单细胞测序 原始数据与处理后数据的准入标准

*版本控制：Version 0.0.1*

*日期：2024/08/28 (YYYY/MM/DD)*

目录

[前言 3](#_Toc175842912)

[1. ID System 6](#_Toc175842913)

[2．元数据、术语与定义 7](#_Toc175842914)

[2.1 元数据 7](#_Toc175842915)

[2.2 术语与定义 7](#_Toc175842916)

[2.2.1 L1层 以FASTQ文件为代表的部分 7](#_Toc175842917)

[2.2.2 L2层 SRA与BAM文件 7](#_Toc175842918)

[2.2.3 L3层 以H5AD为代表的部分 9](#_Toc175842919)

[2.2.4 L4层 经过整合后的数据 10](#_Toc175842920)

[2.2.5 L5层 知识生成与研究成果（形式上的事理图谱） 10](#_Toc175842921)

[3. 控制性质量标准 12](#_Toc175842922)

[3.1 L1层，以FASTQ文件为代表的部分 12](#_Toc175842923)

[3.1.1 测序质量评分标准： 12](#_Toc175842924)

[3.1.2 读取长度与完整性： 12](#_Toc175842925)

[3.1.3. UMI和Barcode的完整性与准确性： 12](#_Toc175842926)

[3.1.4. FASTQ文件格式和完整性： 12](#_Toc175842927)

[3.1.5. 重复序列与污染检查： 13](#_Toc175842928)

[3.2 L2层，以SRA/BAM文件为代表的部分 13](#_Toc175842929)

[3.2.1 对于BAM文件： 13](#_Toc175842930)

[3.2.2 对于SRA文件： 14](#_Toc175842931)

[3.3 L3层 以H5AD,为代表的部分 14](#_Toc175842932)

[3.3.1 基因名的命名和校验： 14](#_Toc175842933)

[3.3.2 表达矩阵数据格式和存储： 15](#_Toc175842934)

[3.3.3和其他注释属性： 15](#_Toc175842935)

[3.3.4 文件完整性和兼容性： 15](#_Toc175842936)

[3.4 L4层 ATLAS级别的数据 15](#_Toc175842937)

[3.4.1数据一致性 15](#_Toc175842938)

[3.5. L5层 知识生成与研究成果 16](#_Toc175842939)

[3.5.1. 唯一性和标识 16](#_Toc175842940)

[3.5.2 数据实体的清晰性和准确性 16](#_Toc175842941)

[3.5.3 逻辑关系和因果推理 16](#_Toc175842942)

[3.5.4 文档化和元数据 17](#_Toc175842943)

[3.5.5 审查和持续改进 17](#_Toc175842944)

# 前言

作为数据标准的标准组件的数据准入的部分的标准，将按照数据流模型的顺序进行内容的组织。在当前的数据主题：单细胞测序的场景下，数据流模型的数据分级如下所示：

**L1 - 原始数据层**

* **原始测序数据**（Fastq文件）
* **空间转录组的切片图像**（如tissue\_hires\_image.png, tissue\_lowres\_image.png）

**L2 - 初级处理数据层**

* **BAM文件**（possorted\_genome\_bam.bam 和 possorted\_genome\_bam.bam.bai）
* **SRA文件**（Sequence Read Archive）
* **对齐的基准点图像**（aligned\_fiducials.jpg）
* **检测到的组织图像**（detected\_tissue\_image.jpg）
* **条形码空间位置信息**（tissue\_positions.csv）

