中国科学院北京基因组研究所（国家生物信息中心）
BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

基因组数据汇交共享及注释

陈梅丽 高级工程师

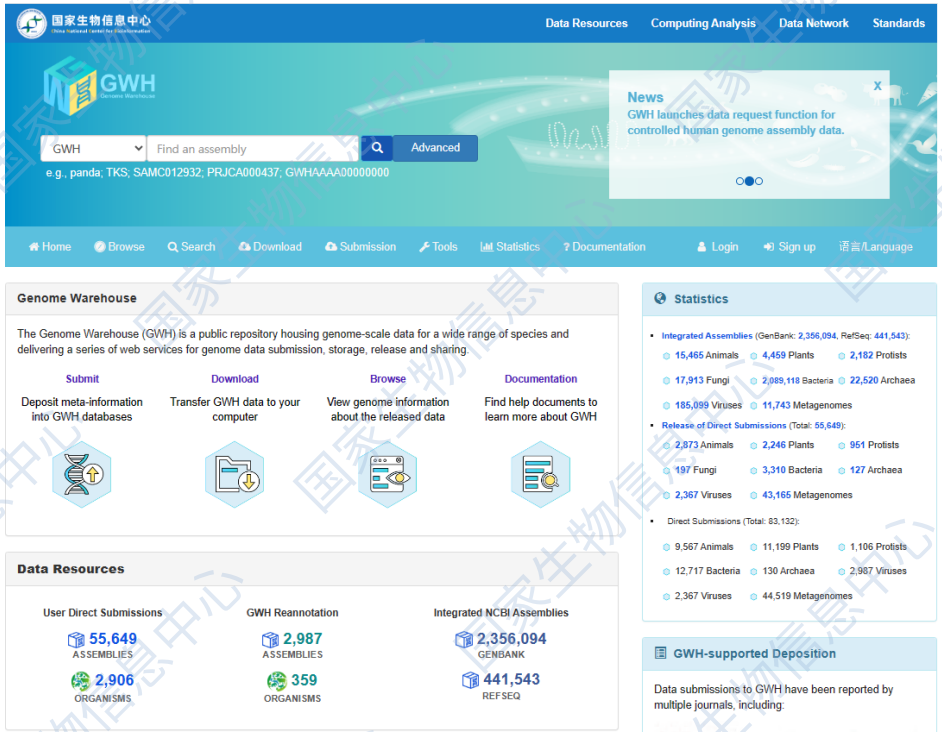
2024-10-26

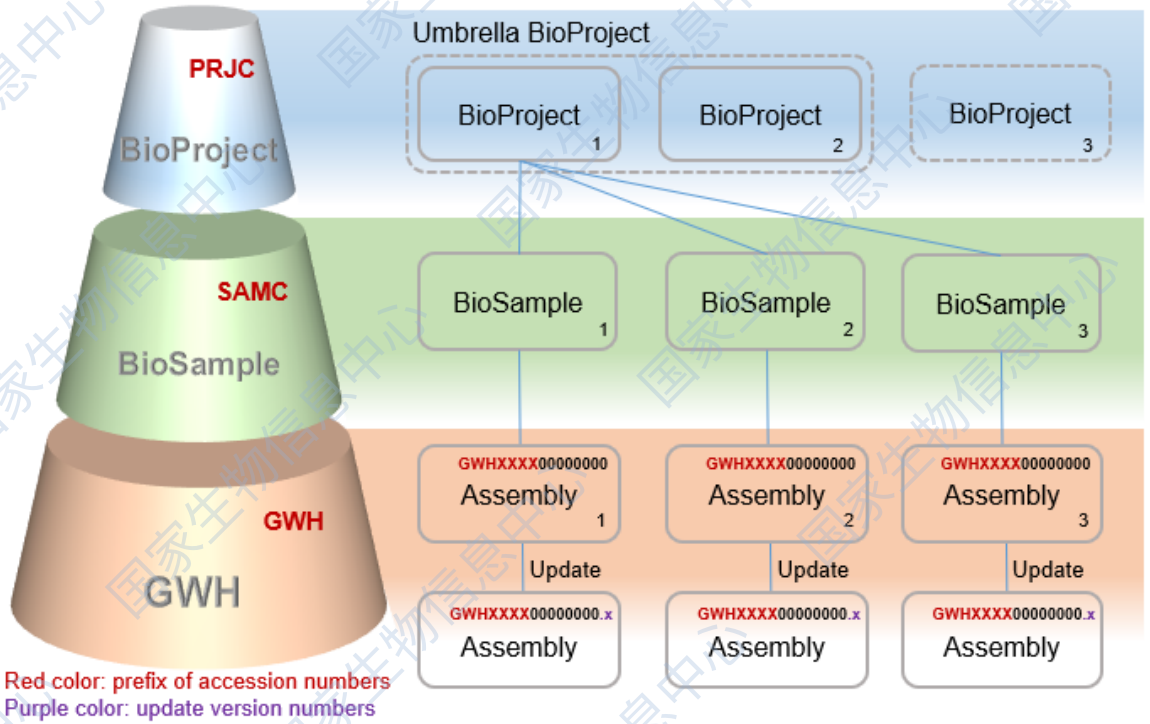
国家基因组科学数据中心
National Genomics Data Center

目录

基因组数据库

围绕基因组大数据汇交与管理，研发基因组数据库（Genome Warehouse，简称GWH），提供基因组组装数据汇交、存储、质控、发布和共享等全链条数据服务





https://ngdc.cncb.ac.cn/gwh/

参考INSDC对基因组组装数据进行结构化的组织和管理

完善的基因组数据管理平台

Genome Data Standards

1 ID System
2 File Format
- 2.1 Submitted data
  - 2.1.1 Data format
  - 2.1.2 Data format description
- 2.2 Downloaded data

3 Metadata
- 3.1 Submitter
- 3.2 Gneral information
- 3.3 Genome assembly files
- 3.4 Sequence assignment
- 3.5 Reference

4 Data Analysis
- 4.1 Initial validation processing
  - 4.1.1 Genome sequence
  - 4.1.2 Genome annotation
  - 4.1.3 Sequence ordering and orientation information
  - 4.1.4 Sequence assignment
- 4.2 Sequence contamination processing

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

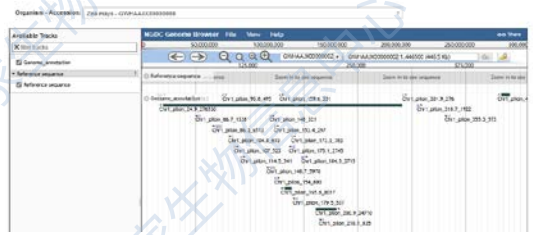
Accession assignment
- Genome assembly
- Genome sequence
- Gene
- Transcript
- Protein

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

Archival management
- Genome
- Assembly
- Download files
- Genome browser
- BIG Search

Release

Visualization



汇交 → 存储 → 质控 → 归档 → 发布 → 共享

与国际数据库保持相同的数据及编码规范

INSDC

International Nucleotide Sequence Database Collaboration

➢ BioProject

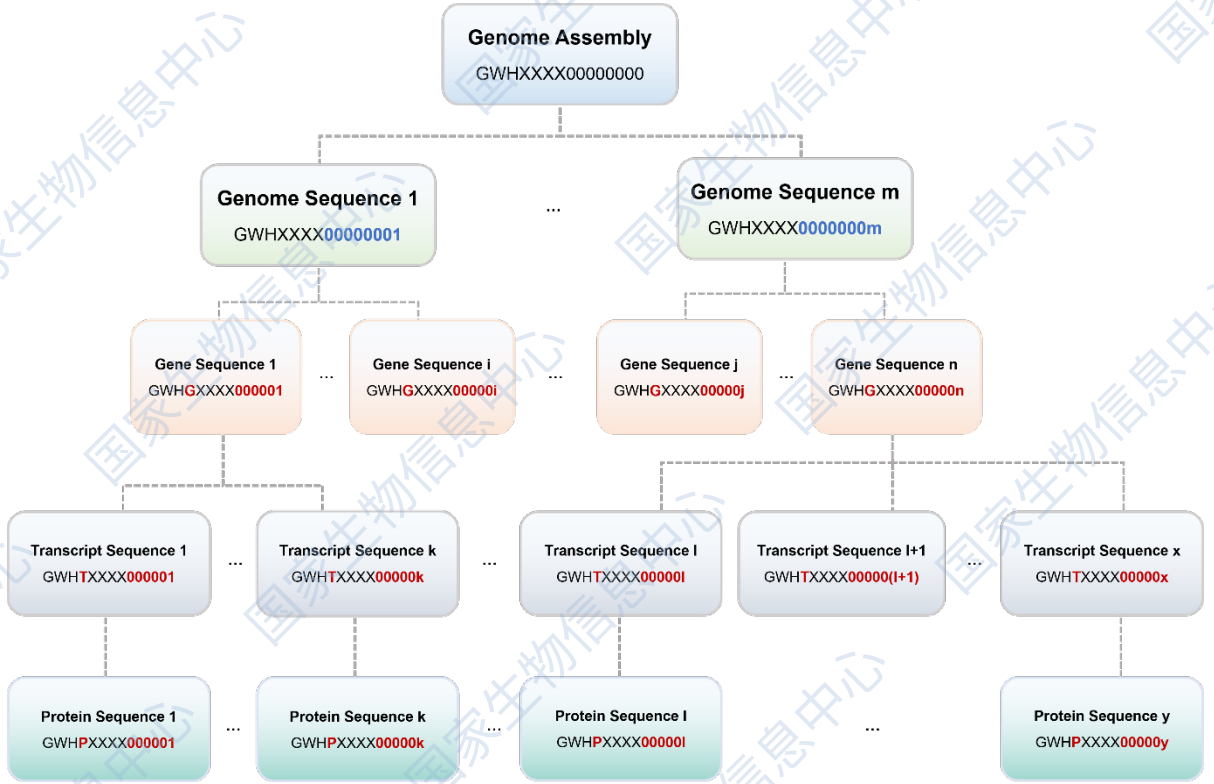
PRJNA – PRJNZ	NCBI
PRJEA – PRJEZ	EBI
PRJDA – PRJDZ	DDBJ
PRJCA – PRJCZ	**NGDC**

➢ BioSample

SAMN	NCBI
SAME	EBI
SAMD	DDBJ
SAMC	**NGDC**

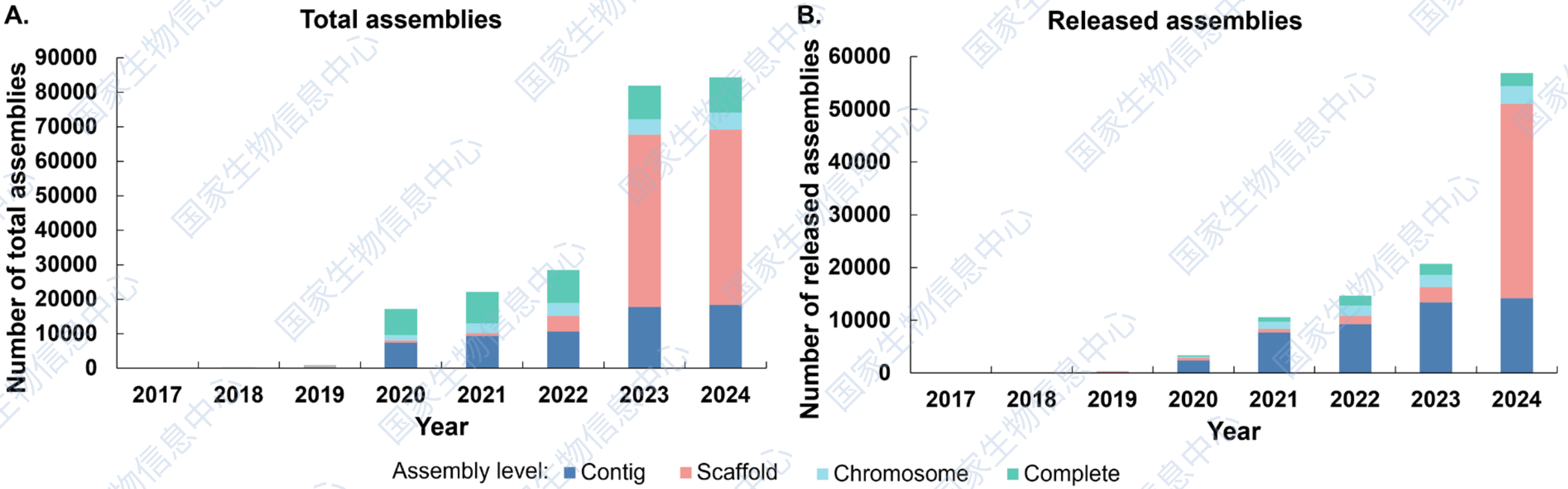
➢ Genome assembly

XXXX00000000	NCBI/EBI/DDBJ
GWHXXXX00000000	**NGDC**



实现国内基因组组装数据统一汇交和安全管理

- 1,482 BioProjects
- 84,422 Genome assembly
- 237,215,095 Sequences
- 44,273 BioSamples
- 4,767 Organisms
- 4,418 Gbases



支撑国家科技项目所产数据的自主管理

Funding agency	Grant number	GWH assembly number
MOST	112	6,143
NSFC	568	18,009
CAS	75	5,635
Other	817	61,649

No. NGDC-2023-0149

科技计划项目数据汇交证明

国家基因组科学数据中心已归档“作物黄萎病菌群体遗传变异机制与流行检测技术研究”项目（资助编号：2018YFE0112500，项目负责人：戴小枫）提交的数据集 160 个，总数据量 849.29 GB。

国家基因组科学数据中心

National Genomics Data Center (NGDC)

2023 年 1 月 31 日

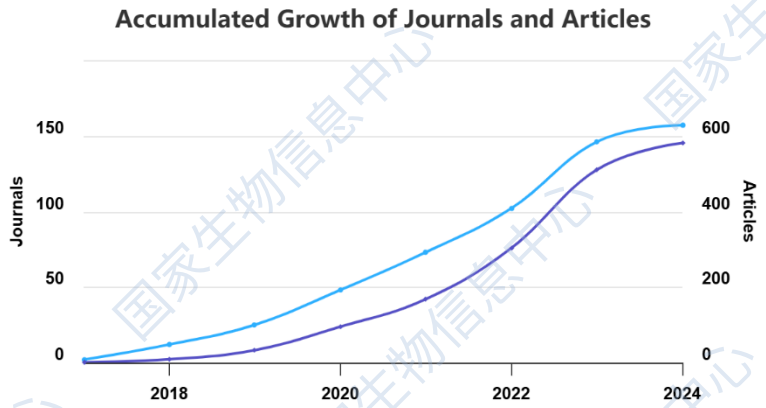
附件清单

接收数据资源清单：

数据集编号	数据集名称	样本数	数据量	数据格式	共享方式	发布时间
CRA009563	Verticillium dahliae GSA data	159	844.08GB	fastq	公开访问	2025/01/11
GWHBQUA00000000	DK001	1	34.88MB	fasta	公开访问	2025/01/14
GWHBQUB00000000	DK002	1	34.10MB	fasta	公开访问	2025/01/14
GWHBQUC00000000	DK003	1	34.33MB	fasta	公开访问	2025/01/14
GWHBQUD00000000	DK004	1	33.83MB	fasta	公开访问	2025/01/14
GWHBQUE00000000	DK005	1	33.49MB	fasta	公开访问	2025/01/14
GWHBQUF00000000	DK006	1	33.28MB	fasta	公开访问	2025/01/14
GWHBQUG00000000	DK008	1	34.07MB	fasta	公开访问	2025/01/14
GWHBQUH00000000	DK009	1	33.17MB	fasta	公开访问	2025/01/14
GWHBQUI00000000	DK010	1	32.48MB	fasta	公开访问	2025/01/14
GWHBQUJ00000000	DK011	1	33.70MB	fasta	公开访问	2025/01/14
GWHBQUK00000000	DK012	1	33.47MB	fasta	公开访问	2025/01/14
GWHBQUL00000000	DK013	1	33.57MB	fasta	公开访问	2025/01/14
GWHBQUM00000000	DK014	1	33.46MB	fasta	公开访问	2025/01/14
GWHBQUN00000000	DK015	1	33.50MB	fasta	公开访问	2025/01/14
GWHBQUO00000000	DK016	1	33.13MB	fasta	公开访问	2025/01/14
GWHBQUP00000000	DK017	1	33.55MB	fasta	公开访问	2025/01/14
GWHBQUQ00000000	DK018	1	33.49MB	fasta	公开访问	2025/01/14
GWHBQUR00000000	DK019	1	33.64MB	fasta	公开访问	2025/01/14
GWHBQUS00000000	DK020	1	33.58MB	fasta	公开访问	2025/01/14
GWHBQUT00000000	DK021	1	33.45MB	fasta	公开访问	2025/01/14
GWHBQUU00000000	DK022	1	32.80MB	fasta	公开访问	2025/01/14

