

国家标准

《单细胞测序 单细胞转录组数据集》
(征求意见稿)

编制说明

《单细胞测序 单细胞转录组数据集》

标准起草组

2025 年 5 月

目录

| | |
|---|----|
| （一）工作简况 | 3 |
| （二）国家标准编制原则、主要内容（如技术指标、参数、公式、性能要求、试验方法、检验规则等）及其确定依据（包括试验、统计数据），修订国家标准时，还包括修订前后技术内容的对比 | 4 |
| （三）主要试验（或验证）的分析、综述报告，技术经济论证，预期的经济效益、社会效益 | 6 |
| （四）与国际、国外同类标准技术内容的对比情况，或者与测试的国外样品、样机的有关数据对比情况 | 10 |
| （五）以国际标准为基础的起草情况，以及是否合规引用或者采用国际国外标准，并说明未采用国际标准的原因 | 11 |
| （六）与有关的现行法律、行政法规及相关标准的关系 | 11 |
| （七）重大分歧意见的处理经过和依据 | 11 |
| （八）涉及专利的有关说明 | 11 |
| （九）实施国家标准的要求，以及组织措施、技术措施、过渡期和实施日期建议等措施建议 | 12 |
| （十）其他应予说明的事项 | 12 |

国家标准《单细胞测序 单细胞转录组数据集》编制说明

(一) 工作简况，包括任务来源、协作单位、起草过程、国家标准主要起草人及其所做的工作等

1、任务来源

本标准根据国标委公布的 2024 年第 一 批国家标准计划项目（国标委发【2024】16号），本项目计划编号为 20240062-T-469，名称为 《单细胞测序 单细胞转录组数据集》。

本标准由全国生化检测标准化技术委员会（SAC/TC 387）提出并归口。

本标准由 深圳华大生命科学研究院、杭州华大生命科学研究院、中国科学院北京基因组研究所（国家生物信息中心）、中山大学、武汉华大生命科学研究院、深圳华大基因科技有限公司、广州基迪奥生物科技有限公司、西北农林科技大学、中国科学院武汉植物园、深圳裕策生物科技有限公司、菁良科技(深圳)有限公司 联合起草。

2、目的和意义

规范单细胞测序场景中单细胞转录组数据及数据元的定义、数据格式要求和数据归档目录要求，以及单细胞转录组研究过程中相关数据元 and 值域的要求，解决组学数据中有关单细胞转录组数据信息的存储、管理、交换与共享过程中的主要问题。单细胞转录组技术已经取得了很大的进展，在国内外已经形成了一定的规模和影响力，有足够的经验和技術基础来制定标准。此外，标准化的单细胞转录组数据集也可以借鉴其他生物信息学数据集的标准化经验，使得标准的制定更加可行和有效。同时本标准主要目的为了单细胞测序行业在数据分析过程中规范化数据格式和归档的要求，为了提高数据的可比性和可重复性。制定标准可以规范数据的质量、格式和注释方式，使得不同实验室产生的数据集可以更好地比较和整合，从而提高数据的可比性和可重复性。促进单细胞转录组技术的发展和應用。标准化的数据集可以更好地支持单细胞转录组技术的应用和发展，为研究人员提供更好的数据资源和分析工具，促进单细胞转录组技术的应用和发展。提高数据的可信性和可靠性。制定标准可以规范数据的质量控制和数据处理流程，从而提高数据的可信性和可靠性，为研究人员提供更加准确和可靠的数据资源，为组学数据的应用提供保障，包括生物医学研究、生物工程和精准医疗等领域。

对于数据安全方面，为了合理制定单细胞转录组数据集的分级分类原则，将单细胞数据的内容、用途和来源进行分类，依照单细胞数据的内容敏感程度、价值、影响情况对不同数据进行敏感级别的划分。根据私有、受控、公开的数据的管理形式划分不同级别，

以确保数据价值可以得到合理的评估以及数据安全可以得到合理的保障。并按照数据集的保密性情况进行分级划分以确保数据集的安全保障。本标准与国内多家单细胞及数据库相关企事业单位联合制定，旨在协调统一数据格式标准，并且会持续联合上述及更多新加入涉及的机构单位共同协调和遵循统一的数据格式标准，为单细胞相关数据产业做出行业贡献和科学数据保障。

