

1 **The Genome Sequence Archive Family: Towards Explosive Data**

2 **Growth and Diverse Data Types**

3

4 Tingting Chen^{1,2,#}, Xu Chen^{1,2,#}, Sisi Zhang^{1,2,#}, Junwei Zhu^{1,2,#}, Bixia Tang^{1,2}, Anke
5 Wang^{1,2}, Lili Dong^{1,2}, Zhewen Zhang^{1,2}, Caixia Yu^{1,2}, Yanling Sun^{1,2}, Lianjiang Chi^{1,3},
6 Huanxin Chen^{1,2}, Shuang Zhai^{1,2}, Yubin Sun^{1,2}, Li Lan^{1,2}, Xin Zhang^{1,2}, Jingfa
7 Xiao^{1,2,4}, Yiming Bao^{1,2,4}, Yanqing Wang^{1,2,*}, Zhang Zhang^{1,2,4,*}, Wenming Zhao^{1,2,4,*}

8

9

10 ¹ *National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of*
11 *Sciences / China National Center for Bioinformation, Beijing 100101, China*

12 ² *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of*
13 *Genomics, Chinese Academy of Sciences, Beijing100101, China*

14 ³ *CAS Key Laboratory of Genomic and Precision Medicine, Beijing Institute of*
15 *Genomics, Chinese Academy of Sciences, Beijing 100101, China*

16 ⁴ *University of Chinese Academy of Sciences, Beijing 100049, China*

17

18 # Equal contribution.

19 * Corresponding author(s).

20 E-mail: zhaowm@big.ac.cn (Zhao W), zhangzhang@big.ac.cn (Zhang Z),

21 wangyanqing@big.ac.cn (Wang Y).

22

23 **Running title:** *Chen T et al / The Genome Sequence Archive Family*

24

25

26

27 Total word counts (from “Introduction” to “Conclusions” or “Materials and
28 methods”): 1728

29 Total figures: 2
30 Total tables: 2
31 Total supplementary figures: 0
32 Total supplementary tables: 0
33 Total supplementary files: 0

34 **Abstract**

35 The Genome Sequence Archive (GSA) is a data repository for archiving raw sequence
36 data, which provides data storing and sharing services for worldwide scientific
37 communities. Considering explosive data growth with diverse data types, here we
38 present the GSA family by expanding into a set of resources for raw data archive with
39 different purposes, namely, GSA (<https://ngdc.cncb.ac.cn/gsa/>), GSA for Human
40 (GSA-Human, <https://ngdc.cncb.ac.cn/gsa-human/>), and
41 Open Archive for Miscellaneous Data (OMIX, <https://ngdc.cncb.ac.cn/omix/>).
42 Compared with the 2017 version, GSA has been significantly updated in data model,
43 online functionalities, and web interfaces. GSA-Human, as a new partner of GSA, is a
44 data repository specialized in human genetics-related data with controlled access and
45 security. OMIX, as a critical complement to the two resources mentioned above, is an
46 open archive for miscellaneous data. Together, all these resources form a family of
47 resources dedicated to archiving explosive data with diverse types, accept data
48 submissions from all over the world and provide free open access to all publicly
49 available data in support of worldwide research activities.

50

51 **KEYWORDS:** Genome Sequence Archive; GSA; GSA-Human; OMIX

52

53 **Introduction**

54 The Genome Sequence Archive [1] (GSA, <https://ngdc.cncb.ac.cn/gsa>) is a public
55 archive of raw sequence data in the National Genomics Data Center (NGDC) [2-4],
56 part of the China National Center for Bioinformation (CNCB). GSA accepts
57 worldwide data submissions, performs data curation and quality control for all
58 submitted data, and provides free open access to all publicly available data without
59 unnecessary restrictions. Since its inception in 2015, GSA has been broadly supported
60 and endorsed by the scientific community, as testified by a total of 324,325
61 experiments, 371,973 runs and 8526 TB files submitted by 1530 users from 385
62 institutions and reported in 634 research articles and 239 scientific journals (as of
63 June 2021). Importantly, GSA serves as one of the core resources in CNCB-NGDC
64 that has stable state funding in biological data management, thus ensuring long-term
65 persistence and preservation of submitted datasets.

