诺禾致源数据结果说明文档

---质控后数据结果

测序得到的原始测序序列，里面含有带接头的、低质量的 reads，会为后续分析增加复杂度。为了保证信息分析质量，需要对 raw reads 进行精细过滤，得到 clean reads，后续分析都基于 clean reads 进行。

数据处理的步骤如下：

(1) 需要过滤掉含有接头序列的 reads；

(2) 当单端测序 read 中含有的 N(N 表示无法确定碱基信息)的含量超过该条 read 长度比例的 10%时,需要去除此对 paired reads；

(3) 当单端测序 read 中含有的低质量(低于 5)碱基数超过该条 read 长度比例的 50%时，需要去除此对 paired reads。

本文件夹下包含每个样本(样本名)的 clean data 数据结果文件：

1.后缀 clean.fq.gz　　经过 QC 处理后的 clean data 序列压缩文件

2.MD5.txt　　clean.fq.gz 的 MD5 值，用于检查文件的完整性

文件名中 1,2 分别代表 Paired-end 两端序列文件

==========

结果文件说明

clean.fq 为高通量测序的过滤后的 clean 数据，结果以 FASTQ 文件格式存储。包含测序序列的序列信息和对应的测序质量信息。FASTQ 文件中每个 read 由四行描述。其格式如下：

@HWI-ST1276:71:C1162ACXX:1:1101:1208:2458 2:N:0:CGATGT
CTGGCTCCGGAGGGGATGGAGGCGGCACTCCCGCCAAGGATGCGTTGGGAAACGACG
TCGTTGCAGTCGAATGGCTCAAAACACACGGGCCCGGTGACCGG
+
BCBFFFFDHHHHHHJJ?EAGIIIAHIJIIGHHHBEDCDDD;>>BD?BDAD<><?BDB@5<BBDCD
DDCDDBDCCDDDCCDDD8?AAB9>B55>BB5904@BB

其中第一行以“@”开头，随后为 Illumina 测序标识符（Sequence Identifiers）和描述文字（选择性部分）；
第二行是碱基序列；
第三行以“+”开头，随后为 Illumina 测序标识别符（选择性部分）；
第四行是对应碱基的测序质量，该行中每个字符对应的 ASCII 值减去 33，即为对应第二行碱基的测序质量值。

Illumina 测序标识别符（Sequence Identifiers）详细信息如下：

HWI-ST1276	Unique instrument name
71	Run ID
C1162ACXX	Flowcell ID
1	Flowcell lane
1101	Tile number within the flowcell lane

1208	'x'-coordinate of the cluster within the tile
2458	'y'-coordinate of the cluster within the tile
2	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
N	Y if the read fails filter (read is bad), N otherwise
0	0 when none of the control bits are on, otherwise it is an even number
CGATGT	Index sequence