



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



国家基因组科学数据中心

National Genomics Data Center

2021年国家基因组科学数据中心 科学计划数据汇交培训会

国家生物信息中心

China National Center for Bioinformation

培训主要内容

- 项目科学数据汇交总体流程
- 生物项目数据库 (BioProject)
- 生物样本数据库 (BioSample)
- 组学原始归档库 (GSA)
- 人类遗传组学原始归档库 (GSA-Human)
- 基因组变异数据库 (GVM)
- 基因组数据库 (GWH)
- BioCode数据库
- OMIX数据库





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



国家基因组科学数据中心

National Genomics Data Center

项目科学数据汇交总体流程

国家生物信息中心

China National Center for Bioinformation

提纲

- 数据汇交背景
- 数据汇交计划制定
- 数据提交与审核
- 出具数据汇交证明



提纲

- **数据汇交背景**
- 数据汇交计划制定
- 数据提交与审核
- 出具数据汇交证明



数据汇交背景

中国生物技术发展中心的专项项目数据汇交工作的通知

中国生物技术发展中心

关于开展重点专项项目科学数据汇交工作的通知

各项目牵头单位、项目负责人：

按照《科学数据管理办法》要求，重点专项项目形成的科学数据，应由项目牵头单位汇交到相关科学数据中心。为更好落实相关文件要求，进一步推动科学数据汇交工作，有关通知如下：

一、汇交原则

1.及时充分。科技计划项目所形成的科学数据应在规定的程序和期限内，按时、完整汇交到科学数据中心，确保数据的及时性和完整性。

2.真实可靠。科技计划项目形成的科学数据应按照科技计划项目实际产生科学数据情况进行汇交，确保所汇交的科学数据的质量，保证数据的真实性和可靠性。

3.科学规范。科技计划项目所汇交的科学数据应按照相关科学数据汇交要求的标准规范加工处理，确保汇交科学数据可发现、可获取、互操作和可重复利用。

二、汇交的责任主体

1.科技计划项目管理方。科技计划项目管理方为中国生

物技术发展中心（以下简称生物中心）。

2.科学数据提交方。科学数据提交方为科技计划项目的牵头法人单位，是科学数据汇交的责任人和义务人。

3.科学数据管理方。科学数据管理方主要指各类科学数据中心，是科学数据管理和开放共享的重要载体。

三、汇交的主要内容

科学数据汇交内容包括项目执行过程中产生的科学数据实体、科学数据描述信息和科学数据辅助工具软件。

1.科学数据实体。科学数据实体是在科技计划形成的原始数据及基于原始数据或研究分析数据所形成的完整数据库或数据文件。

2.科学数据描述信息。科学数据必须提供相应的描述信息，包括数据说明文档及过程资料等。

3.科学数据辅助工具软件。科学数据辅助工具软件是指科技计划形成的用于科学数据处理、加工和分析的专门辅助软件工具等。项目实施中采购的工具软件可不必汇交，但需提供与汇交数据处理相关的工具软件使用说明。

