

数据交付规范

数据交付

1、数据交付给生信部门

A、按照项目编号放置，没有项目编号的由项目管理提供；

B、数据按照以下规则放置：

/固定路径/项目编号/raw_data/测序数据类型/测序平台（种类）/样品/批次/数据文件

示例（Phoenix上路径）：

/export/backup/WHWLZ-20181108A/raw_data

```
├── CustomerProvide-20181120
├── Genome
│   ├── HiC
│   ├── NGS
│   │   ├── XY1-D13800-Q
│   │   │   ├── wuxiapp
│   │   │   └── M2K-01
│   ├── ONT
│   │   ├── XY1_NPL0046
│   │   │   └── 20181112-NPL0046-P1-A1-D1
│   └── Sequel
│       ├── XY1_PBL180159
│       │   └── r54061_20181113_83020
│       │       └── 2_B01
├── RNA
└── Reseq
```

说明：

固定路径：/export/backup ##不同服务器集群上固定路径不完全一致，视情况而定

项目编号：WHWLZ-20181108A ##根据项目不同而更改

数据类型：Genome、RNA、Reseq等 ##分别为基因组、转录组、重测序数据，有其他类型的数据也可以额外增加；客户提供数据，无论类型，统一用“CustomerProvide-20181120”这种格式来放置

测序平台（种类）：HiC、NGS、ONT、Sequel等 ##根据测序平台的不同来命名区分，外包数据需要由项目管理提供准确数据类型

样品名称：XY1_NPL0046、XY1_PBL180159 ##样品名称一般由所测序的样品名称+文库编号（样品编号）组成，三代数据用文库编号，二代数据用样品编号

批次：二代数据的批次由测序服务公司+测序lane两个层级编号组成；自测的MGI数据，用测序仪编号+芯片号做为区分；Sequel数据的批次为run编号，即为“r54061_20181113_83020”；ONT数据的cell编号可以替代测序批次。

C、需要注意的细节：

（1）放到/export/backup下面的数据，需要更改属主为root，更改所有文件、目录权限为755，不允许有777的文件出现。

（2）数据一经放好，路径、文件名不可轻易更改，防止发生前后对不上的情况。如果确实需要更改，注意提前询问相关生信人员数据是否在使用，并将修改后的路径重新发出。

2、数据交付给客户（即数据释放）

原始数据释放

- 统一要求（强制）：

1. 在raw_data下面创建README文件夹，里面统一放置几个文件：

- a、Pacbio Sequel下机数据说明.pdf
- b、ONT下机数据说明.pdf
- c、数据目录结构说明.txt

2. 注意提供MD5校验文件：

- a、二代数据和Sequel数据各自计算MD5，并放置在测序平台这层目录中。
- b、ONT数据的数据中，qc_report和reads下面有各自的md5文件。
- c、MD5文件校验结果写入各自目录的md5_check.log文件中。
- d、交付结果太大的话，考虑到周期问题，可以先行交付。

3. 数据交付记录

- a、数据交付后要做好记录，包含交付数据的项目编号、样品编号、数据类型、文件大小、交付时间等信息。
- b、客户收到数据，确认无误后，签署数据确认函。

- 各平台数据详细要求：

1. Sequel数据：

数据按照上述交付给生信的格式释放：

a、上面的几层目录，需要包含"项目编号/数据类型/测序平台等

b、再按照"样品名_文库编号"的命名规则建一层目录，下面再按照"run编号/cell编号/完整cell数据"的方式来放置，如：

WHWLZ-

201612214C/raw_data/genome/Sequel/Pig_PBL170214/r54061_20170501_090428

```

|
|   └─ 1_A05
|   └─
m54061_170501_091426.adapters.fasta
|   └─
m54061_170501_091426.scrap.s.bam
|   └─
m54061_170501_091426.scrap.s.bam.pbi
|   └─
m54061_170501_091426.sts.xml
|   └─
m54061_170501_091426.subreads.bam
|   └─
m54061_170501_091426.subreads.bam.pbi
|   └─
m54061_170501_091426.subreadset.xml
|   └─
m54061_170501_091426.transferdone
|   └─ tmp-
file-3329685782346117053.txt
└─ 2_B05
```

2. ONT数据：

数据按照上述交付给生信的格式释放：

a、上面的几层目录，需要包含"项目编号/数据类型/测序平台等

b、再按照"样品名_文库编号"的命名规则建一层目录，下面再按照"cell编号/完整cell数据"的方式来放置，如：

WHWLZ-201612214C/raw_data/genome/ONT/Pig_NPL0024

```

|   └─ 20181120-NPL0024-D1
|       └─ GA10000
|           └─ qc_report
|           └─ reads
└─ 20181120-NPL0024-D2
```

c、qc_report下面存放的是fastq（经过）和summary.txt，reads下面是ONT的原始下机数据fast5文件打包后的文件。

3. 二代数据

- a、目前公司的二代数据有很多是外包测序的，每个公司的命名规则、测序质量参差不齐
- b、二代数据因加测产生的批次较多，

综上，有分析任务的项目，二代数据由生信项目负责人对实际分析中用到的数据进行整理。多批次的建议合并后压缩，有多个样品的注意进行区分。纯测序项目由IT中心对数据按照样品进行合并、压缩后交付。

结果文件释放

- 项目结果文件一般由生信部门提供，有以下几部分需要注意：

- a、数据由相关生信人员整理好后，交付给IT中心，在交付前保证交付数据的完整性，并确保已经去除了不需要或者不允许提供给客户的相关文件（如中间文件、程序、脚本）等。
- b、交付到IT中心的数据需要按照规则放到特定目录中，同一任务，IT中心只接收一个目录。
- c、交付的结果中，若有海量小文件或者fa、sam等文件存在，需要对数据进行一定程度的打包、压缩，方便进行数据进行传输与交付。
- d、由生信负责人对交付的数据进行MD5校验文件生成。
- e、交付的数据中，需要包含相关目录、文件的Readme说明文件，由生信提供。

3、数据下载及客户提供数据

1. IT中心只接收可以直接下载链接等，如提供文献、或者NCBI这种的一个Project ID，要求IT中心进行下载，IT中心可以不予接收任务。
2. 客户通过硬盘（公司的或者客户自己的）寄送数据到公司，需由对应的销售在包装内附上相关数据的情况：所属项目、需要拷贝的数据等信息，同时也应提前发邮件说明，以便及时处理数据。否则，数据不予处理。