3、协作单位

计划下达后，由深圳华大生命科学研究院成立了标准编制工作组。

4、标准编制过程和主要工作过程

（1）2022 年 11 月至 2022 年 12 月，标准起草单位组织相关技术人员对《单细胞测序 单细胞转录组数据集》标准项目进行了预研，课题组成员广泛收集了国内外单细胞及转录组数据相关标准、文献，了解了国内外相关技术动态，并且明确了工作思路和进程安排。

（2）2022 年 12 月，标准起草单位组织相关技术人员对《单细胞测序 单细胞转录组数据集》标准项目进行了全国生化检测标准化技术委员会数据及数据库工作组 2022 年第二次工作组会议研制进展的研讨。

（3）2024 年 4 月，收到全国生化检测标准化技术委员会生检标【2024】5号文件《关于下达2024年第一批推荐性国家标准计划的通知》以及该标委会转发的国标委发【2024】16号《国家标准化管理委员会关于下达2024年第一批推荐性国家标准计划及相关标准外文版计划的通知》立项文件，计划编号：20240062-T-469。

（4）2024 年 4 月至 2024 年 11 月，进行《单细胞测序 单细胞转录组数据集》标准的起草研制工作。完成了《单细胞测序 单细胞转录组数据集》标准的草案，并对全国生化检测标准化技术委员会数据及数据库工作组汇报了《单细胞测序 单细胞转录组数据集》标准（草案）情况，与会专家就标准（草案）进行了讨论，提出了宝贵的意见和建议。标准起草小组根据专家意见进行了修改和完善。之后，向全国生化检测标准化技术委员会汇报了标准（草案）情况，委员会专家对标准（草案）提出了宝贵的意见和建议。

（5）2024 年 11 月至 2025 年 5 月，标准起草小组对标准草案进行了进一步修改和完善，并同时完成了编制说明。之后，向全国生化检测标准化技术委员会提交了《单细胞测序 单细胞转录组数据集》标准（征求意见稿）和编制说明。

5、国家标准主要起草人及其所做的工作

（二）国家标准编制原则、主要内容

1、标准编制原则

本文件按照GB/T 1.1—2020给出的规则起草。本文件编制遵循“科学性、实用性、统一性、规范性”的原则。按照全国生化检测标准化技术委员会（SAC/TC 387）相关章程中标准制修订工作程序的要求开展工作。

本标准结合我国单细胞测序行业、生命科学领域生物信息学应用方向数据使用场景，制定本标准时遵循以下原则：

根据市场应用和行业技术的实际情况出发，最大限度的促进我国生命科学领域、单细胞测序领域、生物信息领域的测序生产、质量控制、分析归档等环节的研发及应用等方面的提高与发展。本标准对生命科学领域过程中的数据产出、数据存储、数据访问、数据共享提供技术依据，并为行业未来的技术发展留有一定空间，使得本项目标准具有一定前瞻性和开拓性。

本标准与现行相关法律法规、标准等协调一致。

在确定本标准的数据兼容性、可扩展性、一致性和兼容性、格式验证、数据类型要求、生物数据存储库的要求等内容时，综合考虑行业的需求、以科学研究需要、便利性等方面寻求最大促进行业发展和社会效益角度出发，充分体现了标准在技术上的先进性、科学上的合理性，使标准内容更加全面、完善和易于实施及应用。

根据国情，标准制定坚持面向行业、面向市场的原则。结合我国生命科学领域的实际现状，并以引领生命科学研究过程中的数据格式和描述要求的水平提升为目标而制定，提高我国在这一领域标准的综合水平，使标准适应市场需求，满足领域发展，为科学研究、科研生产领域提供标准指导，引导生命科学相关行业领域采用本标准进行规范化生产、交流，具有一定程度的指导性。

