**微生物公共资源整理方案（草稿）**

胡函

目的: 整理和维护微生物相关的数据资源，促进微生物社区的资源共享,为后续的分析和方法开发提供平台和基础。

**主要特色：**微生物疾病与健康的公共数据集，分析依赖的公共参考数据库

1. **公共数据集**

当前可以偏重人体肠道，健康与适应症的人群。后续做扩展。

**1. 原始数据集维护:**

* 原始数据集以tsv格式整理，表格包含如下字段:

|  |  |
| --- | --- |
| Field | Description |
| ID | 数据集ID (内部排序) |
| ExternalID | 原始数据集ID |
| Description | 对数据集的简要描述 |
| SampleNumber | 样本数 |
| Species | 当前只考虑包含human的数据集 |
| Site | 当前以gut为主 |
| Disease | Health/disease |
| DataType | 16s, metagenomics, etc. |
| DataURL | 直接下载页面，有些数据集需要向作者书面请求，这种情况需要注明 |
| MetadataURL | 原始metadata文件链接 |
| CleanedMetadataURL | 统一字段后的metadataURL |
| Reference | 文章来源 |
| AddedBy | 添加人 |
| OrganizedBy | 整理人 (默认值为NA) |

工作时间预期: 3周1-2人,后期持续更新，表格共享，每人都可以往里面添加新信息

* 维护一个github页面，自动更新当前的数据集统计信息: 涉及样本数，涉及数据集数.

第一阶段可以集中搜集和整理健康人和关键适应症（Cancer, ASD, T1D, T2D, …）数据集，参考的数据源渠道有：microbiomeHD，curatedMetagenomicsData, Disbiome等，或者直接在EBI/Qiita上面搜索关键词。

MicrobiomeHD: <https://zenodo.org/record/569601#.XMbO7Oszbys>

curatedMetagenomicsData: <https://github.com/waldronlab/curatedMetagenomicData>

Disbiome: <https://disbiome.ugent.be/>

1. Metadata清理

* 按照字段规范清理之后的metadata表格存放于S3/FTP上, 提供下载URL.
* 使用脚本进行数据清理，这样便于后期对表格删减内容或者做系统的调整。一套数据集对应一个脚本或者Rmarkdown或者Python notebook, 生成该数据集对应的清理好的metadata表格. 另外有一个脚本对各表格进行合并汇总。
* Metadata可以以现有的数据集项目 (如curatedMetagenomicsData) 作为基础, 定期检查更新。规范的Metadata可以分为多个表格。样本信息表的列名如下:

SampleID, SubjectID, BodySite, StudyCondition (main disease or control), Disease, Age, Gender, BMI, Country

数据集特有字段在通用字段后继续添加。现阶段只规范基本的格式（大小写，无特殊字符，单位需注明）

* 未来可以考虑更多表格

Sample / subject description – 样本信息 (步骤2的表格)

Dataset description – 数据集信息 (步骤1的表格)

Diagnosis – 临床诊断及指标 (免疫因子或者临床指标需要统一单位)

Treatment – 治疗方案和follow up

ExperimentNGS – NGS相关实验设置

ExperimentMS – 质谱相关实验设置

LinkingTable – ID mapping， 新增表格可以在此增加新的ID 字段.

* 使用Postgre对样本Metadata进行管理

工作时间预期: 3个月2人.前期一个月更多涉及流程的优化和表格的调整。后期进行持续更新。

**二．公共数据库**

当前考虑菌群序列分类数据库。

1. 16s数据库 (预期工作2个月):

* 当前16s使用的公共数据库主要为green gene (13-8) 和SILVA数据库。
* ezBioCloud提供了一套数据库供学术机构免费试用。
* 未知君当前有一个程序，可以从NCBI (<ftp://ftp.ncbi.nih.gov/blast/db/>) 上抓取16SMicrobial.tar.gz，生成qiime compatible的数据库。
* Himap提供了从NCBI抓取 16s de novo 序列的程序 (<https://github.com/taolonglab/himap>)
* 后续工作：数据库的更新和完善，比如从metagenomics数据集里面抓取16s序列来自动更新数据库。用户可以使用16s序列得到在各个数据库的分类注释信息，又可以得到在宏基因组数据集的相似菌对应的样本metadata信息。

1. 宏基因组数据库 （预计工作1月）:

* Marker基因数据库：根据reference genome和metagenomics数据集更新当前的marker gene数据库，以提高taxonomical profiling的灵敏度。
  + Metaphlan2的数据库生成比较复杂。需要等待开发者进行更新.
  + mOTUs2的数据库更新: 已有script 添加custom genome （https://motu-tool.org/data/extend\_mOTUs\_DBv3.tar.gz），开发者也提到可以向metagenomics里面添加MAG...

1. 病毒与真菌的marker库（未做调研），需要确定具体的工具。

**三. 评测**

我们可以把各种工具和方法的评测（文献的和我们内部的工作都行）放在这个站上，增加这个平台的credibility.

这块会持续更新。直接放上去的工作可以有:

Tax-credit: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5956843/>

16s database compare: <https://www.biorxiv.org/content/10.1101/565572v1>

CAMI:http://www.nature.com/nmeth/journal/vaop/ncurrent/full/nmeth.4458.html?WT.feed\_name=subjects\_computational-biology-and-bioinformatics