1 **Genome Warehouse: A Public Repository Housing**

2 **Genome-scale Data**

3

4 Meili Chen[1,2,#], Yingke Ma[1,2,#], Song Wu[1,2,3], Xinchang Zheng[1,2], Hongen Kang[1,2,3],

5 Jian Sang[1,2,3,†], Xingjian Xu[1,2,3,††], Lili Hao[1,2], Zhaohua Li[1,2,3], Zheng Gong[1,2,3], Jingfa

6 Xiao[1,2,3], Zhang Zhang[1,2,3], Wenming Zhao[1,2,3], Yiming Bao[1,2,3,*]

7 *[1] National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of*

8 *Sciences / China National Center for Bioinformation, Beijing 100101, China*

9 *[2] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of*

10 *Genomics, Chinese Academy of Sciences, Beijing 100101, China*

11 *[3] University of Chinese Academy of Sciences, Beijing 100049, China*

12

13 [#] Equal contribution.

14 * Corresponding author.

15 E-mail: baoym@big.ac.cn (Bao Y).

16 *[†] Current address: Division of Cancer Epidemiology and Genetics, National Cancer*

17 *Institute, National Institutes of Health, Bethesda, Maryland 20892, USA*

18 *[††] Current address: College of Computer Science Technology, Inner Mongolia*

19 *Normal University, Hohhot, Inner Mongolia 010010, China*

20

21 **Running title:** *Chen M et al / Genome Assembly Data Repository*

22

23 Total letter counts (Title): 63

24 Total letter counts (Running title): 46

25 Total word counts (Abstract): 193

26 Total keywords: 5

27 Total word counts (from "Introduction" to "Conclusions" or "Materials and

28 methods"): 1799

29 Total figures: 3

30   Total tables: 1

31   Total supplementary figures: 0

32   Total supplementary tables: 0

33   Total supplementary files: 0

34

35

Abstract

The Genome Warehouse (GWH) is a public repository housing genome assembly data for a wide range of species and delivering a series of web services for genome data submission, storage, release, and sharing. As one of the core resources in the National Genomics Data Center (NGDC), part of the China National Center for Bioinformation (CNCB, https://bigd.big.ac.cn/), GWH accepts both full genome and partial genome (chloroplast, mitochondrion, and plasmid) sequences with different assembly levels, as well as an update of existing genome assemblies. For each assembly, GWH collects detailed genome-related metadata including biological project and sample, and genome assembly information, in addition to genome sequence and annotation. To archive high-quality genome sequences and annotations, GWH is equipped with a uniform and standardized procedure for quality control. Besides basic browse and search functionalities, all released genome sequences and annotations can be visualized with JBrowse. By December 2020, GWH has received 17,264 direct submissions covering a diversity of 949 species, and has released 3370 of them. Collectively, GWH serves as an important resource for genome-scale data management and provides free and publicly accessible data to support research activities throughout the world. GWH is publicly accessible at https://bigd.big.ac.cn/gwh/.

KEYWORDS: Genome submission; Genome sequence; Genome annotation; Genome warehouse; Quality control

Introduction

59

60 Genome sequences and annotations are fundamental information for a wide range of

61 genome-related studies, including various omics data analysis such as genome [1],

62 transcriptome [2], epigenome [3,4], and genome variation [5,6]. China, as one of the

63 most biodiverse countries in the world, harbors more than 10% of the world's known

64 species [7]. In the past decades, a large number of genome assemblies of featured and

65 important animals and crops in China have been sequenced [1, 8–11], most of which

66 were submitted to International Nucleotide Sequence Database Collaboration (INSDC)

67 members (National Center for Biotechnology Information (NCBI), European

68 Bioinformatics Institute (EBI), and DNA Data Bank of Japan (DDBJ)) [12]. With the

69 rapid growth of genome assembly data, in China for example, large genome data size,

70 slow data transfer rate due to limited international network transfer bandwidth, and

71 language barrier for communication of technical issues have obstructed researchers

72 from efficiently submitting their data to INSDC members. All these call for a

73 centralized genomic data repository within China to complement the INSDC.

74 Here, we report the Genome Warehouse (GWH, https://bigd.big.ac.cn/gwh/), a

75 centralized resource housing genome assembly data and delivering a series of genome

76 data services. As one of the core resources in the National Genomics Data Center

77 (NGDC), part of the China National Center for Bioinformation (CNCB,

78 https://bigd.big.ac.cn/) [13], the aim of GWH is to accept data submissions worldwide

79 and provide an important resource for genome data quality control, data archive, rapid

80 release, and public sharing (*e.g.*, with INSDC) in support of research activities from

81 all over the world. To date, GWH has received a total of 12,366 genome submissions

82 (including 14 international submissions), demonstrating its increasingly important role

83 in global genome data management and sharing.

