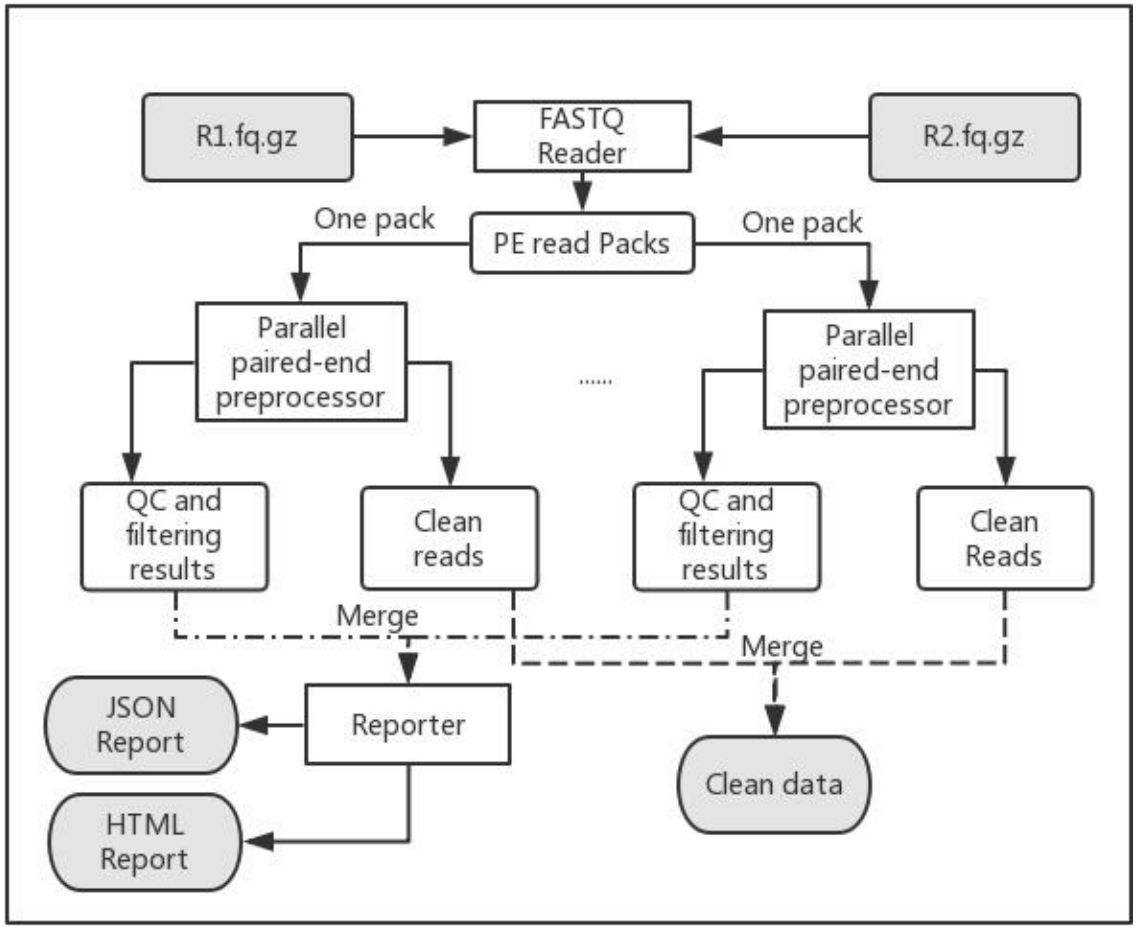
fastp参数说明

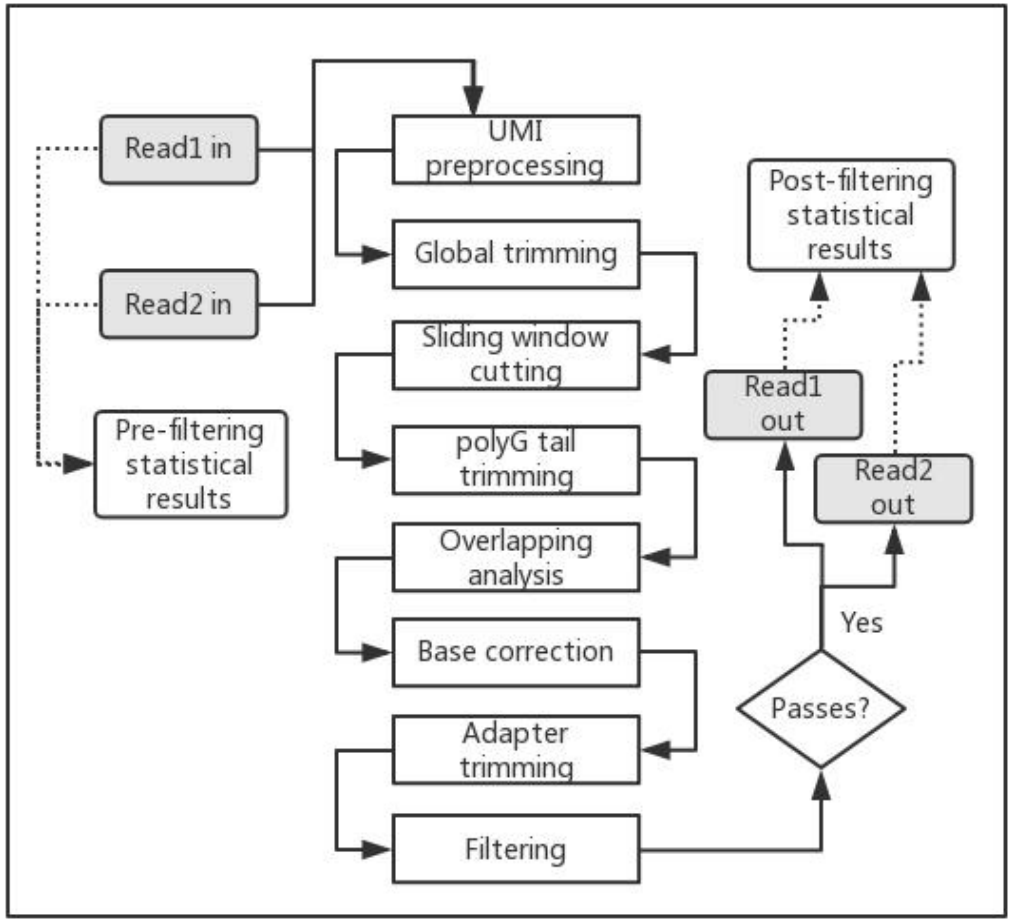
周雪婷



(a) main workflow

□fastp：用于执行FASTQ数据的质量控制，支持单端和双端短读数据，作为一体化的FASTQ预处理器，fastp提供了诸如质量分析，接头修整，读取过滤和基本校正等功能。在处理完所有读取后，这些值将合并，并且报告者将生成HTML和JSON格式的报告

□版本：fastp v0.12.6



(b) paired-end preprocessor

一、输入和输出

/home/lcbio/Softwares/fastp/fastp

-i Input/BS06629/BS06629_R1.fastq.gz

-I Input/BS06629/BS06629_R2.fastq.gz

-o CleanData/BS06629/BS06629_R1.fastq.gz

-O CleanData/BS06629/BS06629_R2.fastq.gz

缩写	解释	数据	示例	原文解释
-i	--in1 read1	input1	Input/BS06629/BS06629_R1.fastq.gz	input file name (string)
-o	--out1	output1	Input/BS06629/BS06629_R2.fastq.gz	read1 output file name (string [=])
-I	--in2	input2	CleanData/BS06629/BS06629_R1.fastq.gz	read2 input file name (string [=])
-O	--out2	output2	CleanData/BS06629/BS06629_R2.fastq.gz	read2 output file name (string [=])

