诺禾致源数据结果说明文档

---原始数据结果

高通量测序（如 Illumina HiSeq XTen/2500/2000/MiSeq 等测序平台）得到的原始图像数据文件经碱基识别(Base Calling)分析转化为原始测序序列（Sequenced Reads），我们称之为 Raw Data 或 Raw Reads

本文件夹下包含每个样本(文库名)的原始数据结果文件：

1.后缀 fq.gz 文件　　高通量测序的原始测序序列压缩文件

2.MD5.txt　　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

1.fq.gz 为高通量测序的原始测序序列，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?EAGIIIAHIJIIGHHHHBEDCDDD;>>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