Data model

84

85 Designed for compatibility with the INSDC data model, each genome assembly in

86 GWH is linked to a BioProject (https://bigd.big.ac.cn/bioproject) and a BioSample

87 (https://bigd.big.ac.cn/biosample), which are two fundamental resources for metadata

88   description in CNCB-NGDC. Full or partial (chloroplast, mitochondrion, and plasmid)

89   genome assemblies with different assembly levels (complete, draft in chromosome,

90   scaffold, and contig) are all acceptable and existing genome assemblies are allowed to

91   be updated. Accession numbers are assigned with the following rules (**Figure 1**): (1)

92   each genome assembly has an accession number prefixed with "GWH", followed by

93   four capital letters and eight zeros (*e.g.*, GWHAAAA00000000); (2) genome

94   sequences have the same accession number format as their corresponding genome

95   assembly, with the exception that the eight digits start from 00000001 and increase in

96   order (*e.g.*, GWHAAAA00000001); (3) genes have similar accession pattern as those

97   of genome sequences, with the addition of letter "G" between the GWH prefix and the

98   four capital letters, and there are six digits at the end instead of eight (*e.g.*,

99   GWHGAAAA000001); (4) transcripts use the letter "T" to replace "G" in accession

100   numbers for genes (*e.g.*, GWHTAAAA000001); (5) proteins use the letter "P" to

101   replace "G" in accession numbers for genes (*e.g.*, GWHPAAAA000001); (6) if the

102   submission is an update of existing submission in GWH, it will be assigned a dot and

103   an incremental number to represent the version (*e.g.*, GWHAAAA00000000.1).

Database components

105   GWH is a centralized resource housing genome-scale data, with the purpose to

106   archive high-quality genome sequences and annotation information. GWH is

107   equipped with a series of web services for genome data submission, release, and

108   sharing, accordingly involving three major components, namely, data submission,

109   quality control, and archive and release (Figure 2).

Data submission

111   GWH not only accepts genome assembly associated data through an on-line

112   submission system but also allows off-line batch submissions. Users need to register

113   first and then to provide complete description on submitted genome sequences.

114   Biological project and sample information should be provided (through BioProject

115   and BioSample, respectively) together with genome assembly sequence, annotation,

116   and associated metadata. Metadata mainly consist of a variety of information about

117 submitter, general assembly, file(s), sequence assignment, and publication (if

118 available). After submission, GWH runs an automated quality control pipeline to

119 check the validity and consistency of submitted genome sequence and genome

120 annotation files. Accession numbers are assigned to assemblies and sequences upon

121 the pass of quality control. The updated assembly data can also be submitted to GWH.

122 It should be noted that compatible with the INSDC members (*e.g.*, NCBI GenBank), it

123 is the responsibility of the submitters to ensure the data quality, completeness, and

124 consistency and GWH does not warrant or assume any legal liability or responsibility

125 for the data accuracy.

126 **Quality control**

127 After metadata and file(s) are received, GWH automatically runs standardized quality

128 control (QC) to check 45 different types of errors in submitted genome sequences and

129 annotations, and to scan for contaminated genome sequences (see details at

130 https://bigd.big.ac.cn/gwh/documents) if needed (Figure 2), which roughly falls into 5

131 QC steps: (1) The component will check the consistency of file(s) according to

132 filename and md5 code. (2) For genome sequences, the component will check the

133 legality of genome sequence ID and sequence content, *e.g.*, unique sequence ID,

134 sequence composition (A/T/C/G or degenerate base), sequence length ($\geq$ 200 bp). (3)

135 For genome annotations, the component will check gene structure completeness and

136 consistency, *e.g.*, unique ID, a exon/CDS/UTR coordinate falling within the

137 corresponding gene coordinate, strand consistency for all features (including

138 gene/transcript/exon/CDS/UTR), codon validity (*e.g.*, valid start/stop codon, no

139 internal stop codon). (4) Finally, it will check the internal consistency of genome

140 sequence and annotation, *e.g.*, sequence ID in genome annotation must match genome

141 sequence ID, a feature coordinate falling within the range of the corresponding

142 genome sequence. (5) Genome sequences will also be scanned to check vectors,

143 adaptors, primers, and indices (collected from UniVec database,

144 ftp://ftp.ncbi.nlm.nih.gov/pub/UniVec/) using NCBI's VecScreen

145 (https://www.ncbi.nlm.nih.gov/tools/vecscreen/). If there is an error, a report will be

146 automatically sent to the submitter by email. To finish a successful submission, the

147 submitter needs to fix all errors and resubmit files until they pass the QC process.