**L3 - 注释和高级分析数据层**

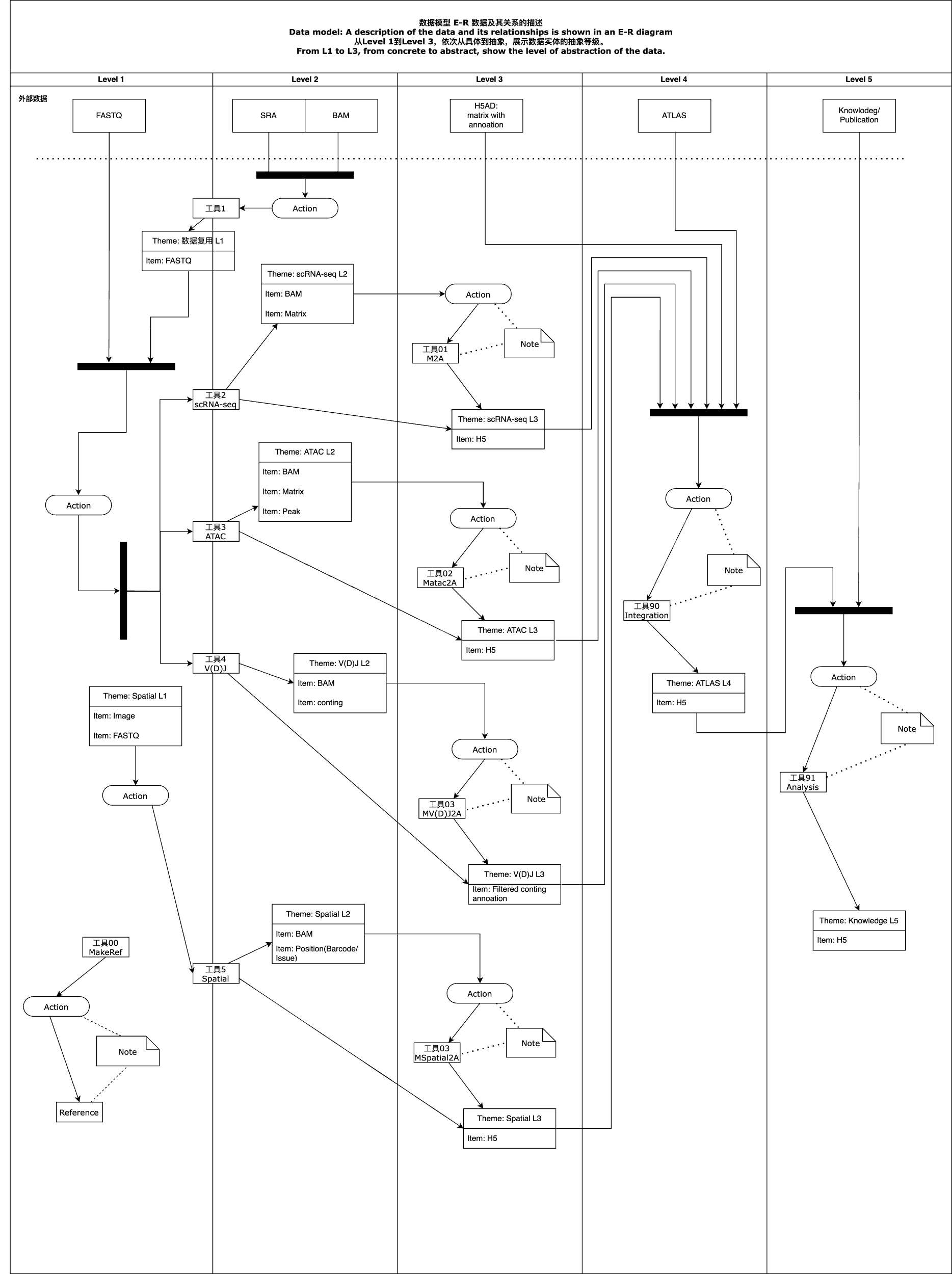
* **带有注释的计数矩阵文件**（如filtered\_feature\_bc\_matrix.h5, raw\_feature\_bc\_matrix.h5）
* **次级分析结果**（analysis 目录下的文件）
* **空间富集分析**（spatial\_enrichment.csv）
* **Loupe Browser 文件**（cloupe.cloupe）
* **注释的峰值与基因关联**（peak\_annotation.tsv）
* **峰值与motif关联**（peak\_motif\_mapping.bed）

**L4 - 综合分析和整合层**

* **多种单细胞数据的整合结果**：例如对不同实验或样本的单细胞数据进行整合分析后产生的综合数据。
* **跨实验分析结果**：例如通过整合多个实验的数据以探索更大范围的生物学问题或趋势。

**L5 - 知识总结和应用层**

* **基于分析结果的论文**：由数据分析结果撰写的科学论文或报告。
* **新知识的总结**：通过分析得出的新的生物学知识、发现和假设。
* **实际应用**：例如基于研究结果提出的治疗策略、药物靶点或其他应用。