二、质量控制和报告

/home/lcbio/Softwares/fastp/fastp

-i Input/BS06629/BS06629_R1.fastq.gz

-I Input/BS06629/BS06629_R2.fastq.gz

-o CleanData/BS06629/BS06629_R1.fastq.gz

-O CleanData/BS06629/BS06629_R2.fastq.gz

-j CleanData/BS06629/fastp_BS06629.json

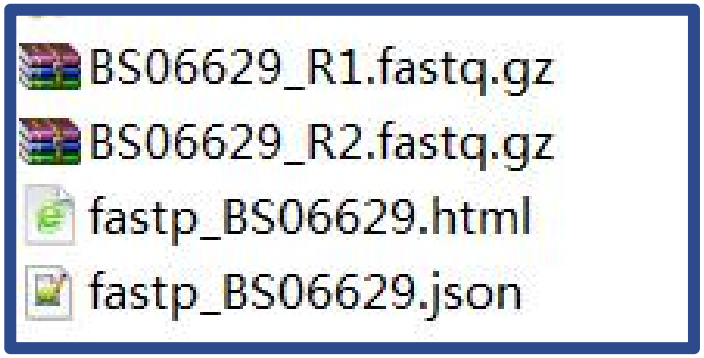
-h CleanData/BS06629/fastp_BS06629.html

-R BS06629 fastp report

缩写	解释	示例	原文解释
-j	--json	（生成）CleanData/BS06629/fastp_BS06629.json	the json format report file name (string [=fastp.json])
-h	--html	（生成）CleanData/BS06629/fastp_BS06629.html	the html format report file name (string [=fastp.html])
-R	--report_title	（生成报告题头）BS06629	should be quoted with ' or ", default is "fastp report" (string [=fastp report])

二：质量控制和报告

- fastp支持通过低质量基本百分比，N基数和读取长度过滤读取，可以通过直接比较数字来评估过滤效果。
- 以JSON和HTML格式报告结果，而JSON报告包含HTML报告中可视化的所有数据。
- JSON报告的格式是手动优化的，人们可以轻松阅读。
- HTML报告是一个单独的网页，所有图形都是使用JavaScript和Web画布动态创建的。



1、压缩：--z 1

```
/home/lcbio/Softwares/fastp/fastp
-i Input/BS06629/BS06629_R1.fastq.gz
-I Input/BS06629/BS06629_R2.fastq.gz
-o CleanData/BS06629/BS06629_R1.fastq.gz
-O CleanData/BS06629/BS06629_R2.fastq.gz
-j CleanData/BS06629/fastp_BS06629.json-hCleanData/BS06629/fastp_BS06629.html
-R BS06629fastpreport
-z 1
```

参数	解释	详解	原文解释
-z	--compression	压缩比例	compression level for gzip output (1 ~ 9). 1 is fastest, 9 is smallest, default is 2. (int [=2])

1、压缩：--z 1

output1	CleanData/BS06629	
output2	Read1 before filtering: (3.04G)	Read2 before filtering: (3.15G)
	total reads: 42000000	total reads: 42000000
	total bases: 6300000000	total bases: 6300000000
	Q20 bases: 6174720971(98.0114%)	Q20 bases: 6129880727(97.2997%)
	Q30 bases: 5970327989(94.7671%)	Q30 bases: 5874909241(93.2525%)
	Read1 after filtering: (2.97G)	Read2 aftering filtering: (3.07G)
	total reads: 41401465	total reads: 41401465
	total bases: 6175319835	total bases: 6175238303
	Q20 bases: 6067183189(98.2489%)	Q20 bases: 6044182561(97.8777%)
	Q30 bases: 5872182296(95.0911%)	Q30 bases: 5803015607(93.9723%)
	Filtering result:	
	reads passed filter: 82802930	
	reads failed due to low quality: 1141902	
	reads failed due to too many N: 0	
	reads failed due to too short: 55168	
	reads with adapter trimmed: 5224060	
	bases trimmed due to adapters: 71749313	
	JSON report: CleanData/BS06629/fastp_BS06629.json	
	HTML report: CleanData/BS06629/fastp_BS06629.html	
	fastp v0.12.6, time used: **712 seconds**	