66 Due to the rapid development of sequencing technologies towards higher
67 throughput and lower cost as well as their wider applications in biomedical research, a
68 large number of multi-omics data have been produced at ever-increasing rates and
69 scales, provoking two major challenges for raw data management in GSA. For one
70 thing, several large-scale sequencing projects (such as Earth BioGenome Project [5],
71 Dog 10K Project [6], Protist 10000 Genomes Project [7]) have been carried out over
72 the past several years, leading to different types of raw sequence data generated
73 around the global and accordingly requiring a suite of web services for massive data
74 submission and deposition. For another, studies on human population genomics and
75 precision medicine have produced millions of personal genome sequences associated
76 with clinical information, requiring controlled access and security management,
77 which is critically vital in promoting human healthcare and precise medical treatment
78 and advancing big-data-driven scientific research, while protecting data privacy.
79 These challenges are particularly crucial in China since it not only features the largest

80 population in the world and rich biodiversity resources, but also has a formidable
81 capacity in genome sequencing throughout the country.

82 To address these challenges, here we provide a family of resources for raw data
83 archive and management, including an updated version of GSA and two newly
84 developed partner resources, namely, GSA for Human (GSA-Human,
85 <https://ngdc.cncb.ac.cn/gsa-human>) and Open Archive for Miscellaneous Data
86 (OMIX, <https://ngdc.cncb.ac.cn/omix>). Specially, we updated GSA with significant
87 improvements on data model, online functionalities and web interfaces. As an
88 important partner to GSA that provides open access to all released data, GSA-Human
89 features controlled-access and security services for human genetics-related data,
90 which is compatible well with the database of Genotypes and Phenotypes (dbGaP) [8]
91 in the National Center for Biotechnology Information (NCBI) [9] and the European
92 Genome-phenome Archive (EGA) [10] in the European Bioinformatics Institute
93 (EBI) [11]. But GSA-human is different from dbGaP and EGA; the former is mainly
94 used to archive and store raw sequence data, while the latter not only archive raw
95 sequence data, but also archive phenotypic data. In addition, OMIX
96 (<https://ngdc.cncb.ac.cn/omix/>), as a critical complement to the above two resources,
97 is an open archive for miscellaneous data that are unsuitable to store in GSA,
98 GSA-Human or other databases at CNCB-NGDC. Together, all these resources form
99 a family of resources dedicated to archiving explosive data with diverse types.

100 **Archival resources**

101 GSA, built based on the INSDC (International Nucleotide Sequence Database
102 Collaboration) [12] data standards and structures, is a public data repository for
103 archiving raw sequence reads. Over the past several years, GSA has been frequently
104 and considerably updated since its establishment in 2015, with significant
105 improvements in data structure, online submission, quality control, and web
106 functionalities (**Table 1**). First, data structure has been significantly changed (Figure
107 1); BioProject (<https://ngdc.cncb.ac.cn/bioproject/>) and BioSample

108 (<https://ngdc.cncb.ac.cn/biosample/>) have been separated from GSA, serving as
109 independent meta-information databases and acting as an organizational framework to
110 provide centralized access to descriptive metadata about research projects and
111 samples, respectively. Second, to help users submit massive data with different types,
112 more sequencing platforms, sample types, and file formats were acceptable, and
113 importantly, batch submission of multiple experiments and runs was enabled in the
114 updated version of GSA. In addition, to provide users with convenient services for
115 uploading raw sequence files, GSA not only provides an FTP server but also equips
116 with Aspera (<https://www.ibm.com/products/aspera>) to realize high-speed data
117 transmission. Third, GSA was greatly enhanced by improving the expert curation
118 process and integrating an automated quality control pipeline, with the aim to provide
119 value-added services for archiving high-quality data. Fourth, multiple web
120 functionalities for bilingual support (both English and Chinese), online
121 documentation, data statistics and visualization charts, were updated/newly added.
122 Taken together, the updated version of GSA is more efficient and friendly in big
123 omics-data submission, deposition and management.

124 GSA-Human, established in April 2018, is a data repository specialized in the
125 secure management of human genetics-related data. It accepts submissions of various
126 studies, including disease, cohort, cell line, clinical pathogen and human associated
127 metagenome. GSA-Human uses the “individual” to organize its metadata and
128 sequence reads and provides two different data access mechanisms: open access and
129 controlled-access. Open access means that all data are public for global researchers,
130 whereas controlled-access means that data can be downloadable only after being
131 authorized by the Data Access Committee (DAC) that is responsible for
132 authorizing/declining data access to data requester. Therefore, GSA-Human provides
133 a series of data services including access control, data request, access
134 authorization/decline, and security management.