适用于两类情况：

情况一，用于规范从公共数据库获取的数据实体；用于规范由操作实体提供的数据实体。

情况二，用于规范数据处理过程中产生的数据实体。

实施和监测：

* 所有数据在进入数据存储库之前，必须通过自动化的质量控制流程进行检查和验证。
* 定期对数据质量控制流程进行审查和更新，以适应新的科技进展和研究需求。
* 实施定期的培训和评估，确保相关人员对控制性质量标准的理解和执行到位。

声明：本标准规范了数据准入阶段的最低标准，如注册元数据、新增字段等需求，按元数据：注册-审核-发布的标准进行规范。

**注意：本标准未将“人遗”、“伦理”纳入，默认所有引入的数据实体均无此风险。**

**注意：仍需单独的文件对数据实体的归属、以及数据实体在数据处理过程中产生的次生数据实体的归属作出界定，存在法律风险。**

# 1. ID System

用于唯一标识数据实体

待定，需要进一步设计。

# 2．元数据、术语与定义

## 2.1 元数据

参照《单细胞组学 单细胞测序科研数据汇交的元数据》执行。

## 2.2 术语与定义

### 2.2.1 L1层 以FASTQ文件为代表的部分

**FASTQ文件**：

存储生物测序数据的文本格式，每个记录包含一个序列及其相应的质量评分。

**质量评分（Quality Score）**：

使用ASCII码表示的序列每个碱基的错误概率，常见标准为Phred质量评分。

**测序读取（Read）**：

从DNA或RNA样本通过测序得到的单个碱基链。

**UMI（Unique Molecular Identifier）**：

独特的分子标识符，用于标记单个分子，有助于消除PCR和测序过程中的重复。

**Barcode序列**：

用于在单细胞测序中识别不同细胞的短序列标签。

**切片图像（Tissue Section Image）**：

空间转录组分析中使用的组织切片的高分辨率图像，显示了样本的细胞结构和组织形态。

### 2.2.2 L2层 SRA与BAM文件

* **BAM文件：**

二进制版本的序列对齐/映射格式（Binary Alignment/Map），用于存储大规模的测序数据。BAM文件是SAM（序列对齐/映射格式）文件的压缩形式，通常用于存储处理过的测序数据。

* **SRA文件：**

序列读取档案（Sequence Read Archive）格式，是一种公共数据库接受的格式，用于存档原始测序数据。SRA可以包含来自多种测序平台的数据，如Illumina、Ion Torrent、PacBio等。

* **HEAD部分：**

BAM或SAM文件中的头部部分，包含了文件的元数据，如版本信息、参考序列和对齐程序使用的参数等。

* **CB标签（Cell Barcode）：**

用于单细胞测序中标记来自同一细胞的读取（reads）的标签。

* **UMI标签（Unique Molecular Identifier）：**

用于标记单个分子的唯一序列，有助于区分测序过程中产生的PCR重复。

* **RG标签（Read Group）：**

用于区分同一样品的不同实验条件或运行的读取，包含了关于样品库、测序平台和日期等信息。

* **Spot ID：**

在SRA格式中，Spot ID代表一个测序反应产生的单个数据点或事件的唯一标识符。

* **SAM文件：**

序列对齐/映射格式（Sequence Alignment/Map），是一种文本格式，用于存储生物测序数据的对齐信息。

### 2.2.3 L3层 以H5AD数据格式的文件为代表的部分

* **H5AD文件：**

一种基于HDF5的文件格式，专门用于存储单细胞基因表达数据。H5AD是anndata库使用的标准格式，广泛用于单细胞分析软件如scanpy。

* **基因名的命名空间：**

指用于识别基因的一组标准化名称。常见的命名空间包括Ensembl Gene ID和NCBI Gene ID，这些名称有助于确保基因的唯一和准确识别。这里指存储在H5AD的var部分的基因集注释的信息。

* **（基因）表达（计数）矩阵：**

存储在H5AD文件中的核心数据结构，记录了单细胞测序实验中每个细胞对应每个基因的表达水平。

* **标准化前的数据：**

指未经过任何数学转换处理的原始计数数据，通常是整数类型（INT）。

* **标准化数据：**

对原始计数数据进行某种形式的数学处理（例如TPM、CPM、FPKM等）后的数据，通常为浮点数（Float32）类型。

* **obs属性：**

在H5AD文件中，obs是存储有关观测的元数据的表格，例如细胞的注释信息。

* **Cell Barcode：**

用于标识单细胞测序中单个细胞的短序列。

* **参考基因组版本：**

基因表达分析中用作比对的基因组版本，如GRCh38或mm10，这里必须详细到具体的修订版本的版本号，如NCBI的build56、Ensemble的release 110。

