交付数据格式

示例：

```
    raw_data/genome/ONT/Sample_NPL0001
                        ├── 20181120-NPL0001-D1
                        │   └── GA10000
                        │       ├── qc_report.md5.tsv
                        │       ├── qc_report
                        │       │   ├── fastq
                        │       │   │   ├── 20181120-NPL0024-D1.fastq.gz
                        │       │   │   └── 20181120-NPL0024-
D1.sequencing_summary.txt.gz
                        │       │   └── stat
                        │       │       ├── 20181120-NPL0024-
D1.summary4stat.tsv.gz
                        │       │       ├── distribution_stat.json
                        │       │       ├── nx0_stat.json
                        │       │       └── qc_stat.json
                        │       └── reads
                        │           ├── 0.tar
                        │           ├── 1.tar
                        │           ├── ... ...
                        │           └── 51.tar
                        └── 20181120-NPL0001-D2
```

各级目录说明

以上述示例中的GridION数据为例，对ONT数据交付的各个目录进行说明如下：

- **Sample_NPL0001**:样品目录，“Sample”是样品名称，“NPL0001”是文库编号。
- **20181120-NPL0001-D1**：ONT的cell编号，也可以代表批次编号。“20181120”是上机时间，“NPL0001”是文库编号，“D1”是ONT测序仪的内部编号+槽号。这个编号是实验人员上机指定的，具有唯一性。
- **GA10000**：测序仪的槽编号，由测序仪自动生成。GridION的槽编号有GA10000~GA50000，分别代表芯片槽1号到5号；PromethION的槽编号有1-A1-E1、1-E1-D1、2-A1-D2等。
- **qc_report**：过滤后的fastq数据及相关统计信息，其中子目录的信息如下。
- **fastq**：ONT的下机fastq数据。
  ONT系列的测序仪在basecalling的过程因为软件bug，可能会出现一些异常，导

致生成的fastq以及summary.txt文件出现随机文本错误，对后续的质控、统计、组装等分析造成影响。未来组根据已有经验，对原始下机的fastq和summary.txt进行过滤，去除了异常的reads以及错误序列，使得后续分析可以正常进行。

> 补充说明：
> 这个过滤的步骤，只会去除异常的错误reads，没有对低质量以及短reads进行过滤。这些异常的reads在完整的一个cell的reads总数中占比极少，对数据量、长度分布等指标基本上不会造成影响。

- **stat**：对qc_report中的数据进行统计，统计结果文件记录在这个目录下面的文件中。
- **reads**：原始下机的fast5文件，各个目录以tar的方式进行了打包。

目录内文件说明：

- qc_report.md5.tsvqc_report.md5.tsv：

  记录qc_report下面的所有数据文件的MD5值，用于文件传输完成后的完整性校验。

- 20181120-NPL0024-D1.fastq.gz：

  过滤了异常reads后剩余的fastq文件，用gzip进行了压缩。

- 20181120-NPL0024-D1.sequencing_summary.txt.gz

  过滤了异常reads后剩余的sequencing_summary.txt文件，用gzip进行了压缩。

- 20181120-NPL0024-D1.summary4stat.tsv.gz

  sequencing_summary.txt中包含的信息比较多，summary4stat.tsv中存放的是提取了其中用于数据质控的几列相关信息，分别是read_id、read长度、read的平均质量分数。

