**filter用户手册**

|  |  |  |
| --- | --- | --- |
| 文件状态：  [ ] 正在修改  [√] 草稿  [ ] 正式发布 | 文件标识： | BIS\_Tool\_Manual\_02 |
| 当前版本： | 4.0 |
| 作 者： | 汪亮 |
| 完成日期： | 2012-06-25 |

修订历史

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 版本 | 编写/修订说明 | 修订人 | 修订日期 | 备注 |
| 1 | 1.0 | 全部 | 汪亮 | 2012.06.20 |  |
| 2 | 2.0 | 3.3、4.2、4.4、6 | 汪亮 | 2012.06.25 |  |
| 3 | 2.0 | 6 | 汪亮 | 2012.06.28 |  |
| 4 | 4.0 | 2、3.3 | 汪亮 | 2012.07.05 |  |

***注：****本文档为产品设计人员编写。主要面向软件使用者*

|  |  |
| --- | --- |
| *内容* | *字体* |
| ***一级标题*** | 三号，宋体，加粗 |
| ***二级标题*** | 四号，宋体，加粗 |
| ***三级标题*** | 小四，宋体，加粗 |
| ***正文*** | 小四，宋体 |
| ***表头*** | 五号，加粗，居中，表正上方 |
| ***表*** | 五号，宋体，居中 |
| ***网址*** | 五号 |

目录

[1. 软件名称 4](#_Toc328401356)

[2. 功能描述 4](#_Toc328401357)

[3. 特点 4](#_Toc328401358)

[3.1 适用范围 4](#_Toc328401359)

[3.2 优点 4](#_Toc328401360)

[3.3 局限性 4](#_Toc328401361)

[4. 用法 5](#_Toc328401362)

[4.1使用命令 5](#_Toc328401363)

[4.2参数 5](#_Toc328401364)

[4.4输入 7](#_Toc328401365)

[4.5输出 7](#_Toc328401366)

[6. 原理 9](#_Toc328401367)

[7. 开发人员 10](#_Toc328401368)

# 1. 软件名称

全称：FASTQ数据预处理工具

简称：filter

# 2. 功能描述

filter是工具包SOAPnuke中的工具之一，它对测序仪的下机原始数据FASTQ文件（小RNA数据除外）进行预处理。包括过滤低质量reads，过滤adapter，屏蔽index，截取数据等等。

# 3. 特点

## 3.1 适用范围

1. 第二代测序仪HiSeq-2000 FASTQ下机原始数据
2. 测序类型：DNA Pair-End(PE)测序；RNA Single-End(SE)测序

## 3.2 优点

1. 涵盖数据类型广：包括RNA-seq、RNA-ref、BS、MeDIP、CHIP、RNA denovo以及DNA测序下机产生的FASTQ原始数据（小RNA数据除外）。
2. 用C++语言开发，考虑现状，利用CPU，减少IO，采用多线程调用，分块读取，对IO和运行时间等大大优化；
3. 模块化设计，各项功能利用函数封装，便于拓展升级。

## 3.3 局限性

1. 支持的最大读长为256；适用于Hiseq-2000下机数据等。
2. 去除duplication的算法中，不能区分N和G，即会把N和G判断为相同的碱基，是由压缩算法导致（这一点是嵌入原生产流程代码）

# 4. 用法

## 4.1使用命令

SOAPnuke filter [options]

其中，filter为命令，options为参数选项（见4.2参数说明）

例如：

SOAPnuke filter –f adapter1.list -r adapter2.list -1 reads1.fq.gz -2 reads2.fq.gz –o ./

SOAPnuke filter –f AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC –r AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTA -1 reads1.fq -2 reads2.fq –o ./

## 4.2参数

必选参数

-f, --adapter1 : <s> adapter sequence or adapter list file

-r, --adapter2 : <s> adapter sequence or adapter list file[for PE reads]

-1, --fq1 : <s> fq1 file

-2, --fq2 : <s> fq2 file [for PE reads]

可选参数

The next two options only for adapter sequence:

-M, --misMatch : <i> the max mismatch number when match the adapter (default: [1])

-A, --matchRatio: <f> adapter's shortest match ratio(default: [0.5])

-l, --lowQual : <i> low quality threshold (default: [5])

-q, --qualRate : <f> low quality rate (default: [0.5])

-n, --nRate : <f> N rate threshold (default: [0.05])

-p, --polyA : <f> filter poly A, percent of A, 0 means do not filter (default: [ 0 ])