135 OMIX, as a new member of the archival resources in CNCB-NGDC, aims to meet
136 users' needs for submitting various types of data other than sequences. It collects not
137 only raw data from transcriptome, epigenome, and microarray, but also functional
138 data such as lipidome, metabolome, proteome, and even data like clinical information,
139 demographic data, questionnaire and so on. With the concise interface and simplified
140 submission process, OMIX enables data submission and deposition very easy. Of
141 note, similar to GSA-Human, OMIX has a data security management strategy for
142 human genetic data. Any controlled-access dataset in OMIX can be accessed only
143 with the permission of the original data submitter/owner.

144 **Data submission and retrieval**

145 Data submission to the GSA family is aided by a series of web services, including
146 BIG Single Sign-On (SSO; <https://ngdc.cncb.ac.cn/sso/>) that is a user access control
147 system and BIG Submission portal (BIG Sub; <https://ngdc.cncb.ac.cn/gsub/>) that is a
148 unified one-stop portal providing submission services for a variety of database
149 resources in CNCB-NGDC. To submit data to the GSA family, user needs to register
150 an account and log into any database via SSO that can help user gain access to
151 multiple independent systems with a single ID and password.

152 Overall, the GSA family provides a suite of services for data retrieval, download
153 and access. Public data in these resources can be retrieved via BIG Search
154 (<https://ngdc.cncb.ac.cn/search/>), a scalable text search engine that performs more
155 powerful data retrieval and analytical capabilities. All released data are publicly
156 accessible and downloadable via FTP and HTTP, but controlled data in GSA-Human
157 and OMIX require access permission. To access the controlled data, requester needs
158 to create a request and send required documents for data access. Once the request has
159 been reviewed and approved, the requester gains the access to the data.

160 **Data statistics**

161 The GSA family has received a large number of data submissions with explosive
162 growth in data and users, thus exhibiting their important roles in raw data

163 management (**Figure 1** and **Table 2**). The volume of archived data has increased by
164 more than 40 times, compared to the 200 TB archived in the previous release of GSA
165 [1]. Till June 2021, GSA and GSA-Human have collected 324,325 Experiments,
166 371,973 Runs and more than 8.5 PB of data submitted from 1530 submitters of 385
167 organizations (Figure 2A). In particular, GSA-Human has archived 61,225 individuals
168 and housed 4.9 PB of raw sequence data within less than one year, clearly showing
169 that human genetic data are growing at an unprecedented rate and scale. More
170 importantly, GSA-Human has received a total of 721 access requests from 485
171 requesters, with 178 requests approved till June 2021. Regarding the trend of archived
172 data over time, it is observed that it took about three years to accumulate the first PB
173 of data and currently reaches to 8.5 PB in just over two and a half years, with a
174 formidably dramatic decrease in days for data accumulation (Figure 2B). Strikingly,
175 the third PB volume took only 30 days, principally contributed by a large-scale
176 sequencing project [13] with 344 TB of data archived. Meanwhile, the number of
177 species involved is also on a rapid increase, from 80 in December 2016 to more than
178 1000 at present. Also, albeit newly established, OMIX has collected 160 files of 801
179 GB.

180 Currently, the GSA family has more than 5377 registered users and has been
181 visited by 648,274 unique IPs from 111 countries/regions, with a total of 35,010,529
182 page views and an average of 4 TB of downloads per day. Data housed in these
183 resources have been reported in more than 239 scientific
184 journals(<https://ngdc.cncb.ac.cn/gsa/statistics?active=journals>), including Cell,
185 Genome Research, Genomics Proteomics Bioinformatics, Nature, Plant Cell and
186 PNAS. More importantly, with frequent updates and improvements in the past several
187 years, GSA has been recognized as one of the certified repositories in
188 FAIRsharing.org and re3data.org, and therefore meets the requirement as a supported
189 repository by Elsevier, Taylor & Francis, and Wiley. More detailed statistics can be
190 found online at <https://ngdc.cncb.ac.cn/gsa/standards>.