四、汇交的基本流程

（一）已完成验收的项目

1.制定科学数据汇交计划。项目承担单位自行选择1个科学数据中心（清单见附件1）作为数据汇交终端，并按要求编制“科学数据汇交计划”（格式见附件2）。数据汇交计划

- 汇交原则
- 汇交的责任主体
- 汇交的主要内容
- 汇交的基本流程



数据汇交背景

• 汇交的主要内容

科学数据实体

- 项目形成的**原始数据**及基于原始数据或**研究分析数据**所形成的**完整数据库或数据文件**



- ✓ **自产**多组学原始数据
- ✓ **自产**多组学的分析后数据

科学数据描述信息

- 提供相应的**描述信息**，包括数据说明文档及过程资料等



- ✓ 多组学原始及分析后数据的项目、样本、实验、数据文件等描述元数据信息

科学数据辅助工具

- 项目形成的用于科学数据处理、加工和分析的专门**辅助软件工具**等

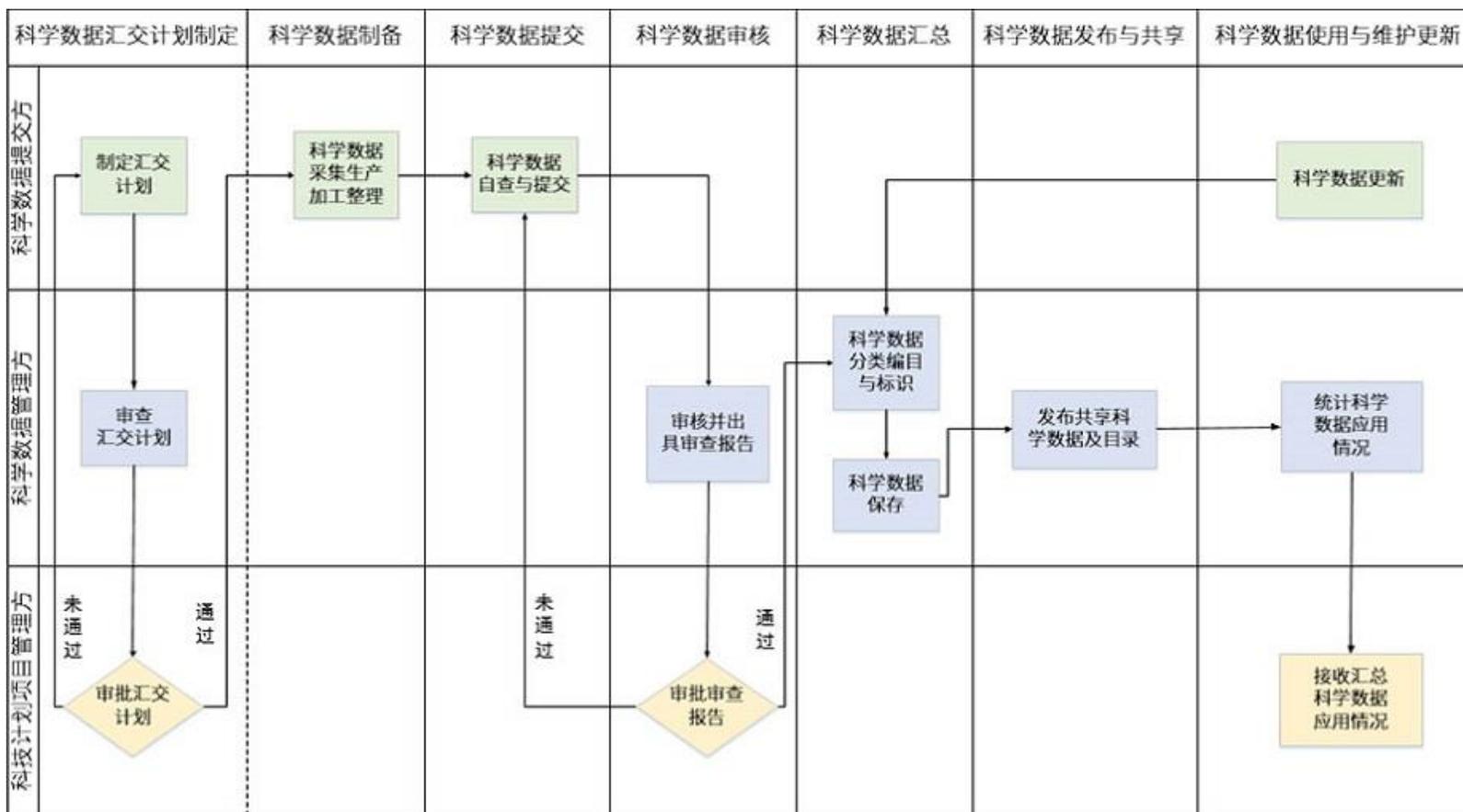


- ✓ 辅助软件工具源代码等相关信息
- ✓ 采购的工具软件不必汇交



数据汇交背景

- 汇交的流程：<https://www.escience.org.cn/data-archive/progress>



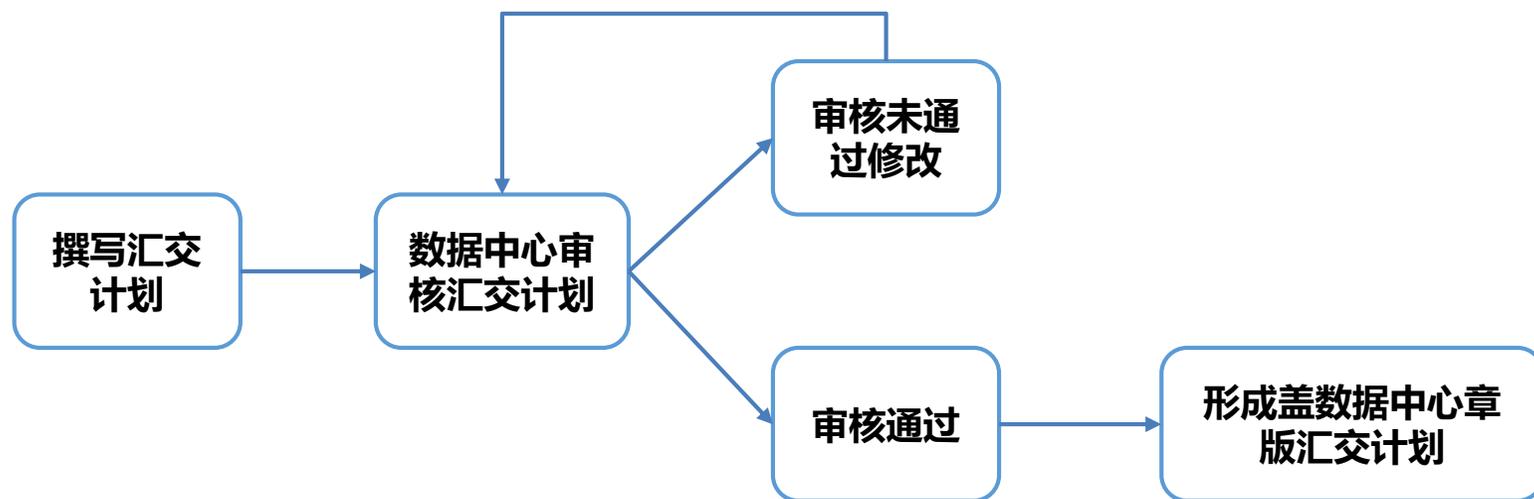
提纲

- 数据汇交背景
- **数据汇交计划制定**
- 数据提交与审核
- 数据汇交证明出具



数据汇交计划制定

- 中心汇交计划审核流程



数据汇交计划制定

- 中心可接收的数据类型

基因组学数据

- 组学原始测序数据
- 基因组组装
- 基因组变异
- 芯片数据
- 生信分析后数据

功能基因组学数据

- 代谢组（包括脂质组）
- 蛋白质组

调研型数据

- 临床信息
- 人口统计学数据
- 问卷调查
- 影像等

算法工具类

- 源代码
- 使用说明
- 文章等



数据汇交计划制定

