tsv-utils之指定的文件列进行注释: subset

一、tsv-utils subset介绍

功能描述:

tsv-utils subset 根据列表中的元素,对文件中的指定列进行操作,取交集或者补集。

命令行接口:

```
1  $ tsv-utils subset
2
3  Usage: tsv-utils subset [options] <tsv> <list>
4
5  Options:
6  -c INT target col for select, default: 1
7  -r subset option: -r for supplementary set, default: intersection.
8  -k query using key id.
```

可选参数:

```
1 -c 整数 指定目标列,默认为第一列;
2 -r 设定该参数则为选取补集,默认为取交集;
3 -k 设定使用命令行参数字符串选取列;
```

二、使用场景实例及其用法

经典使用案例

1. 16S扩增子数据分析: 去除线粒体/叶绿体污染以及其它污染序列后获取的序列标识符, 从原始 zotu 表抽取无污染的 zotu, 并生成新的 zotu 表。

示例演示

示例文件: zotu_table.txt, zotu_identifiers.txt

```
1 | $ cat zotu_identifiers.txt | head -n6
2 ZOTU_19
3 ZOTU_5
4 ZOTU_9
5 ZOTU_12
6 ZOTU_7
7
  ZOTU_18
8
9 $ cat zotu_table.txt | head -n6
10 #OTU ID A-1 A-2 B-1
                           B-2
                                  C-1
                                        C-2
                                  1829
11 ZOTU_1 0
                    87
                           278
                                         3608
12 ZOTU_2 223 447 1268 1583
                                  52
                                        69
              0
                    162
                                  1116 2021
13
   ZOTU_3 0
                           159
14
   ZOTU_4 0
              0
                      99
                            50
                                  1250
                                        2172
15 ZOTU_5 0
               0
                      1
                                  1216
                                         2143
```

示例1: 根据ZOTU列表文件生成子 ZOTU 表

```
1 $ tsv-utils subset -c 1 zotu_table.txt zotu_identifiers.txt | head -n6
2 #OTU ID A-1 A-2
                       B-2 C-1
                                  C-2
                 B-1
3 ZOTU_1 0
            0
                  87
                        278
                             1829
                                   3608
4 ZOTU_2 223
                                  69
            447
                 1268
                       1583 52
                 162
5 ZOTU_3 0
                       159 1116 2021
            0
6 ZOTU_4 0
            0
                 99
                       50
                            1250 2172
7 ZOTU_5 0
            0
                 1
                        8
                             1216 2143
```

示例2:根据ZOTU字符串抽取 ZOTU 丰度表,使用-k 参数通过字符串传递列表;

```
1 $ tsv-utils subset -c 1 -k zotu_table.txt ZOTU_1,ZOTU_2
2 #OTU ID A-1 A-2
                  B-1
                        B-2 C-1
                                    C-2
3 ZOTU_1 0
             0
                   87
                               1829
                                     3608
                         278
4 ZOTU_2 223
                   1268
                         1583
                               52
                                    69
             447
```

示例3:根据ZOTU字符串不包含列表标识符的 ZOTU 丰度信息,使用 -r 求补集;

1	<pre>\$ tsv-utils n6</pre>	subset -c	1 -r	zotu_t	able.txt	zotu_i	dentifiers.txt head -
2	#OTU ID A-1	A-2	B-1	B-2	C-1	C-2	
3	ZOTU_608	29	38	0	0	0	0
4	ZOTU_629	0	0	18	8	0	0
5	ZOTU_654	0	0	0	0	5	17
6	ZOTU_656	22	5	0	0	0	0
7	ZOTU_674	10	21	0	0	0	0

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: Friday, August 28, 2020