1、压缩：--z 5

output1	CleanData/BS06629	
output2	Read1 before filtering: (3.04G)	Read2 before filtering: (3.15G)
	total reads: 42000000	total reads: 42000000
	total bases: 6300000000	total bases: 6300000000
	Q20 bases: 6174720971(98.0114%)	Q20 bases: 6129880727(97.2997%)
	Q30 bases: 5970327989(94.7671%)	Q30 bases: 5874909241(93.2525%)
	Read1 after filtering: (2.71G)	Read2 aftering filtering:(2.80G)
	total reads: 41401465	total reads: 41401465
	total bases: 6175319835	total bases: 6175238303
	Q20 bases: 6067183189(98.2489%)	Q20 bases: 6044182561(97.8777%)
	Q30 bases: 5872182296(95.0911%)	Q30 bases: 5803015607(93.9723%)
	Filtering result:	
	reads passed filter: 82802930	
	reads failed due to low quality: 1141902	
	reads failed due to too many N: 0	
	reads failed due to too short: 55168	
	reads with adapter trimmed: 5224060	
	bases trimmed due to adapters: 71749313	
	JSON report: CleanData/BS06629/fastp_BS06629.json	
	HTML report: CleanData/BS06629/fastp_BS06629.html	
	fastp v0.12.6, time used: **1786 seconds(29min)**	

1、压缩：--z 9

output1	CleanData/BS06629	
output2	Read1 before filtering:	Read2 before filtering:
	total reads: 42000000	total reads: 42000000
	total bases: 6300000000	total bases: 6300000000
	Q20 bases: 6174720971(98.0114%)	Q20 bases: 6129880727(97.2997%)
	Q30 bases: 5970327989(94.7671%)	Q30 bases: 5874909241(93.2525%)
	Read1 after filtering:	Read2 aftering filtering:
	total reads: 41401465	total reads: 41401465
	total bases: 6175319835	total bases: 6175238303
	Q20 bases: 6067183189(98.2489%)	Q20 bases: 6044182561(97.8777%)
	Q30 bases: 5872182296(95.0911%)	Q30 bases: 5803015607(93.9723%)
	Filtering result:	
	reads passed filter: 82802930	
	reads failed due to low quality: 1141902	
	reads failed due to too many N: 0	
	reads failed due to too short: 55168	
	reads with adapter trimmed: 5224060	
	bases trimmed due to adapters: 71749313	
	JSON report: CleanData/BS06629/fastp_BS06629.json	
	HTML report: CleanData/BS06629/fastp_BS06629.html	
	fastp v0.12.6, time used: **8678 seconds(144min)**	

2、接头修剪（Adapter trimming）：-A

/home/lcbio/Softwares/fastp/fastp

-i Input/BS06629/BS06629_R1.fastq.gz -I Input/BS06629/BS06629_R2.fastq.gz -o CleanData/BS06629/BS06629_R1.fastq.gz -O CleanData/BS06629/BS06629_R2.fastq.gz

-j CleanData/BS06629/fastp_BS06629.json-hCleanData/BS06629/fastp_BS06629.html

-R BS06629fastpreport -z 1

-A

参数	解释	详解	原文解释
-A	--disable_adapter_trimming	默认启用了接头处理，但是可以使用-A 命令来关掉	adapter trimming is enabled by default. If this option is specified, adapter trimming is disabled
-a	--adapter_sequence	-a 参数来输入你的接头	the adapter for read1. For SE data, if not specified, the adapter will be auto-detected. For PE data, this is used if R1/R2 are found not overlapped. (string [=auto])

2、接头修剪（Adapter trimming）：-A

Adapters

Adapter or bad ligation of read1

The input has little adapter percentage (~0.569814%), probably it's trimmed before.

