

fastx-utils之质量控制：filter

一、fastx-utils filter 介绍

功能描述：

`fastx-utils filter` 对fastq文件进行质量控制, 过滤不能满足条件的序列.

命令行接口：

```
1 $ fastx-utils filter
2
3 Usage: fastx-utils filter [options] <in1.fq> <in2.fq> <prefix>
```

```
1 Options:
2   -q INT      < Qvalue for filter. [5].
3   -e FLOAT    Discard reads with bases qvalue low than -q >%<FLOAT>. [0.5].
4   -n FLOAT    Discard reads with N >%<FLOAT> [0.1].
```

可选参数：

| | | |
|---|-------|---------------------------|
| 1 | -q 整形 | 设定Q值阈值，默认5. |
| 2 | -e 浮点 | 去掉小于Q值的碱基占比>0.5的序列，默认 50% |
| 3 | -n 浮点 | 去掉N碱基占比超过一定比例的序列，默认 10%. |

二、使用场景实例及其用法

示例演示：

示例文件： `A1.1.fastq`, `A1.2.fastq`

```
1 $ cat A1.1.fastq | head -n 6
```

```

1 @HISEQ:483:HLJ2LBCXY:1:2215:13265:62834 1:N:0:CACCGG
2 CCTATGGGATGCAGCAGTGGGGAATCTTGCCCAATGGGCGCAAGCCTGAGGCAGCGACGCCGCTGAGGGATGACGG
  CCTTCGGGTTGTAAACCTCGTTCAGCAGGGACGAAGCGAAAGTGACGGTACCTGCAGAAGAAGCACCGGCTAACTAC
  GTGCCAGCAGCCGCGGTAATACGTAGGGTGCAAGCGTTGTCCGAATTATTGGGCGTAAAGAGCTCGTAGGCGGCTT
  GTCGCGTCTGCT
3 +
4 IHIIIGGIIIEHIIIIHIIIIIIHIIIIIIIIIIIIHEHIIIIIIIIIIHFIIIIIHIIIIIGIIIIIIIIIIID
  HHIIIIIIIIIIH?
  HHFHHIIIIHIIIIIIHIIIIIIHIIHHIIHHIIIIIIIIIGIIIIHHHHIHHHHIIHIIIIIIHIIHHHHIHHI
  HIGIIIIHIIIIHHHHGHCAH.?GHHFD-EAFGHIHHIDHHGFHHIIGFFHEHHHHCH,,CHHIGIHHCHII
5 @HISEQ:483:HLJ2LBCXY:1:1210:15779:30764 1:N:0:CACCGG
6 CCTATGGGGGACACAGTGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCCATGCCGCGTGGGTGAAGAAGG
  CCTTAGGGTTGTAAACCCCTTTCAGCGGGGAAGATAATGACGGTACCCGCAGAAGAAGTCCCGGCTAACTCCGTGCC
  AGCAGCCGCGGTAATACGGAGGGGACTAGCGTTGTTCCGAATTACTGGGCGTAAAGCGCACGTAGGCGGATTTGTAA
  GTCAGGGGTGAA
7
8 $ cat A1.2.fastq | head -n 6

```

```

1 @HISEQ:483:HLJ2LBCXY:1:2215:13265:62834 2:N:0:CACCGG
2 GGACTACCCGGGTATCTAATCCTGTTGCTCCCCACGCTTTCGCTCCTCAGCGTCAGCAACGGCCCAGAGACCCGCC
  TTCGCCACCGGTGTTCTCCTGATATCTGCGCATTCCACCGCTACACCAGGAATTCGGTCTCCCCTACCGCGCTCT
  AGCCTGCCCGTATCCACTGCAGGCCCGGAGTTAAGCTCCGGGTTTTACAGCAGACGCGACAAGCCGCCTACGAGCT
  CTTTACGCCCACTA
3 +
4 HH?
  GHIHHIIIGIDHHIEDF@HHIIIIIGGHHIIIIIIIIIIIGIIIIIIIIIIIGIHHIIIIIIIIIIHIIHHHI
  HHFHHIHCHEIIIEF<GHFHGHIIIIHHGEFHIIIIHIIIIH@GGHEHHIIIGIIIIHHHHH@HHGHDHHHHHHI
  IHCHIIHHH?GEHHHHHHHEHIIII=HDE@6-8?E?GGIHG=@-GHHIHHH-@D,+>D:>@E++;E-6@@@C+--
  @@-6@-C:C-@@
5 @HISEQ:483:HLJ2LBCXY:1:1210:15779:30764 2:N:0:CACCGG
6 GGACTACGGGGGTATCTAATCCTGTTGCTCCCCACGCTTTCGACCTCAGCGTCAGTACCGGACCACTGAGCCGCC
  TTCGCACTGGTGTCTTCTAATATCTACGAATTCACCTCTACACTAGGAATTCACCTCACCTCTTCCGGACTCG
  AGATCCGCAGTATCAAAGGCAGTTCGAGGTTGAGCCTCGGGATTTACCCCTGACTTACAAATCCGCCTACGTGCG
  CTTTACGCCCACTA

```

运行命令：

```
1 $ fastx-utils filter -q 5 -e 0.5 -n 0.1 A1.1.fastq A1.2.fastq A1.qc
```

```
1 $ ls
```

```
1 A1.1.fastq A1.2.fastq A1.qc_R1.fastq A1.qc_R2.fastq
```

本文材料为 **BASE (Biostack Applied bioinformatic SEies)** 课程 **Linux Command Line Tools for Life Scientists** 材料，版权归 上海逻捷信息科技有限公司 所有。

Last Update: 2020-08-10 11:56 AM