支撑科技论文发表

期刊数: 157　　文章数：581



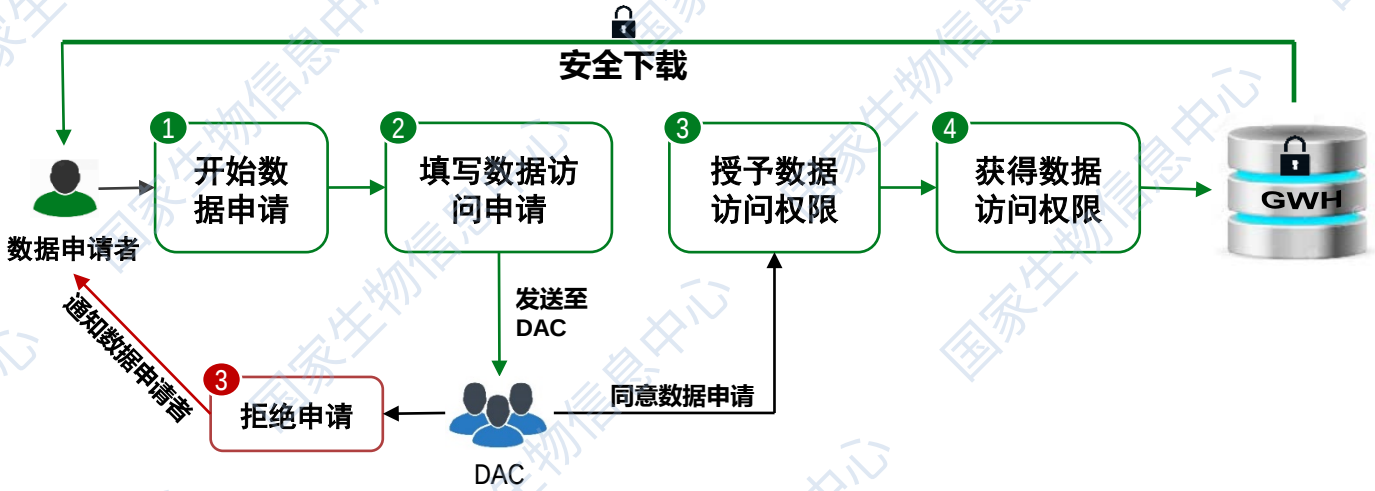
ID	Journal	Article title	Published Date	GWH
1	Scientific Data	Chromosome-level genome assembly of Helwingia omeiensis: the first genome in the family Helwingiaceae	2024-07	GWHEQHK00000000
2	Industrial Crops and Products	Leguminous industrial crop guar (Cyamopsis tetragonoloba): The chromosome-level reference genome de novo assembly	2024-06	GWHERDQ00000000
3	The Plant Journal	In depth exploration of the genomic diversity in tea varieties based on a newly constructed pangenome of Camellia sinensis	2024-06	GWHESSY00000000.1 GWHESSW00000000.1 GWHESSX00000000.1
4	DNA Research	High integrity Pueraria montana var. lobata genome and population analysis revealed the genetic diversity of Pueraria genus	2024-06	GWHESXN00000000.1 GWHESXM00000000.1
5	Horticulture Research	The jacktree genome and population genomics provides insights for the mechanisms of the germination obstacle and the conservation of endangered ornamental plants	2024-06	GWHCBFM00000000
6	Scientific Data	Chromosome-level genome assembly of the cottony cushion scale Icerya purchasi	2024-06	GWHERBG00000000
7	Scientific Data	Haplotype-resolved chromosome-level genome assembly of Huyou (Citrus changshanensis)	2024-06	GWHEQVQ00000000.1
8	Science Advances	Multiple independent origins of the female W chromosome in moths and butterflies	2024-06	GWHCBIK00000000 GWHCBIK00000000.1
9	Scientific Data	Chromosome-scale genome assembly of oil-tea tree Camellia crapnelliana	2024-06	GWHERAW00000000
10	Scientific Data	Haplotype-resolved chromosome-level genome assembly of Ehretia macrophylla	2024-06	GWHEQHN00000000

人类基因组数据资源安全共享

- 两种访问方式
  - 受控访问 (Controlled-access)
  - 公开访问 (Open-access)

  注：共享前通过人遗备份和事先报告
- 受控访问数据，采用申请审核制
  - 数据管理委员会（Data Access Committee, DAC）
  - 审核数据访问申请
  - 授予访问权限

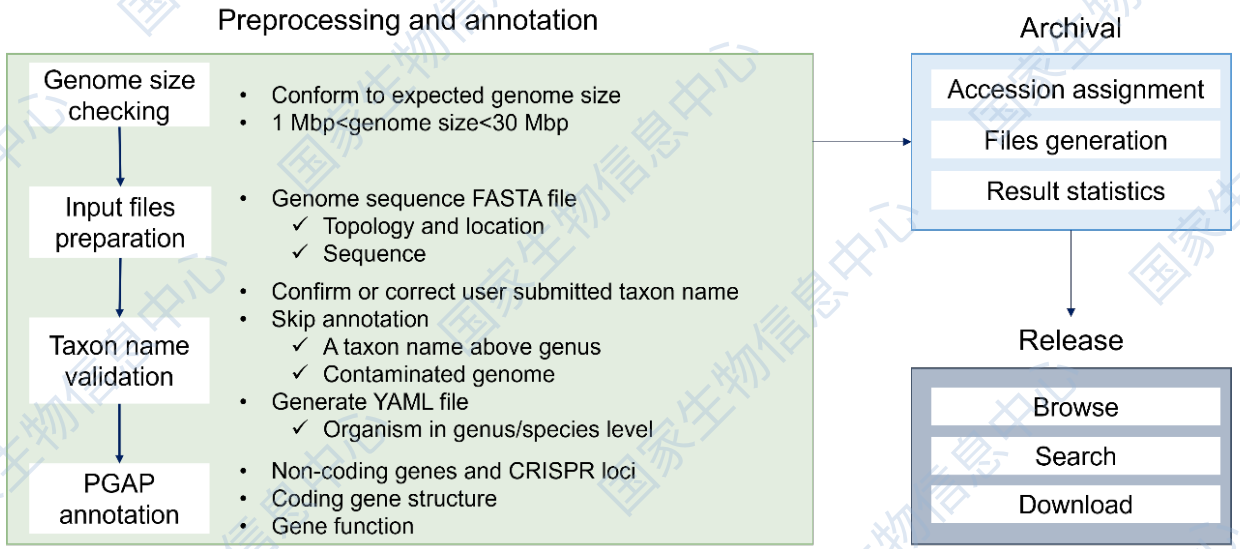
✓ 7,663	个人类基因组数据	✓ 42	次下载申请
✓ 1,187	个已公开发布	✓ 17	次获授权（含境外9次）
✓ 1,073	个可公开访问	✓ 4,594	次完成下载
✓ 114	个受控访问	✓ 2,896	Gb下载量



GWH人类基因组受控数据申请审核流程

高质量原核生物基因组增值再利用

基因组注释信息缺乏限制了基因组功能相关研究，GWH库选取了用户汇交的高质量基因组数据进行有效再利用，建立了原核生物（细菌和古细菌）基因组重注释资源，助力科学研究



GWH库原核生物基因组重注释和共享流程

重注释资源特色

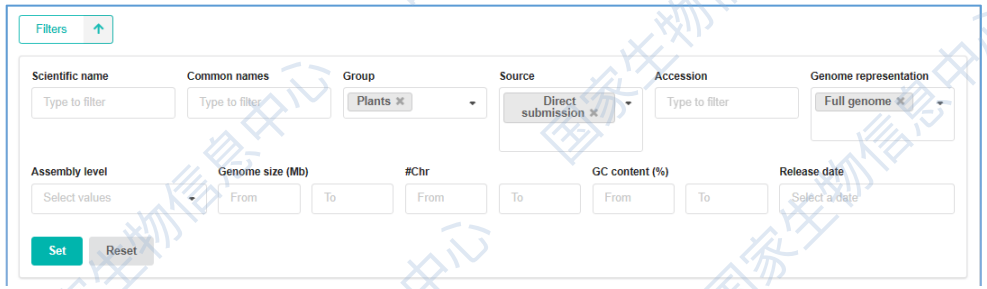
- 注释**标准统一规范**：552个物种，3,688个基因组
- 显著提升原核生物基因组注释率：1.5% -> **88.4%**
- 注释**内容丰富**：编码蛋白基因、tRNA、rRNA、小的ncRNA、CRISPR元件、假基因
- 注释**信息完整**：基因结构、蛋白产物、基因名称和基因功能等

国际数据的本地化安全管理与共享

为了给用户提供便捷的综合数据服务，GWH库同步镜像NCBI GenBank和RefSeq数据库中的基因组数据资源，并实现日更新

获取感兴趣的基因组

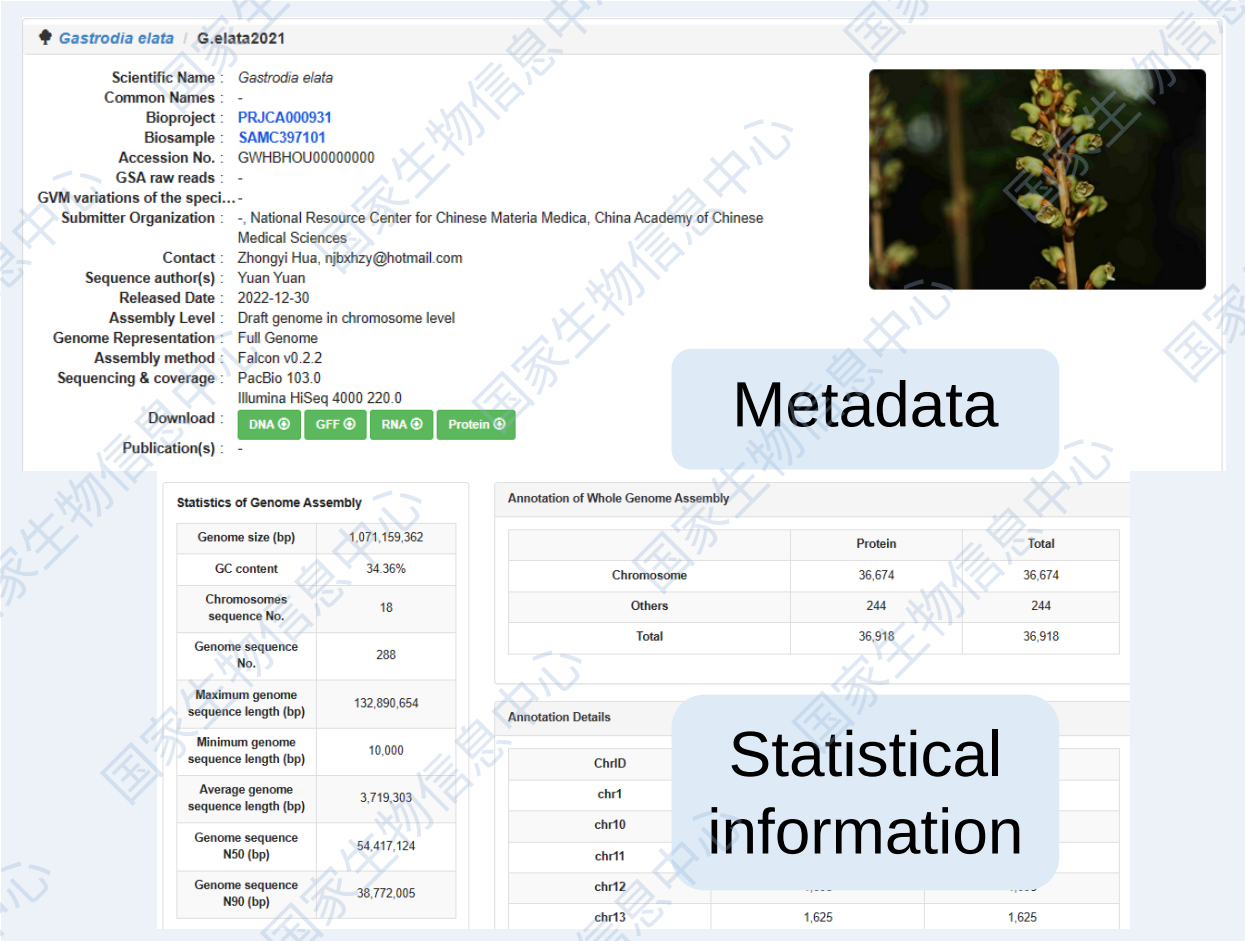
- 多条件关联查询实现个性化筛选



- 近20个条件自由组合的高级检索

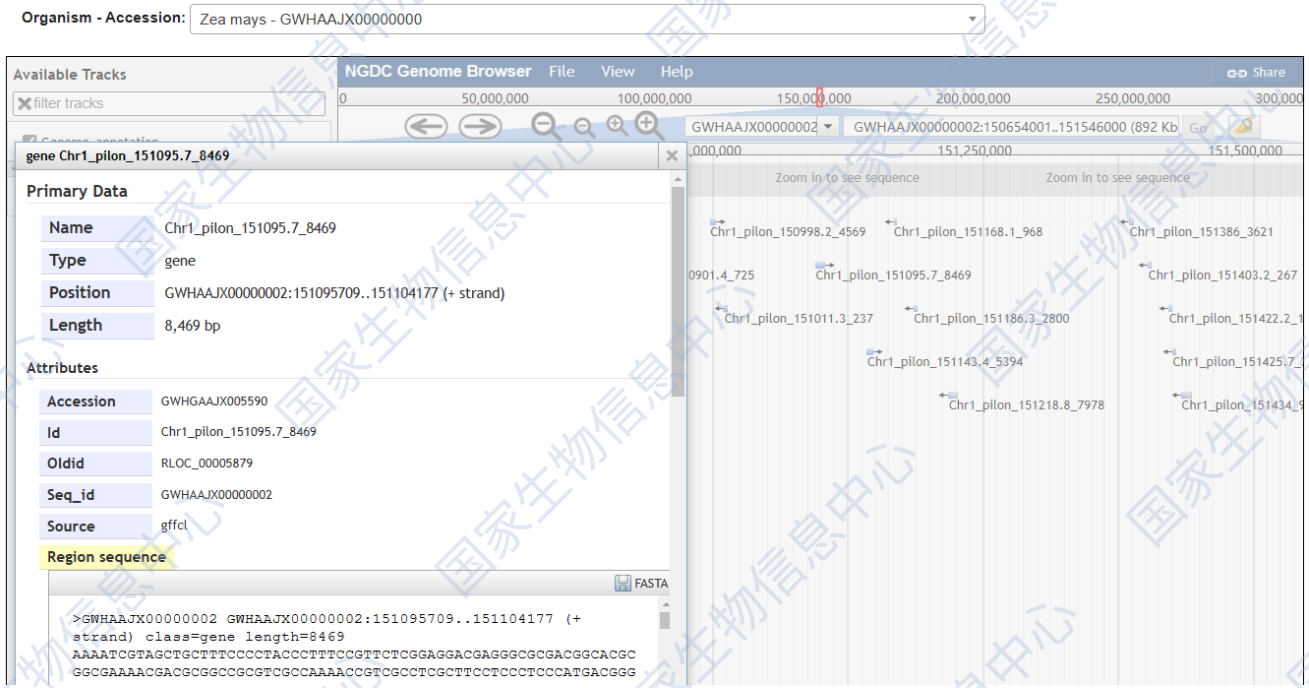
https://ngdc.cncb.ac.cn/gwh/search/advanced

- 详细页面提供丰富的元信息和统计信息



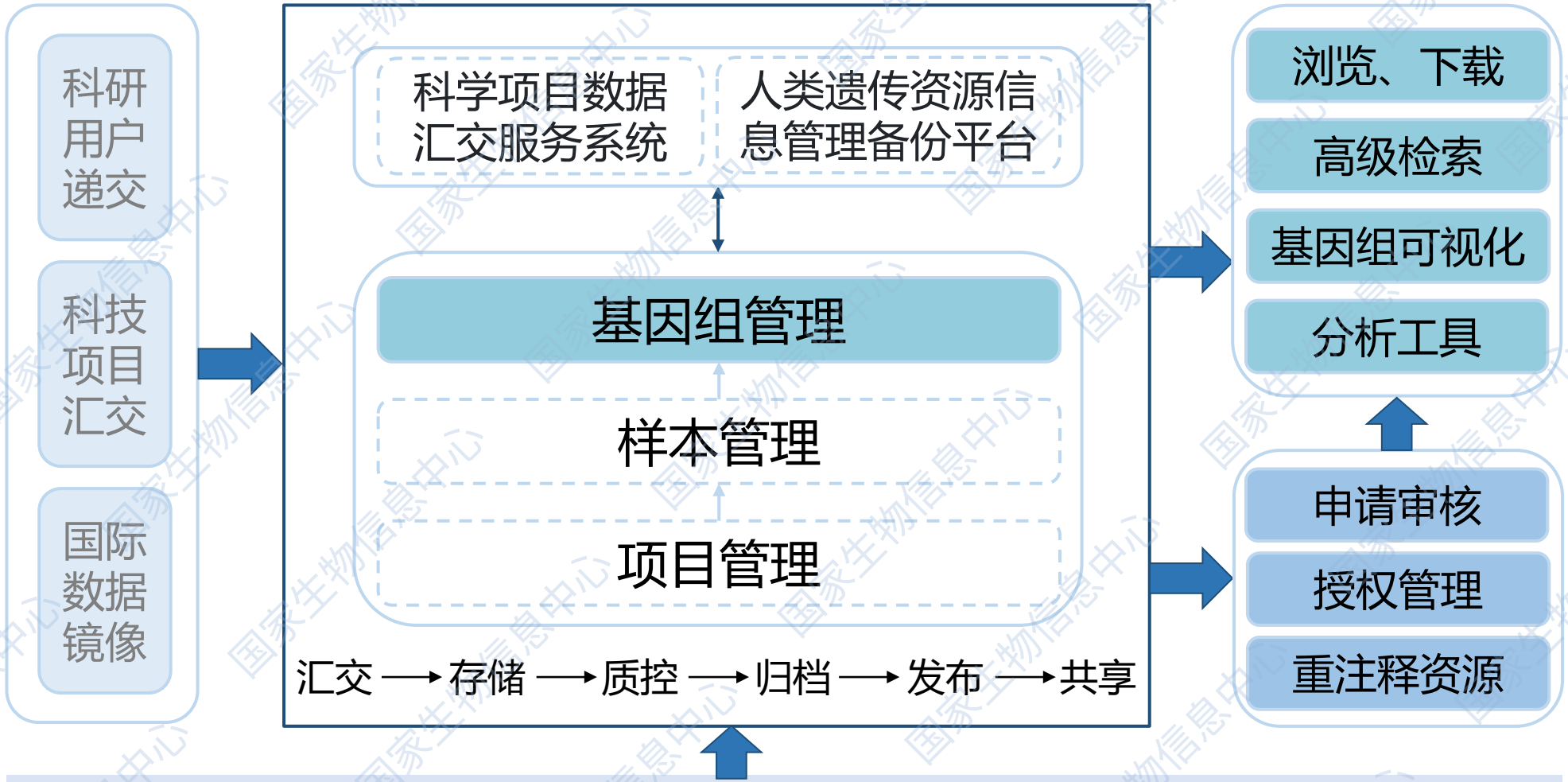
数据自助分析工具

- 提供用户友好的序列和注释**可视化**界面



- 综合的交互式在线BLAST分析平台

基因组数据管理体系


科研用户递交
科技项目汇交
国际数据镜像
科学项目数据汇交服务系统
人类遗传资源信息管理备份平台
基因组管理
样本管理
项目管理
汇交 → 存储 → 质控 → 归档 → 发布 → 共享
浏览、下载
高级检索
基因组可视化
分析工具
申请审核
授权管理
重注释资源
服务能力建设
高性能计算
高性能存储
全闪存储
网络带宽
Web服务器
数据库服务器
数据备份
网络安全