• 中心可接收的数据类型

数据库简称	描述信息	接收数据格式	备注
GSA	非人类的组学原始测试下机数据的归档库	Fastq/Bam/Sff/Ab1/Solid_Native/PacBio_HDF5/PacBio_Sequel_Native/Oxford_Nanopore_Native/Bnx/10xGenomics/Helicos_Native/Complete_Genomics_Native	工作邮箱: gsa@big.ac.cn QQ 群: 548170081 帮助文档: GSA使用说明 GSA-Human使用说明
GSA-Human	人类的组学原始测试下机数据归档库		
OMIX	分析后的转录组、表观组等基因组学二级数据	结果文件, 如abundance measurements/raw count/txt/bed等	工作邮箱: gsa@big.ac.cn QQ 群: 516418530
	芯片数据	xls/xlsx/csv等	
	代谢组 (包括脂质组)	RAW文件: raw/wiff/scan/baf/mzXML/mzML 或tar格式打包上传	
	蛋白组	Result/search文件: prideXML/mzIdentML/XML/txt 或tar打包上传	
	调研性数据 (临床信息、人口统计学数据、问卷等)	txt等或tar打包上传	
	其他 (所有符合科技要求的科学研究类科学数据)	txt等或tar打包上传	
GWH	基因组组装数据归档库	拼接文件格式: fasta	工作邮箱: gwh@big.ac.cn QQ 群: 541196594 帮助文档: GWH使用说明
		注释文件格式: gff/tbl	
GVM	基因组序列变异数据归档库	VCF/GVCF	工作邮箱: gvm@big.ac.cn QQ 群: 468638108 帮助文档: GVM使用说明
BioCode	软件工具和方法的数据库	源代码等	工作邮箱: biocode@big.ac.cn QQ群: 537473603



