Sample
Experiment
Run
Run
GSA-Human
Collection
Individual
Study
Series
Image
Image
OBIA
Core Resource

A


9000
7500
6000
4500
3000
1500
0
Number of Datasets
Raw sequencing
Functional genomics
Microarray
Proteomics
Metabolomics
Clinical information
Biomarker
Imaging
Neurobiological Data
Flow cytometry
Spectra Data
Others
Biomedical image
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Year

B

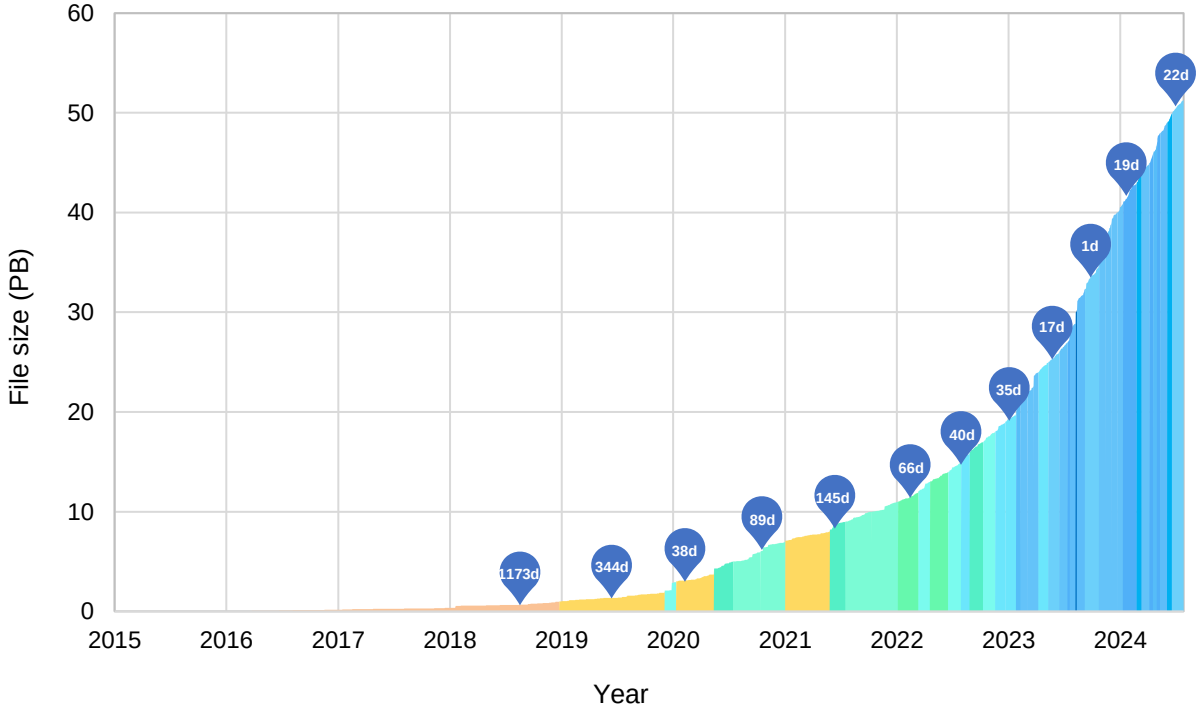

60
50
40
30
20
10
0
File size (PB)
1173d
344d
38d
89d
145d
66d
40d
35d
17d
1d
19d
22d
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Year

Table 1 Data submissions within the GSA Family

Data Submitted Item	GSA	GSA-Human	OMIX	OBIA	Total
No. of datasets	13,290	5,540	5,106	4	23,940
No. of open-access datasets	13,290	1,300	2,244	/	16,834
No. of controlled-access datasets	/	4,240	2,862	4	7,106
No. of BioProjects	11,184	4,887	3,199	1	19,271
No. of Individuals	/	527,400	/	937	528,337
No. of Samples	711,985	766,260	/	/	1,478,245
No. of Experiments	755,486	851,004	/	/	1,606,490
No. of Runs	816,376	1,038,289	/	/	1,854,665
No. of Files	1,491,007	1,725,876	26,445	1,938,309	5,181,637
Total submitted file size (TB)	16,212.83	36,467.77	81.88	0.35	52,762.83
No. of support publications	2268	923	684	/	3,726

Note: All statistics regarding data submissions were derived from the GSA family as of July 2024. "/" means not applicable. TB, terabyte. Due to articles referencing serial accessions from multiple databases simultaneously, the total article count is lower than the sum of individual article counts in each database.

Table 2 Data sharing within the GSA family

Data Sharing Item		GSA	GSA-Human	OMIX	OBIA	Total
All published datasets	No. of datasets	8,992	3,169	2,651	4	14,816
	Sharing times	4,111,830	187,345	69,397	7	4,368,572
	Total downloads file size (TB)	5,924.34	7,322.85	37.25	0.67	13,285.11
Open-access datasets	No. of datasets	8,992	923	1,532	/	11,447
	Download times	4,111,830	185,190	69,148	/	4,366,168
	Total downloads file size (TB)	5,924.34	504.37	14.83	/	6,443.54
Controlled-access datasets	No. of datasets	/	2,246	1,119	4	3,369
	No. of requests	/	6,432	647	23	7,102
	No. of requesters	/	3,281	547	16	3,844
	Approved times	/	2,155	249	7	2,411
	Total downloads file size (TB)	/	6,818.48	22.42	0.67	6,841.57

Note: All statistics regarding data sharing were derived from the GSA family as of July 2024. "/" means not applicable. TB, terabyte.