国际核酸数据库联盟INSDC对标库

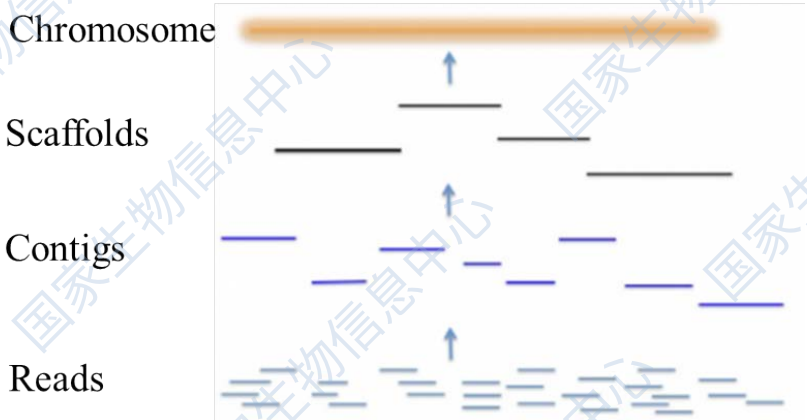
国家基因组科学数据中心	数据资源	美国NCBI	欧洲EBI	日本DDBJ
基因组数据库 GWH	重注释	RefSeq	Ensembl	/
基因序列库 GenBase	用户递交	GenBank：全基因组WGS	ENA: Assembled/annotated sequence	DDBJ (Annotated/Assembled Sequences)
	用户递交	GenBank：传统GenBank		

目录

基因组数据汇交共享过程

基因组组装的一些概念说明

- 基因组组装水平
  - ✓ 完成图（complete genome）：全部是完整的染色体，不含gap区，不含未定位的染色体片段
  - ✓ 染色体级别的草图（draft genome in chromosome level）：将scaffolds定位到染色体上，形成完整的染色体，可含gap区，可含未定位的染色体片段
  - ✓ scaffold：根据reads、基因等的连接关系，将contigs拼接成更长的序列，允许包含gap序列
  - ✓ contig：根据reads之间的重叠区域对片段进行拼接形成较长的连续序列



基因组组装层次

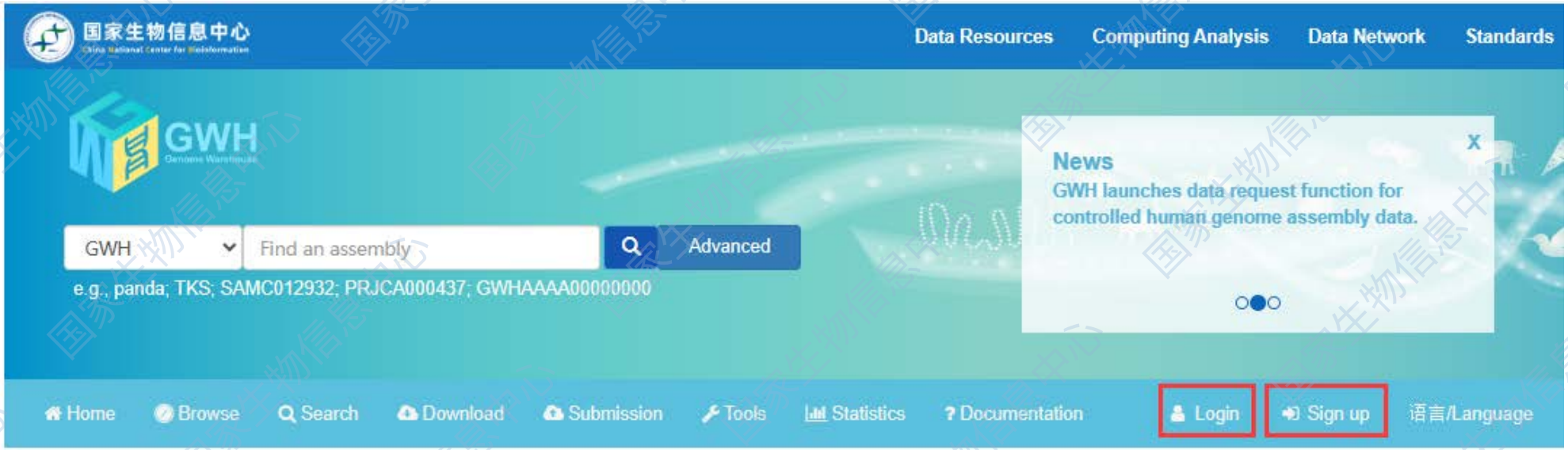
- Gap区：指的是在组装过程中未能得到准确填充的缺口，即序列中出现连续的N
  - ✓ 测序技术限制
  - ✓ 基因组序列复杂性
  - ✓ 区域间无序性

基因组组装的一些概念说明

- 基因组组分（genome representation）
  - ✓ 全基因组（full genome）：包含完整的基因组（即使是基因组草图），真核基因组组装可不含细胞器（即仅有核基因组）或者原核基因组组装可能不包括质粒
  - ✓ 部分基因组（partial genome）：特定选取的基因组的一个子集，如只有外显子或真核生物的单条染色体，或者只有基因组的非重复区域

准备工作：1. 账户准备——注册和登录

- 访问GWH库：https://ngdc.cncb.ac.cn/gwh/，进行注册和登录



SSO单点登录：中心任意一个网站注册和登录一次，就可以被授权访问中心其它数据库

准备工作：1. 账户准备-注册和登陆

- 登陆跳转到如下页面，填写相关信息，点击login登陆
- 未注册用户需要先注册再登陆

准备工作：2. 元信息文件准备

创建新的提交

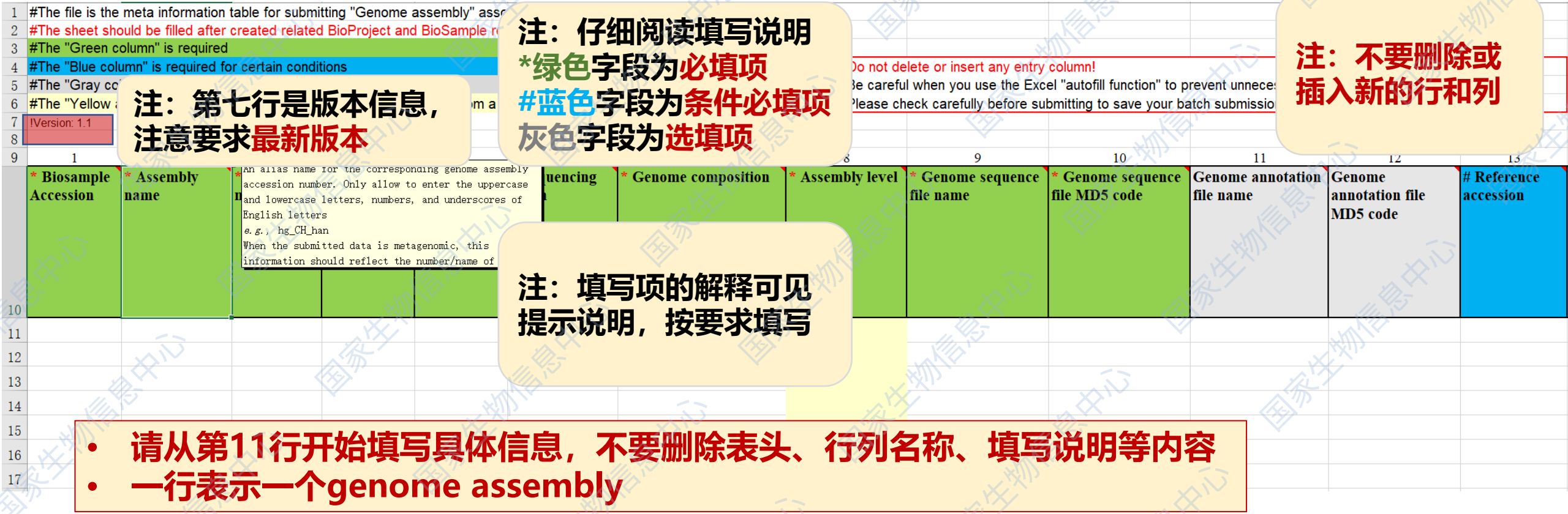
元信息文件下载位置

请注意：GWH接受与基因组相关的数据提交，但我们对提交数据的准确性并不负责，提交者应对数据的准确性负责。

1. 下载GWH批量提交模板文件 GWH_Assembly_en.xlsx / GWH_Assembly_ch.xlsx， 填写并仔细检查后上传
2. 批量提交模板的相关解释和示例，请参见GWH_Assembly_en.xlsx / GWH_Assembly_ch.xlsx。
3. 下载GWH从属信息模板文件 chr.csv / plasmid.csv / organella.csv 及其 说明文档，填写并仔细检查后通过FTP方式上传(可选)。
4. GWH提交相关常见问题解答，请参阅此 pdf。
5. 更多信息，请参阅 帮助文档。

准备工作：2. 元信息文件准备

下载后打开最新版本的GWH-batchsubmission-English.xlsx模板文件



注：第七行是版本信息，注意要求最新版本

注：仔细阅读填写说明
*绿色字段为必填项
#蓝色字段为条件必填项
灰色字段为选填项

注：不要删除或插入新的行和列

注：填写项的解释可见提示说明，按要求填写

- 请从第11行开始填写具体信息，不要删除表头、行列名称、填写说明等内容
- 一行表示一个genome assembly

准备工作：3. 数据文件准备

- 基因组组装序列文件（必需，fasta格式）

```
>seq_1 description line
ATCGATCGATCGATCGATCGATCGATCGATCG
ATCGATCGATCGAT
```

Sequence ID命名要求：
（1）以字母开头
（2）字母、数字、横线-、下划线_、点.、冒号:、星号*和#组成
如上面序列的sequence ID为seq_1

一个基因组只有：
- 1个序列文件
- 1个基因组注释文件，同时含编码、非编码的注释请合并

- 基因组注释（非必需，gff/tbl格式）
  - gff格式（九列）

1. Sequence ID;
2. Annotation information source (null value is '.');
3. Feature type, eg: gene, transcript;
4. Start position of a feature on the sequence (from 1);
5. End position of a feature on the sequence (no more than sequence length);
6. Score (null value is '.');
7. Strand;
8. Frame of a CDS (0, 1, 2);
9. Attributes, additional information about each feature;

For example,

```
##sequence-region ctg123 1 1497228
ctg123 .  gene 1000 4000 .  +  .  ID=gene00001;Name=EDEN
ctg123 .  TF_binding_site 1000 1012 .  +  .  Parent=gene00001
ctg123 .  mRNA 1300 3920 .  +  .  ID=mRNA00001;Parent=gene00001
ctg123 .  exon 1300 1500 .  +  .  ID=exon00001;Parent=mRNA00001
ctg123 .  exon 2000 3950 .  +  .  ID=exon00002;Parent=mRNA00001
ctg123 .  CDS 1300 1500 .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 .  CDS 2000 3901 .  +  0  ID=cds00002;Parent=mRNA00001
ctg123 .  UTR 3902 3920 .  +  0  ID=UTR00001;Parent=mRNA00001
```

准备工作：3. 数据文件准备

- 序列从属和命名说明文件（组装水平为complete或者draft in chromosome时必需，csv格式）

染色体序列从属和命名说明文件

Sequence ID	Chromosome name	Complete	Circular
seq_1	1	true	false

细胞器序列从属和命名说明文件

Sequence ID	Type	Complete	Circular
seq_2	Mitochondrion	true	false

质粒序列从属和命名说明文件

Sequence ID	Plasmid name	Complete	Circular
seq_3	pBR322	true	false

- Sequence ID：必须是来自于基因组序列文件中的Sequence ID，即序列文件中 > 号后面的非空格字符串
- 完整度：“Complete=true” 表示该序列代表确定的染色体、细胞器或质粒。在染色体序列从属说明文件中，“Complete=false”表示与特定的染色体相关，但尚未定位到染色体上的特定位置，同理细胞器和质粒
- 是否环化：“Cirular=true”表示该条序列为可环化（序列首尾可含gap），“Cirular=false”表示该条序列为线性或为环化序列的片段（序列首尾不可含gap）

提交准备页面

详细的提交流程说明

提交准备内容

https://ngdc.cncb.ac.cn/gwh/submit/prepartion

提交流程：1. 创建新的提交

一个批量提交要求关联同一个BioProject，同一篇文章

方法二、通过中心统一汇交入口BIG Sub，登录GWH库的提交系统：https://ngdc.cncb.ac.cn/gsub/

方法一、通过GWH库登录提交系统：https://ngdc.cncb.ac.cn/gwh/submit/submission

提交流程：1. Meta信息在线提交

第一步：提交者信息

注：Submitter的邮箱将用于后续提交过程的信息联系，如有需要请同时填写需抄送的备用邮箱，数据发布后不公开

提交流程：1. Meta信息在线提交

第二步：基本信息

注：标题不对外公开，仅限于用户区分提交列表中的多个提交

注：请注意阅读数据发布正常和免责声明

注：受控访问方式仅限于人类基因组数据资源

提交流程：1. Meta信息在线提交

第三步：序列作者、联系人及文章信息

序列作者：数据发布时仅公开姓名，不公开email
序列联系人：数据发布时公开姓名和email

注：此处为发表该提交的基因组关联的文章信息

注：请确认提供的pubmed ID号对应的详细信息是否准确无误，其它方式需提供详细的文章作者列表

提交流程：1. Meta信息在线提交

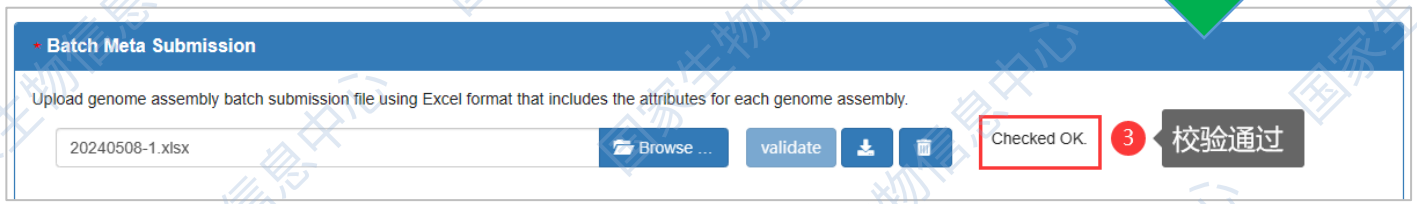
第四步：上传组装元信息文件（一次批量提交可提交多个基因组），并在线校验

①选择本地文件并上传
②上传成功后，在线校验

校验失败

报错信息：
①文本文件下载
②在线可视化显示

校验通过



注：

- Error的报错必须修改后重新提交
- Warning的报错则用户根据实际数据情况自行判断

提交流程：1. Meta信息在线提交

第五步：预览提交信息，确认无误后提交

第1步 提交者信息 | 第2步 基本信息 | 第3步 文章信息 | 第4步 Meta | 第5步 预览

信息概览

批量提交号: Batch0023947
标题： test
提交状态： Unfinished Step 4

本提交的数据将于 **2024-05-23**（8 天后）或引用文章公开发表后公开发布，这取决于哪个日期先到。
请注意：GWH数据的发布将触发关联的BioProject和BioSample一起发布。

FTP提交

您可以使用以下信息将文件传输到GWH的FTP站点:

- **主机地址**: submit.big.ac.cn
- **用户名**: chenml
- **密码**: *与登录GWH/NGDC时的用户密码相同, 例如:123456*
- **路径**: /GWH/Batch0023947

FTP服务器登录说明

请注意：使用FTP客户端软件（如FileZilla client）登录FTP客户端。
提交文件后，请在您的提交列表中点击"**完成上传**"按钮 (**https://ngdc.cncb.ac.cn/gwh/submit/submission**)，通知我们审核您的提交。

提交者信息

- **提交者信息**: Meili Chen
- **电子邮箱**: chenml@big.ac.cn
- **提交单位**: The CAS Key Laboratory of Genome Sciences and Information,Beijing Institute of Genomics, Chinese Academy of Sciences

基本信息

- **BioProject编号**: PRJCA000494
- **发布日期**: 2024-05-23
- **受控访问**: False

文章信息

- **序列作者**: mei null test
- **发表状态**: published
- **文章标题**: [The laboratory in programs for enteric infection control].
- **文章作者**:
- **通讯作者**: mei null test (test@qq.com)