对标准的结构编排、编写格式和内容表达方法等按 GB/T 1.1-2020 等系列标准的规定进行编写，使标准规范化。

2、确定国家标准主要内容

本标准全文分为8章节和2个附录。标准的主要内容如表1所示：

表1《单细胞测序 单细胞转录组数据集》主要内容

| 章节 | 名称 | 内容简要 |
|----|----|--|
| 1 | 范围 | 明确规定了单细胞测序领域在单细胞转录组数据和相应元数据格式和归档目录的要求。 |

| | | |
|----|-----------|-------------------------------------|
| 2 | 规范性引用文件 | 本文件没有规范性引用文件。 |
| 3 | 术语和定义 | 对本标准过程中的术语及定义的描述 |
| 4 | 缩略语 | 适用于本文文件的缩略语。 |
| 5 | 数据及数据文件要求 | 对于单细胞转录组数据集的数据及其文件在格式、文件类型等方面的具体要求。 |
| 6 | 数据元目录 | 对于单细胞转录组数据集数据元的目录进行规范化要求。 |
| 7 | 数据归档目录 | 对于单细胞数据归档目录结构及要求的描述。 |
| 8 | 数据安全的管理 | 对于单细胞转录组数据的分级分类原则和分级分类方式的要求。 |
| 9 | 附录 A | 本附录对于数据集的数据元目录、实验测序信息及生物信息分析部分进行描述。 |
| 10 | 附录 B | 本附录对数据元值域代码进行描述。 |

（三）主要试验（或验证）的分析、综述报告，技术经济论证，预期的经济效益、社会效益和生态效益

1、验证分析

本标准在编制过程中，牵头单位同时开展了标准验证工作。验证工作按照本文件规定的文件、生物数据归档存储库的要求等方面对业务流程中所涉及的相关内容验证。结果显示本文件中对数据格式等方面内容符合国内及行业生命科学、单细胞转录组学产业发展情况。

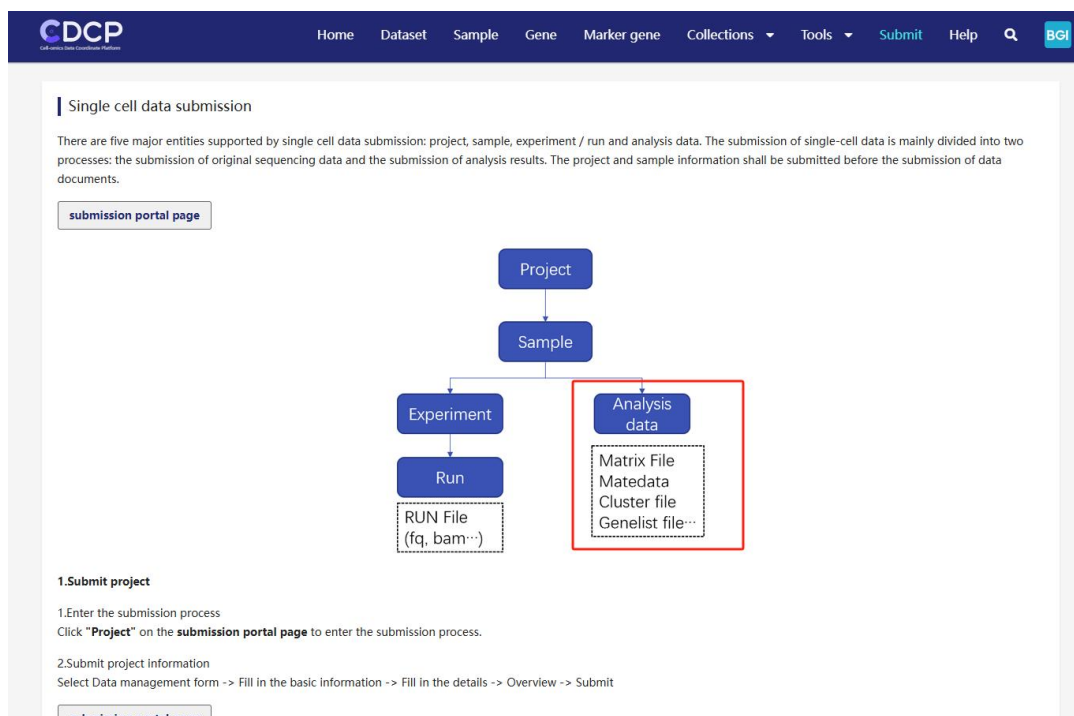
牵头单位对行业内常见的数据文件类型，包括表达矩阵文件、元数据文件、聚类文件、基因列表文件等数据文件类型进行了各参编单位使用的生命科学数据库包括 Single Cell Portal、Human Cell Atlas（HCA）、Cell-omics Data Coordinate Platform（CDCP）、GEO 的验证，得出了如下表的验证情况。

| 编号 | 文件类型 | 验证情况 | 单细胞科学数据库 |
|----|--------|------|---------------------------------|
| 1 | 表达矩阵文件 | 符合 | Single Cell Portal/HCA/CDCP/GEO |
| 2 | 元数据文件 | 符合 | Single Cell Portal/HCA/CDCP/GEO |

| | | | |
|---|--------|----|------------------------------------|
| 3 | 聚类文件 | 符合 | Single Cell Portal/HCA/CDCP/GEO |
| 4 | 基因列表文件 | 符合 | CDCP |