191 **Future directions**

192 The explosive volume of raw data submitted to the GSA family is still on the increase,
193 posing significant challenges to handle and share such big data [14]. Nowadays,
194 CNCB-NGDC, hosting a suite of database resources including the GSA family, is
195 going to be enhanced by national big data infrastructure, with stable governmental
196 funding investment in upgrading storage, computing and network resources, thus
197 providing fundamental support in raw data archive and management of the GSA
198 family. In addition, our future efforts will be made in continuous optimization of data
199 models and curation processes in evolution of users' needs, establishment of cloud
200 infrastructure for big data storage, and development of a variety of tools to facilitate
201 big data submission and high-speed transfer. To make effective use of human genetic
202 data and promote precision healthcare and treatment, efforts will also be devoted to
203 optimizing procedures and mechanisms to enable data sharing with controlled access
204 and security by conforming to applicable regulations and ethical norms. We also
205 advocate worldwide collaborations in developing data standards, tools and approaches
206 towards global biodiversity & health big data sharing (BHBD alliance;
207 <http://bhbd-alliance.org/>).

208 **CRedit author statement**

209 Tingting Chen: Investigation, Methodology, Data Curation, Writing - Original Draft.
210 Xu Chen: Software. Sisi Zhang: Investigation, Methodology, Data Curation, Writing -
211 Original Draft. Junwei Zhu: Software. Bixia Tang: Software. Anke Wang: Writing -
212 Original Draft, Software. Lili Dong: Data Curation. Zhewen Zhang: Data Curation.
213 Caixia Yu: Data Curation. Yanling Sun: Data Curation. Lianjiang Chi: Software.
214 Huanxin Chen: Resources. Shuang Zhai: Resources. Yubin Sun: Resources. Li Lan:
215 Resources. Xin Zhang: Resources. Jingfa Xiao: Writing - Review & Editing. Yiming
216 Bao: Conceptualization, Writing - Review & Editing, Funding acquisition. Yanqing
217 Wang: Conceptualization, Investigation, Methodology, Software, Writing - Review &
218 Editing, Project administration. Zhang Zhang: Conceptualization, Writing - Review &

219 Editing, Funding acquisition. Wenming Zhao: Conceptualization, Methodology,
220 Writing - Review & Editing, Supervision, Funding acquisition.

221 **Competing interests**

222 The authors have declared no competing interests.

223 **Acknowledgments**

224 We sincerely thank Prof. Jingchu Luo and Prof. Weimin Zhu for their valuable
225 suggestions and a number of users for their contributions to the data submission. We
226 also thank Changrui Feng and Zhuojing Fan for their assistance in drawing the figures
227 of the manuscript. This work was supported by grants from Strategic Priority
228 Research Program of Chinese Academy of Sciences [XDB38060100 and
229 XDB38030200 to Y.B.; XDB38050300 to W.Z.; XDB38030400 to J.X.;
230 XDA19050302 to Z.Z.]; National Key Research and Development Program of China
231 [2016YFC0901603 to W.Z.; 2017YFC1201202 to Y.W.; 2020YFC0847000 and
232 2018YFD1000505 to W.Z.; 2016YFE0206600 to Y.B.; 2017YFC0907502 to Z.Z.];
233 The 13th Five-year Informatization Plan of Chinese Academy of Sciences
234 [XXH13505-05 to Y.B.]; Genomics Data Center Construction of Chinese Academy of
235 Sciences [XXH-13514-0202 to Y.B.]; Open Biodiversity and Health Big Data
236 Programme of IUBS [to Y.B.]; The Professional Association of the Alliance of
237 International Science Organizations [ANSO-PA-2020-07 to Y.B.]; National Natural
238 Science Foundation of China [32030021 and 31871328 to Z.Z.]; International
239 Partnership Program of the Chinese Academy of Sciences [153F11KYSB20160008 to
240 Z.Z.].