148 **Archive and release**

149 GWH will assign a unique accession number to the submitted genome assembly upon

150 the pass of quality control, allot accession numbers for each genome sequence, gene,

151 transcript, and protein, generate and backup downloadable files of genome sequence

152 and annotation in FASTA, GFF3, and TSV formats. Data generation is performed

153 with in-house-writing scripts based on submitted genome sequence and annotation

154 files. In order to ensure the security of submitted data, a copy of backup data is stored

155 on a physically separate disk. GWH will release sequence data on a user-specified

156 date, unless a paper citing the sequence or accession number is published prior to the

157 specified release date, in which case the sequence will be released immediately. For

158 the released data, GWH will generate web pages containing two primary tables:

159 genome and assembly. The former shows species taxonomy information and genome

160 assemblies, and the latter contains general information of the assembly (including

161 external links to other related resources), statistics of genome assembly and its

162 corresponding annotation. All released data are publicly available at GWH FTP site

163 (ftp://download.big.ac.cn/gwh/). GWH provides data visualization for both genome

164 sequence and genome annotation using JBrowse [14]. It offers statistics and charts in

165 light of total holdings, assembly levels, genome representations, citing articles,

166 submitting organizations, sequencing platforms, assembly methods, and downloads.

167 GWH provides user-friendly web interfaces for data browse and query using BIG

168 Search [13], in order to help users find any released data of interest. For a released

169 genome assembly, GWH also provides machine-readable APIs (Application

170 Programming Interfaces) for publicly sharing and automatically obtaining information

171 on its associated BioProject, BioSample, genome, and assembly metadata and file

172 paths.

Global sharing of SARS-CoV-2 and coronavirus genomes

During the COVID-19 outbreak, GWH, in support of the 2019 Novel Coronavirus Resource (2019nCoVR) [15, 16] has received worldwide submissions of more than a thousand SARS-CoV-2 genome assemblies with standardized genome annotations [17], and has released 134 of them. To expand the international influence of data, 62 of the released sequences have been shared, with the submitters' permission, in GenBank [18] through a data exchange mechanism established with NCBI. In this model, GWH accessions are represented as secondary accessions in NCBI GenBank records, which are retrievable by the NCBI Entrez system. This model sets a good example for data sharing among different data centers.

In addition, GWH offers sequences of the Coronaviridae family to facilitate researchers to reach the data conveniently and thus to study the relationship between SARS-CoV-2 and other coronaviruses. To promote the data sharing and make all relevant information of the Coronaviridae readily available, GWH integrates genomic and proteomic sequences as well as their metadata information from NCBI [19], China National GeneBank Database (CNGBdb) [20], National Microbiology Data Center (NMDC) [21] and CNCB-NGDC. Duplicated records from different sources are identified and removed to gain a non-redundant dataset. As of December 31, 2020, the dataset has 83,095 nucleotide and 575,438 protein sequences of the Coronaviridae. Filters are implemented to narrow down the required Coronaviridae sequences using multiple conditions, including country/region, host, isolation source, length, and collection date. Both the metadata and sequences of the filtered results can be selected and downloaded as a separate file. The daily updated sequences and all sequences can also be downloaded from FTP (ftp://download.big.ac.cn/Genome/Viruses/Coronaviridae/).

Data statistics

By December, 2020, GWH has received 17,264 direct submissions covering a broad diversity of species (**Table 1**) with different assembly levels (Figure 3). These genome assemblies link to 301 BioProjects and 16,538 BioSamples, and are

202 submitted by 231 submitters from 61 institutions (including 5 international submitters

203 from 2 countries). There are a total of 3370 released submissions, which were

204 reported in 83 articles from 44 journals. GWH has over 135,000 visits from 153

205 countries/regions, with ~891,000 downloads. The amount of data, visits, and

206 downloads in the GWH has been on the dramatic increase over the past years, clearly

207 showing its great utility in genome-scale data management.

Summary and future directions

209 Collectively, GWH is a user-friendly portal for genome data submission, release, and

210 sharing associated with a matched series of services. The rapid growth of genome

211 assembly submissions demonstrates the great potential of GWH as an important

212 resource for accelerating the worldwide genomic research. With the aim to fully

213 realize the findability, accessibility, interoperability, and reusability (FAIR) of

214 genome data [22], GWH has made ongoing efforts, including but not limited to,

215 improvement of web interfaces for data submission, presentation, and visualization,

216 continuous integration of newly sequenced genomes, and development of useful

217 online tools to help users analyse genome data (such as BLAST [23]). Therefore, we

218 will put in more efforts to provide genome annotation services, especially for bacteria

219 and archaea genomes, with the particular consideration that uniform standardized

220 annotation determines the accuracy of downstream data analysis. Besides, we will

221 expand the Coronaviridae dataset to other important pathogens to improve the ability

222 of public health emergency response. Finally, we plan to share and exchange all

223 public genome assembly data with the INSDC members to provide comprehensive

224 data for researchers globally.