### 2.2.4 L4层 经过整合后的数据

* **控制词汇：**

为了保证数据格式和术语的一致性而设立的一组标准化词汇。

应用规则：所有注释信息的字段必须与控制词汇保持一致。如出现非控制词汇字段，需通过注册-审核-发布流程进行标准化。

* **数据格式一致性：**

确保不同数据层级之间在数据类型和结构上的统一性。

要求：L4层的数据格式必须与L3层保持一致，包括数据类型（如INT或Float32）和数据结构，以支持自动化处理和数据比较。

* **元数据：**

定义：描述数据集的详细信息，包括数据来源、处理步骤、整合方法及最小数据元等。

详细要求：每个数据集都必须有详细的元数据记录，明确数据的许可信息和使用限制。

* **最小数据元：**

定义：描述数据集中最小的、不可分割的数据单位。

应用：元数据中必须明确最小数据元，以便于数据的精确理解和处理。

### 2.2.5 L5层 知识生成与研究成果（形式上的事理图谱）

* **数据实体（Data Entity）**

数据实体是直接从实验、观测或其他数据收集过程中获得的原始数据或处理后的数据。这些实体可以是数字化的测序数据、实验测量结果、或通过某种算法转换后的数据集。

数据实体具有可度量、可重复和可验证的特性，通常存储在数据库或数据集中，形式可以是表格、数值矩阵或其他结构化数据格式。

* **知识实体（Knowledge Entity）**

知识实体是对数据实体进行分析和解释后得到的信息，它反映了对原始数据的理解和洞见。知识实体可以是科学假设、模型预测、统计推断或从数据中提炼出的规律和模式。

知识实体通常基于数据实体，通过应用科学方法和逻辑推理形成，具有抽象和解释性质。知识实体的有效性依赖于其生成的数据和分析过程的质量。

* **数据实体独立性**

数据实体应被视为相互独立的单元，每个实体封装了关于特定现象或对象的特定信息集。在数据库或数据管理系统中，应通过独立的记录和唯一标识符来维护数据实体的独立性，即使这些数据实体来源于同一个现实世界对象。

* **数据实体的多重描述**

数据实体可能以多种方式描述同一真实世界对象，反映从不同角度或使用不同方法获得的信息。确保数据管理系统能够接受并存储同一对象的多重数据实体，同时通过链接或关联机制说明它们之间的联系。

* **知识实体间的逻辑关系**

知识实体之间的联系基于逻辑关系，这些关系可以构建成模型，反映实体之间的因果、关联或演化关系。使用形式化的逻辑表达和模型建构工具（如使用图数据库、逻辑编程或因果推理模型），明确并记录知识实体之间的逻辑联系。

* **专家经验**

专家经验是指在特定领域通过长期研究和实践积累的深入理解和技能。这种经验包括对特定现象、数据或问题的直觉、洞察力和预测能力。

* **数据实体与知识实体之间的联系**

数据实体与知识实体之间的联系是通过数据分析、模型建立和逻辑推理建立的。这种联系表明了如何从原始或加工的数据实体中提取或推导出知识实体。这种联系可以通过数据流图、算法描述或编码实践（如编写的分析脚本和应用的统计方法）具体化和数字化。

* **数据一致性**

验证所有数据实体在录入系统前的完整性和一致性，确保它们正确反映了原始信息。

* **知识验证**

对于每个知识实体，应有明确的来源、依据和推导过程。重要的知识实体应通过实验、额外的数据分析或专家评审进行验证。

# 3. 控制性质量标准

## 3.1 L1层，以FASTQ文件为代表的部分

### 3.1.1 测序质量评分标准：

* **Phred版本**：指定使用Phred-33或Phred-64质量评分（根据测序平台），并确保所有工具和流程与此标准兼容。
* **质量评分要求**：
  + 平均Phred质量评分不低于30，保证低于1/1000的碱基出错概率。
  + 至少80%的读取必须具有Q30以上的质量评分，以确保数据的可靠性。
  + 使用ASCII码表示质量，确保Phred+33或Phred+64质量评分的正确解读。

### 3.1.2 读取长度与完整性：

* **读取长度**：所有读取长度必须符合预设的范围，例如150bp，不得有意外短读或不完整的序列。
* **读取完整性**：每个序列的读取应完整，无任何技术性截断或损坏。