元信息属性

每个基因组组装特有的元信息属性说明

显示第 1 至 1 项结果，共 1 项

BioSample编号	物种拉丁名	组装名称	组装水平	基因组序列文件名	基因组注释文件名	参考基因组编号	已有的GWH基因组编号	Chromos
SAMC016108	-	wgs_MAG_1	complete	JRJ_duck1.0.fa	N/A	N/A	N/A	

显示 10 项结果

完成提交

提交流程：1. Meta信息在线提交

Meta信息在线提交完成后返回提交列表

创建新的提交

请注意：GWH接受与基因组相关的数据提交，但我们对提交数据的准确性并不负责，提交者应对数据的准确性负责。

我的提交 已删除的提交

搜索:

7 个提交记录

组装编号	批量提交号	提交标题	BioProject编号	发布日期	创建日期	最后更新日期	提交状态	操作
GWHAZPK01000000 2 个组装编号，显示全部	Batch0018094	SARS-CoV-2_human_CHN_Wuhan_S5_2020	PRJCA004241	2027-07-26	2024-04-23	2024-04-23	SUCCESSFUL	审稿人页面 立即发布 修改发布日期 删除
	Batch0023947	test	PRJCA000494	2024-05-23	2024-04-23	2024-04-23	Waiting for files	完成上传 删除

提交流程：2. 数据文件FTP提交

◘ 登录操作步骤：

①主机地址：submit.big.ac.cn

②用户名：与登录GWH/NGDC时的用户名相同, 例如：chenml@big.ac.cn

③密码：与登录GWH/NGDC时的用户密码相同, 例如：123456

④点快速连接（默认端口21）

◘ 文件FTP上传操作

- 选定本地磁盘路径和文件
- GWH提交指定的FTP路径： /GWH/$Batch_ID，例如：/GWH/Batch0023947
- 本地文件直接拖拽至指定的FTP路径

提交流程：3. 完成提交后触发自动审核

- 元信息和数据文件完成提交后，返回提交列表，点击对应提交的“完成上传”（Finish upload）按钮

审核内容

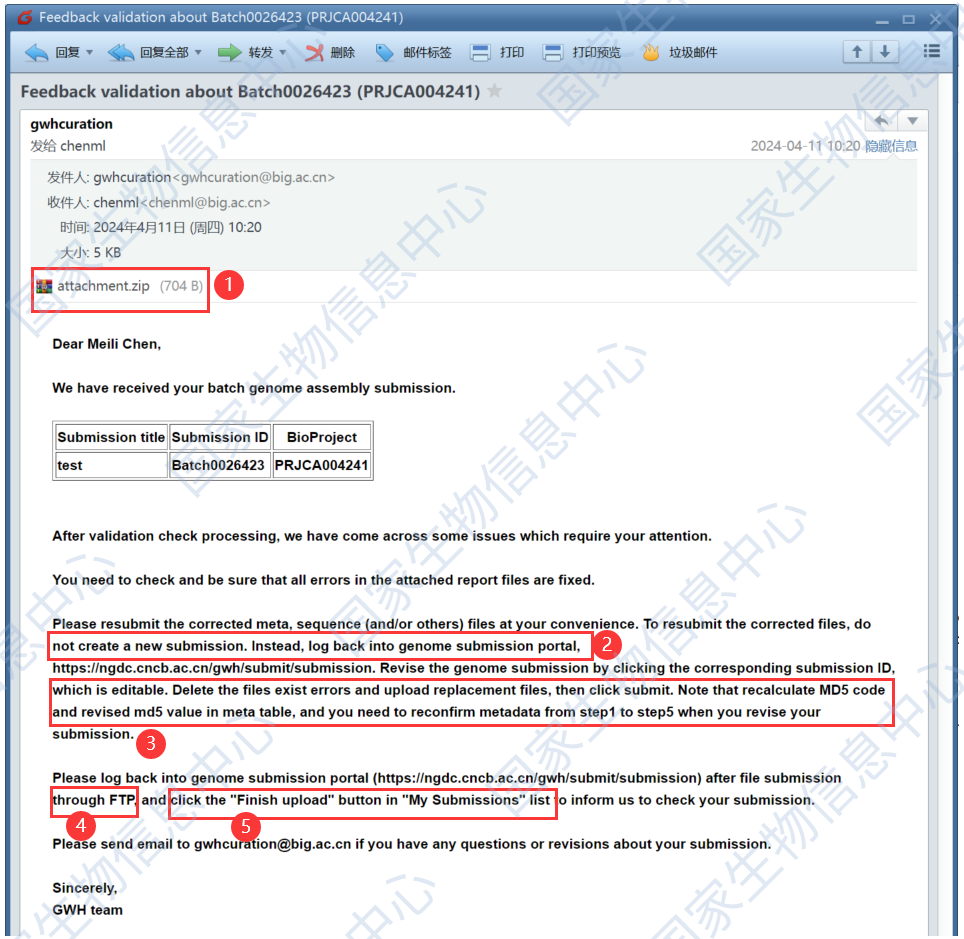
- 基因组序列文件内容的合法性和有效性
- 基因组注释文件内容的合法性和有效性
- 基因组序列和注释内容一致性审核
- 基因组序列接头、载体等污染审核

数据提交状态为“Waiting for files”（即等待上传文件）

注：请务必点击该操作按钮来触发系统自动质控和归档
注意查收邮件，系统自动反馈质控报错/通过

提交流程：3. 系统审核及反馈

➢ 审核结果报错



①查收邮件，根据附件的错误报告**修改后提交**

②在原来的提交号中修改并提交

③如数据文件有修改，请注意重新计算文件MD5码

④FTP中重新上传修改后的数据文件

⑤返回提交列表，点击“完成上传”（Finish upload）按钮

提交流程：3. 系统审核及反馈

➢ 审核通过

- 邮件通知审核通过
- 附上正式Accession号
- 发表文章时的引用模板

数据提交状态为“Successful”（即归档成功）

① 显示分配的Accession号（X表示大写字母，x表示版本号）

✓ 基因组包含多条序列：GWHXXXX00000000.xx，第一版本GWHXXXX00000000.1，其等价于GWHXXXX00000000

✓ 基因组只含一条单序列：GWHXXXXxx000000，第一版本GWHXXXX01000000，如GWHAZPK01000000

② 显示当前数据的状态为“Successful”（即归档成功）

③ 支持操作：生成审稿人链接、立即发布数据、修改发布日期

提交注意事项1：元信息文件

□ Assembly name（第2列）

- **常规基因组**：即基因组拼接Accession号的别称，只允许输入英文字母大小写、数字、下划线，且要求唯一性。例如：hg38
- **元基因组**：此信息需体现Bin或MAG的编号/名称，即以*_bin_1、 *_bin_2...或*_MAG_1、 *_MAG_2...编号结尾，同一套元基因组来源的前缀一致
- **单倍型基因组**：此信息需体现单倍型类型及关系（编号/名称），即以*.pri/*.alt或*.hap1/*.hap2/*.hap3/*.hap4或*.pat/*.mat...等编号结尾，同一套单倍型基因组来源的前缀一致

□ Assembly method与Program version or release date （第3和4列）

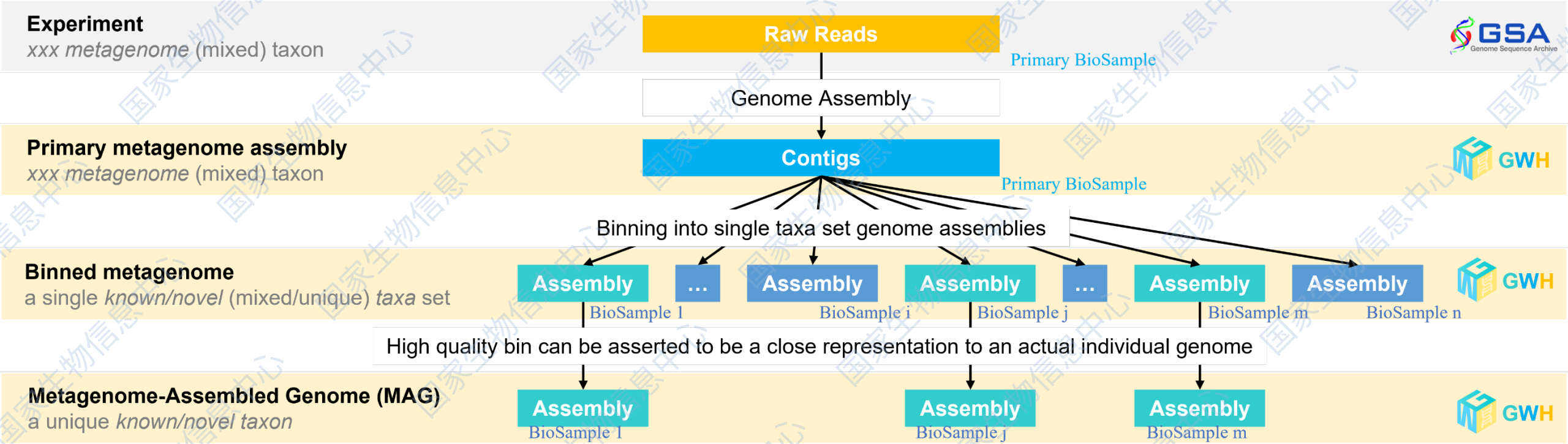
需要配对出现，多个值用分号;隔开

□ Assembly level （第8列）

对于基因组组装水平为scaffold或者chromosome时，与序列间隙gap有关的列（21-24）为必填项，如Linkage evidence、Gap type等

提交注意事项2：元基因组数据

- 测序原始数据应先提交到GSA，且为必需提交（关联Primary Biosample）
- 元基因组组装（关联Primary Biosample）提交到GWH
- 由binning分箱组成的contigs（Binned/MAG类型，关联各自的BioSample）提交到GWH



元基因组数据结构

提交注意事项2：元基因组数据

元信息文件需要关注1、2、25~36列，其中：

□Metagenome assembly data type（第25列）：

元基因组组装数据类型，可选Primary/Bin/MAG。

- **Primary类型**：是指元基因组测序原始数据经过组装的结果，即Binning分析前的组装结果，其（第1列BioSample关联的）物种名称是xxx metagenome；
- **Bin类型**：是指元基因组测序原始数据组装结果进行Binning分析后鉴定的分箱基因组，其（第1列BioSample关联的）物种名称名称不能是xxx metagenome，而是注释得到的物种名称；
- **MAG类型**：元基因组组装的基因组，是指将基因组测序原始数据组装后，进行Binning分析得到的被认为是一个能够确定到species/subspecies水平的基因组的结果。

□GSA experiment accession（第26列）：

元基因组组装所使用的测序原始数据对应的experiment accession编号。多个值用分号；隔开。例如：CRX1108622

□BioSample for Metagenome Primary Assembly （第27列）：

元基因组组装（即Binning分析前）关联的测序原始样本BioSample（多物种的混合样），其物种名称是xxx metagenome，通过该列可以识别出哪些元基因组数据是来自于同一套样本的。如果是多个宏基因组样本测序后混合再组装，则可以关联对应多个值，用分号；隔开。例如:SAMC3777840; SAMC3777841

提交注意事项2：元基因组数据

举例：提交从海洋元基因组中鉴定到MAG水平的XX菌

1. 在BioProject库中创建一个该元基因组的生物项目，获得BioProject accession编号
2. 为海洋元基因组创建BioSample编号（Primary BioSample），对应的物种是Marine metagenome
3. 为鉴定出的XX细菌创建BioSample编号（BioSample），对应的物种是XX菌
4. 将测序原始数据提交到GSA （Primary BioSample），获得Experiment accession编号

5. 填写GWH的元信息模版文件

①第一列的BioSample Accession填写XX菌对应的biosample编号（BioSample）。如果是XX病毒，那么此数据需要上传至GenBase数据库。添加新物种可以联系gsa@big.ac.cn

②第二列的Assembly name按照*_MAG_1、*_MAG_2...编号结尾的填写规则，如：mm240313_MAG_1

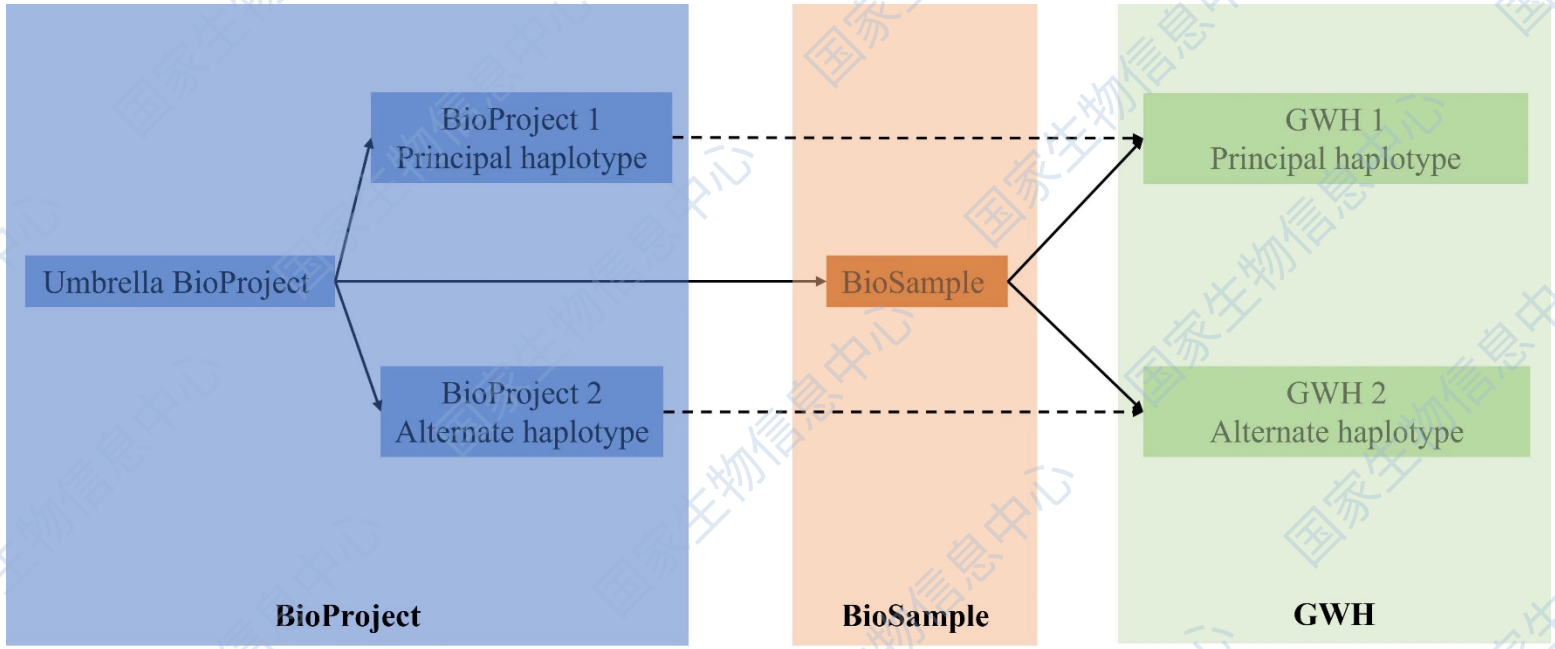
③第25列的Metagenome assembly data type列选择MAG

④第26列的GSA experiment accession列填写上述experiment accession编号

⑤第27列的BioSample for Metagenome Primary Assembly列填写Marine metagenome对应的biosample编号（Primary BioSample）

提交注意事项3：单倍型数据

- **常规基因组**：BioProject→BioSample→Genome是1:1:1的关系
- **单倍体基因组**：由于同一个个体的不同单倍型是来自于同一个生物样本，BioSample→Genome则是1：n的关系，即共享相同的BioSample，为了区别不同的单倍型基因组，则除了原来BioSample关联的BioProject外，每个基因组有分别独自关联的BioProject、与BioSample关联的BioProject之间形成伞状结构关系，形成关联



单倍型的类型可以是：

- a. Principal haplotype / Alternate haplotype
- b. Haplotype 1 / Haplotype 2 / Haplotype 3 / Haplotype 4
- c. Maternal haplotype / Paternal haplotype
- d. Diploid
- e. Polyploid
- f. Haploid-with-alt-loci
- g. Unresolved-diploid

单倍型基因组数据结构关系（注：BioProject的伞状结构关系是在BioProject创建时指定生成的）

提交注意事项3：单倍型数据

元信息Step2的BioProject，step4的元信息文件需要注意1、2、37、38列，其中：

□ Step2的BioProject Accession：

BioSample关联的BioProject Accession

注：两个或多个单倍型基因组隶属于一个总的伞形的BioProject（即提交页面中step2中填写的BioProject accession），以此来形成不同单倍型基因组数据间的关联，由此总的BioProject可以将Haplotype BioProject联系起来形成伞状结构

□ BioSample Accession（第1列）：

由于多倍体基因组是来自于同一样品分别组装形成的两个或多个单倍型基因组，它们应该来自于同一个样品，所以它们必须与相同的BioSample相关联，即相同的“Biosample Accession”

□ Haplotype BioProject（第38列）：

由于多倍体基因组可组装成单独的两个或多个单倍型基因组，为了以示区分和保留数据间的关联关系，需要创建不同的BioProject与单倍型基因组形成关联，即不同的“Haplotype BioProject”

提交注意事项3：单倍型数据

举例：提交Principal haplotype和Alternate haplotype数据

1. 创建**两个**分别的BioProject用于区分Principal和Alternate的单倍型基因组组装数据，**在Title中标注是关联哪个单倍型基因组**。例如PRJCA001233和PRJCA008205
2. 再创建**一个**总的BioProject来关联这两个不同的BioProject，形成伞状结构关联关系，创建方式为在BioProject的第二步基本信息的关联项目中添加上已创建好的两个分别的BioProject。例如将新创建的PRJCA009877与第1步中已创建的PRJCA001233和PRJCA008205形成BioProject关联（如图）
3. 创建来自于同一样品的BioSample，需要将其与总的BioProject进行关联，如PRJCA009877
4. GWH中step2中填写总的BioProject，如PRJCA009877
5. 元信息表Biosample Accession列填写来自于同一样品的BioSample
6. Assembly name列按照*.pri/*.alt编号结尾的填写规则，前缀一致
7. Assembly type列分别选择Principal haplotype和Alternate haplotype
8. Haplotype BioProject列填写用于区分Principal和Alternate的两个分别的BioProject（第1步创建的）

常见错误类型

后缀/名称等格式问题：

- 染色体名称的规范

Error1: An chrmosome name should be assigned to a positive integer (eg: 1) or a capital letter (eg: A) or a captital letter with positive integer (eg: A1) or a positive integer with a letter (eg: 1a or 1A).