CRediT author statement

226 **Meili Chen:** Methodology, Software, Investigation, Data Curation, Writing - Original

227 Draft, Project administration. **Yingke Ma:** Software, Writing - Original Draft. **Song**

228 **Wu:** Software, Data Curation. **Xinchang Zheng:** Data Curation. **Hongen Kang:**

229 Software. **Jian Sang:** Investigation, Data Curation. **Xingjian Xu:** Software. **Lili Hao:**

230 Investigation. **Zhaohua Li:** Data Curation. **Zheng Gong:** Data Curation. **Jingfa Xiao:**

231 Writing - Review & Editing. **Zhang Zhang:** Writing - Review & Editing. **Wenming**

232 **Zhao:** Writing - Review & Editing. **Yiming Bao:** Conceptualization, Writing -

233 Review & Editing, Supervision.

Competing interests

235 The authors have declared no competing interests.

Acknowledgments

ORCID

255 ORCID: 0000-0003-0102-0292 (Chen Meili)

256 ORCID: 0000-0002-9460-4117 (Ma Yingke)

257 ORCID: 0000-0002-0923-639X (Wu Song)

258 ORCID: 0000-0001-5739-861X (Zheng Xinchang)

259 ORCID: 0000-0002-9581-1329 (Kang Hongen)

260   ORCID: 0000-0003-4953-3417 (Sang Jian)

261   ORCID: 0000-0002-4466-3821 (Xu Xingjian)

262   ORCID: 0000-0003-3432-7151 (Hao Lili)

263   ORCID: 0000-0002-2673-0103 (Li Zhaohua)

264   ORCID: 0000-0001-7285-2630 (Gong Zheng)

265   ORCID: 0000-0002-2835-4340 (Xiao Jingfa)

266   ORCID: 0000-0001-6603-5060 (Zhang Zhang)

267   ORCID: 0000-0002-4396-8287 (Zhao Wenming)

268   ORCID: 0000-0002-9922-9723 (Bao Yiming)

269

270

References

[1] Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, et al. Pan-genome of wild and cultivated soybeans. Cell 2020;182:162-76.e13.

[2] Guan Y, Chen M, Ma Y, Du Z, Yuan N, Li Y, et al. Whole-genome and time-course dual RNA-Seq analyses reveal chronic pathogenicity-related gene dynamics in the ginseng rusty root rot pathogen *Ilyonectria robusta*. Sci Rep 2020;10:1586.

[3] Li R, Liang F, Li M, Zou D, Sun S, Zhao Y, et al. MethBank 3.0: a database of DNA methylomes across a variety of species. Nucleic Acids Res 2018;46:D288–D95.

[4] Xiong Z, Li M, Yang F, Ma Y, Sang J, Li R, et al. EWAS Data Hub: a resource of DNA methylation array data and metadata. Nucleic Acids Res 2020;48:D890–D5.

[5] Song S, Tian D, Li C, Tang B, Dong L, Xiao J, et al. Genome Variation Map: a data repository of genome variations in BIG Data Center. Nucleic Acids Res 2018;46:D944–D9.

[6] Tang B, Zhou Q, Dong L, Li W, Zhang X, Lan L, et al. iDog: an integrated resource for domestic dogs and wild canids. Nucleic Acids Res 2019;47:D793–D800.

[7] McBeath J, McBeath JH. Biodiversity conservation in China: policies and practice. Journal of International Wildlife Law & Policy 2006;9:293–317.

[8] Fan H, Wu Q, Wei F, Yang F, Ng BL, Hu Y. Chromosome-level genome assembly for giant panda provides novel insights into Carnivora chromosome evolution. Genome Biol 2019;20:267.

[9] Xia Q, Zhou Z, Lu C, Cheng D, Dai F, Li B, et al. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science 2004;306:1937–40.

[10] Lin T, Xu X, Ruan J, Liu SZ, Wu SG, Shao XJ, et al. Genome analysis of *Taraxacum kok-saghyz Rodin* provides new insights into rubber biosynthesis. Natl Sci Rev 2018;5:78–87.