241 **ORCID**

242 ORCID: 0000-0003-1296-3093 (Tingting Chen)

243 ORCID: 0000-0001-6102-1751 (Xu Chen)

244 ORCID: 0000-0002-3852-4796 (Sisi Zhang)

245 ORCID: 0000-0003-4689-3513 (Junwei Zhu)

- 246 ORCID: 0000-0002-9357-4411 (Bixia Tang)
- 247 ORCID: 0000-0002-2565-2334 (Anke Wang)
- 248 ORCID: 0000-0003-0953-6306 (Lili Dong)
- 249 ORCID: 0000-0002-9422-822X (Zhewen Zhang)
- 250 ORCID: 0000-0002-3882-9979 (Caixia Yu)
- 251 ORCID: 0000-0002-3175-3625 (Yanling Sun)
- 252 ORCID: 0000-0003-4836-0577 (Lianjiang Chi)
- 253 ORCID: 0000-0003-1293-4495 (Huanxin Chen)
- 254 ORCID: 0000-0002-2084-7132 (Shuang Zhai)
- 255 ORCID: 0000-0003-3810-7156 (Yubin Sun)
- 256 ORCID: 0000-0002-4761-2245 (Li Lan)
- 257 ORCID: 0000-0002-2300-1036 (Xin Zhang)
- 258 ORCID: 0000-0002-2835-4340 (Jingfa Xiao)
- 259 ORCID: 0000-0002-9922-9723 (Yiming Bao)
- 260 ORCID: 0000-0002-7985-7941 (Yanqing Wang)
- 261 ORCID: 0000-0001-6603-5060 (Zhang Zhang)
- 262 ORCID: 0000-0002-4396-8287 (Wenming Zhao)
- 263

264 **References**

- 265 [1] Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, et al. GSA: Genome Sequence
266 Archive. *Genomics Proteomics Bioinformatics* 2017;15:14–8.
- 267 [2] Song S, Zhang Z. Database Resources in BIG Data Center: Submission,
268 Archiving, and Integration of Big Data in Plant Science. *Mol Plant* 2019;12:279–81.
- 269 [3] National Genomics Data Center Members and Partners. Database Resources of the
270 National Genomics Data Center in 2020. *Nucleic Acids Res* 2020;48:D24–D33.
- 271 [4] CNGB-NGDC Members & Partners. Database Resources of the National
272 Genomics Data Center, China National Center for Bioinformation in 2021. *Nucleic*
273 *Acids Res* 2021;49:D18–D28.
- 274 [5] Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al.
275 Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci U*
276 *S A* 2018;115:4325–33.
- 277 [6] Tang B, Zhou Q, Dong L, Li W, Zhang X, Lan L, et al. iDog: an integrated
278 resource for domestic dogs and wild canids. *Nucleic Acids Res* 2019;47:D793–D800.
- 279 [7] Miao W, Song L, Ba S, Zhang L, Guan G, Zhang Z, et al. Protist 10,000 Genomes
280 Project. *The Innovation* 2020;1.
- 281 [8] Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, et al. NCBI's Database
282 of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* 2014;42:D975–9.
- 283 [9] Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, et al. Database
284 resources of the National Center for Biotechnology Information. *Nucleic Acids Res*
285 2021;49:D10–D7.
- 286 [10] Lappalainen I, Almeida-King J, Kumanduri V, Senf A, Spalding JD, Ur-Rehman
287 S, et al. The European Genome-phenome Archive of human data consented for
288 biomedical research. *Nat Genet* 2015;47:692–5.
- 289 [11] Cantelli G, Cochrane G, Brooksbank C, McDonagh E, Flicek P, McEntyre J, et
290 al. The European Bioinformatics Institute: empowering cooperation in response to a
291 global health crisis. *Nucleic Acids Res* 2021;49:D29–D37.
- 292 [12] Cochrane G, Karsch-Mizrachi I, Takagi T, International Nucleotide Sequence
293 Database Collaboration. The international nucleotide sequence database collaboration.
294 *Nucleic Acids Res* 2016;44:D48–50.
- 295 [13] Li J, Xu C, Lee HJ, Ren S, Zi X, Zhang Z, et al. A genomic and epigenomic atlas
296 of prostate cancer in Asian populations. *Nature* 2020;580:93–9.
- 297 [14] Zhang Z, Song S, Yu J, Zhao W, Xiao J, Bao Y. The elements of data sharing.
298 *Genomics Proteomics Bioinformatics* 2020;18:1–4.

299

300

301 **Figure legends**

302 **Figure 1 Data model of the GSA family**

303 GSA data structure has been significantly changed. BioProject and BioSample have
304 been separated from GSA, serving as independent meta-information databases and
305 acting as an organizational framework to provide centralized access to descriptive
306 metadata about research projects and samples, respectively. GSA-Human is used to
307 archive human genetic resources data and OMIX is used for various non-sequencing
308 types of data management.

