fastx-utils之去除重复序列：uniques

一、fastx-utils uniques介绍

功能描述：

`fastx-utils uniques` 根据序列去除重复。

命令行接口：

```
$ fastx-utils uniques

Usage: fastx-utils uniques [options] <in.fa/q>

Options:
  -l STR   sequence label, default: [Uniq]
  -m       print membership to stderr.
```

可选参数：

```
-l    字符串     序列标签，默认为 'Uniq'；
-m               输出标准错误；
```

二、使用场景实例及其用法

使用场景经典案例：

1. 16S扩增子数据分析： 对合并后的PE数据进行去除重复序列，并进行计数；

示例演示：

示例文件： `A1.fastq`

```
$ zcat A1.fastq.gz | head -n 8
```

```
@HISEQ:483:HLJ2LBCXY:1:1101:7924:2136
TGGGGAATATTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGAGTGATGAAGGCCCTAGGGTTGTAAAGCCCT
TTCGGCGGGGAAGATAATGACGGTACCCGCAGAAGAAGCCCCGGCTAACTTCGTGCCAGCAGCCGCGGTAATACGAAGGG
GGCTAGCGTTGCTCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGCTTTCTAAGTCGGGGGTGAACTCCTGGAGCTTAA
CTCCAGAACTGCCTTCGATACTGGAGAGCTCGAGTCCGGGAGAGGTGAGTGGAACTGCGAGTGTAGAGGTGAAACTCGTA
GATATTCGCAAGAACACCAGTGGCGAAGGCGGCTCACTGGCCCGGTACTGACGCTGAGGTGCGAAAGCGTGGGGAGCAAA
CAGG
+
<DEG/CEHHIHEHCHECFHICHHHHHDHIICHHIEHCHCEGHHIICH?EEHFHHHIECEE?HE?
1GHHHHDHHHHHFFGHHHCHHHII?H?HHEHCHF1CGHC0DEHEHHD<<FEHHHIGHIC??
HHCGHHHHIHH@FFEEHIHDHHCHHIHIHE@GH-G??EDEGEHIHIICHHC?EHH-
KKKGK=KKKKKKKKKKKKKKKKKKHE@@G@CEC6+>=>8.-C@B8-HHIHHACHCBFA.GA?
HHEHHFHHHCIHE@CGBAFGIHIHHHCEHHCC,HEHIHHGDFHFHEGCHEEHIH@HFGHGIIHHHHHECEHGIIHEHHCE
C/C/DGIHHGCC@HGD00C01GCHCCDE?C1D1GCC?DH=HFHHG@CEDEGIHIHFE=EHEHHHHCHHIHHHG?
EHEG@<1
@HISEQ:483:HLJ2LBCXY:1:2108:14400:12517
TGGGGAATATTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGGGTGATGAAGGCCCTAGGGTTGTAAAGCCGT
TTCGGCGGGGAAGATAATGACGGTACCCGCAGAAGAAGCCCCGGCTAACTTCGTGCCAGCAGCCGCGGTAATACGAAGGG
GGCTAGCGTTGCTCGGAATTACGGGGCGTAAAGCGCACGTAGGCGACTCTTTAAGTCGGGGGTGAAATCCTGGAGCTCAA
ACCCAGAACTGCCTTCGATACTGGGGAGCTCGAGTTCGGGAGAGGTGAGTGGAACTGCGAGTGTAGCGGTGAAATTCGTA
GATATTCGCAAGAACACCAGTGGCGAAGGCGGCTCACTGGCCCGATACTGACGCTGAGGTGCGAAAGCGTGGGCAGCAAA
CAGG
+
HHHIIGHIIIIIIGIIIIHHGIHIIIIIHIGIIHIGIIIIIIIIIIIIIHICHHIIIIIIIIIIHIIIHHHIIIIIHG/<
D@GDHIIIHH?HEFHHHHIEFEHEHHHHHHDHHIIIIIH?BAHHDHHHFHHIIIIIGHII@@FHHHDHHH?AHHA?
GHHDHGIIHIIHHIIHIII<HHIIHI.87<DHHID?HHKKKKKKKKKK$K<KKKKKKKKKKKKKKKKK-
4AA.B8.88.-88..8.HHAB6.EHCHGCB@@A=HHHG.?GCF.</DCEDGEH?
IIHFHG@HHHEHHHDHFHCGCE.CC0IIIIHHCGHHEHIHHHHDCHGEIHHIIHHHHIIIHECHH@HHDHH@F?
IIHHHHHHIIIHHHHIHIHHF@C<<0<IIIHIHHEIHD1HHHHHIIHHF
```

运行命令： 去除`A1.fastq`文件中的重复序列, 使用 `-l Uniq` 重新命令.

```
$ fastx-utils  uniques -l Uniq A1.fastq.gz | head -n 6
>Uniq_1
TGGGGAATATTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGAGTGATGAAGGCCCTAGGGTTGTAAAGCCCT
TTCGGCGGGGAAGATAATGACGGTACCCGCAGAAGAAGCCCCGGCTAACTTCGTGCCAGCAGCCGCGGTAATACGAAGGG
GGCTAGCGTTGCTCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGCTTTCTAAGTCGGGGGTGAACTCCTGGAGCTTAA
CTCCAGAACTGCCTTCGATACTGGAGAGCTCGAGTCCGGGAGAGGTGAGTGGAACTGCGAGTGTAGAGGTGAAACTCGTA
GATATTCGCAAGAACACCAGTGGCGAAGGCGGCTCACTGGCCCGGTACTGACGCTGAGGTGCGAAAGCGTGGGGAGCAAA
CAGG
>Uniq_2
TGGGGAATATTGGACAATGGGCGCAAGCCTGATCCAGCCATGCCGCGTGGGTGATGAAGGCCCTAGGGTTGTAAAGCCGT
TTCGGCGGGGAAGATAATGACGGTACCCGCAGAAGAAGCCCCGGCTAACTTCGTGCCAGCAGCCGCGGTAATACGAAGGG
GGCTAGCGTTGCTCGGAATTACGGGGCGTAAAGCGCACGTAGGCGACTCTTTAAGTCGGGGGTGAAATCCTGGAGCTCAA
ACCCAGAACTGCCTTCGATACTGGGGAGCTCGAGTTCGGGAGAGGTGAGTGGAACTGCGAGTGTAGCGGTGAAATTCGTA
GATATTCGCAAGAACACCAGTGGCGAAGGCGGCTCACTGGCCCGATACTGACGCTGAGGTGCGAAAGCGTGGGCAGCAAA
CAGG
>Uniq_3
TGGGGAATATTGGACAATGGGCGAAAGCCTGATCCAGCCATGCCGCGTGGATGAAGGAGGCCCTAGGGTTGTAAAGTCCT
TTCGATGGTGAAGATAATGACGGTAACCATACAAGAAGCCCCGGCTAATTTCGTGCCAGCAGCCGCGGTAATACGAAAGG
GGCTAGCGTTGTTCGGAATTACTGGGCGTAAAGCGCACGTAGGCGGCCGTGTAAGTTGGGGGTGAGATCCCGGGGCTCAA
CCCCGGAACTGCCTCCAAAACTACATAGCTAGAGGATGTGAGAGGACAGTGGAATTCCGAGTGTAGAGGTGAAATTCGTA
GATATTCGGAAGAACACCGGTGGCGAAGGCGACTGTCTGGCACATTTCTGACGCTGAGGTGCAAAAGCGTGGGGAGCAAA
CAGG
```

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