和趣求真 乘实生信

ZhangLab-小课堂 第二期

周润

2022.04.08

数据处理中常见问题

1. 基因组微调时需对局部片段反转
2. 根据ID提取序列
3. 合并多个文件中相同ID的行（sort　join　common）
4. 提取文件特定的列，不受顺序影响（cut　awk）
5. 文件的列层次不齐，查找信息容易看错列

gene_id	TT_1	TT_2	TT_4	U9Y_1	U9Y_2	U9Y_3	U9Y_4	U9G_1	U9G_2	U9G_3	U9G_4	t1
MtU9_01T0130300.1_POD	0.377334		0.0		0.785681		0.0	0.0	1.006567			1.
MtU9_01T0189600.1_4CL	41.939201		29.870426		38.588322		85.741333		31.457666			
MtU9_01T0218900.1_LAC	0.0	0.0	0.482831		0.30263	2.59863	0.0	0.0	0.925288			
MtU9_01T0300700.1_F5H	0.12684	0.241333		0.564839			0.256943		0.0	0.0	0.	
MtU9_01T0341500.1_F5H	0.655076		0.0	0.0	2.12442	3.838678		5.30347	6.101371			
MtU9_01T0351500.1_PAL	61.998138		72.147202		65.299919		61.153698		79.179634			
MtU9_01T0357500.1_4CL	95.782249		60.887871		74.345497		39.336777		45.55975			
MtU9_01T0360200.1_POD	119.083633		68.66993		122.353622		32.89946		74.758736			
MtU9_01T0365900.1_LAC	0.621534		0.095871		0.374017		0.353872		0.45476	1.		
MtU9_02T0052900.1_PAL	3.117974		3.094423		2.043212		1.237714		2.19183	6.		