[11] Li C, Song W, Luo Y, Gao S, Zhang R, Shi Z, et al. The HuangZaoSi *m*aize genome provides insights into genomic variation and improvement history of maize. Mol Plant 2019;12:402–9.

[12] Arita M, Karsch-Mizrachi I, Cochrane G. The international nucleotide sequence database collaboration. Nucleic Acids Res 2021;49:D121–D4.

[13] Members C-N, Partners. Database resources of the National Genomics Data Center, China National Center for Bioinformation in 2021. Nucleic Acids Res 2021;49:D18–D28.

[14] Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol 2016;17:66.

[15] Zhao WM, Song SH, Chen ML, Zou D, Ma LN, Ma YK, et al. The 2019 novel coronavirus resource. Yi Chuan 2020;42:212–21.

[16] Song S, Ma L, Zou D, Tian D, Li C, Zhu J, et al. The global landscape of SARS-CoV-2 genomes, variants, and haplotypes in 2019nCoVR. Genomics, Proteomics & Bioinformatics 2020. [DOI: https://doi.org/10.1016/j.gpb.2020.09.001]

313   [17] Shean RC, Makhsous N, Stoddard GD, Lin MJ, Greninger AL. VAPiD: a
314   lightweight cross-platform viral annotation pipeline and identification tool to facilitate
315   virus genome submissions to NCBI GenBank. BMC Bioinformatics 2019;20:48.
316   [18] Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I.
317   GenBank. Nucleic Acids Res 2020;48:D84–D6.
318   [19] Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, et al. Database
319   resources of the National Center for Biotechnology Information. Nucleic Acids Res
320   2021;49:D10–D7.
321   [20] Chen FZ, You LJ, Yang F, Wang LN, Guo XQ, Gao F, et al. CNGBdb: China
322   National GeneBank DataBase. Yi Chuan 2020;42:799–809.
323   [21] Wu L, Sun Q, Desmeth P, Sugawara H, Xu Z, McCluskey K, et al. World data
324   centre for microorganisms: an information infrastructure to explore and utilize
325   preserved microbial strains worldwide. Nucleic Acids Res 2017;45:D611–D8.
326   [22] Zhang Z, Song S, Yu J, Zhao W, Xiao J, Bao Y. The elements of data sharing.
327   Genomics Proteomics Bioinformatics 2020;18:1–4.
328   [23] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al.
329   Gapped BLAST and PSI-BLAST: a new generation of protein database search
330   programs. Nucleic Acids Res 1997;25:3389–402.
331

332 **Figure legends**

333 **Figure 1   Data model in GWH**

334 Genome assembly accession number is prefixed with "GWH", followed by four

335 capital letters (represented by XXXX) and 8 zeros. For genome sequence accessions,

336 eight digits increase in order. For gene sequence, transcript sequence, and protein

337 sequence accessions, G, T, and P are followed by the GWH prefix, respectively, with

338 six digits at the end that increase in order.

339 **Figure 2   Major components in GWH data processing workflow**

340 **Figure 3   Statistics of genome assembly in GWH (as of December 31, 2020)**

341 **Tables**

342 **Table 1   Total data holdings in GWH**

| Status	Type	Animals	Plants	Fungi	Bacteria	Archaea	Viruses	Metagenomes	Others	Total
Released	Assembly	187 (5.55%)	210 (6.23%)	13 (0.39%)	220 (6.53%)	73 (2.17%)	701 (20.80%)	1957 (58.07%)	9 (0.27%)	3370
	Species	72 (19.41%)	139 (37.47%)	12 (3.23%)	106 (28.57%)	11 (2.96%)	19 (5.12%)	3 (0.81%)	9 (2.43%)	371
Unpublic	Assembly	6783 (48.82%)	926 (6.66%)	5 (0.04%)	68 (0.49%)	13 (0.09%)	939 (6.76%)	4702 (33.84%)	458 (3.30%)	13,894
	Species	22 (3.67%)	549 (91.50%)	5 (0.83%)	7 (1.17%)	2 (0.33%)	6 (1.00%)	5 (0.83%)	4 (0.67%)	600
Total	Assembly	6970 (40.37%)	1136 (6.58%)	18 (0.10%)	288 (1.67%)	86 (0.50%)	1640 (9.50%)	6659 (38.57%)	467 (2.71%)	17,264
	Species	92 (9.69%)	675 (71.13%)	16 (1.69%)	110 (11.59%)	13 (1.37%)	24 (2.53%)	7 (0.74%)	12 (1.26%)	949