Sequence	Count
A	155300
AG	147425
AGA	142781
AGAT	136618

Adapter or bad ligation of read2

The input has little adapter percentage (~0.569064%), probably it's trimmed before.

Sequence	Count
A	155188
AG	147381
AGA	142763
AGAT	136691

- 单端数据，通过组装高频读取尾部来检测接头序列
- 双端数据，通过找到每对的重叠来检测接头序列 ，并将重叠区域外的基数视为接头内容。

3、Base校正（Base correction）：-c

参数	解释	详解	原文解释
-c	--correction	该校正功能默认没有开启使用-c 参数可以启用，对于一些对噪声容忍度低的应用，比如液体活检，建议开启	enable base correction in overlapped regions (only for PE data), default is disabled

- 对于双端端数据，如果可以检测到具有良好重叠的一对读数，则可以比较重叠区域内的碱基。如果读取质量很高，它们通常是反向互补。

- 如果在重叠区域内发现一些不匹配，fastp将尝试纠正不匹配。 fsatp仅用不平衡的质量得分校正不匹配的碱基对，这意味着其中一个具有高质量得分（> Q30）而另一个具有低质量得分（<Q15）。为了减少错误校正，如果总错配数小于阈值T（T = 5），则fastp仅执行校正。

4、滑动窗口质量切割（Sliding window quality cutting）：-5、-3、-W、-M

/home/lcbio/Softwares/fastp/fastp
-i Input/BS06629/BS06629_R1.fastq.gz -I Input/BS06629/BS06629_R2.fastq.gz -o CleanData/BS06629/BS06629_R1.fastq.gz -O CleanData/BS06629/BS06629_R2.fastq.gz
-j CleanData/BS06629/fastp_BS06629.json
-hCleanData/BS06629/fastp_BS06629.html
-R BS06629fastpreport -z 1
-3 -W 16 -M 20

参数	解释	详解	阈值	原文解释
-5	--cut_by_quality5	参数开启在 5’端，也就是 read 的开头的剪裁		enable per read cutting by quality in front (5'), default is disabled (WARNING: this will interfere deduplication for both PE/SE data)
-3	--cut_by_quality3	-3 参数开启在 3’端，也就是 read 末尾的剪裁		enable per read cutting by quality in tail (3'), default is disabled (WARNING: this will interfere deduplication for SE data)
-W	--cut_window_size	指定滑动窗大小，默认是 4	划窗大小：2~10	the size of the sliding window for sliding window trimming, default is 4 (int [=4])
-M	--cut_mean_quality	指定要求的平均质量值，默认是 20，也就是 Q20	质量值：1 ~ 30，suggest：15 ~ 20	the bases in the sliding window with mean quality below cutting_quality will be cut, default is Q20 (int [=20])