-d, --rmdup : <b> remove PCR duplications

-i, --index : <b> remove index

-c, --cut : <f> reserve read number in each fq file (unit:M, 0 means not cut reads)

-t, --trim : <s> trim some bp of the read's head and tail, they means: read1's head and tail and read2's head and tail(default: [ 0,0,0,0 ])

-S, --small : <b> filter the small insert size

the next two options only for filter the small insert size

-O, --overlap : <i> minimun match length (default: [ 10 ])

-P, --mis : <f> the maximum miss match ratio (default: [ 0.1 ])

-Q, --qualSys : <i> quality system 1:illumina, 2:sanger (default: [ 1 ])

-L, --read1Len : <i> read1 max length (default: all read1's length are equal, and auto acquire)

-I, --read2Len : <i> read2 max length (default: all read2's length are equal, and auto acquire)

-G, --sanger : <b> set clean data quality system to sanger (default: illumina)

-a, --append : <s> the log's output place : console or file (default: [console])

-o, --outDir : <s> output directory, directory must exists (default: current directory)

-h, --help : <b> help

-v, --version : <b> show version

## 4.4输入

1. 测序下机FASTQ格式数据（PE为一对文件，SE为一个FASTQ文件）；

2. 测序所用接头（adapter）序列或者adapter文件。

adapter序列例如：

AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC

adapter文件格式如下：

#reads\_id reads\_len reads\_start reads\_end adapter\_id adapter\_len adapter\_start adapter\_end align\_len mismatch

FCD0CP1ABXX:7:1101:17857:2175#CGATGTAT/1 90 0 33 iPE-3+ 34 0 33 34 4

FCD0CP1ABXX:7:1101:18889:2548#CGATGTAT/1 90 79 89 iPE-3+ 34 0 10 11 1

FCD0CP1ABXX:7:1101:19063:2746#CGATGTAT/1 90 79 89 iPE-3+ 34 0 10 11 2

## 4.5输出

**1. Clean FASTQ文件**

过滤后的FASTQ文件，与原始数据文件格式相同，在小RNA中，一般不用这种方式输出结果。

@FCD0L72ACXX：5：1101：1104：2164#AAGTCTCT/1

CAGTGGATGGTGGAGAGTTGGGTGCACAAGTGTTTAGGGAGAGTTTGATCCTAAC

+

a\_acceccgg`egbfbffghiidfefhiiicfcggafhiafcfcffhegaefghi

每个以@开头的四行数据，为一条read，PE测序中，read必须是成对出现：

1. 第一行：序列名称
2. 第二行：序列
3. 第三行：序列名称(+号，无需关注)
4. 第四行：质量值列

**2. 统计文件**

1. Basic\_Statistics\_of\_Sequencing\_Quality.txt

基本信息统计，包括对原始fq和输出clean文件的统计。

1. Statistics\_of\_Filtered\_Reads.txt

被过滤掉的reads，按过滤的类型分类统计。

1. Base\_distributions\_by\_read\_position.txt

对原始FASTQ文件进行的统计，对每一个位点上的ATGCN的含量统计。

1. Base\_quality\_value\_distribution\_by\_read\_position.txt

对原始FASTQ文件进行的统计，统计每一个位点上的质量值分布情况，Q2~Q41碱基数，以及平均值，中位数等。

1. Distribution\_of\_Q20\_Q30\_bases\_by\_read\_position.txt

对原始FASTQ文件进行的统计，统计每一个位点上质量值≥Q20或≥Q30的碱基含量。

输出内容：3，4，5项也统计一下过滤后的数据

# 6. 原理

* 1. 去除含有adapter的reads（adapter和read最小需要匹配的碱基个数和adapter长度成比例）
  2. 去除质量值小于10（默认值，可更改）的碱基占总碱基数50%以上的reads
  3. 去除N的比例大于5%的reads
  4. 去除poly A.
  5. 去除序列ID中index 序列
  6. 截取指定数据量(比如1M条reads) clean data的碱基量
  7. 去除序列相同的reads，针对denovo数据而言
  8. 去除短插入片段 。针对PE测序reads

若read1和read2的overlap >=10bp, mismatch <=10%，则认为是段插入片段。此项对于DNA denovo 数据默认不做

* 1. 若输出结果为FASTQ格式时，将质量值转换为sanger体系的质量值.
  2. 输出Clean data和Raw data

# 7. 开发人员

生物信息核心技术与云计算实验室 标准化部（BIS）

陈永胜 [chenyongsheng@genomics.cn](mailto:chenyongsheng@genomics.cn)

陈浩森 chenhaosen@genomics.cn