- 序列ID的规范

Error2: In Genome sequence file, the sequence ID must start from letters (eg. >Contig21, the sequence ID is Contig21).

- 基因组文件名称规范

Error3: Acceptable genome sequence file suffix have .fasta, .fa, .fsa, .gz and .bz2. Allowed compressed format is .gz or .bz2, rather than packing compression in .tar.gz or .tar.bz2, etc. Please submit the sequencing data fq file to the GSA website.

- 基因组注释文件名称规范

Error4: Acceptable genome annotation file is .gff or .tbl format. Allowed compressed format is .gz or .bz2, rather than packing compression in .tar.gz or .tar.bz2, etc.

- 序列从属信息文件名称规范

Error5: Assignment information file name has invalid suffix. Acceptable file suffix have .csv, .txt. Please modify the assignment information file name.

常见错误类型

基因组序列和注释文件内容问题：

- Completed 基因组不应存在Gap区域

Error6: Completed genome must not contain any Gap region. The genome contain N sequence more than 10 bp.

- 基因组序列重复提交

Error7: This record has identical nucleotide sequence to a previous submission, explain whether this submission is: an inadvertent duplicate of the previous submission (in this case, you should delete the new submission) or an update to the previous submission.

- 序列中N碱基超过40%

Error8: Sequence Chr01 has more than 40% Ns. Ambiguous nucleotides shouldn't be over 40%.

- 基因组大小异常

Error9: The submitted genome is smaller/larger than expected based on the taxid. The expected total ungapped genome sequence length is ranged [T1, T2]. The current geneome size is {size}.

- 基因组注释的gff文件中，tab分割符/缺失ID属性信息/层级关系（gene-mRNA-exon-CDS）混乱/无起始密码子等

Error10: The 9th column in gff file is missing ID or Parent information. Tab is used as the column interval. In gff3 file, the hierarchy validation failed.

报错修正文档1:常见问题解答

1、注册邮箱激活失败
2、GWH接受数据的要求
3、数据FTP形式上传注意事项
4、提交数据注释文件
5、元信息文件中Assembly level填写说明
6、元信息文件中Assignment相关信息填写说明
7、序列文件规范
8、注释文件规范
9、数据释放时间
10、杂志认定
11、获取审稿人链接
12、报错文件（.err和.user）的打开方式
13、部分具体报错信息分析
14、追加注释文件
15、修改数据meta信息
16、文章已发表操作
17、数据update
18、提交者邮箱
19、多个biosample的选择
20、Linkage evidence信息
21、对于提交的MAG数据
22、参考文档

报错修正文档2：常见错误类型及解决方案

示例

报错：SEQ_FEAT.NoStop

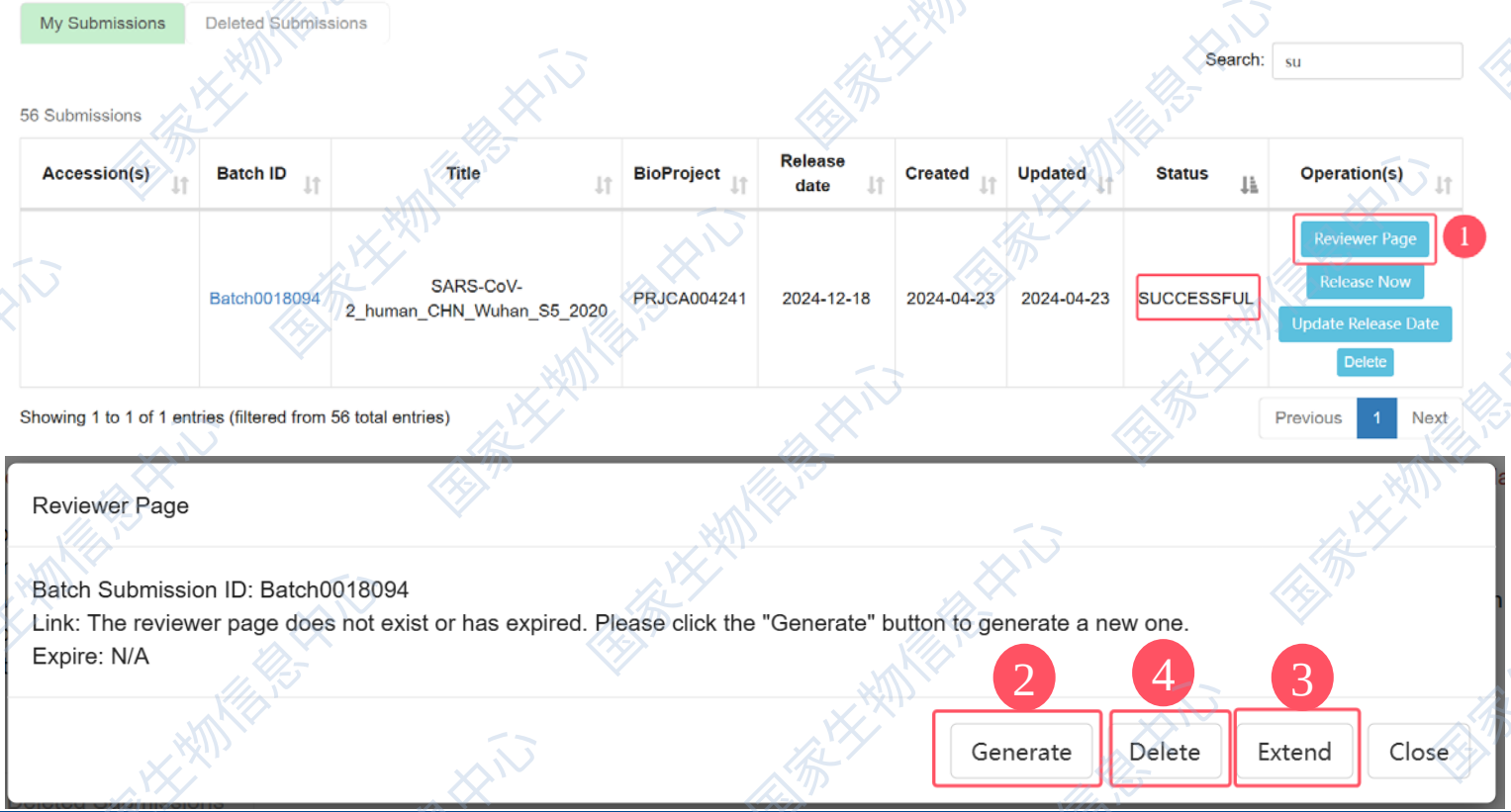
解释：CDS 在其 3 ' 端不以终止密码子结束。

建议：

1. 是否 CDS 位置标记错误，如是则考虑延长 CDS 序列位置，直至出现终止密码子
2. 是否序列在 3' 端存在不完整情况，如是则可以将 3' 端标记为不完整
3. 是否为假基因，如果确定为假基因并且无法找到终止密码子，可以在 GFF 文件对应的 gene 行的第九列添加 pseudo属性值，代表该序列无法被正常翻译

数据发布前审稿人链接自助生成和延期

对于通过质控归档的数据，在发布前（状态为Successful）可自助生成和延期审稿人链接。



① 在提交列表中点击后进入审稿人链接管理

② 首次生成审稿人链接

③ 延期已生成的审稿人链接

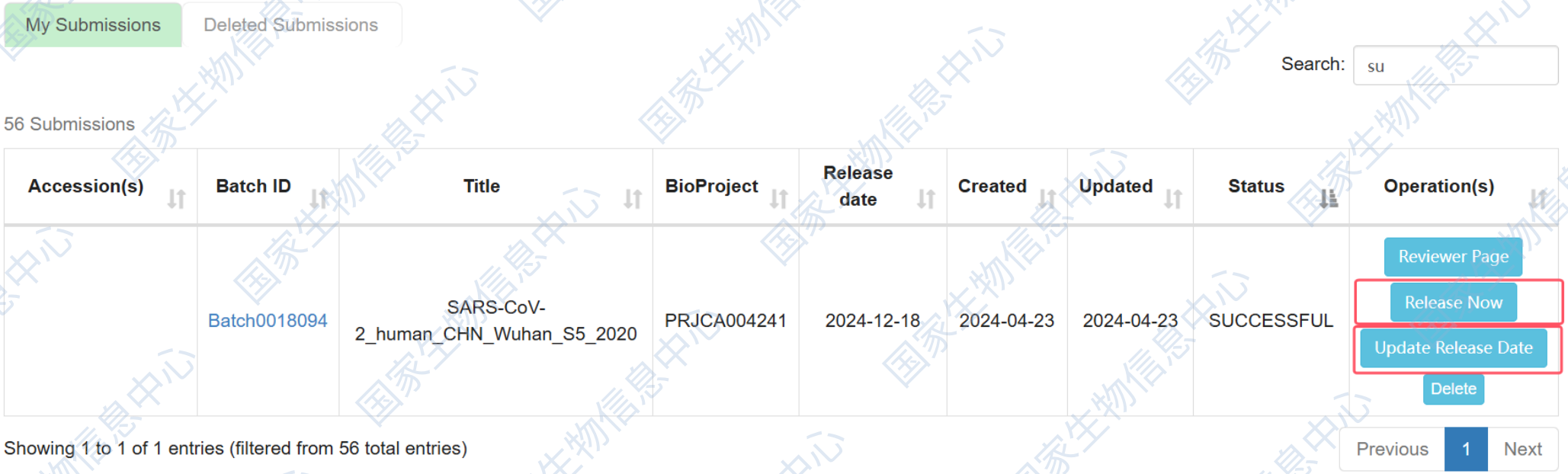
④ 删除已生成的审稿人链接使其失效

注：请注意妥善保管审稿人链接，获取到链接的用户都可以访问数据，尤其是申请将数据文件一起共享的

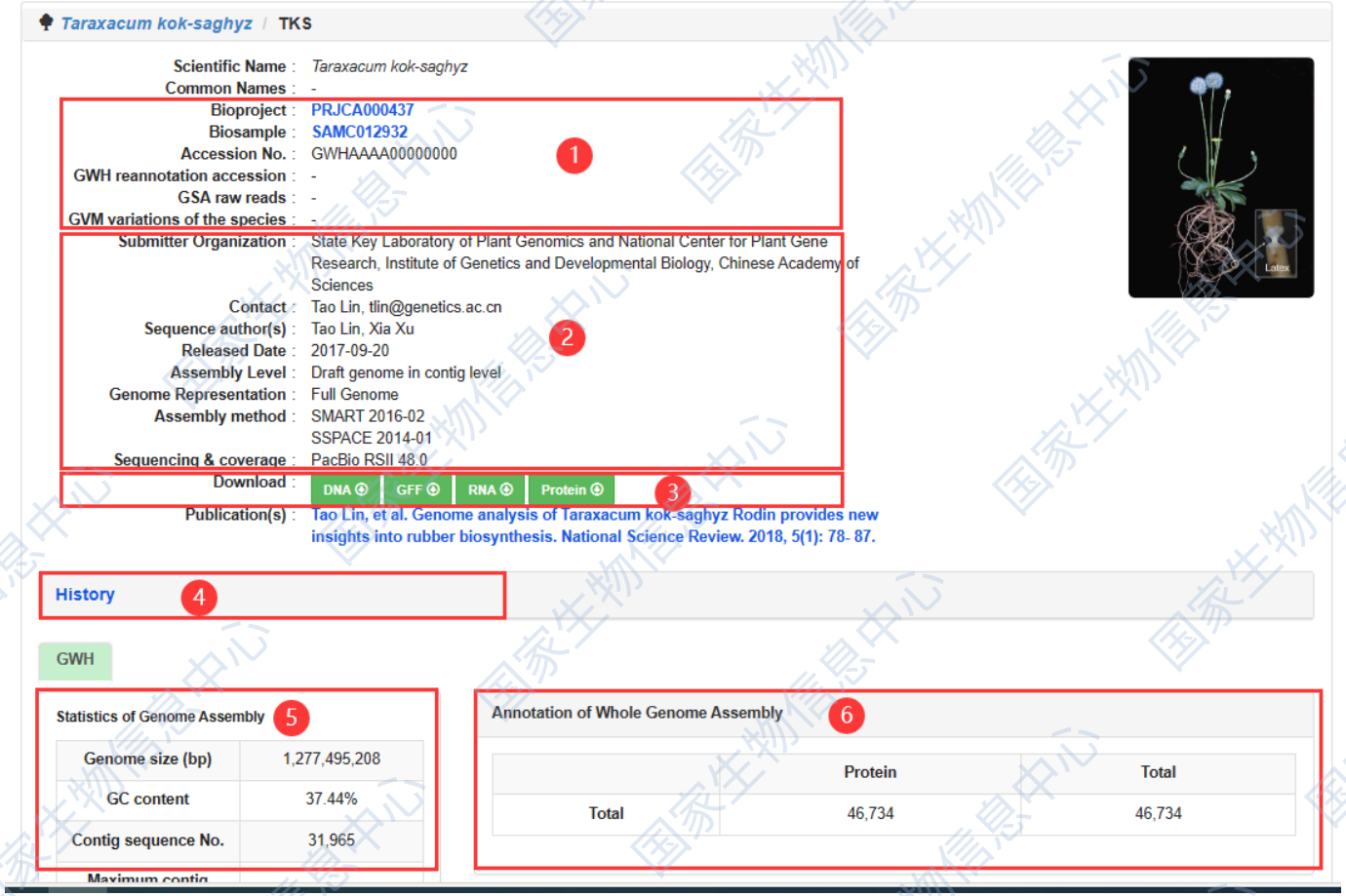
https://ngdc.cncb.ac.cn/gwh/submit/submission