提纲

- 数据汇交背景
- 数据汇交计划制定
- **数据提交与审核**
- 数据汇交证明出具



数据提交与审核

- BIG Sub统一递交入口: <https://bigd.big.ac.cn/gsub>

The screenshot shows the BIG Sub submission portal interface. At the top, there is a navigation bar with links for Databases, Tools, Standards, Publications, and About. Below this is the BIG Sub logo and the text 'BIG Submission Portal'. There are language selection buttons for '中文' and 'English', and links for 'Home', 'Documentation', 'Login', and 'Register'. The main content area is titled '请选择您要提交的数据类型' (Please select the data type you want to submit). Below this title, there is a brief description of the system: '生物数据递交系统(BIG Submission, BIG Sub)是国家基因组科学数据中心生物数据统一递交入口, 为用户提供一站式数据递交服务。' (The biological data submission system (BIG Submission, BIG Sub) is the unified submission portal for biological data at the National Genomics Science Data Center, providing a one-stop data submission service for users.)

The main content area displays seven data submission options, each with a title, description, and icon:

- Genome Sequence Archive (GSA)**: 组学原始数据汇交、存储、管理与共享系统 (Genomics original data submission, storage, management and sharing system)
- GSA for Human**: 人类遗传资源组学原始数据库 (Human genetic resource genomics original database)
- Genome Warehouse (GWH)**: 多物种基因组组装序列和注释数据归档库 (Multi-species genome assembly sequences and annotation data archive)
- BioProject**: 收集与共享生物学研究项目信息的资源库 (Resource for collecting and sharing biological research project information)
- BioSample**: 收集与共享实验相关的生物样本信息的资源库 (Resource for collecting and sharing biological sample information related to experiments)
- Genome Variation Map (GVM)**: 基因组序列变异信息汇交、管理与检索的资源库 (Resource for genome sequence variation information submission, management and retrieval)
- BioCode**: 整合开源生物信息软件工具的数据库 (Database integrating open-source bioinformatics software tools)
- Database Commons**: 生物数据库收集、分类、检索系统 (Biological database collection, classification, and retrieval system)



提纲

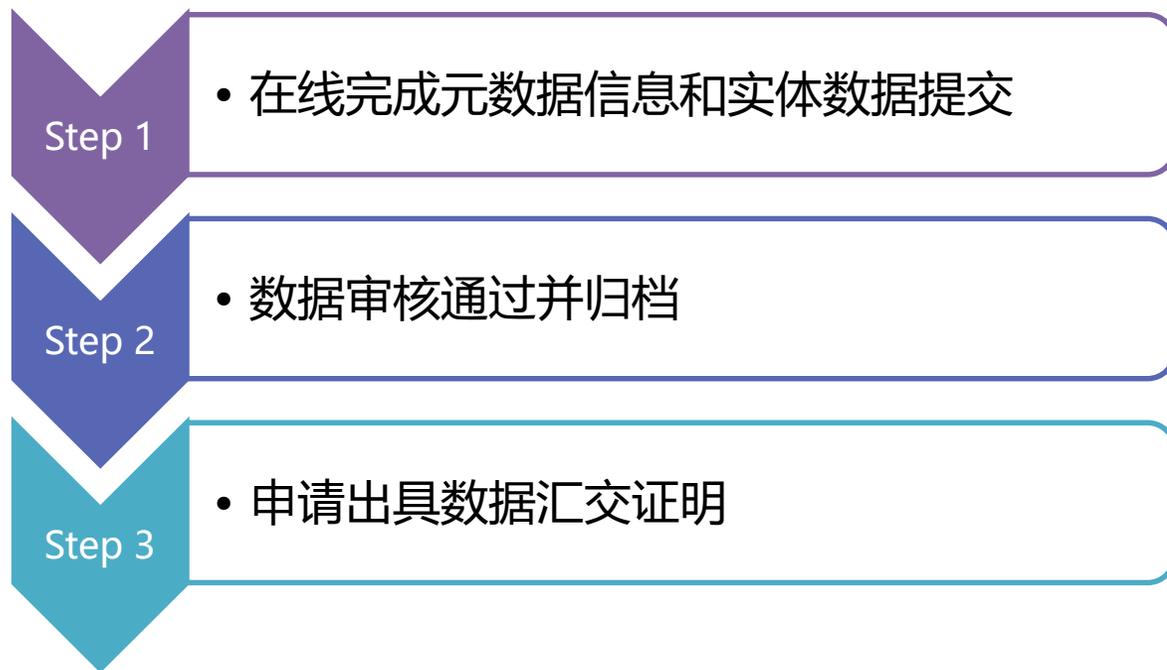
- 数据汇交背景
- 数据汇交计划制定
- 数据提交与审核
- **数据汇交证明出具**