343
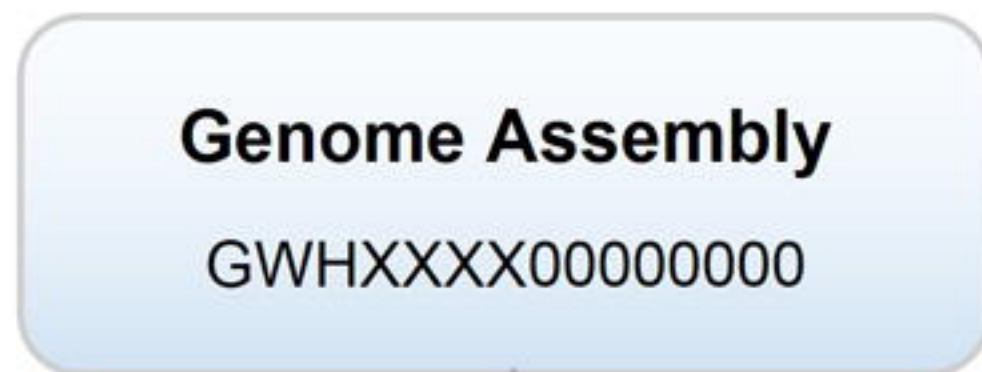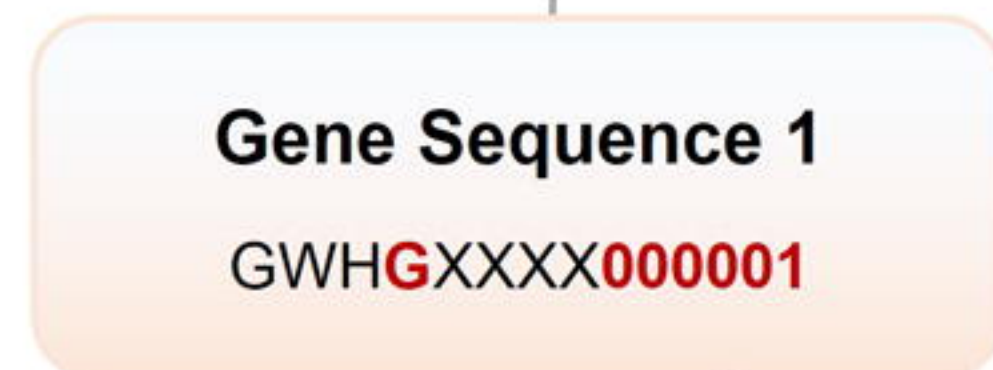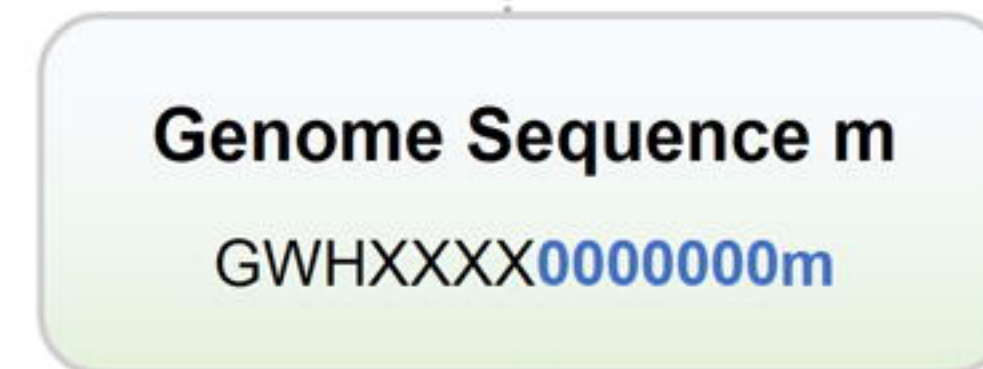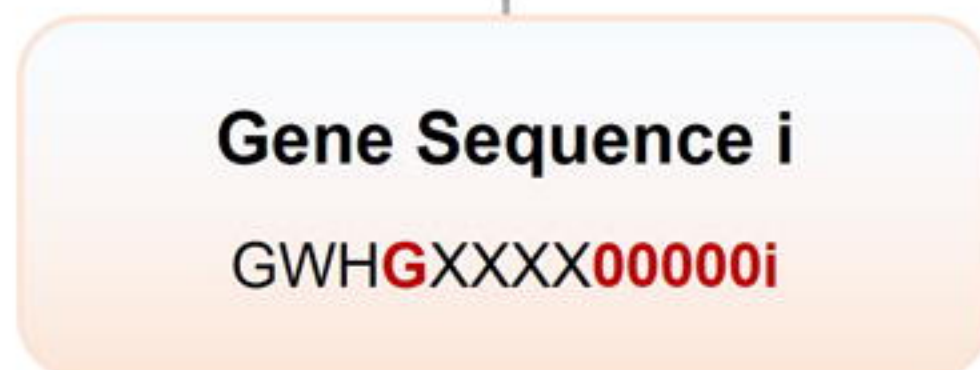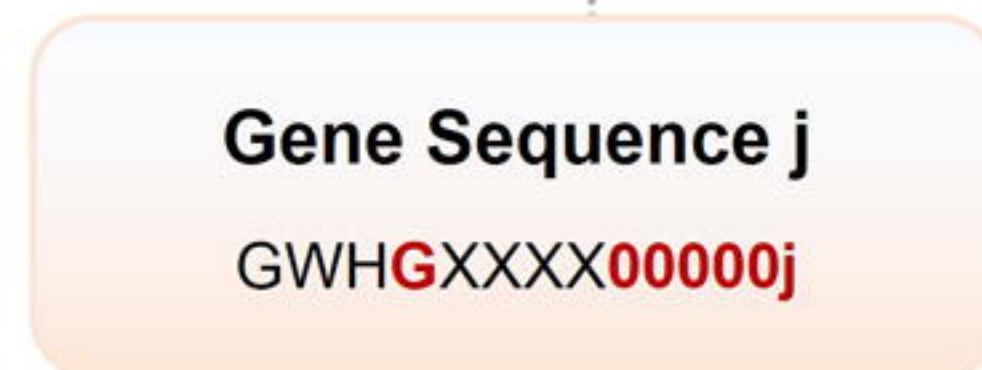