309

310 **Figure 2 Data statistics of the GSA family**

311 **A.** Number of runs accumulated from 2016 to 2021, with five major species indicated.

312 **B.** Trend of submitted data volume in association with days involved. All statistics
313 were derived from GSA and GSA-Human as of June 2021.

314

315 **Tables**

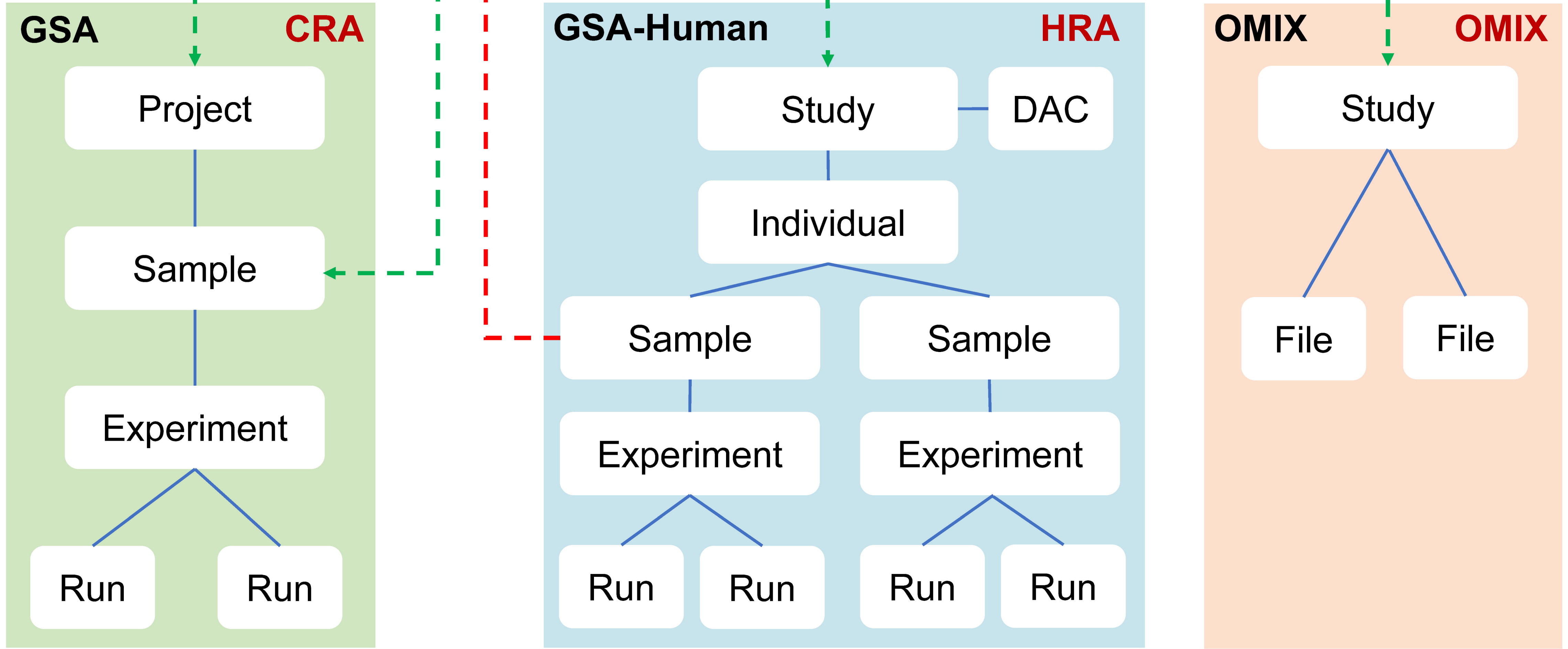
316 **Table 1 Comparison between GSA in 2017 and the GSA family in 2021**

317 **Table 2 Data items of the GSA family**

BioProject

BioSample

bioRxiv preprint doi: <https://doi.org/10.1101/2021.06.29.449849>; this version posted July 1, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



