

针对目前完成的数据集我们主要检查三个部分，我会把检查过的数据集错误标注在“关于现有数据集全面核查情况的记录”中。下面我会列出每个数据集需要人工检查的部分，以及每一个错误对应数据集需要修改的内容：

1, unstructureData

这一部分对应数据集展示的首页，需要检查的部分主要是展示出来的几个关键 keys，包括：

- **题名 metadata['title']**
题名是否对应的是文章标题，而不是数据集标题
- **GSE 号 metadata['accessionNumber'] / PMID metadata['pubmedID']**
这两部分是否是和数据及内容相对应的，正确的编号
- **摘要 metadata['abstract']**
是否完整，不能复制多或者复制少了
- **摘要图 metadata['figureURL']**
对于明确表示有 graphic abstract 的文章，我们需要把这张图放在展示页面；如果没有 graphic abstract，那么放文章第一张图。之前的数据集大部分人没有放图，或者放的是 cluster 图，这部分需要修改的比较多。当摘要图模糊时，更换链接，在文章页面访问原图，使用原图链接；或者访问杂志网站，使用杂志提供的图片链接
- **物种 metadata['taxonomyID']**
大部分为人 Homo sapiens/鼠 Mus musculus，其余会显示 others，检查与文中所用实验对象是否一致
- **组织 metadata['tissue']**
是否对应文中实验取材来源，以及是否是词表中包含的关键字
- **建库方法 metadata['libraryPreparationMethod']**
是否对应文中和数据库中的处理方法（一般在文中 method 和数据集 sample 中的 protocol 位置），以及拼写是否对应词表中的正确格式。
- **杂志 metadata['journal']/出版日期 metadata['publicationDate']**
作者 metadata['authors']/关键词 metadata['keywords']
这四部分由内置函数获取，一般不会出错。

除了上述提到的部分以外，其余部分也全部需要检查，如 doi 是否正确，genomebuild 是否与文中一致等。总之数据的 unstructureData 部分填写的字段全部需要与文中和 GEO 上保持一致，方可验收通过。

2, cellAnnotation

这一部分对应数据集展示的 Dimensional Reduction 中的 clusterName 及对于细胞的其余注释，对应 cellAnnotation 中的“meta_”一类字段。需要检查：

1, 由人员手动填入的 Cluster 信息与原文插图或者文后所提供文档中信息是否一致；如果原文后有提供 Cluster 信息（包含 clusterName, tsne, umap）的话，这部分信息需要以文中提供的为准。如果文章没有提供相应可填入 cluster 信息的文档或者表格，使用函数计算得到的生成名。其余情况可视为填写错误，如有不确定的情况及时在工作群内反馈。

2, cell_meta 的信息是否缺少（以 GEO 中的 GSM 所提供的信息，以及文章中所提供的 description/design 这种对于样本补充信息的注释表格为准。）

, 3, 内容格式混乱的需要调整 (例如大小写差异)

3, 表达矩阵及分 part

1) 数据使用完整性以及正确性

根据文章所提供的公开数据, 应该保证生成以下矩阵:

只提供了 rawcounts: rawcounts 矩阵, 生成 TPM 矩阵

只提供了 normalize: normalize 矩阵 (声明 normalize 方法应与文章中一致), 由 normalize 方法计算 TPM 矩阵 (如果 normalize 方法不明确, 记录并复制 normalize 矩阵作为 TPM 矩阵)

提供了 rawcounts 和 normalize: rawcounts 矩阵, normalize 矩阵, 由 **normalize** 矩阵生成的 TPM。

根据文章中声明的细胞数量, 结合 GEO 中的 sample 信息, 做到矩阵中不缺漏数据, 不混杂非单细胞数据。

2) 分 part

以文中进行的聚类分析作为是否分 part 的依据。如果文中进行了两次独立的聚类分析, 那么应该将数据拆分成 2 个 part; 如果文中进行了 2 次聚类分析, 但第二次是对第一次中某一个 cluster 进行的再聚类, 那么应该只算 1 个 part; 如果文章进行了多次聚类分析, 但某一组数据被重复使用, 那么这组数据只保留一次, 不出现重复数据。

对于分了 part 的数据, 每个 part 都需要执行上面提到的检查项, 还需要额外检查 description, 即对于分 part 标准的叙述是否清晰。

4, not scRNA-seq 数据集

记录并提交报告, 等待清除。