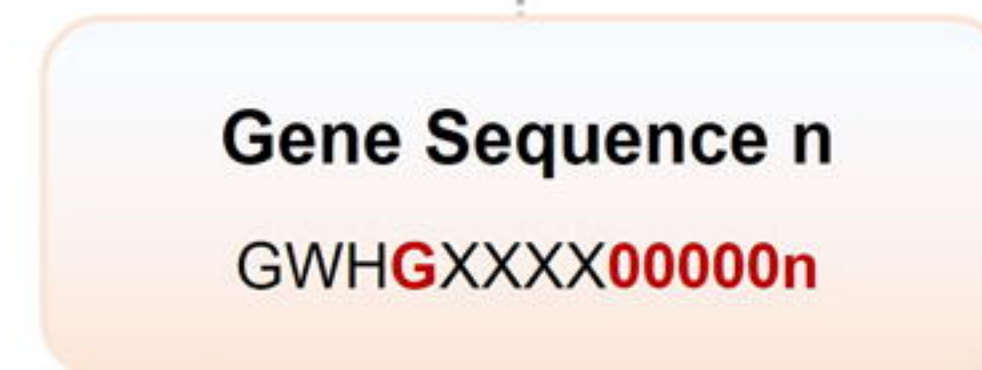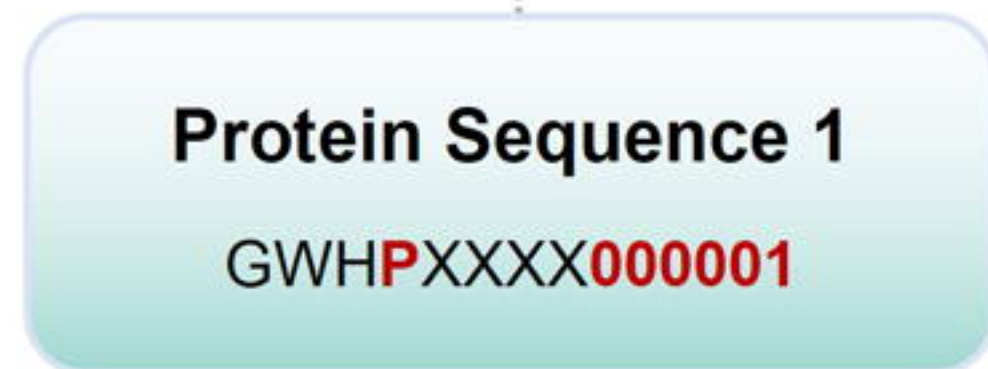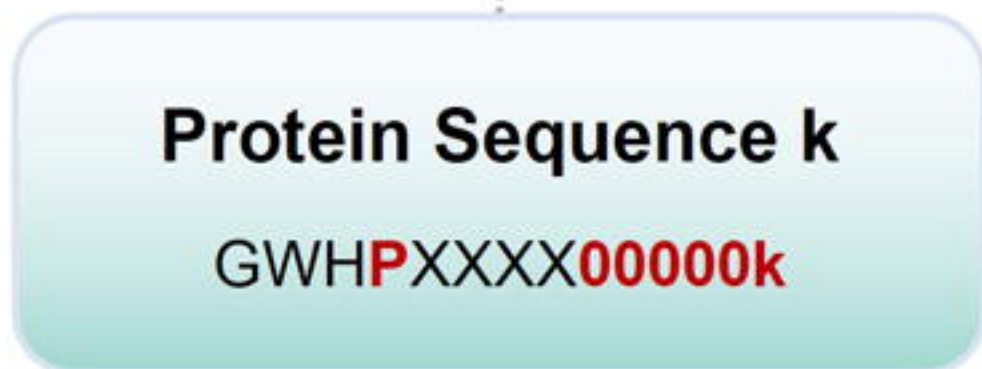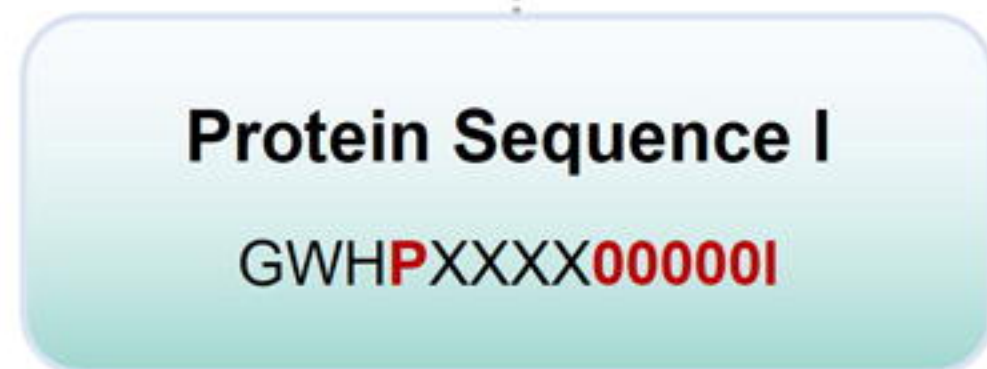
Data Submission