部分验证过程的符合证明，如下列各图所示：

- (1) 表达矩阵文件、元数据文件、聚类文件、基因列表文件在 CDCP 数据库中的要求：



- (2) 表达矩阵文件、元数据文件、聚类文件文件在 HCA 数据库中的要求：

HUMAN CELL ATLAS DATA PORTAL

Exploring Projects
Requesting Access to Controlled Access Data
Exploring Biological Network and Atlas Data
[Accessing HCA Data and Metadata](#)
Exporting HCA Data to Terra
Exploring Project Matrices
Accessing Metadata via TDR

Downloading Metadata in a Data Manifest

Once you have downloaded the selected data files, you can download all the metadata associated with the cross-project data files.

This metadata, also called a "Data Manifest" is in TSV file format and lists all the details about your selected data such as donor information and disease-state; however, the manifest is not the actual data file itself.

To download the metadata from the Export Selected Data page:

1. Under Download a File Manifest with Metadata for the Selected Data, select Request File Manifest

DATA EXPLORER

Choose Export Method

Download Study Data and Metadata (Curl Command)
Obtain a curl command for downloading the selected data.

[Request curl Command](#)

Current Query
All Data
Selected Data Summary

CONTENTS

- Finding Data
- Preparing Data For Export
- Downloading Data With A Curl Command
- Downloading Metadata In A Data Manifest

HCA Data Portal Data Matrix Overview

Cell-by-gene matrices (commonly referred to as "count matrices" or "expression matrices") are files that contain a measure of gene expression for every gene in every cell in your single-cell sample(s). These matrices can be used for downstream analyses like filtering, clustering, differential expression testing, and annotating cell types.

This overview describes the HCA Data Portal matrix types, how to download them, and how to link them back to the HCA metadata.

Overall, three types of matrices are currently available for HCA Data Portal data:

- HCA Data Portal-generated matrices (Loom file format) for projects
- HCA Data Portal-generated matrices (Loom file format) for individual library preparations within a project
- Contributor-generated matrices (variable file format) provided by the project-contributor

HCA Data Portal-Generated Matrices

Each HCA Data Portal project that is processed with uniform pipelines has two types of HCA Data Portal-generated matrices available for download:

(3) 表达矩阵文件、元数据文件、聚类文件文件在 Single Cell Portal 数据库中的要求:

Single Cell Portal

Single Cell Portal > Setting up an SCP study > Adding study files to enable data visualization

Articles in this section

Upload data in AnnData mode

Key study files depend on unique NAMES to enable visualizations

Expression matrix files

Cluster files

Expression matrix files

9 months ago • Updated

Follow

Matrix files contain the expression measurements of a study; in RNA-Seq, this would be the expression of your cells. The values are used throughout the study in many visualizations. Although the units of the expression data in a processed matrix file are up to the author of the study, we recommend some variant of $\log_2(\text{TPM} + 1)$. You can specify your units in the "Expression axis label" field so that, when viewing expression, the axes are correctly labeled. We also require upload of a raw matrix of measurements (like a count matrix) to encourage data sharing and enable data reuse. **Please note, cell names in the matrix file must match cell names in other study files for metadata and cluster files to support data visualization.**

Single Cell Portal

Single Cell Portal > Setting up an SCP study > Adding study files to enable data visualization

Articles in this section

Upload data in AnnData mode

Key study files depend on unique NAMES to enable visualizations

Expression matrix files

Cluster files

Cluster files

9 months ago • Updated

Follow

Cluster files create the different ordinations plots of cells in the study. You are welcome to create as many of these plots as you would like. Each plot is created by loading a different file. These plots can be 2D or 3D and can contain metadata specific to just the plot (and not shared with other plots). These files/plots do not need to contain all your cells, giving you the freedom to plot subsets of your study for targeted visualization.