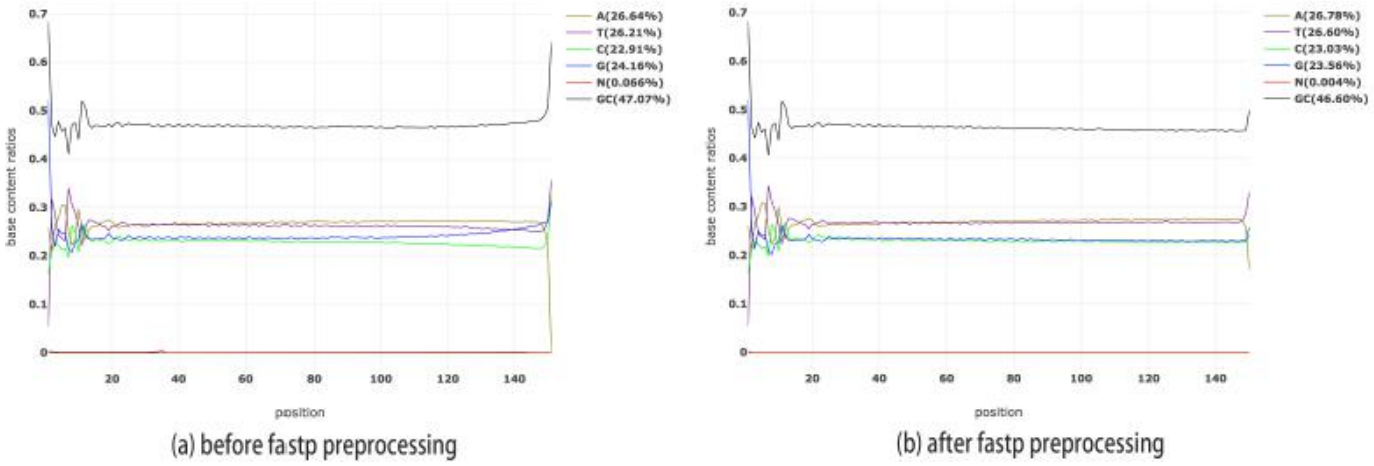
4、滑动窗口质量切割（Sliding window quality cutting）：-5、-3、-W、-M

- 为了提高读取质量，fastp支持滑动窗口方法来删除每个读取头部和尾部的低质量基础。窗口可以从5'滑动到3'，或从3'滑动到5'，并评估窗口内的平均质量分数；
- 指定滑动窗大小，默认是 4；
- 指定要求的平均质量值，默认是 20；
- 如果平均质量低于阈值，则窗口中的基数将被标记为丢弃，窗口将向前移动一个基数，否则算法将停止。

5、polyG尾部修剪（polyG tail trimming）：-g

参数	解释	详解	原文解释
-g	--trim_poly_g	开启该功能，可以指定-g 参数	force polyG tail trimming, by default trimming is automatically enabled for Illumina NextSeq/NovaSeq data
-G	--disable_trim_poly_g	如果想强制关闭该功能，则可以指定-G 参数	disable polyG tail trimming, by default trimming is automatically enabled for Illumina NextSeq/NovaSeq data

原因：PolyG是Illumina NextSeq和NovaSeq系列中常见的问题，因为它们基于双色化学。这样的系统使用两个不同的灯（即红色和绿色）来表示四个碱基，一个仅检测到红光信号的碱基将被称为C，一个仅检测到绿光信号的碱基将被称为T，检测到红色和绿色光，它将被称为A，对于没有检测到光的基极，它将被称为G。但是，由于合成测序（SBS）进入后续周期，每个周期的信号强度DNA簇变得更弱。这个问题会导致一些T和C在读取尾部被错误地解释为G，这个问题叫做polyG tail。



（a）在fastp预处理之前，和（b）在fastp预处理之后

6、全局剪裁：-f、-t、-F、-T

参数	解释	详解	原文解释
-f	--trim_front1	默认启用read1的头部的剪裁/裁剪前面有多少个基数为Read1，不履行为0（int = 0）	trimming how many bases in front for read1, default is 0 (int [=0])
-t	--trim_tail1	默认启用read1 的尾部的剪裁（ trim_tail1 (--trim_tail1) should be 0 ~ 30, suggest 0 ~ 4 ）	trimming how many bases in tail for read1, default is 0 (int [=0])
-F	--trim_front2	默认启用read2的头部的剪裁	trimming how many bases in front for read2. If it's not specified, it will follow read1's settings (int [=0])
-T	--trim_tail2	默认启用read2 的尾部的剪裁	trimming how many bases in tail for read2. If it's not specified, it will follow read1's settings (int [=0])