- Register account
- Register BioProject
- Register BioSample
- Create a submission
- Fill meta-data
- Upload files

Quality Control

- **Filename & md5 code**
- **Genome sequence**
  - Sequence ID
  - Sequence content
- **Genome annotation**
  - Gene structure completeness
  - Gene structure consistency
- **Data internal consistency**
- **Sequence contamination** (optional)

Archive & Release

Accession assignment
- Genome assembly
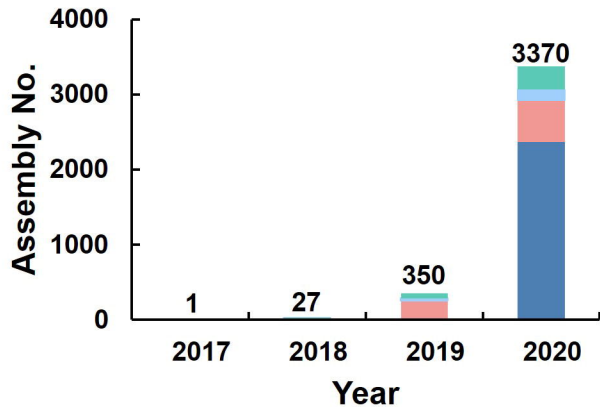- Genome sequence
- Gene
- Transcript
- Protein

Generation of downloadable files & backup
- Genome sequence
- Genome annotation
- Gene feature
- RNA sequence
- CDS sequence
- Protein sequence

Release & sharing
- Genome
- Assembly
- Download files
- Genome browser
- BIG Search

A **Total Assembly**
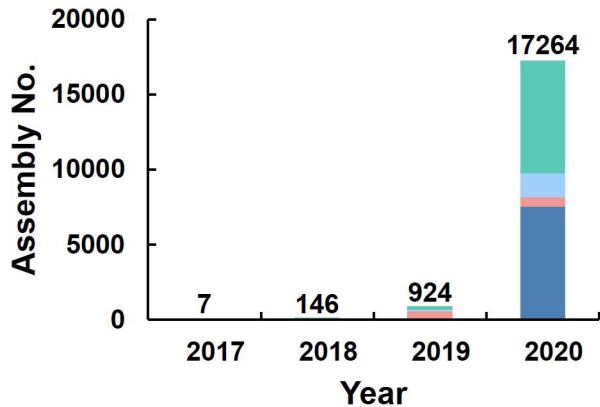
B **Released Assembly**

Assembly level: ■ Contig ■ Scaffold ■ Chromosome ■ Complete