出具数据汇交证明



• 中心数据汇交证明出具流程



反馈邮箱: data@big.ac.cn

科技计划项目科学数据汇交证明

汇交项目名称: 基于多组学谱特征的前列腺癌分子分型研究
项目编号: 2016YFC0902200
负责审核部门名称: 国家基因组科学数据中心
接收数据资源清单:

序号	归档编号	数据集名称	实体数据量		数据格式	共享方式
			大小	单位		
1	HRA000099	PRJCA001124	344.66	TB	Fastq and BAM	受控访问

根据《科技计划项目科学数据汇交工作方案(试行)》要求, 国家基因组科学数据中心已收到项目承担单位提交的以上数据, 审查合格。
本凭证一式三份。

负责审核部门负责人: _____
负责审核部门单位代章: _____
日期: 2020年9月4日



培训主要内容

- 项目科学数据汇交总体流程
- **生物项目数据库 (BioProject)**
- 生物样本数据库 (BioSample)
- 组学原始归档库 (GSA)
- 人类遗传组学原始归档库 (GSA-Human)
- 基因组变异数据库 (GVM)
- 基因组数据库 (GWH)
- BioCode数据库
- OMIX数据库





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



国家基因组科学数据中心

National Genomics Data Center

生物项目数据库 (BioProject)

国家生物信息中心

China National Center for Bioinformation

BioProject主网页和提交入口

The screenshot shows the BioProject website interface. At the top, there is a navigation bar with 'CNCB-NGDC' and links for 'Databases', 'Tools', 'Standards', 'Publications', and 'About'. Below this is the 'BioProject Biological Project Library' logo and a search bar with a dropdown menu set to 'BioProject' and a search button. A language selector shows '中文' and 'English'. A secondary navigation bar includes 'Home', 'Browse', 'Umbrella', 'Search', 'Statistics', 'Documentation', 'Login', and 'Register'. The main content area is divided into several sections:

- 欢迎来到生物项目数据库**: A welcome message in Chinese stating that the BioProject database is a resource for biological research project information, covering various types of projects like genomics, transcriptomics, and epigenomics. It includes a '新建BioProject' button.
- BioProject增长趋势**: A line graph showing the growth trend of projects from 2016-09 to 2021-05. The Y-axis represents the number of projects (0 to 4300), and the X-axis represents time. Two lines are shown: 'Total Projects' (blue line with diamond markers) and 'Released Projects' (black line with square markers). Both lines show a steady upward trend, with Total Projects reaching approximately 3400 and Released Projects reaching approximately 2500 by 2021-05.
- 帮助和支持**: A section providing contact information for help and support, including an email address (gsa@big.ac.cn) and a QQ group (548170081). It also includes a message of appreciation for user support and a commitment to system improvement.
- 最新发布项目**: A table listing the most recently published projects.

