**filtersRNA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\_01 |
| 当前版本： | 2.0 |
| 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 | 3.0 | 6 | 汪亮 | 2012.07.02 |  |
| 4 | 4.0 | 3.3、4 | 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*

*FiltersRNA为一个预处理工具包，本文档只介绍其中过滤部分；其他用法（比如pnggetter）详见其他用户手册。*

|  |  |
| --- | --- |
| *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：小RNA FASTQ数据预处理工具

简称：filtersRNA

# 2. 功能描述

filtersRNA是工具包SOAPnuke中的工具之一，对测序仪的下机原始数据small RNA的FASTQ文件进行预处理。包括过滤低质量reads，过滤adapter，屏蔽index，Sanger质量体系转换等等。

# 3. 特点

## 3.1 适用范围

1. 第二代测序仪HiSeq-2000 下机的小RNA原始数据
2. 测序类型：SE50
3. 测序数据样本类型：小RNA（small RNA）

## 3.2 优点

1. 涵盖数据类型广：包括small RNA，RNA-seq、RNA-ref、BS、MeDIP、CHIP、RNA denovo以及DNA测序下机产生的FASTQ原始数据；
2. 用C++语言开发，多线程调用，分块读取，对内存IO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. 只支持FASTQ格式的输入数据。

# 4. 用法

## 4.1使用命令

SOAPnuke filtersRNA [options]

其中，filtersRNA为命令，options为参数选项（见4.2,4.3）。

例如：

SOAPnuke filtersRNA –f sample.fq.gz

SOAPnuke filtersRNA –f sample.fq.gz -3 TCGTATGCCGTCTTCTGCTTG -5 GTTCAGAGTTCTACAGTCCGACGATC -o ./

## 4.2参数

必选参数：

-f, --fq :<s> fastq file

可选参数：

-3,--adapter3 <string>:3' adaptor sequence (default: TCGTATGCCGTCTTCTGCTTG)

-5,--adapter5 <string>:5' adaptor sequence (default: GTTCAGAGTTCTACAGTCCGACGATC)

-o, --outDir <string> : out directory (default: current directory)

-x, --outPfx <string> : out file prefix (default: clean)

-s, --strict <switch> : filter low quality reads strictly (default: off)

-z, --minSize <int> : small insert size (default: 18)

-p, --polyA <float> : filter poly A, percent of A, 0 means do not filter, (default: 0.7)

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

-q, --fastq <switch> : out file type: on:fastq, off:fasta (default: off)

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

-G, --sanger <switch> : output sanger quality score system fq. (defaul: off illumina)

-u, --untrim <switch> : do not trim 3' adapter (default: off)

-w, --unlowQ <switch> : do not filter low quality reads (default: off)

-L, --readLen <int> : Max read length in fq file (default: 49)

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

-c, --cut <int> : reserve read number in each fq file (K reads(1024 reads), 0 means not cut reads)

-a, --append <string> : logger's appender: console or file (defualt: console)

-h, --help <switch> : help

-v, --version <switch> : version information

-C, --continuous <int> : mini 5' adapter continuous alignment length (default: 6)

-A, --alignRate <float> : mini alignment rate when find 5' adapter: alignment/tag (default: 0.8)

-l, --miniAlign <int> : mini alignment length when find 3' adapter (default: 5)

-E, --errorRate <float> : Max error rate when find 3' adapter (mismatch/match) (dfault: 0.4)

-M, --misMatch <int> : Max mismatch number when find 3' adapter (default: 4)

## 4.4输入

1. 测序下机FASTQ格式数据（\*.fq或\*.fq.gz）；

FASTQ格式例如：

@FCC047NABXX:2:1101:1205:2150#AAGTCGCN/1

TTGNTCGGTGCTGACGCAATGTGATTTCATCTCGTATGCCGTCTTCTGC

+

bbaBbdddcdeeeaeeeaeeeeeaeRddcdeeeeeeeeeeedeeea]bb

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，不用关注)
4. 第四行：质量值列

2. 测序所用接头（adapter）序列。

adapter序列例如：

AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC

## 4.5输出

**1. Clean FASTA文件**

过滤后的FASTA文件，常用于小RNA测序。格式如下：

>seqName 953571

ACGTACGTACGTAAAACGGT

1. 第一行：“>序列名称 序列数量”
2. 第二行：reads的碱基序列

**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。

1. Length\_distribution.txt

Clean文件中，对tag的长度分布的统计。

# 6. 原理

1. 检测并去除低质量read;

满足以下条件之一即判断为低质量reads：

1. 整条read中，N碱基的个数>2；
2. 第1-30号碱基中，N碱基的个数>0；
3. 第1-30号碱基中，质量值q<10的个数>2或者质量值q<13的个数>3（默认条件）；第1-30号碱基中质量值q<10的个数>4，或者质量值q<13的个数>6（放宽条件）。
4. 检测并去除3’接头缺失的reads；

将3’接头与read中的reads序列进行比对。比对过程中N不考虑，当接头处于某一位置时，接头与序列重叠部分错配数<=4个(默认值)，以及错配数/正确比对数<0.4(默认值)，此时认为是找到3’接头。

1. 检测并去除插入序列空载的reads

对于有3’接头reads，去除3’接头后，如果tag的长度≤2，则认为插入序列空载，去除该reads。

1. 检测并去除5’接头污染的reads。

将5’接头与read中的reads序列进行比对，连续正确比对数>＝6，且tag中 80%的base可以比对到5’接头上，则认为该read被5’接头污染。去除该read。

1. 检测并去除polyA 的reads。
2. 检测并去除小片段reads;

若read长度小于18nt，认为是小片段，被去除

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.

# 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)

蒋帅 jiangshuai@genomics.cn