### 3.1.3. UMI和Barcode的完整性与准确性：

* **UMI和Barcode一致性**：每个读取必须包含与测序技术规定一致的UMI和Barcode序列。例如，使用10x Genomics平台时，Barcode和UMI的长度和位置应符合该平台的规范。
* **序列准确性**：UMI和Barcode的错误率应通过校对软件进行验证，确保其准确性低于1%。

### 3.1.4. FASTQ文件格式和完整性：

* **文件格式**：
  + 确保每个FASTQ记录遵循标准格式：'@'后跟序列标识符和描述，然后是原始序列，'+'可选的描述，最后是序列的质量评分。
  + 文件应采用压缩格式（如gzip）存储，以减少存储空间并提高传输效率。
* **数据完整性**：
  + 执行文件完整性校验，如MD5哈希检查，确保数据在传输或存储过程中未被破坏。
  + GC含量统计应与预期的参考基因组或文库制备过程中的预期相符，波动范围控制在±10%内。

### 3.1.5 重复序列与污染检查：

* **重复序列分析**：使用专门的软件（如Picard Tools）分析数据中的重复序列，不得超过设定的20%比例。
* **交叉污染检测**：
  + 评估样品是否有外源性DNA污染，污染率不得超过5%。
  + 使用专门的工具（如FastQC的Overrepresented sequences模块）进行污染分析。

### 3.1.6 空间转录组切片图像完整性

* **图像元数据完整性：**
  + 确保图像数据对应的元数据的完整性，因测序平台有所区别。具体指标如图像唯一的slide编号，图像位置标识信息以及对应的样本类型信息，以能够实现图像的复用为标准。
* **图像清晰度和分辨率**：
  + 确保每个组织切片图像的清晰度和分辨率符合空间转录组分析的需求，具体参数应根据使用的分析平台和研究目标明确规定。

## 3.2 L2层，以SRA/BAM文件为代表的部分

### 3.2.1 对于BAM文件：

3.2.1.1HEAD部分的完整性：

* 所有BAM文件必须包含完整的HEAD部分，其中详细记录了参考基因组信息、对齐软件版本及其参数设置。
* 必须清晰记录序列化过程中使用的所有重要参数，如使用的索引文件版本和对齐选项。
  + - 1. CB和UMI标签的应用：
* 每个读取（read）必须含有CB（细胞条形码）和UMI（独特分子标识符）标签，以确保单细胞数据可以准确追踪到特定细胞和分子。
* 标签的格式和内容应符合预定义的规范，以保证数据处理的一致性和准确性。
  + - 1. RG标签的使用：
* 每个BAM文件必须包含RG（读取组）标签，以标识不同的样品处理批次或实验条件。
* RG标签应包含详尽的样本信息，如样本ID、测序平台和日期，以便于数据的追踪和重用。
  + - 1. 文件格式和完整性验证：
* 所有BAM文件必须通过格式验证工具（如Samtools或Picard）检查，确保无格式错误或损坏。
* 必须实施文件的MD5哈希校验或采取其它校验措施，以验证文件在传输或存储过程中未被篡改。

### 3.2.2 对于SRA文件：

3.2.2.1 SRA格式规范：

* SRA文件必须清晰标识其包含的数据类型，无论是FASTQ还是BAM，以便于正确处理和使用。
* 对于FASTQ数据，必须包含CB和UMI标签信息；对于BAM数据，必须保留完整的标签信息，如CB、UMI和RG。

3.2.2.2 Spot ID的记录：

* 所有SRA文件条目必须包含Spot ID，确保每个数据点的唯一性和可追溯性。
* Spot ID应与原始数据生成时的记录保持一致，以支持数据的质量追踪和问题调查。

3.2.2.3数据完整性和验证：

* 通过SRA Toolkit或其他官方工具验证SRA文件的完整性和格式准确性。
* 确保SRA文件在上传到公共数据库之前符合所有相关的数据提交标准和质量要求。

## 3.3 L3层 以H5AD,为代表的部分

### 3.3.1 基因名的命名和校验：

* 所有基因名称必须对应至最少一个公认的命名空间（如Ensembl Gene ID或NCBI Gene ID），以确保一致性和准确性。
* 在数据整合或上传前，必须执行基因名称与对应参考基因组版本的比对校验，以消除因命名错误带来的数据偏差。