- distribution_stat.json

  记录各类型reads长度分布规律。

- nx0_stat.json

  记录各类型reads的N10~N90的相关统计情况。

- qc_stat.json

综合展示各类型reads的相关统计信息。

文件数据格式解读

- **fastq格式说明**（20181120-NPL0024-D1.fastq.gz）：

```
@0db0be86-28ea-4537-8e06-fbfb2d47ff60 runid=8b28251aab26537572894331345l7b570589fc8a sampleid=20181120-NPL0024-D1 read=1682 ch=479 start_time=2018-11-11T01:11:32Z
CATCGGTATGCTTTACGTTCAGTACATTATTGCTAATACTGTTAACTAAAACCTTTATTGGAGTGATAAAGTAGTGCCAAACTTCTCTATCTACGTCAGTAATAAGGGAGAACCATACGTAATAAATTTAA
+
""((&$%$$#++.'$#&*-)$(,%)&(,)-/./3/9/)*3**.-*).+8662017+%/)3(')),-202-+,*(&%&$(,)(-'+()*...-(&(+*$'*,3)-('')(%'&$"$*((*()+1-,1*07*.1
@59278f02-c78d-4ed9-8522-63205aed5c9c runid=8b28251aab26537572894331345l7b570589fc8a sampleid=20181120-NPL0024-D1 read=1528 ch=220 start_time=2018-11-11T01:10:57Z
CTCGGTATTACTTCGTTCAGTTACGTATTGCTTAATTAGATCTCTTGCCCGTTGATTAATTTAGTCAATGGTACATCCATCAATGTTCTGTCCCGGTGCTCCATTTTTCATAAATAAATTGAATCCTATATGGTCAT
+
"/1%++&%*)/070()10,1640223)0'$',*-(-)&$&)')'-0-(-5+(03,*+l*0,::0789.2.(*$$$"*/,,.&&%*%07340.2''',+),.8;73<<:,8768;;4885382/8540/02)*.1,,$(
@29a3ab74-88cc-4c2a-875f-3c07084b8296 runid=8b28251aab26537572894331345l7b570589fc8a sampleid=20181120-NPL0024-D1 read=1980 ch=509 start_time=2018-11-11T01:11:46Z
ATATTGTACTTCGTTCAGTTACGTATTGCTATAAGAGGAGAAGTCATAACTACCTCCCCATGTTACCTCCTTATTTGTTGATTATTCTACGAGTTCTTTTGATTTATATTGGAG
+
"+(-,),-028/.472.090/104+3,*,6552+))-,(('')%(),525/0/4(*.76,.+):,+45461)/$7.'%,.*(*,&,4'($&#%*5873:;9143;/--,5-4+-.
```

说明：

a. 上述示例图中，包含三条reads，基本上就是fastq文件的标准格式

b. ONT下机的fastq中，reads id包含很多元素，分别代表的含义：

> 以“@0db0be86-28ea-4537-8e06-fbfb2d47ff60 runid=8b28251aab26537572894331345l7b570589fc8a sampleid=20181120-NPL0024-D1 read=1682 ch=479 start_time=2018-11-11T01:11:32Z”这个ID为例

@0db0be86-28ea-4537-8e06-fbfb2d47ff60：

ONT下机数据中，每条reads都有一个唯一的ID，“0db0be86-28ea-4537-8e06-fbfb2d47ff60”就是这条reads的ID。

runid=8b28251aab26537572894331345l7b570589fc8a：

ONT每个cell测序的时候都会产生一个runid,同一个cell的每条reads的runid都是一致的，不同cell的runid不一致。

sampleid=20181120-NPL0024-D1：

sampleid，和每个cell的cell编号相同。

read=1682：

这条reads是对应channel上通过的第几条链。

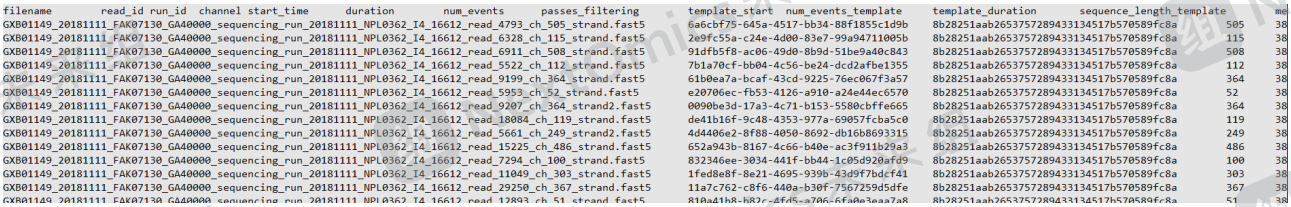
ch=479：

测序芯片上的channel编号

start_time=2018-11-11T01:11:32Z：

这条reads开始生成的时间。

- **sequencing_summary.txt格式说明**(20181120-NPL0024-D1.sequencing_summary.txt.gz)



