taxon-utils之将物种分类归并到指定更高层分类等级：bin

一、atlas-utils bin介绍

功能描述：

`atlas-utils bin` 可以将`NCBI`物种分类等级归并到指定更高等级, 比如都归并到`phylum`水平.

命令行接口：

```
$ taxon-utils  bin

Usage: taxon-utils bin [options] <taxon.map> <taxon>

Options:
  -l STR   taxon levels [phylum|order|class|family|order|genus|species>]
default: [species]
```

可选参数：

```
  -l  字符串        指定物种分类水平，默认为种水平
```

二、使用场景实例及其用法

使用场景经典案例：

宏基因组数据分析: 将所有的`reads`都`bing`到指定分类等级。

示例演示：

示例文件在: `data` 目录

示例文件： `classify.txt.gz`, `taxon.map.gz`

```
$ zcat classify.txt.gz | head -n  6
```

```
C       A01050:204:HF7FGDSXY:4:1101:12943:1016  25457   150|149 0:22 25457:4 0:7 25458:5 0:78 |:| 3:4 3343:5 8109:2 25457:5 25460:1 25457:6 25693:5 0:48 25458:2 0:1 3343:2 14694:7 3343:10 3:5 0:12
C       A01050:204:HF7FGDSXY:4:1101:12337:1031  18331   150|149 0:60 31294:5 0:25 18331:3 0:23 |:| 0:93 18331:3 0:19
C       A01050:204:HF7FGDSXY:4:1101:16866:1047  8364    150|150 0:17 20367:2 35825:1 0:8 3:4 20616:5 3:1 4523:5 3:3 0:9 3:5 0:19 27173:2 0:12 27173:2 0:21 |:| 0:2 2878:3 24660:5 28576:3 30576:2 3:13 0:9 2483:17 27170:3 2483:2 27170:5 3:6 8364:3 3:7 8364:3 27172:2 8364:5 27172:6 171:5 0:15
C       A01050:204:HF7FGDSXY:4:1101:22742:1047  23158   150|150 0:82 23158:4 0:30 |:| 0:116
C       A01050:204:HF7FGDSXY:4:1101:28664:1063  559     150|150 0:22 559:3 0:7 171:5 0:79 |:| 0:116
C       A01050:204:HF7FGDSXY:4:1101:20473:1094  25986   150|150 0:45 25986:3 0:68 |:| 0:116
```

注意事项: `classify` 文件为 `Kraken2` 分类的结果, 第一列为标识符: `C`: 可以分类的序列, `U`: 不能分类的序列, 第三列为分类的 `Taxonomy ID`, 可以为 `NCBI` 分类号或者使用 `GTDB` 的自定义分类号, 正常我们只需要第二列和第三列。

```
$ zcat taxon.map.gz | head -n 6
```

```
1       1       no rank root    root
2       1       superkingdom    Archaea root
3       1       superkingdom    Bacteria        root
4       3       phylum  4572-55 Bacteria
5       3       phylum  AABM5-125-24    Bacteria
6       3       phylum  AB1-6   Bacteria
```

示例文件：

将所有的分类 `bin` 到 `class` 分类水平， 输入文件要求两列， 第一列为标识符， 第二列为分类编号；

```
$ zcat classify.txt.gz | grep -P "^C" | cut -f2,3 | head -n 6
```

```
A01050:204:HF7FGDSXY:4:1101:12943:1016  25457
A01050:204:HF7FGDSXY:4:1101:12337:1031  18331
A01050:204:HF7FGDSXY:4:1101:16866:1047  8364
A01050:204:HF7FGDSXY:4:1101:22742:1047  23158
A01050:204:HF7FGDSXY:4:1101:28664:1063  559
A01050:204:HF7FGDSXY:4:1101:20473:1094  25986
```

```
$ zcat classify.txt.gz | grep -P "^C" | cut -f2,3 | head -n 6 | taxon-utils bin  -l  class  taxon.map.gz -
```

```
A01050:204:HF7FGDSXY:4:1101:12943:1016  265     Eremiobacteria
A01050:204:HF7FGDSXY:4:1101:12337:1031  283     Gemmatimonadetes
A01050:204:HF7FGDSXY:4:1101:16866:1047  171     Actinobacteria
A01050:204:HF7FGDSXY:4:1101:22742:1047  175     Alphaproteobacteria
A01050:204:HF7FGDSXY:4:1101:28664:1063  171     Actinobacteria
A01050:204:HF7FGDSXY:4:1101:20473:1094  171     Actinobacteria
```

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