项目编号	描述信息
PRJCA004947 (2021-05-06)	Tension wood formation in <i>N. cadamba</i>
PRJCA002948 (2021-05-06)	Ancient Xinjiang mitogenomes reveal intense admixture with high geneti...
PRJCA004927 (2021-05-06)	Circadian rhythms of thyroid stimulating hormone in zebrafish
PRJCA004933 (2021-05-06)	16s Amplicon project
PRJCA004943 (2021-05-05)	DDRAD-seq of brown-marbled grouper (<i>Epinephelus fuscoguttatus</i>)
PRJCA004942 (2021-05-05)	microecology in lung injury

BioProject主网页:

<https://bigd.big.ac.cn/bioproject/>

BIG Sub递交入口:

<https://bigd.big.ac.cn/gsub>

问题反馈邮箱:

gsa@big.ac.cn



培训主要内容

- 项目科学数据汇交总体流程
- 生物项目数据库 (BioProject)
- **生物样本数据库 (BioSample)**
- 组学原始归档库 (GSA)
- 人类遗传组学原始归档库 (GSA-Human)
- 基因组变异数据库 (GVM)
- 基因组数据库 (GWH)
- BioCode数据库
- OMIX数据库





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



国家基因组科学数据中心

National Genomics Data Center

生物样本数据库 (BioSample)

国家生物信息中心

China National Center for Bioinformation

BioSample主网页和提交入口

CNCB-NGDC Databases Tools Standards Publications About

BioSample
Biological Sample Library

BioSample find a biosample accession 中文 English

e.g., SAMC013103; human

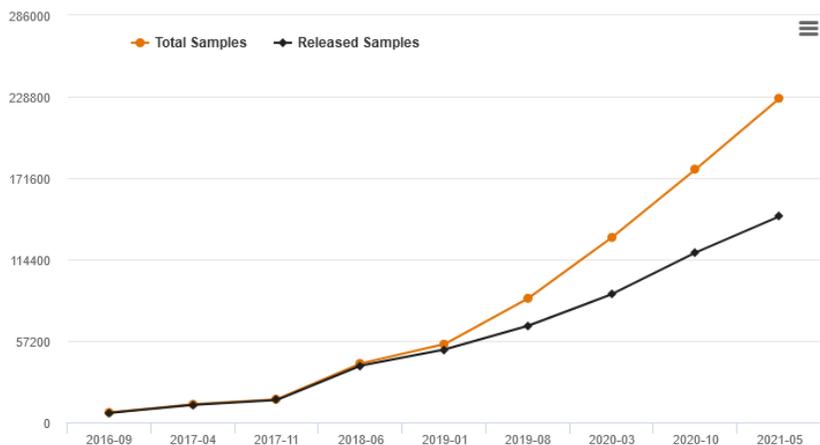
Home Browse Search Statistics Documentation Login Register

欢迎来到生物样本数据库

生物样本数据库 (BioSample) 是收集与共享生物样品信息的资源库, 提供生物样品的结构化描述信息递交和发布, 涵盖的样品类型包括人、动物、植物、微生物 (含环境微生物)、病毒等, 提供批量数据上传及离线数据递交服务。

[新建BioSample](#)

BioSample增长趋势



帮助和支持

如果您在数据上传过程中遇到问题, 或发现任何系统报错, 请随时联系我们。

Email: gsa@big.ac.cn

QQ群: 548170081

非常感谢您对BioSample数据库的支持, 欢迎随时提出宝贵意见或建议, 我们将根据您的需求, 不断完善和改进系统功能。

最新发布样本

样本编号	描述信息
SAMC353765 (2021-05-06)	Compression wood
SAMC353764 (2021-05-06)	Compression wood
SAMC353763 (2021-05-06)	Compression wood
SAMC353762 (2021-05-06)	Tension wood
SAMC353761 (2021-05-06)	Tension wood
SAMC353760 (2021-05-06)	Tension wood

BioSample主网页:

<https://bigd.big.ac.cn/biosample/>

BIG Sub递交入口:

<https://bigd.big.ac.cn/gsub>

问题反馈邮箱:

gsa@big.ac.cn