Format: This is a tab-delimited file with 3 required columns (with the option of more) and 2 required header rows.

Single Cell Portal > Setting up an SCP study > Adding study files to enable data visualization > Metadata file

Search

Articles in this section

Metadata file overview

9 months ago · Updated

Follow

A metadata file provides metadata describing cells in the study. The metadata provided in this file will be interpreted as either "group" (categorical/factor) or "numeric" (continuous) data. These metadata are available throughout the visualization portal.

Metadata will be available to paint cell plots. Categorical metadata will paint the cells with discrete color panels (as discrete groups, each with their own color). Continuous metadata will paint cells as a gradient of color. These metadata not only determine color on plots of cells but also are used when viewing genes across cells.

(4) 表达矩阵文件、元数据文件、聚类文件文件在 GEO 数据库中的要求:

NCBI GEO Gene Expression Omnibus

NCBI » GEO » Info » Submitting high-throughput sequence data to GEO

Submitting high-throughput sequence data to GEO

- Submission instructions [YouTube](#)
 - Metadata spreadsheet **REQUIRED**
 - Processed data files **REQUIRED**
 - Raw data files **REQUIRED**
- Tutorial video
- Data file compression
- Single-cell studies
- NanoString GeoMx Digital Spatial Profiling (DSP)
- Organizing your submission
- Uploading your submission
- General information
 - Data provisions, standards and administration
 - Categories of sequence submissions accepted by GEO

SKA. You can get that information for your SUB ID on the [Submission Portal](#) page.

- Processed data files**

GEO requires that submitters deposit the processed data that support the findings of their study. The processed data should have a quantitative component, such as gene abundances or other count data. Please do not submit alignment files (e.g., BAM, SAM, BED) as processed data, as these are considered intermediary files and do not include a quantitative component. When standard alignments are the only processed data available, please [write to us](#) to inquire about whether your data are suitable for submission to GEO.

Processed data format and content will depend on the data type: RNA-seq processed data can include raw and/or normalized counts (FPKM, TPM, etc) of sequencing reads for the features of interest (protein-coding genes, lncRNA, miRNA, circRNA, etc).

ChIP-Seq and ATAC-seq processed data can include peak files with quantitative data, tag density files, etc. Common formats include WIG, bigWig, bedGraph. Please leave files in native format and do not paste peak data into Excel.

Methylation data are often provided as average beta values.

Processed data guidelines:

 - Processed data may be formatted either as a **matrix table** or individual files for each sample.
 - If processed data for all samples is submitted in a matrix table, column headers should match the library name for each sample listed in the SAMPLES section of the metadata spreadsheet.
 - Provide **complete** data with values for all features (e.g., genes) and all samples. Do not submit lists of genes identified with differential expression.
 - Features (e.g., genes, transcripts) in processed data files should be traceable using public accession numbers or chromosome coordinates. The reference assembly used (e.g., hg19, mm9, GCF_000001405.13) should be provided in the metadata spreadsheet.
 - If you provide WIG, bedGraph, GFF, or GTF files, please refer to the UCSC file format FAQ for format requirements.
- Raw data files**

Raw data are a required part of GEO submissions. The raw data files should be the original files containing reads and quality scores, as generated by the sequencing instrument. Edited files may not be processed correctly by SRA.