filename	read_id	run_id	channel	start_time	duration	num_events	passes_filtering	template_start	num_events_template	template_duration	sequence_length_template	me
GXB01149_20181111_FAK07130_GA40000_sequencing_run_20181111_NPL0362_I4_16612_read_4793_ch_505_strand.fast5	6a6cbf75-645a-4517-bb34-88f1855c1d9b	8b28251aab26537572894331345l7b570589fc8a				505						38
GXB01149_20181111_FAK07130_GA40000_sequencing_run_20181111_NPL0362_I4_16612_read_6328_ch_115_strand.fast5	2e9fc55a-c24e-4d00-83e7-99a94711005b	8b28251aab26537572894331345l7b570589fc8a				115						38
GXB01149_20181111_FAK07130_GA40000_sequencing_run_20181111_NPL0362_I4_16612_read_6911_ch_508_strand.fast5	91dfb5f8-ac06-49d0-8b9d-51be9a40c843	8b28251aab26537572894331345l7b570589fc8a				508						38
GXB01149_20181111_FAK07130_GA40000_sequencing_run_20181111_NPL0362_I4_16612_read_5522_ch_112_strand.fast5	7b1a70cf-bb04-4c56-be24-dcd2afbe1355	8b28251aab26537572894331345l7b570589fc8a				112						38
GXB01149_20181111_FAK07130_GA40000_sequencing_run_20181111_NPL0362_I4_16612_read_9199_ch_364_strand.fast5	61b0ea7a-bcaf-43cd-9225-76ec067f3a57	8b28251aab26537572894331345l7b570589fc8a				364						38
GXB01149_20181111_FAK07130_GA40000_sequencing_run_20181111_NPL0362_I4_16612_read_5953_ch_52_strand.fast5	e20706ec-fb53-4126-a910-a24e44ec6570	8b28251aab26537572894331345l7b570589fc8a				52						38
GXB01149_20181111_FAK07130_GA40000_sequencing_run_20181111_NPL0362_I4_16612_read_9207_ch_364_strand2.fast5	0090be3d-17a3-4c71-b153-5580cbffe665	8b28251aab26537572894331345l7b570589fc8a				364						38
GXB01149_20181111_FAK07130_GA40000_sequencing_run_20181111_NPL0362_I4_16612_read_18084_ch_119_strand.fast5	de41b16f-9c48-4353-977a-69057fcba5c0	8b28251aab26537572894331345l7b570589fc8a				119						38
GXB01149_20181111_FAK07130_GA40000_sequencing_run_20181111_NPL0362_I4_16612_read_5661_ch_249_strand2.fast5	4d4406e2-8f88-4050-8692-db16b8693315	8b28251aab26537572894331345l7b570589fc8a				249						38
GXB01149_20181111_FAK07130_GA40000_sequencing_run_20181111_NPL0362_I4_16612_read_15225_ch_486_strand.fast5	652a943b-8167-4c66-b40e-ac3f911b29a3	8b28251aab26537572894331345l7b570589fc8a				486						38
GXB01149_20181111_FAK07130_GA40000_sequencing_run_20181111_NPL0362_I4_16612_read_7294_ch_100_strand.fast5	832346ee-3034-441f-bb44-1c05d920afd9	8b28251aab26537572894331345l7b570589fc8a				100						38
GXB01149_20181111_FAK07130_GA40000_sequencing_run_20181111_NPL0362_I4_16612_read_11049_ch_303_strand.fast5	1fed8e8f-8e21-4695-939b-43d9f7bdcf41	8b28251aab26537572894331345l7b570589fc8a				303						38
GXB01149_20181111_FAK07130_GA40000_sequencing_run_20181111_NPL0362_I4_16612_read_29250_ch_367_strand.fast5	11a7c762-c8f6-440a-b30f-7567259d5dfe	8b28251aab26537572894331345l7b570589fc8a				367						38
GXB01149_20181111_FAK07130_GA40000_sequencing_run_20181111_NPL0362_I4_16612_read_12893_ch_51_strand.fast5	810e41b8-b82c-4fd5-a706-6fa0e3eaa7a8	8b28251aab26537572894331345l7b570589fc8a				51						38

说明：

a. sequencing_summary.txt是纯文本文件，包含的信息比较多，上图没有完全展示出来

b. 每一列代表的含义分别是：

filename：表示read从该fast5文件中提取
read_id：read编号
run_id：表示read在该run中进行的测序
channel：flowcell上四个紧挨着的纳米孔称为一个channel，表示read是从该channel中产出
start_time：表示该read开始测序的时间（s），从整个cell上机开始计时为0
duration：表示该read测序的持续时间（s）
num_events：表示该read在事件删除前的事件数目，一个事件代表5个碱基叠加在一起的一次信号
passes_filtering：根据mean_qscore_template值是否大于7，大于7的为True，否则为False
template_start：表示模板链开始测序的时间（s）
num_events_template：表示模板链在事件删除前的事件数目
template_duration：表示模板链测序的持续时间（s）
sequence_length_template：表示模板链在碱基化后的碱基数目
mean_qscore_template：表示模板链的平均质量值分数
strand_score_template：表示模板链的对数似然分数，归一化为模板链长度

---

pass reads相关说明

- **概念**

  pass reads也称为High Quality Reads（HQ），是指在ONT的测序中，平均质量分数较高的reads。一般是以sequencing_summary.txt中mean_qscore_template的数值大于等于7为标准的。