7、质量值：-Q、-q、-u、-n、-L、-l

参数	解释	详解	原文解释
-Q	--disable_quality_filtering	默认启用质量筛选。如果指定此选项，则禁用质量筛选	quality filtering is enabled by default. If this option is specified, quality filtering is disabled
-q	--qualified_quality_phred	参数来指定合格的 phred 质量值，比如-q 15 表示质量值大于等于Q15 的即为合格	the quality value that a base is qualified. Default 15 means phred quality >=Q15 is qualified. (int [=15])
-u	--unqualified_percent_limit	指定最多可以有多少百分比的质量不合格碱基，比如-q 15 -u 40 表示一个 read 最多只能有 40%的碱基的质量值低于Q15，否则会被扔掉	how many percents of bases are allowed to be unqualified (0~100). Default 40 means 40% (int [=40])
-n	--n_base_limit	使用-n 可以限定一个 read 中最多能有多少个 N	if one read's number of N base is >n_base_limit, then this read/pair is discarded. Default is 5 (int [=5])
-L	--disable_length_filtering	还默认启用了 read 长度过滤，但也可以使用-L 参数关闭	length filtering is enabled by default. If this option is specified, length filtering is disabled
-l	--length_required	使用-l 参数指定最低要求一个 read 有多长，比如-l 30 表示低于 30 个碱基的 read 会被扔掉	reads shorter than length_required will be discarded, default is 15. (int [=15])

7、质量值：-Q、-q、-u、-n、-L、-l

小贴士：

phred ：测序仪器下机数据的 fastq 文件中, 每条序列都对应了相同长度的质量值, 反映出每个碱基的准确性和可靠性, (现在主流用的) 计算公式为:Q = -10log10p而这个 p 值就是 Phred 计算出来的, 表示一个碱基被识别错误的可能性, Phred 一开始是一个软件 (或者说计算方法), 对测序仪器识别到的荧光强度 进行评估。

8、分子标识符预处理（UMI preprocessing）：-U

参数	解释	详解	原文解释
-U	--umi	超低频突变检测应用，该功能默认没有启用，需要使用-U 参数开启	enable unique molecular identifer (UMI) preprocessing

- 独特的分子标识符（UMI）技术，用于降低背景噪声并提高在深度测序应用（即ctDNA测序）中**检测超低频突变的灵敏度**。 UMI方法可用于消除重复并生成高质量的共识读取。

9、输出分裂（Output splitting）：-s、-S、-d

参数	解释	详解	原文解释
-s	--split	通过使用此选项(2~999)限制总的拆分文件号，将顺序号前缀添加到输出名称(**0001.out.fq，0002.out.fq.**)，默认禁用(int[=0])	split output by limiting total split file number with this option (2~999), a sequential number prefix will be added to output name (0001.out.fq, 0002.out.fq...), disabled by default (int [=0])
-S	--split_by_lines	通过使用此选项限制每个文件的行数**(>=1000)**，输出名称将添加一个顺序号前缀(0001.out.fq，0002.out.fq.)，默认禁用(long[=0])	split output by limiting lines of each file with this option(>=1000), a sequential number prefix will be added to output name (0001.out.fq, 0002.out.fq...), disabled by default (long [=0])
-d	--split_prefix_digits	序列号填充的数字(1~10)，默认值为4，因此文件名将填充为0001.xxx，0以禁用填充(int[=4])	the digits for the sequential number padding (1~10), default is 4, so the filename will be padded as 0001.xxx, 0 to disable padding (int [=4])

NGS数据的并行处理已成为一种新趋势；两种拆分模式：按文件行拆分和按文件编号拆分，后者更复杂，因为fastp必须评估输入文件的总行数，这对于GZIP压缩数据尤其重要。

10、过渡表达序列分析：-p

参数	解释	详解	原文解释
-p	--overrepresentation_analysis	过表达分析程度	enable overrepresented sequence analysis.