数据发布自助管理

对于通过质控归档的数据，在发布前（状态为Successful）可自助立即发布和修改发布日期。



数据发布和共享：数据页面



https://ngdc.cncb.ac.cn/gwh/Assembly/1/show

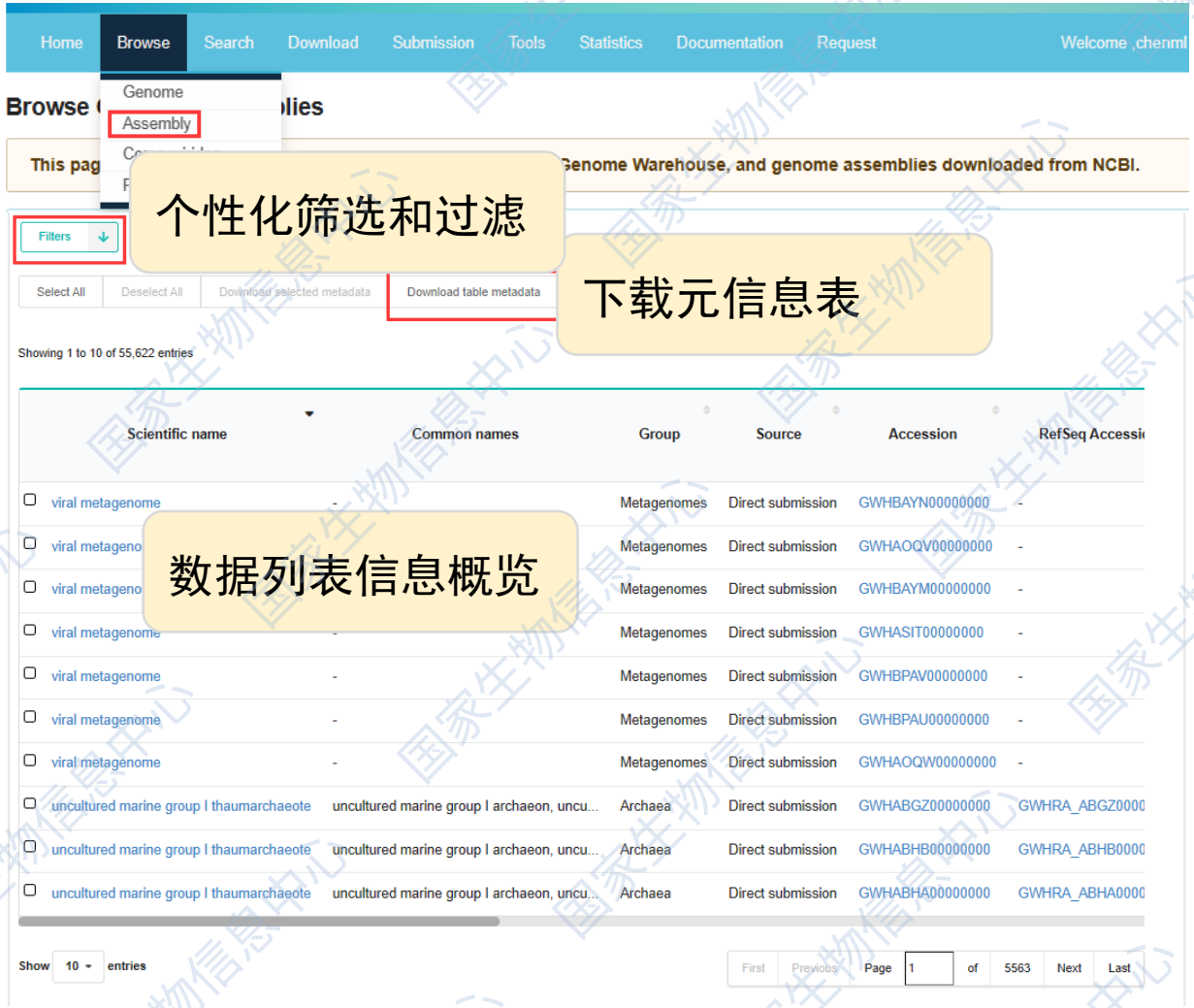
数据发布触发机制

- 到期发布：数据到期
- 提前发布：引用Accession的文章公开发表
- 关联数据发布：关联的BioProject和BioSample一并发布

基因组组装信息概览：

① 关联数据资源信息
② 组装基本元信息
③ 归档文件下载链接
④ 历史版本列表
⑤ 基因组序列的统计信息
⑥ 基因组注释的统计信息

数据发布和共享：浏览和搜索



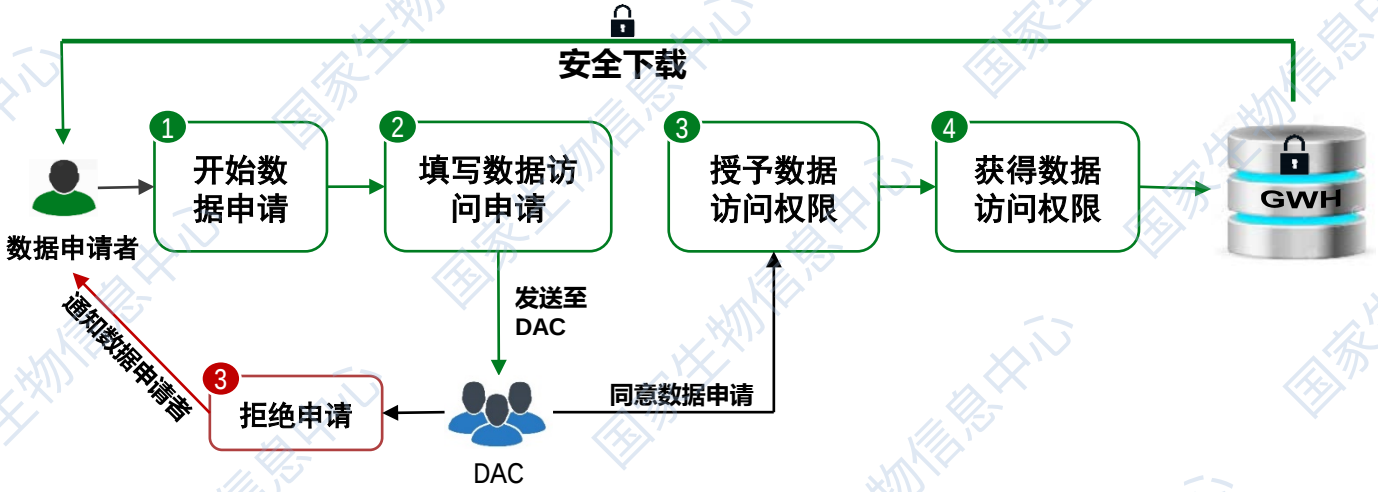
https://ngdc.cncb.ac.cn/gwh/browse/assembly

数据发布和共享：人类基因组数据

- 两种访问方式
  - 受控访问 (Controlled-access)
  - 公开访问 (Open-access)

  注：共享前通过人遗备份和事先报告

- 受控访问数据，采用申请审核制
  - 数据管理委员会（Data Access Committee, DAC）
  - 审核数据访问申请
  - 授予访问权限



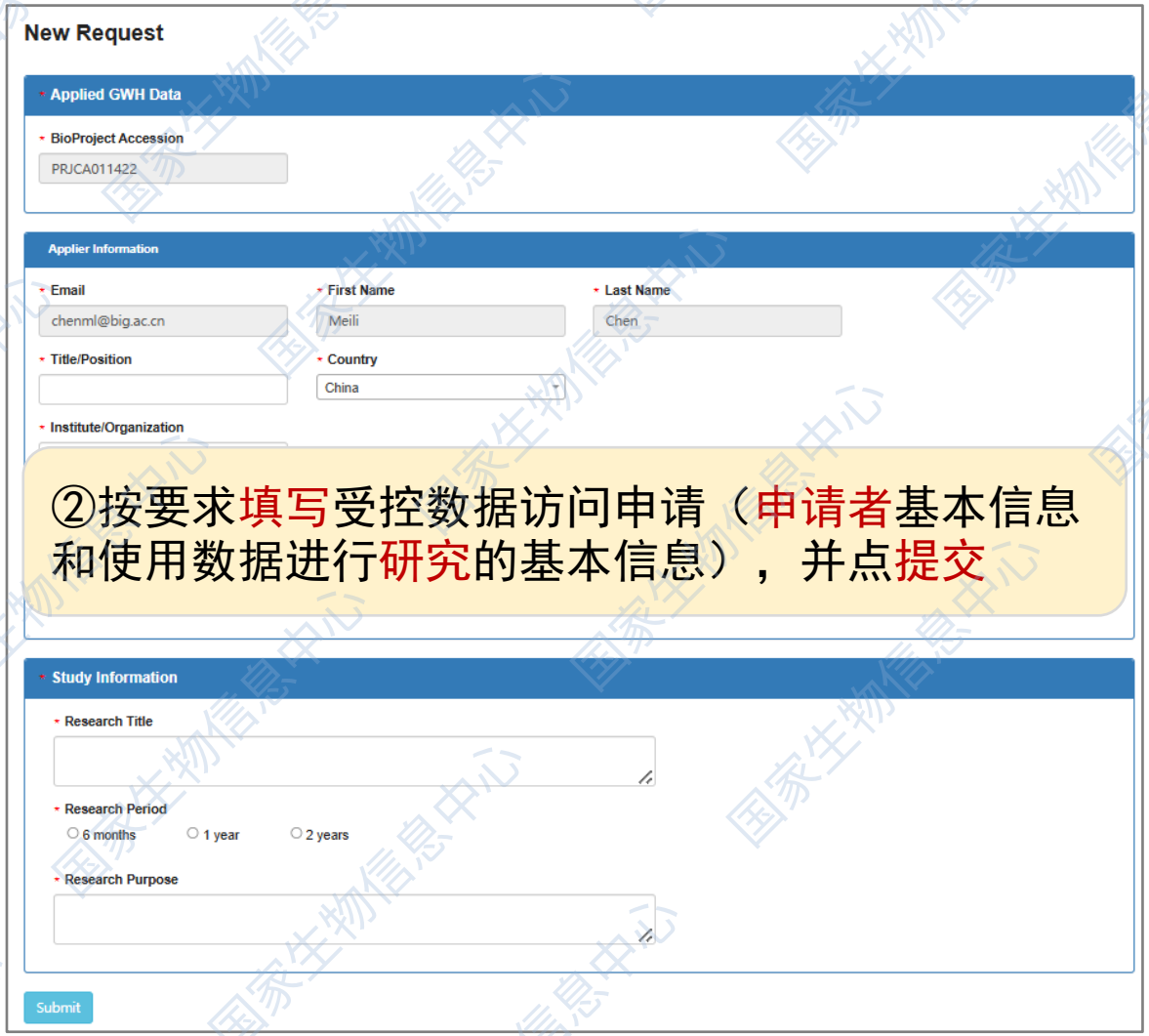
GWH人类基因组受控数据申请审核流程

人类基因组数据的发布，要求完成人类遗传资源信息管理备份和卫健委人类遗传资源服务管理系统的事先报告

人类基因组受控数据访问申请

- 访问申请步骤：

①从受控数据访问页面的下载行点击“注册和申请”



②按要求填写受控数据访问申请（申请者基本信息和使用数据进行研究的基本信息），并点提交

- DAC审核结果
  - 通过审核，分配下载权限
  - 拒绝申请

注：数据要在有效期内完成下载，过期后下载权限自动取消，FTP下载帮助文档见 https://ngdc.cncb.ac.cn/gwh/files/request/GWH_Controlled_Data_Download_Guide_CN.pdf

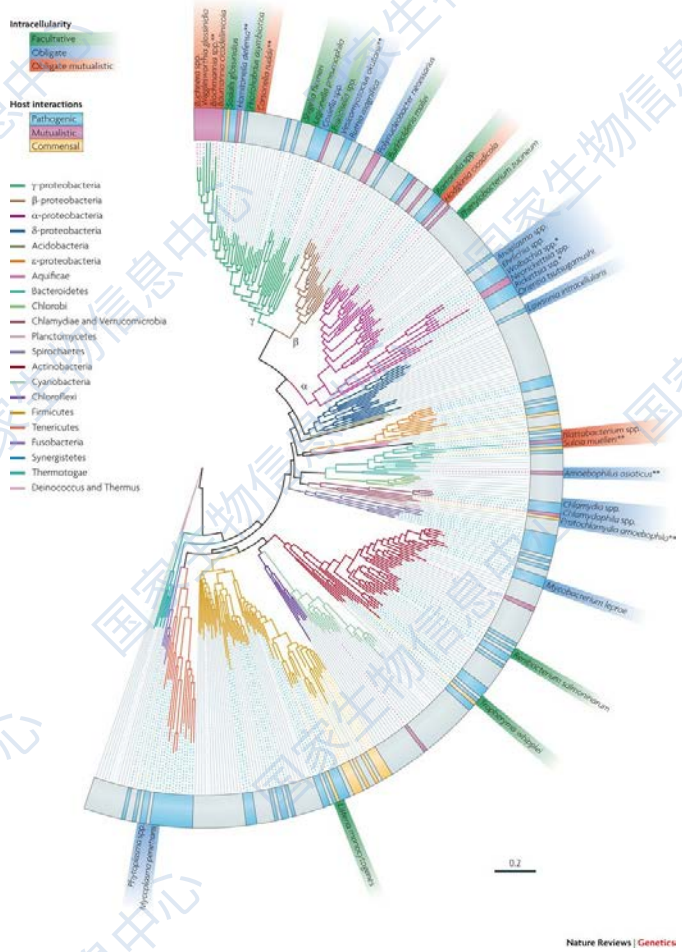
基因组和基因序列汇交的异同

数据库	数据汇交范围	数据类型	数据组织结构和关联
基因组数据库GWH	大基因组 • 真核生物全基因组 • 原核生物全基因组 • 元基因组组装数据	• **基因组**组装序列 • 注释数据（非必需）	**涉及**BioProject和BioSample
基因序列库GenBase	小基因组和基因片段 • 病毒基因组 • 细胞器基因组 • 质粒基因组 • 所有物种基因片段	• **基因片段**序列 • 注释数据（非必需）	**可以不涉及** BioProject 和 BioSample

目录

原核生物基因组研究的重要价值

- **生物学基础研究**
  - **基因组结构与功能**：研究原核生物基因功能与调控机制，为理解基本生物学过程提供模型
  - **进化研究**：通过比较不同原核生物基因组，可以揭示基因组演化的规律
- **生物技术**
  - **基因编辑**：了解原核生物的基因组结构可以为CRISPR等基因编辑技术的发展提供理论支持
  - **工程应用**：原核生物基因组的解析为基因工程和合成生物学提供了基础，如改造细菌以提高代谢产物的产量，应用于生物燃料和药物的生产
- **医学应用**
  - **病原体基因组**：对病原性细菌的基因组研究有助于理解其致病机制，推动新抗生素和疫苗的研发
  - **微生物组研究**：原核生物基因组的分析有助于揭示人类微生物组的组成及其与健康的关系



Nat Rev Genet **11**, 465–475 (2010)

基因组注释是原核生物基因组研究的基础

基因组注释是应用生物信息学方法和工具，识别基因组序列上的各种元素，包括编码基因、非编码RNA、重复序列和调控元件等，并推断它们的生物学功能。

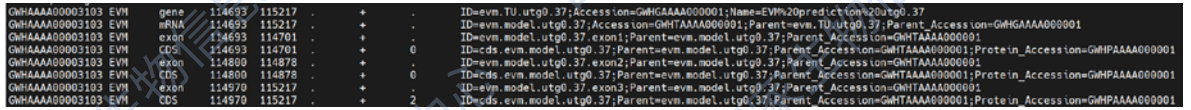
基因组注释的潜在应用：

基因组自动化重注释的必要性

汇交数据质量和价值有待提升：

- 基因组注释依赖生物信息学算法和工具，对用户的专业化程度要求较高
- 基因组数据注释比例低、质量参差不齐，导致重复利用度低
- 注释信息不完整、缺乏基因功能等重要信息
  - 仅有蛋白编码基因，无tRNA、rRNA、lncRNA等
  - 仅有基因结构，无基因名称、蛋白产物、基因功能等

GWH库中基因组注释信息汇交比例



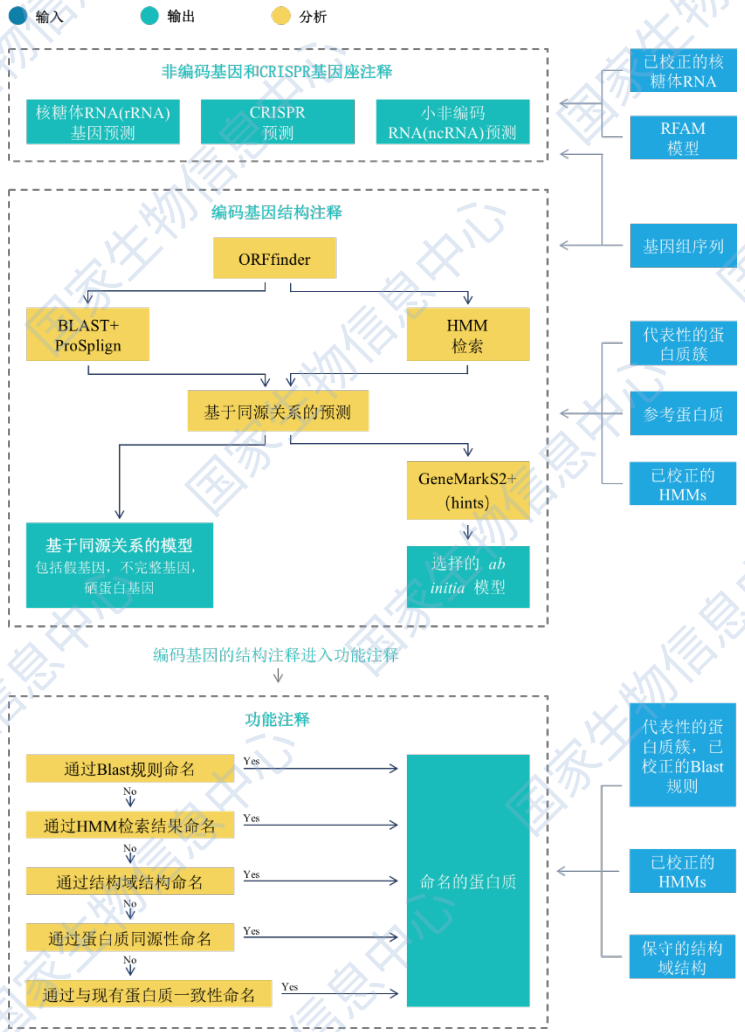
GWH库中基因组注释信息例子

目标：创建高效、准确的自动化重注释系统，促进原核生物基因组的深入研究与应用

原核生物基因组重注释流程—PGAP

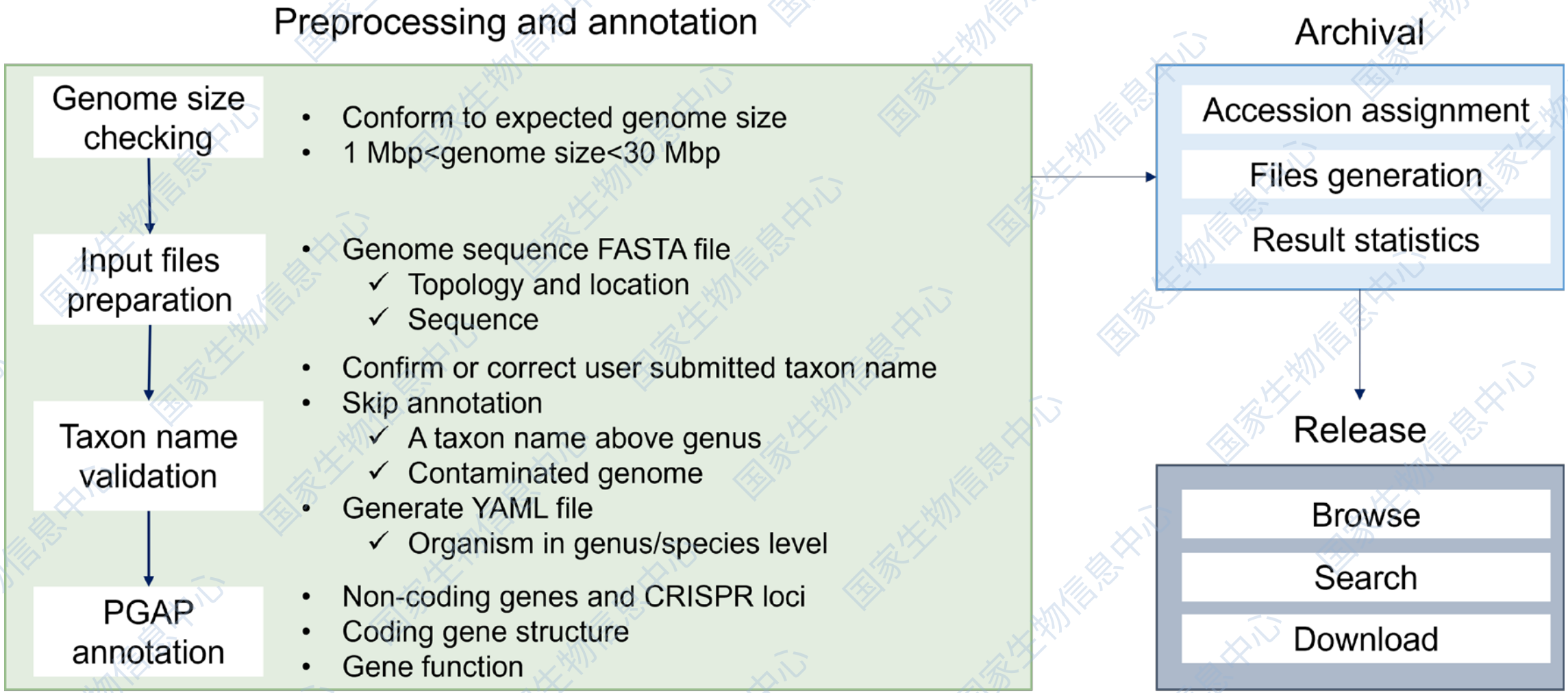
PGAP（Prokaryotic Genome Annotation Pipeline）：是NCBI提供的一款细菌、古细菌基因组和质粒自动化注释流程，包括蛋白质编码基因和结构性RNA、tRNA、小RNA、假基因、插入序列、转座子和其他移动单元等功能基因组单元的多层次预测。通过挖掘现有的蛋白质信息而进行调节，以建立新的核心蛋白簇，并根据来自于所提交细菌基因组的不断增加的大量可用证据，反复改善其注释功能。可产生高品质的注释，并遵守UniProt命名原则，旨在满足NCBI对于细菌基因组序列提交的INSDC标准。