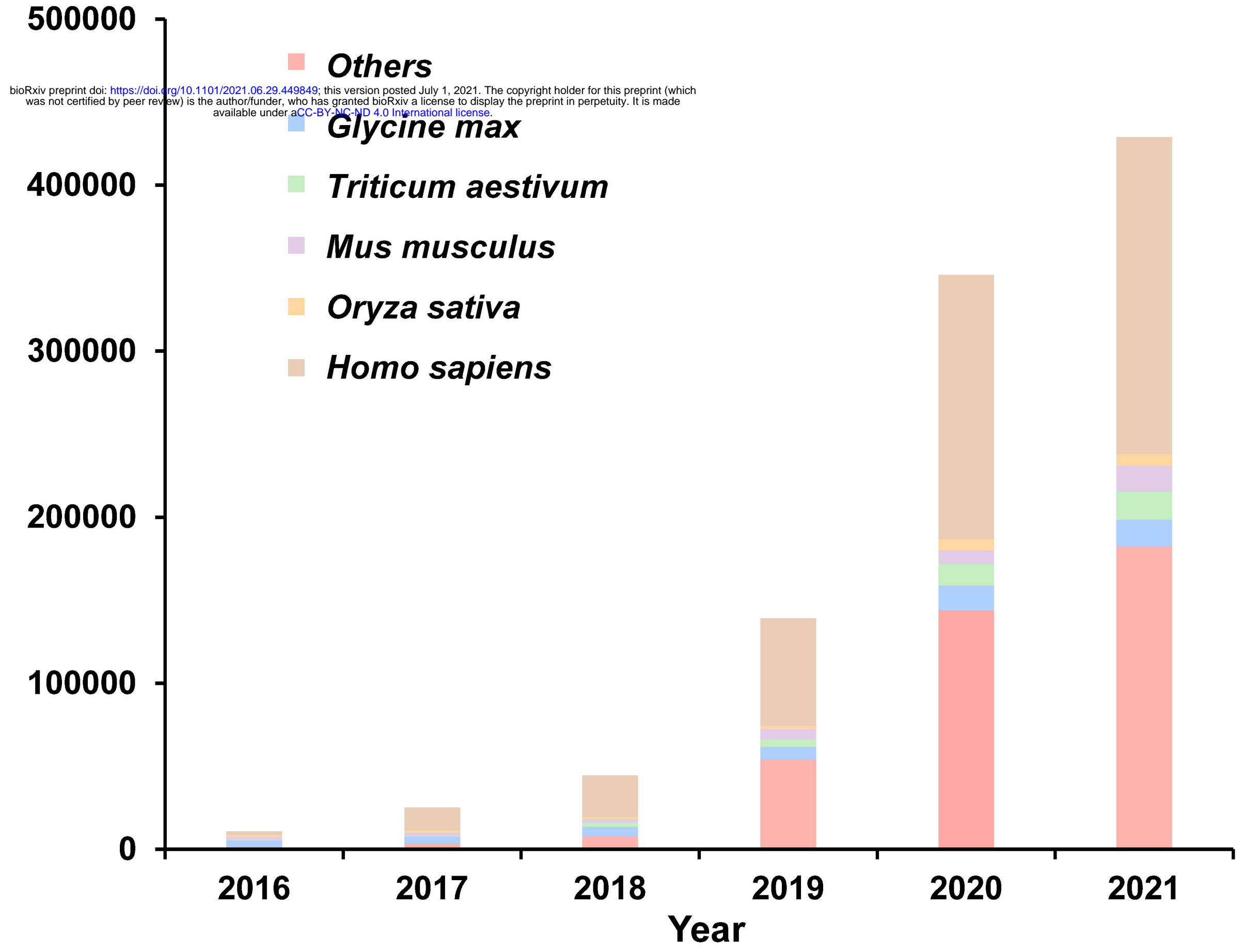
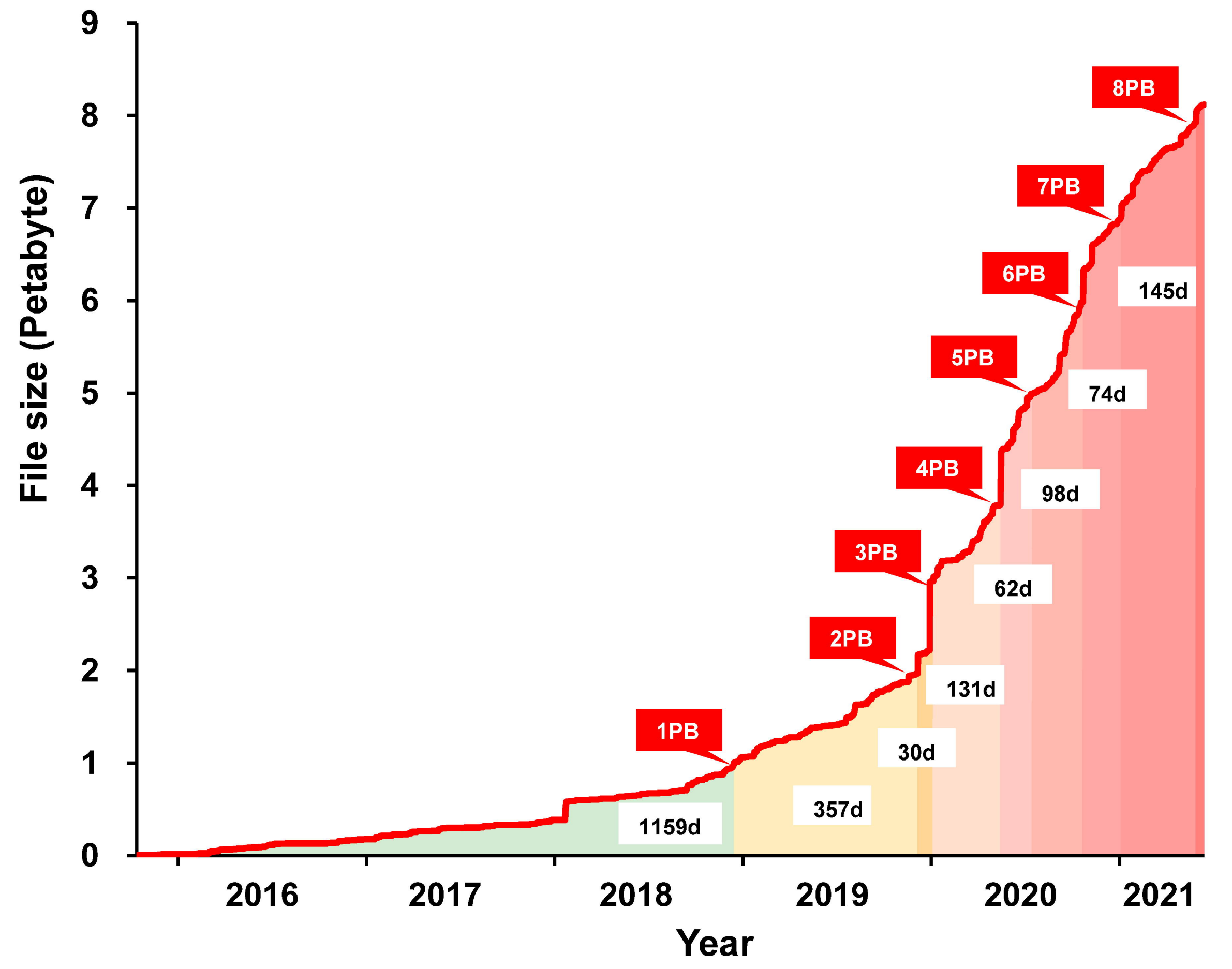
A**B**

Table 1 Comparison between GSA in 2017 and the GSA family in 2021

Category	2017	2021
Archival resources	GSA	GSA, GSA-Human, OMIX
Number of supported sample types*	7	11
Batch submission	NA	Available
Data statistics	NA	Available
Supported languages	English	English, Chinese
Controlled access	NA	Available
Data transfer	FTP	FTP, Aspera
Number of supported sequencing platforms*	49	66
Number of supported data formats*	9	13
Quality control*	Metadata	Metadata, Data

* More details are available at <https://ngdc.cncb.ac.cn/gsa/standards>.

Table 2 Data items of the GSA family

Item*	GSA	GSA-Human	OMIX	Total
BioProjects	2398	537	83	2920
Individuals	/	61,225	/	61,225
BioSamples	241,360	125,715	/	367,075
Experiments	178,670	145,655	/	324,325
Runs	195,298	176,675	/	371,973
File size (Tbyte)	3545	4980	0.888	8526
Registered users		4610		

* All statistics were derived from the GSA family as of June 2021.