### 3.3.2 表达矩阵数据格式和存储：

* 原始的计数数据必须存储为整数型（INT）数据，以反映未处理的原始读数。
* 标准化后的表达矩阵必须存储为浮点数（Float32）类型，确保数据处理后的精度和一致性。
* 所有表达数据必须在H5AD文件内明确标记为“原始数据”或“标准化数据”，并在文件的元数据中记录所使用的标准化方法和参数。

### 3.3.3和其他注释属性：

* 所有数值型注释，如计数、百分比或其他量化数据，必须使用浮点数（Float32）格式存储，以保持数据精度的一致性。
* 非数值型注释，如细胞类型、实验条件等，必须统一存储为字符串格式（object），以支持数据的一致性和可查询性。
* 细胞ID的命名必须遵循明确的规范，如10X Genomics数据必须保留原始的Cell Barcode格式。

### 3.3.4 文件完整性和兼容性：

* H5AD文件必须包含完整的文件结构描述文档，说明各数据层的内容和格式，以便用户和自动化工具的理解和使用。
* 文件必须通过兼容性检验，确保能够被主流的单细胞分析工具无误读取。

## 3.4 L4层 ATLAS级别的数据

### 3.4.1数据一致性

* + - 1. Gene ID的明确版本要求：
* 所有gene ID必须明确指定到具体的release版本或build版本（例如，Ensembl Release 104或UCSC hg38）。
* 必须验证所有gene ID在指定命名空间中的存在性和一致性，以防止版本冲突或错误映射。
  + - 1. 原始数据的保存：
* 即使在图谱级别的数据中，也建议保留原始的、未经标准化的计数矩阵。这不仅有助于数据验证和重新分析，也增强了数据的透明度和可追溯性。
* 对于只包含选定基因集的情况，应提供完整原始测序数据及完整的进行基因集筛选的，并可自动化完成筛选的可执行文件。

3.4.1.3数据格式和术语一致性

* 文件中的注释信息的字段应当与控制词汇保持一致。
* 对于文件中出现的非控制词汇字段、非控制数据实体，由用户申请进行控制词汇/数据实体的注册-审核-发布流程。
* 确保L4层的数据格式与L3层保持一致性，例如，在数据类型（如INT或Float32）和数据结构上应保持统一，以便于自动化处理和跨层级数据比较。

3.4.1.4元数据的详细记录和公开：

* 对于每个数据集，提供详细的元数据，描述最小数据元、数据的来源、处理步骤、数据整合方法及其它相关信息。
* 元数据应包括数据访问和使用的条件，如许可信息和任何使用限制。

## 3.5. L5层 知识生成与研究成果

### 3.5.1. 唯一性和标识

* **数据实体唯一性**：确保每个知识实体在系统中拥有唯一的标识符。
* **版本控制**：为每个知识实体实施严格的版本控制机制。

### 3.5.2 数据实体的清晰性和准确性

* **结构化表示**：所有知识实体都应采用结构化的表示方式，如预定义的数据模型或标准化的格式，以减少歧义和提高信息的可读性。
* **数据清洗和验证**：对收集到的数据进行严格的清洗和验证，确保所有数据都是准确、完整的。对于从不同来源整合而来的数据，需要特别注意消除冗余和解决冲突。

### 3.5.3 逻辑关系和因果推理

* **逻辑模型的构建**：在事理图谱中，数据实体之间的逻辑关系应当明确并且易于理解。使用形式化的逻辑表示方法（如本体语言OWL）来定义和存储这些关系。
* **因果推理验证**：对于包含因果推断的知识实体，需要提供足够的证据支持其因果关系的正确性，这可能包括统计数据、实验结果或专家评审。

### 3.5.4 文档化和元数据

* **详细的文档**：为每个知识实体提供详尽的文档，描述其生成的背景、使用的数据、分析方法、以及任何相关的假设或限制条件。
* **元数据管理**：每个知识实体都应有完整的元数据，包括关于其创建、修改、用途和数据源的详细信息。

### 3.5.5 审查和持续改进

* **定期审查**：定期对知识实体和事理图谱进行审查，以确保其持续符合最新的科研标准和实践。
* **改进机制**：建立反馈和改进机制，鼓励用户报告问题，并根据用户反馈对知识实体进行更新和完善。