Nucleic Acids Res. 44(14):6614-24 (2016)



（图片改自：https://www.ncbi.nlm.nih.gov/genome/annotation_prok/process/）

GWH库自动化重注释过程



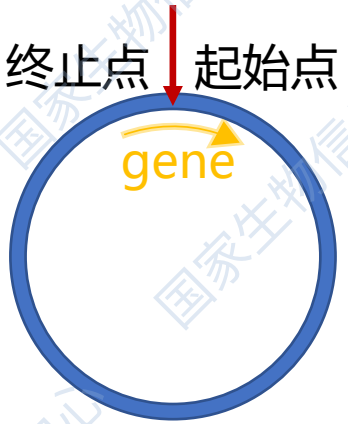
基因组序列文件预处理

- 格式：Fasta
- 辅助信息
  - Topology和序列定位，信息重要性
    - ✓可环化信息确定了基因组序列头尾是否可以含有N序列，基因是否可以跨越头尾
    - ✓定位信息确定了序列对应的genetic code，决定了蛋白翻译的准确性

注：未定义的默认为topology=linear, location=chromosome

例子：

>contig001 [location = plasmid] [plasmid-name = pABC01] [topology=circular]
cgaaagcctaatttcatttttcctcaattttagcttaaaagatttttttggaagcgggaaaatggaatgagcgcagcga
aaaaaagccaatgtggatatctttcgcctcgtttttttctatccacaatcattatcgacaacattttcacaaaactaactgc
tgtggacaatatttcttcaacagtttgtatagtgtgtgga…



基因组大小校验

- 将提交的基因组组装大小与物种的预期基因组大小范围进行比较，以识别可能由以下错误导致的异常值：
  - ✓ 不正确的物种分配
  - ✓ 宏基因组作为生物体基因组提交
  - ✓ 目标基因组组装子集未标记为部分基因组组成
  - ✓ 存在其他序列的严重污染

参考：https://ncbi.nlm.nih.gov/genbank/genome-size-check/
https://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/species_genome_size.txt.gz

- 如不存在预期基因组大小范围的物种，则要求基因组大小为1Mb~30Mb

物种名称校验

物种信息的准确性决定了注释过程中参考模型选取的准确性，PGAP提供物种校验功能，通过与参考库中的代表性基因组进行序列比对，基于平均核苷酸一致性(average nucleotide identity, ANI)计算方法，校验用户提交的物种名称是否与基因组序列匹配

- 对于大多数物种，高置信度定义
  - ANI阈值=96%，最小覆盖阈值=80%
- 校验结果
  - 可注释：确认/可校正
  - 跳过注释：存在污染/无法确定

注：要求提供的物种名称是genus/species水平的

物种名称校验失败

```
ANI report for assembly: <fasta_file_name>
Submitted organism: Staphylococcus aureus (taxid = 1280, rank = species, lineage = Bacteria; Firmicutes
Predicted organism: Staphylococcus aureus (taxid = 1280, rank = species, lineage = Bacteria; Firmicutes
Submitted organism has type: Yes
Status: CONTAMINATED
Confidence: HIGH
99.045 (54.5 80.3) 4972758 assembly  Ochrobactrum quorumnocens (GCA_002278035.1, ASM227803v1)
99.450 (31.5 94.1) 11348628 assembly  Staphylococcus aureus (GCA_006364675.1, ASM636467v1)
99.450 (31.5 94.3) 10960368 assembly  Staphylococcus aureus subsp. aureus (GCA_006094915.1, ASM609491v1
```

提交物种名称： *Staphylococcus aureus*

基于ANI预测物种名称： *Staphylococcus aureus*， ANI=99. 450% query_coverage=31.5%

置信度：高

校验状态：存在污染（*Ochrobactrum quorumnocens*），ANI=99.045% query_coverage=54.5%

target_coverage=80.3%

结果：跳过注释

PGAP注释过程

- 蛋白编码基因结构注释
  - 基于序列比对证据的有参预测

  （1）ORFfinder预测开放阅读框（ORFs）

  （2）ORF与蛋白质隐马尔可夫模型（HMMs）库进行搜索比对

  （3）ORF与特定谱系的参考蛋白质集进行比对分析，预测蛋白编码基因结构

  （4）噬菌体相关蛋白的注释：与噬菌体参考蛋白集进行比对
  - 从头预测： GeneMarkS-2+，并为依靠HMM证据的ORFs选择起始位点
  - 假基因的注释：除一些特殊情况外，移码突变或内部终止密码子会被注释为假基因
  - Partial gene的注释：无法找到完整的证据起点或终点的基因被注释为partial gene
  - 当partial gene靠近序列末端或gap时，将被翻译；如果在序列中间则标记假基因

PGAP注释过程

- 非编码基因和移动元件注释
  - **结构性RNA（5S、16S 和 23S rRNA）和sRNA：infernal**
  - **tRNA：** tRNAscan-SE
  - **CRISPR：** PILER-CR和CRISPR识别工具（CRT）
- 蛋白编码基因功能注释
  - 蛋白质家族模型：分配匹配的名称和属性（包括基因符号、相关文献以及EC编号）
  - 蛋白质命名：遵循国际蛋白质命名指南

PGAP软件安装要求

- 需要提前安装的软件
  - Python（Conda环境）： https://www.anaconda.com/download/success
  - Docker（或者Singularity、Podman）：
    - https://docs.docker.com/get-started/get-docker/
    - https://docs.sylabs.io/guides/3.5/admin-guide/installation.html
    - https://podman.io/
  - Git： https://git-scm.com/downloads
  - >100 GB硬盘
  - >4 GB内存

PGAP软件和参考库数据安装

- 安装PGAP运行封装脚本
  - curl -OL https://github.com/ncbi/pgap/raw/prod/scripts/pgap.py
  - 或： wget -O pgap.py https://github.com/ncbi/pgap/raw/prod/scripts/pgap.py
  - 增加执行权限： chmod +x pgap.py
- 下载Docker镜像
  - docker pull ncbi/pgap
  - docker pull ncbi/pgap-utils
- 下载测试基因组
  - wget https://s3.amazonaws.com/pgap-data/test_genomes-2024-07-18.build7555.tgz

注释测试基因组

- **运行PGAP的命令**
  - python3 pgap.py --no-self-update --ignore-all-errors --no-internet -r -o mg37_results -g PATH_TO_TEST_GENOMES/test_genomes/MG37/ASM2732v1.annotation.nucleotide.1.fasta -s 'Mycoplasmoides genitalium’
- **重要参数说明：**
  - --no-self-update 不自动更新程序（国内网络访问GitHub常出问题）
  - --ignore-all-errors 忽略序列质控错误，获得注释草图
  - --no-internet 整个流程运行时，不联网只使用本地数据，避免访问国内无法获取的数据导致注释失败

PGAP注释结果输出文件

- *.fasta - 你提供的核苷酸FASTA文件
- ani-tax-report.txt - ANI报告的文本格式
- ani-tax-report.xml - 供机器处理的ANI报告，XML格式
- annot-gb.ent - GenBank格式的ASN.1文件（Seq-entry）
- annot.faa - 包含所有蛋白质的FASTA文件
- annot.fna - 包含所有核苷酸的FASTA文件（注意：该文件可能是您输入的核苷酸FASTA文件规范化的结果）
- annot.gbk - 注释的平面文件格式
- annot.gff - 注释的GFF3格式
- annot.sqn - Seq-submit格式的ASN.1文件
- annot_cds_from_genomic.fna - 包含所有编码区核苷酸序列的FASTA文件
- annot_translated_cds.faa - 包含所有编码区翻译序列的FASTA文件
- annot_with_genomic_fasta.gff - 将注释的GFF格式与FASTA格式的核苷酸序列相结合的文件
- checkm.txt - 该基因组的CheckM输出结果
- cwltool.log - CWL工具日志，可用于分析失败的原因
- fastaval.xml - 输入FASTA文件的验证结果XML文件

PGAP网络资源

- 源代码：https://github.com/ncbi/pgap
- Docker镜像主页：https://hub.docker.com/r/ncbi/pgap
- Wiki（使用说明）： https://github.com/ncbi/pgap/wiki
- 软件和流程的详细解释：
  - 概述：https://www.ncbi.nlm.nih.gov/refseq/annotation_prok/
  - 注释流程：https://www.ncbi.nlm.nih.gov/refseq/annotation_prok/process
  - 注释标准： https://www.ncbi.nlm.nih.gov/refseq/annotation_prok/standards
  - 基因组的筛选： https://www.ncbi.nlm.nih.gov/assembly/help/anomnotrefseq/
  - 软件升级信息：https://www.ncbi.nlm.nih.gov/refseq/annotation_prok/release_notes/

原核生物基因组注释案例

- 第一步：“*Acinetobacter calcoaceticus*”基因组序列数据准备

```
>GWHAAAE00000001 [location=Chromosome] [topology=circular]
cgaaagcctaatttcatttttcctcaattttagcttaaaagatttttttggaagcgggaaaatggaatgagcgcagcgaaaaaaagccaatgtggatatc
tttcgcctcgtttttttctatccacaatcattatcgacaacattttcacaaaactaactgctgtggacaatatttcttcaacagtttgtatagtgtgtgga
taagttcctcgcaaccattgcaaccactcgcttattctgatattatatttgtgttttaactcttgataacaaattggctgccaatccattatccacaaac
tgtggataagttgtggaaagttttttcacagggtgtgcagtaatttgtccacatcttgtgaagaatgtcgaaaagacgtttttctactatattatatgtt
ttcaacatttcattaacgaatggactcatccatttgctctttttttgtgttctataacaggttggacaagcaaatattgctggtaaaggagggacgaacc
gccattatgaaaaacatatcggatctttggaatcaagctttggggcagatcgaaaaaaaattgagcaagcccagctttgaaacatggatgaaatcgacaa
aggctcattcattgcagggcgatacgctgatcatcaccgcaccgaacgagttcgccagagactggcttgaatcaagatacctgcacctgatcgccgatac
gatctacgatctgacaggagaagaattgagcattaaatttgtcattcctcagaatcaaaatgaagaagattttatgcctaagtctccaatcaaaaaaatg
tcgaaagaagaaccggctgatttttccgcaaaacatgctgaatcccaaatatacattcgatacgttcgttatcggttcaggaaaccgattcgcccatgcgg
cgtctttggcagtggccgaagccccggcgaaagcttacaatccgctgtttatttatggggagtcggacttggaaagactcacttaatgcatgcgatcgg
gcactatgtcatcgatcacaatccatctgcaaaagtggttttatttgtcatctgagaaatttacaaatgaattcattaactcgatccgagacaataaagct
gtcgattttcgcaatcgctatcgaaatgttgacgttcttttgatagatgatattcaatttttagccggaaaagaacaaacgcaagaggaatttttccata
cgtttaatacgctgcatgaagaaacaaagcagatcgtcatttccagcgaccggcctccaaaagagatcccaacgcttgaagaccgtttgcgctcccgttt
tgaatggggattgatcacagacatcacgcctcctgatctggaaacaagaattgcgattttaagaaagaaagcaaaagcagaaggacttgatatcccgaat
gaagtcatgctttatattgccaatcagatcgacagcaatatcagggagctggaaggggcattaatcagggttgtcgcatattcttccctgatcaataaag
acatcaacgccgatctggctgccgaagctttgaaagatatcattccgtcttcaaagccgaaagtcattacgatcaaagacatccaaagaatcgtcggcca
gcagttcaatatcaagctggaagatttcaaagcgaagaaacggacaaaatcggtagcttttccgaggcagatcgctatgtatctatcaagagaaatgaca
gattcttctcttccgaagatcggcgaagagtttggcggacgcgaccacacgactgtcatccatgcccatgagaaaatatcaaaactgctgagcgatgatg
aacagctccagcagcagattaaagaaattaaagagcagctgagataagtggtgagcgggaaagtgtgaataacttgaacatagtcttacacagtctgtcc
acatgtggataggctgtatttccgttcgttttttacacttatccacaaatccacagcccctactagtacttctgctatttttataaaacataataattaa
tagattcccgcaaggaggatgattatgaagtttacaattcaaaaagaccgcctagtcgaaggtgtccaagatgtgttaaaagctgtttcttcaagaacga
cgataccgatcttaaccggtattaaaattgtggcctctgatgaaggtgtctctctgacaggaagcgattccgatatttcgatcgaatcgtttatcccgaa
agaagacggcgatttagaggtcgtgacaattgaacagcccggcagcattgtgcttcaagcccgttttttcagtgaaattgtcaaaaagctgccgatgtca
acggtggaaatcgaggttcaaaatcaatacttaacgattatccgctccggcaaagcagagtttaacttaaacggtttagatgcgagcgaatatccgcttt
tgccgcaaattgaagagcatcacgcttttcaaattccgaccgatctgctaaaaaacctgatccgccaaaccgttttcgcagtgtccacctcagaaacacg
cccaatcttgacaggtgtaaactggaatgtcgcgagcggtgaattaatatgcactgcgacggatagtcaccgtctagcgctaagaaaagctaagctcgac
attcacgaagacagttcttacaatgtcgtcatcccaggaaaaagcttaaccgagctgagcaagatccttgatgaccaccaggagcttgtcgatattgtaa
ttaccgaaacacaagtgttgtttaaaacaaaaaacgtgctgtttttctccagactgcttgacggaaactacccggatacaaaccgcctgattcctcagga
```

- 第二步：校验物种大小
  - 提交的物种名称： *Acinetobacter calcoaceticus*
  - 提交的基因组大小： 4,198,078 bp
  - 预期物种的基因组大小：[3,200,000　4,700,000]
  - **结论：符合预期**


```
▼<genome_size_response>
  ▼<input>
    <species_taxid>471</species_taxid>
    <length>4198078</length>
  </input>
  <organism_name>Acinetobacter calcoaceticus</organism_name>
  <species_taxid>471</species_taxid>
  <size_source>species</size_source>
  <genome_count>42</genome_count>
  <expected_ungapped_length>3950000</expected_ungapped_length>
  <minimum_ungapped_length>3200000</minimum_ungapped_length>
  <maximum_ungapped_length>4700000</maximum_ungapped_length>
  <length_status>within_range</length_status>
</genome_size_response>
```


NCBI的API接口：
https://api.ncbi.nlm.nih.gov/genome/v0/expected_genome_size?species_taxid=471&length=4198078

- 第三步：运行PGAP的物种名称校验模块
  - 运行命令

python3 ./pgap.py **--taxcheck-only** --no-self-update --use-version 2024-07-18.build7555 -r -o ./taxcheck --ignore-all-errors --no-internet -g ./GWHDODT00000000.genome.fasta –s “*Acinetobacter calcoaceticus*”

  - 输出结果文件：ani-tax-report.txt
  - 结论：物种名称不匹配，高置信度的匹配物种名称为 “*Acinetobacter pittii*”
  - 处理方式：将物种名称修改为“*Acinetobacter pittii*”后再注释

```
ANI report for assembly: GWHDODT00000000.genome.fasta
Submitted organism: Acinetobacter calcoaceticus (taxid = 47
proteobacteria; Moraxellales; Moraxellaceae; Acinetobacter;
Best match: Acinetobacter pittii (taxid = 48296, rank = spec
a; Moraxellales; Moraxellaceae; Acinetobacter; Acinetobacte
Submitted organism has type: Yes
Status: MISASSIGNED
Confidence: HIGH
Table legend:
ANI  : ANI value between this assembly and the type listed
(Coverages) : query-coverage and subject-coverage of this a
NewSeq : the count of bases best assigned to this type asse
Assembly : Release-id of the type-assembly (this value matc
Organism : Organism of this type-assembly
(assembly_accession, assembly_name) : of this type-assembly

ANI    (Coverages)   NewSeq  Assembly  Organism      (assem
96.771 (84.5 90.5)  1041226   1530028  Acinetobacter pittii
96.840 (82.6 90.5)  2441700    596368  Acinetobacter pittii
```