- 在FASTQ数据中，某些序列甚至整个读数可能过多。对这些过量表达序列的分析可以概述一些测序假象，如重复PCR，polyG尾和接头污染。
- fastp对所有读数进行均匀读数以评估过度表达的序列以消除部分分布偏差。

Before filtering: read2: overrepresented sequences

Sampling rate: 1 / 20

overrepresented sequence	count (% of bases)	distribution: cycle 1 ~ cycle 150
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA AAAA	201 (0.006381%)	
AGATCGGAAGAGCGTCGTGT	25370 (0.161079%)	
ATCGGAAGAGCGTCGTGTAG	37 (0.000235%)	
GATCGGAAGAGCGTCGTGTA	30 (0.000190%)	
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG GGGG	1485 (0.047143%)	
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG GGGG	1131 (0.053139%)	

11、增加复杂度：-y

参数	解释	详解	原文解释
-y	--low_complexity_filter	增加碱基复杂度、使AGCT碱基平衡	enable low complexity filter. The complexity is defined as the percentage of base that is different from its next base (base[i] != base[i+1]).

四、全外显子测序质控代码

```
/home/lcbio/Softwares/fastp/fastp
-i Input/BS06629/BS06629_AH3YWKDSXX_S47_L004_R1_001.fastq.gz
-I Input/BS06629/BS06629_AH3YWKDSXX_S47_L004_R2_001.fastq.gz
-o CleanData/BS06629/BS06629_R1.fq.gz
-O CleanData/BS06629/BS06629_R2.fq.gz
-z 1
-t 0
-T 0
-j CleanData/BS06629/fastp_BS06629.json
-h CleanData/BS06629/fastp_BS06629.html
-R BS06629fastpreport
-M 20
-3
-y
-w 16
-p
-c
```

四、全外显子测序质控代码

参数	解释	
-i	Input/BS06629/BS06629_AH3YWKDSXX_S47_L004_R1_001.fastq.gz	原始数据R1
-I	Input/BS06629/BS06629_AH3YWKDSXX_S47_L004_R2_001.fastq.gz	原始数据R2
-o	CleanData/BS06629/BS06629_R1.fq.gz	输出CleanDataR1
-O	CleanData/BS06629/BS06629_R2.fq.gz	输出CleanDataR2
-z	1	压缩比例为1（最快）
-t	0	不需要启动R1的尾部的剪裁
-T	0	不需要启动R2的尾部的剪裁
-j	CleanData/BS06629/fastp_BS06629.json	--json
-h	CleanData/BS06629/fastp_BS06629.html	--html
-R	BS06629	--report_title
-M	20	指定要求的平均质量值，默认是 20，也就是 Q20
-y		启用增加复杂度、使AGCT碱基平衡
-W	16	指定滑动窗大小为16
-p		启用过度表达序列分析
-c		该校正功能默认没有开启使用-c 参数可以启用，对于一些对噪声容忍度低的应用，比如液体活检，建议开启

Read1 before filtering:	Read1 after filtering:
total reads: 42000000	total reads: 41342364
total bases: 6300000000	total bases: 6164607961
Q20 bases: 6174720971(98.0114%)	Q20 bases: 6061591363(98.3289%)
Q30 bases: 5970327989(94.7671%)	Q30 bases: 5867574749(95.1816%)
Read2 before filtering:	Read2 aftering filtering:
total reads: 42000000	total reads: 41342364
total bases: 6300000000	total bases: 6164612619
Q20 bases: 6129880727(97.2997%)	Q20 bases: 6039756936(97.9746%)
Q30 bases: 5874909241(93.2525%)	Q30 bases: 5800017410(94.0857%)
Filtering result:	
reads passed filter: 82684728	
reads failed due to low quality: 1174290	
reads failed due to too many N: 0	
reads failed due to too short: 109346	
reads failed due to low complexity: 31636	
reads with adapter trimmed: 5222756	
bases trimmed due to adapters: 71569842	
reads corrected by overlap analysis: 3482785	
bases corrected by overlap analysis: 4981501	
JSON report: CleanData/BS06629/fastp_BS06629.json	
HTML report: CleanData/BS06629/fastp_BS06629.html	
fastp v0.12.6, time used: **459** seconds (7min)	

Thanks for Your Attention !

Any Questions ?