经验证，相关数据库所要求的文件类型，均符合本标准规定的文件类型及描述要求。

2、经济与社会效益

在单细胞测序领域中，建立单细胞转录组数据集的国家标准对社会效益和产业应用发展具有深远的影响。首先，本次国家标准的实施有利于促进单细胞测序技术的发展，推动其在肿瘤、生殖、免疫等生命科学和医学领域的应用。这不仅能够推动我国基础科学研究的进步，还能提升我国在生命科学研究的整体创新能力和国际竞争力。其次，单细胞转录组数据集的国家标准对于提高数据的可比性和可重复性至关重要。制定标准可以规范数据的质量、格式和归档方式，使得不同实验室产生的数据集可以更好地比较和整合，从而提高数据的可比性和可重复性。这对于科研人员来说是一个巨大的进步，因为它意味着他们可以在一个统一的数据标准下进行研究，从而更容易地验证和比较不同研究的结果。此外，国家标准的制定和实施促进了单细胞转录组技术的发展和應用。标准化的数据集可以更好地支持单细胞转录组技术的应用和发展，为研究人员提供更好的数据资源和分析工具。这有助于推动单细胞测序技术在更广泛的领域中的应用，包括但不限于早期胚胎发育、干细胞、癌症、免疫等研究领域。国家标准还提高了数据的可信性和可靠性。通过规范数据的质量控制和数据处理流程，可以为研究人员提供更加准确和可靠的数据资源，为组学数据的应用提供保障，包括生物医学研究、生物工程和精准医疗等领域。这对于推动个性化医疗的发展尤为重要，因为精准医疗依赖于高质量、可靠的数据来设计个性化治疗方案。标准化和规范化可以提高数据的质量和可靠性，促进数据共享和交流，提高数据的可比性，同时也遵守国际标准和规范，为数据共享提供了保障。综上，建立单细胞转录组数据集的国家标准对于提升我国在生命科学领域的研究水平、推动相关技术的应用发展、提高数据的质量和可靠性、促进国际合作和数据共享等方面都具有重要的作用。这些好处不仅体现在科研领域，也对公众健康和医疗产业的发展产生了积极的影响。

（四）与国际、国外同类标准技术内容的对比情况，或与测试的国外样品、样机的有关数据对比情况

1)、国外标准情况

在国外Human Cell Atlas, HCA组建了人类细胞图谱计划，系统的描绘了人体细胞图谱，对人体中的近37万亿个细胞进行分类和测序，并且通过这些，加深对疾病诊断、监测、治疗的理解。HCA以“项目（Projects），样本（Biomaterials），实验（Protocols、Processes），数据文件（Files）”的数据标准，作为HCA计划数据归档分析的基础。但，此项工作尚未形成ISO等国际标准。

同时在国外,欧洲生物信息中心EBI,构建了EBI Single Cell Expression Atlas 数据标准,兼容HCA计划的标准,也整合国际其他同类数据库或其他单细胞科研项目 Fly Cell Atlas。EBI作为欧洲生命科学基础设施ELIXIR的核心节点,在欧洲推广和应用这项数据标准,但也没有形成ISO等国际标准。

2)、国内标准情况

在国内,由于起步较晚,对于单细胞转录组数据集标准的研究和推广还相对较少。此前已经形成诸如GB/T 31074-2014 科技平台数据元设计与管理、GB/T 34798-2017核酸数据库序列格式规范、GB/T 35890-2018 高通量测序数据序列格式规范等的核算序列规范标准,但未有直接的单细胞转录组数据集的标准。高通量测序数据序列格式规范中规定了测序数据序列的格式及其相关标准规范,而单细胞转录组数据集的标准则是在此基础上约定了单细胞转录组的分析结果生成的数据集相关的标准规范。

同时在行业标准中,先后形成了诸如WS/T 305-2009 卫生信息数据集元数据规范、WS/T 306-2009 卫生信息数据集分类与编码规则、WS 363-2011 卫生信息数据元目录、WS 364-2011 卫生信息数据元值域代码、WS 370-2012 卫生信息基本数据集编制规范、WS 371-2012 基本信息基本数据集个人信息、WS 372-2012 疾病管理基本数据集、WS 375-2012 疾病控制基本数据集等的数据集标准,但也未有直接的单细胞转录组数据集标准。

在2021年,由深圳国家基因库和深圳华大生命科学研究院牵头,建立了T/SZAS 39-2021单细胞转录组学数据集 (深圳市标准化协会)团体标准。但并没有直接的国家及行业标准。

(五) 以国际标准为基础的起草情况,以及是否合规引用或者采用国际国外标准,并说明未采用国际标准的原因

国际上无同类标准。

(六) 与有关的现行法律、行政法规及相关标准的关系

本标准与现行相关法律法规、标准等协调一致。

(七) 重大分歧意见的处理经过和依据

无重大分歧意见。

(八) 涉及专利的有关说明

本标准不涉及专利情况。

（九）实施国家标准的要求，以及组织措施、技术措施、过渡期和实施日期建议等措施建议
无。

（十）其他应予说明的事项

无。

标准起草组

2025 年 5 月