• 第四步：PGAP注释所需的所有输入文件

（1）基因组序列文件（.fasta格式，含拓扑结构和定位信息）

```
>GWHAAAE00000001 [location=Chromosome] [topology=circular]
cgaaagcctaatttcatttttcctcaattttagcttaaaagatttttttggaagcgggaaaatggaatgagcgcagcgaaaaaaagccaatgtggatatc
tttcgcctcgtttttttctatccacaatcattatcgacaacattttcacaaaactaactgctgtggacaatatttcttcaacagtttgtatagtgtgtgga
taagttcctcgcaaccattgcaaccactcgcttattctgatattatatttgtgttttaactcttgataacaaattggctgccaatccattatccacaaac
tgtggataagttgtggaaagtttttttcacagggtgtgcagtaatttgtccacatcttgtgaagaatgtcgaaaagacgtttttctactatattatatgtt
ttcaacatttcattaacgaatggactcatccatttgctctttttttgtgttctataacaggttggacaagcaaatatgctggtaaaggagggacgaacc
gccattatgaaaaacatatcggatctttggaatcaagctttggggcagatcgaaaaaaaattgagcaagcccagctttgaaacatggatgaaatcgacaa
aggctcattcattgcagggcgatacgctgatcatcaccgcaccgaacgagttcgccagagactggcttgaatcaagatacctgcacctgatcgccgatac
gatctacgatctgacaggagaagaattgagcattaaatttgtcattcctcagaatcaaaatgaagaagatttttatgcctaagtctccaatcaaaaaaatg
tcgaaagaagaaccggctgattttccgcaaaacatgctgaatcccaaatatacattcgatacgttcgttatcggttcaggaaaccgattcgcccatgcgg
cgtctttggcagtggccgaagccccggcgaaagcttacaatccgctgtttatttatggggagtcggacttggaaagactcacttaatgcatgcgatcgg
gcactatgtcatcgatcacaatccatctgcaaaagtggtttatttgtcatctgagaaatttacaaatgaattcattaactcgatccgagacaataaagct
gtcgattttcgcaatcgctatcgaaatgttgacgttcttttgatagatgatattcaatttttagccggaaaagaacaaacgcaagaggaatttttccata
cgtttaatacgctgcatgaagaaacaaagcagatcgtcatttccagcgaccggcctccaaaagagatcccaacgcttgaagaccgtttgcgctcccgttt
tgaatggggattgatcacagacatcacgcctcctgatctggaaacaagaattgcgattttaagaaagaaagcaaaagcagaaggacttgatatcccgaat
gaagtcatgctttatattgccaatcagatcgacagcaatatcagggagctggaaggggcattaatcagggttgtcgcatattcttccctgatcaataaag
acatcaacgccgatctggctgccgaagctttgaaagatatcattccgtcttcaaagccgaaagtcattacgatcaaagacatccaaagaatcgtcggcca
gcagttcaatatcaagctggaagatttcaaagcgaagaaacggacaaaatcggtagcttttccgaggcagatcgctatgtatctatcaagagaaatgaca
gattcttctcttccgaagatcggcgaagagtttggcggacgcgaccacacgactgtcatccatgcccatgagaaaatatcaaaactgctgagcgatgatg
aacagctccagcagcagattaaagaaattaaagagcagctgagataagtggtgagcgggaaagtgtgaataacttgaacatagtcttacacagtctgtcc
acatgtggataggctgtatttccgttcgttttttacacttatccacaaatccacagcccctactagtacttctgctatttttataaaacataataattaa
tagattcccgcaaggaggatgattatgaagtttacaattcaaaaagaccgcctagtcgaaggtgtccaagatgtgttaaaagctgtttcttcaagaacga
cgataccgatcttaaccggtattaaaattgtggcctctgatgaaggtgtctctctgacaggaagcgattccgatatttcgatcgaatcgtttatcccgaa
agaagacggcgatttagaggtcgtgacaattgaacagccccggcagcattgtgcttcaagcccgtttttttcagtgaaattgtcaaaaagctgccgatgtca
acggtggaaatcgaggttcaaaatcaatacttaacgattatccgctccggcaaagcagagtttaacttaaacggtttagatgcgagcgaatatccgctttt
tgccgcaaattgaagagcatcacgcttttcaaattccgaccgatctgctaaaaaacctgatccgccaaaccgttttcgcagtgtccacctcagaaacacg
cccaatcttgacaggtgtaaactggaatgtcgcgagcggtgaattaatatgcactgcgacggatagtcaccgtctagcgctaagaaagctaagctcgac
attcacgaagacagttcttacaatgtcgtcatcccaggaaaaagcttaaccgagctgagcaagatccttgatgaccaccaggagcttgtcgatattgtaa
ttaccgaaacacaagtgttgtttaaaacaaaaaacgtgctgttttctccagactgcttgacggaaactacccggatacaaaccgcctgattcctcagga
```

（2）基因组元信息文件（.yaml格式）

```yaml
organism:
    genus_species: 'Acinetobacter pittii'
    strain: ''
contact_info:
    last_name: 'Chen'
    first_name: 'Meili'
    email: 'chenml@big.ac.cn'
    organization: 'BIG, CAS'
    department: 'NGDC'
    phone: '+8610-84097298'
    fax: '+8610-84090000'
    street: 'Building 104, West Beichen Road'
    city: 'Beijing'
    state: 'Beijing'
    postal_code: '100101'
    country: 'China'
authors:
    - author:
        last_name: 'Chen'
        first_name: 'Meili'
consortium: 'Genome Warehouse'
locus_tag_prefix: 'dodt0001'
```

（3）输入文件信息（.yaml格式，如input.yaml）

```yaml
fasta:
    class: File
    location: GWHDODT00000000.genome.fasta
submol:
    class: File
    location: submol.yaml
```

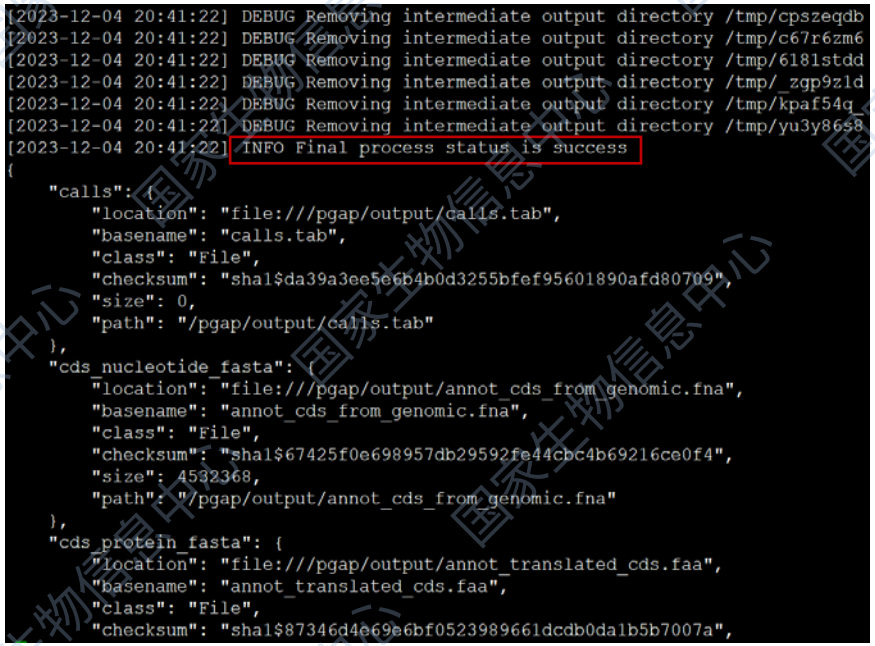
- 第五步：运行PGAP进行基因组注释
  - 运行命令

python3 ./pgap.py --no-self-update --use-version 2024-07-18.build7555 -r -o ./annotation ./input.yaml --ignore-all-errors --no-internet

  - 注释成功/失败判断：运行日志cwltool.log中，**成功**则输出“INFO Final process status is success”，**失败**则输出“Permanent Fail“及失败原因

```
annot_cds_from_genomic.fna  annot.gbk  annot_translated_cds.faa    cwltool.log
annot.faa                   annot.gff  annot_with_genomic_fasta.gff  fastaval.xml
annot.fna                   annot.sqn  checkm.txt                  GWHDODT00000000.genome.fasta
```

注释成功输出结果文件列表



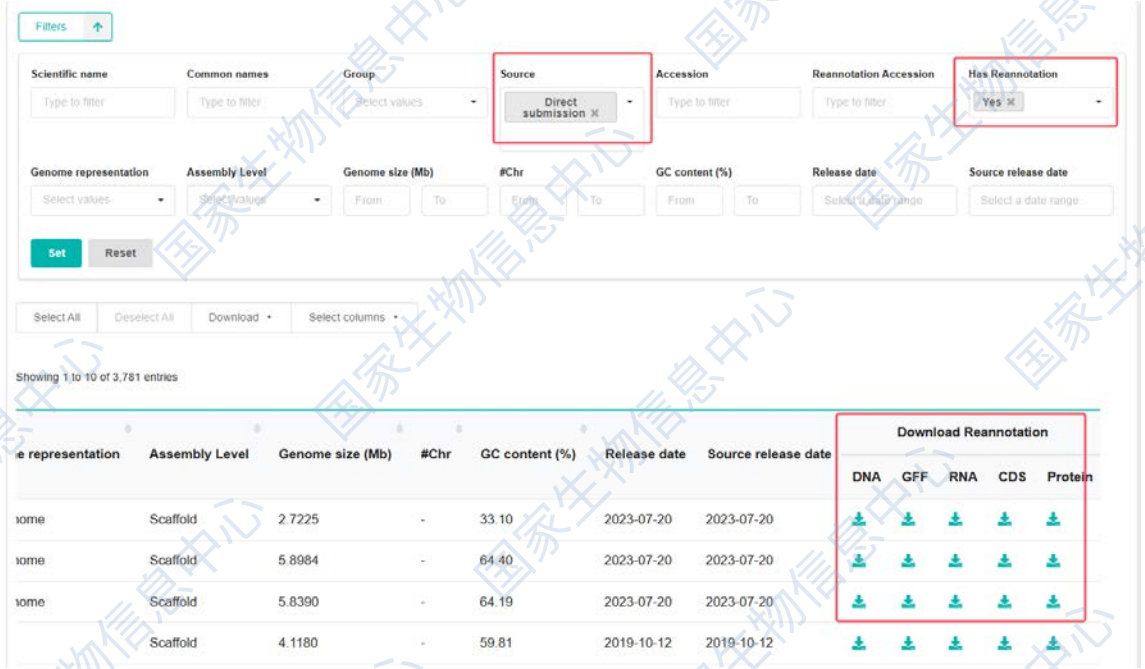
cwltool.log输出文件截图

- 输出的注释文件annot.gff

```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
##sequence-region   1 265398
##species https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=48296
GWHDODT00000001 Local   region  1       265398  .       +       .       ID=GWHDODT00000001:1..265398;Dbxref=taxon:48296
GWHDODT00000001 .       gene    121     333     .       +       .       ID=gene-dodt0001_000001;Name=dodt0001_000001;gb
GWHDODT00000001 Protein Homology        CDS     121     333     .       +       0       ID=cds-dodt0001_000001;Parent=g
GWHDODT00000001 .       pseudogene      330     1316    .       -       .       ID=gene-dodt0001_000002;Name=dodt0001_0
GWHDODT00000001 Protein Homology        CDS     330     1316    .       -       0       ID=cds-dodt0001_000002;Parent=g
GWHDODT00000001 .       gene    1614    2549    .       +       .       ID=gene-dodt0001_000003;Name=yadG;gbkey=Gene;ge
GWHDODT00000001 Protein Homology        CDS     1614    2549    .       +       0       ID=cds-dodt0001_000003;Parent=g
GWHDODT00000001 .       gene    2546    3319    .       +       .       ID=gene-dodt0001_000004;Name=yadH;gbkey=Gene;ge
GWHDODT00000001 Protein Homology        CDS     2546    3319    .       +       0       ID=cds-dodt0001_000004;Parent=g
GWHDODT00000001 .       gene    3316    4128    .       +       .       ID=gene-dodt0001_000005;Name=queF;gbkey=Gene;ge
GWHDODT00000001 Protein Homology        CDS     3316    4128    .       +       0       ID=cds-dodt0001_000005;Parent=g
GWHDODT00000001 .       gene    4152    4802    .       +       .       ID=gene-dodt0001_000006;Name=dodt0001_000006;gb
GWHDODT00000001 Protein Homology        CDS     4152    4802    .       +       0       ID=cds-dodt0001_000006;Parent=g
GWHDODT00000001 .       gene    4929    6071    .       +       .       ID=gene-dodt0001_000007;Name=rodA;gbkey=Gene;ge
GWHDODT00000001 Protein Homology        CDS     4929    6071    .       +       0       ID=cds-dodt0001_000007;Parent=g
GWHDODT00000001 .       gene    6091    7092    .       +       .       ID=gene-dodt0001_000008;Name=mltB;gbkey=Gene;ge
GWHDODT00000001 Protein Homology        CDS     6091    7092    .       +       0       ID=cds-dodt0001_000008;Parent=g
GWHDODT00000001 .       gene    7382    7993    .       +       .       ID=gene-dodt0001_000009;Name=dodt0001_000009;gb
GWHDODT00000001 Protein Homology        CDS     7382    7993    .       +       0       ID=cds-dodt0001_000009;Parent=g
GWHDODT00000001 .       gene    8567    9016    .       +       .       ID=gene-dodt0001_000010;Name=dodt0001_000010;gb
GWHDODT00000001 Protein Homology        CDS     8567    9016    .       +       0       ID=cds-dodt0001_000010;Parent=g
GWHDODT00000001 .       gene    9060    9377    .       -       .       ID=gene-dodt0001_000011;Name=dodt0001_000011;gb
GWHDODT00000001 Protein Homology        CDS     9060    9377    .       -       0       ID=cds-dodt0001_000011;Parent=g
GWHDODT00000001 .       gene    9379    10095   .       -       .       ID=gene-dodt0001_000012;Name=dodt0001_000012;gb
annot.gff
```

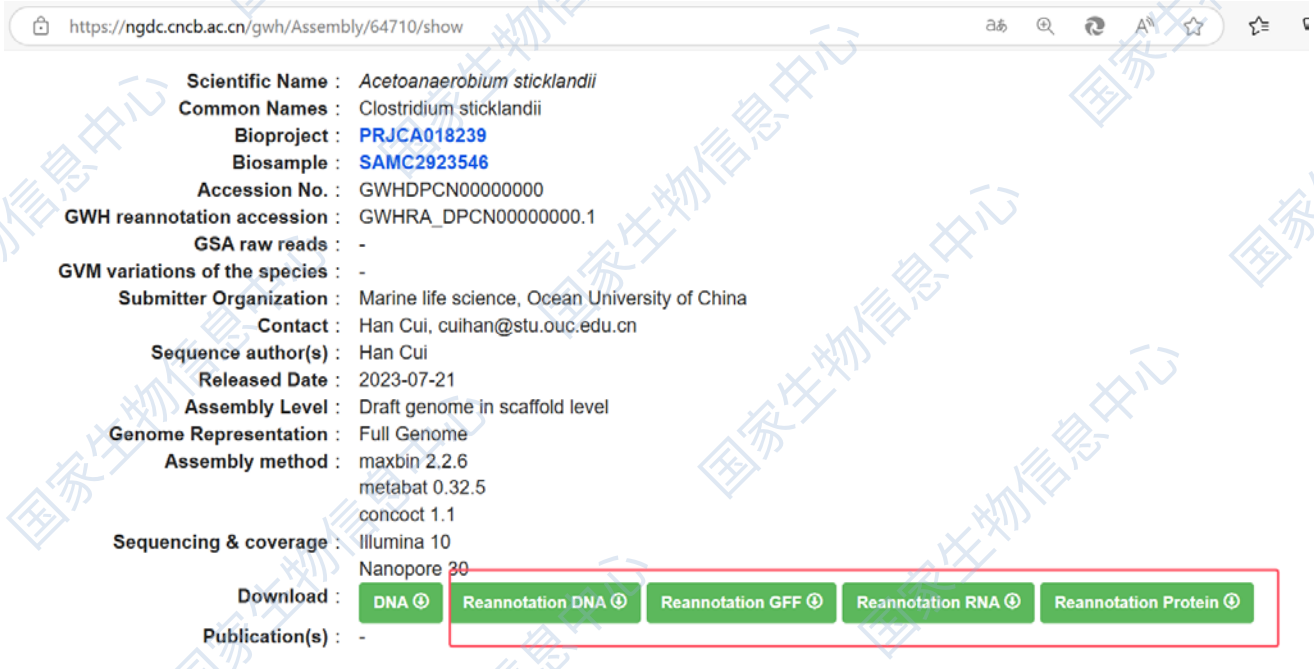
GWH库原核基因组重注释信息获取

- 浏览页面



https://ngdc.cncb.ac.cn/gwh/browse/assembly

- Assembly详细页面



https://ngdc.cncb.ac.cn/gwh/Assembly/64710/show

致谢

- GWH库全体成员

鲍一明 研究员

陈梅丽 高级工程师
GWH组组长

赵学彤 工程师
数据审编

马英克 助理研究员
基因组重注释资源建设

韩镇先 工程师
数据库开发和维护

CSCB 中国细胞生物学学会
CHINESE SOCIETY FOR CELL BIOLOGY

国家基因组科学数据中心
National Genomics Data Center

中国科学院北京基因组研究所（国家生物信息中心）
BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

中国细胞生物学学会功能基因组信息学与系统生物学分会

中国科学院
CHINESE ACADEMY OF SCIENCES

中华人民共和国科学技术部
Ministry of Science and Technology of the People's Republic of China

100 years IUBS Unifying biology through diversity

THANKS

欢迎将数据递交到国家基因组科学数据中心！

https://ngdc.cncb.ac.cn/gwh

联系方式：

邮箱： gwhcuration@big.ac.cn

gwh@big.ac.cn

电话：010-84097298

QQ群： 183915274