  与pass reads相对应的为fail reads，也称为Low Quality Reads（LQ），是平均质量分数小于7的reads。

  pass reads可以直接用于进行组装，也可以对pass reads过滤掉短片段后再进行

组装。

- **提取方式**：
  pass reads的提取要结合sequencing_summary.txt来进行。未来组的生信工程师编写了一个简单的小程序用于提取pass reads，并发布在了Github上。程序下载链接及使用说明：https://github.com/FlyPythons/ontbc
  该程序还可以用来拆分加bacode混测的ONT数据的拆分。

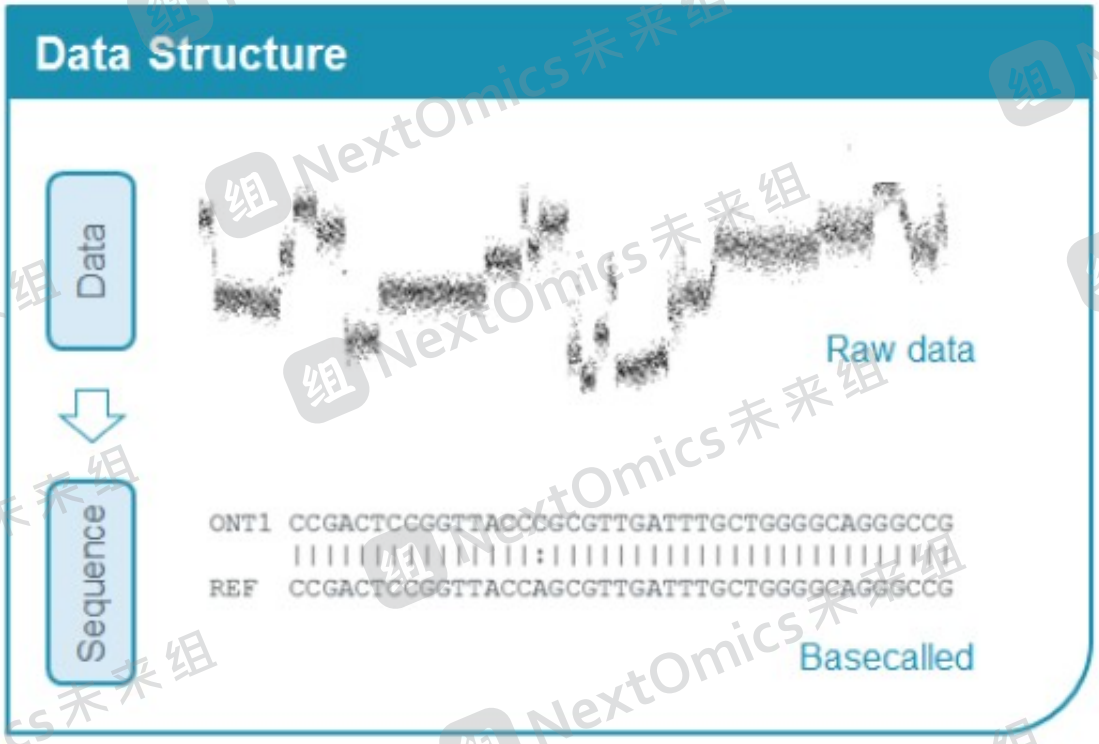
---

fast5相关说明

- fast5文件是一种HDF5格式的文件（ https://www.hdfgroup.org/ ），是一种高压缩格式的二进制文件。
- fast5文件包含了ONT测序过程中所获取到的全部电位信息，可用于后续相关的生信分析（如甲基化分析等）。
- 一个fast5文件包含的内容，就是一条reads的全部信息。每个cell的数据统计中有多少条reads，就有多少个fast5文件，所以每个cell产生的fast5文件数量是海量的，可以是数十万、数百万甚至千万级别的。海量的文件对于数据传输的效率影响极大，为了保障数据的快速传输，我们对fast5采用了tar打包。
- fast5文件可以通过HDF5系列工具进行查看，windows下用的是HDFView，linux下使用的是HDFView或者HDF5。

ONT basecalling相关说明

- **概念**：
  在ONT的测序平台中，将通过纳米孔的DNA或RNA链产生的电位信号转化为相应的碱基序列的过程，称为basecalling。



- **工具**:

  ONT官方提供了多个工具，可用于不同平台、不同场景下的basecalling工作。

  1. Albacore：官方提供的工具，采用神经网络算法，适用于常规的单节点或者集群环境。目前对外提供可执行程序，源码仅对开发者开放。
  2. Guppy：官方提供的工具，与Albacore使用相同的算法，安装在GridION和PromethION测序仪的系统中，是这两个型号测序仪默认的basecalling工具。Guppy可以调用GPU进行加速，运行效率高于Albacore。源码及安装包都仅对开发者开放。