linux文本处理：awk，sed，grep, cut，不适用对csv格式操作

文本处理的时候为了一个小操作写Python/R脚本有点小题大作，且难以复用

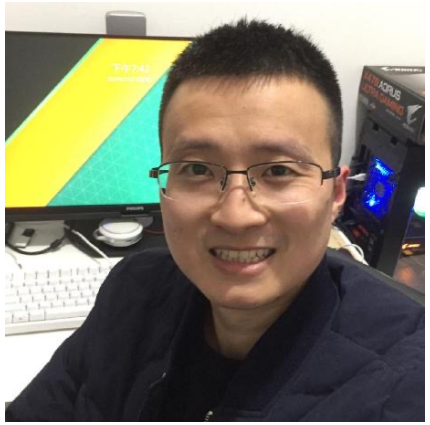
熟用小工具可以为您节省大量（不）编写 Python/R 脚本的时间。

和趣求真 乘实生信

Linux命令行超实用小工具

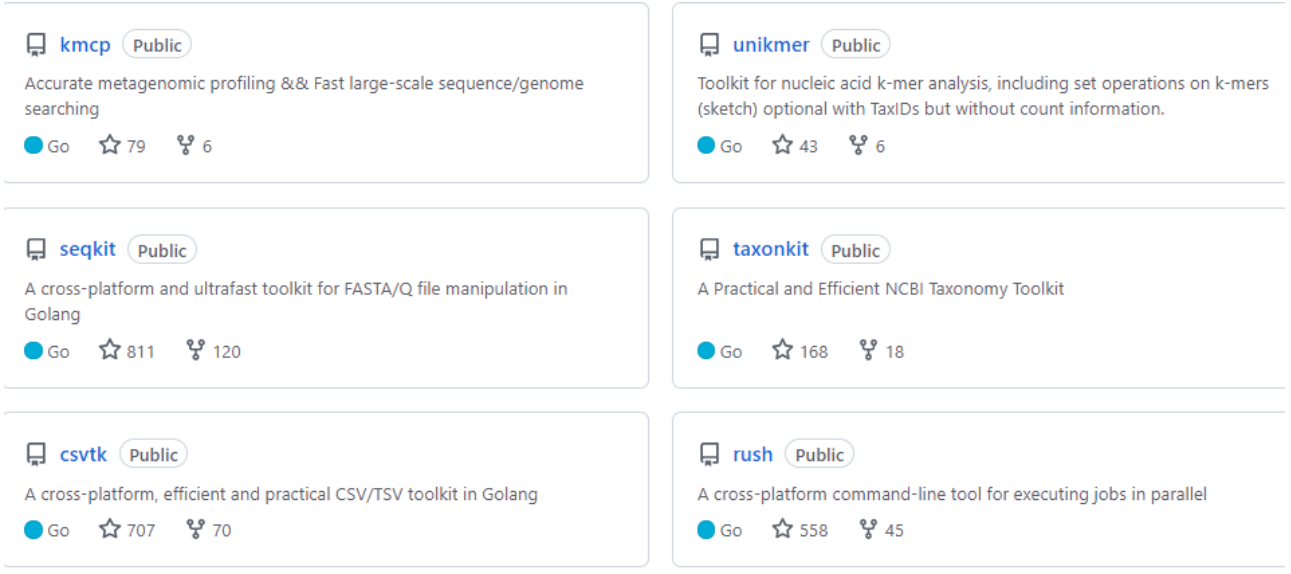
Seqkit：fasta/fq文件处理万能工具 （2016）

Csvtk:文本处理万能工具



Wei Shen 沈伟

博士后
重庆医科大学第二附属医院病毒性肝炎研究所
https://github.com/shenwei356



https://zhang.hzau.edu.cn

Seqkit 能干什么呢

1 seq 对序列取反向，互补,大小写 （适用于基因组局部微调）
2 subseq 根据gff或者bed文件提取特定区域的序列
3 Stats 可对多个fastq，fa的序列进行统计（常用于pacbio，RNA-seq，illumina的原始数据的统计）
4 sample 模拟数据集的时候随机抽取序列
5 grep 根据基因ID提取序列
6 rmdup 去除ID或序列来去除重复的序列
7 replace 可根据正则对name/seq进行修改
8 格式转换 fq2fa fx2tab

-
-
-

csvtk 的强大之处

1. Join 合并多个文件中相同ID的行
2. Cut 选择特定列
3. Pretty 增强文件可读性，更美观
4. 可以跟sed， awk等基础命令行结合使用

⋮

实战：

示例一：对基因组2号染色体取反向互补，并对染色体号重命名

echo "chr02" | seqkit grep -f - genome.fa | sed 's/chr02/Chr02/' | seqkit seq - -r -p - > Chr02_re.fa

```
(base) [rzhou@login01 test]$ cat genome.fa
>chr01
ATCGGATCGATGCTAGC
>chr02
CATGCATCGATCGATCGCATCGATCA
(base) [rzhou@login01 test]$ echo "chr02" | seqkit grep -f - genome.fa | sed 's/chr02/Chr02/' | seqkit seq -  -r -p -
[WARN] flag -t (--seq-type) (DNA/RNA) is recommended for computing complement sequences
>Chr02
TGATCGATGCGATCGATCGATGCATG
```

实战：

示例二：对基因的ID加标签

```
less  IRGSP_CESA.protein.fa | awk '{if(/>/){print$1}else{print}}' | seqkit replace -p '(.+)' -r '{kv}' -k IRGSP_CESA.id ->IRGSP_CESA.proteion.fa
```

```
>Os01t0750300-01 Cellulose synthase catalytic subunit,
>Os03t0808100-01 Similar to Cellulose synthase-5.
>Os03t0837100-01 Similar to Cellulose synthase-6.
>Os05t0176100-01 Similar to Cellulose synthase BoCesA1.
>Os07t0208500-01 Similar to Cellulose synthase-4.
>Os07t0252400-01 Similar to Cellulose synthase-8.
>Os07t0424400-01 Similar to Cellulose synthase-7.
>Os09t0422500-01 Cellulose synthase A catalytic subunit
>Os10t0467800-01 Secondary wall-specific cellulose synt
```

```
>Os01t0750300-01_CESA4
>Os03t0808100-01_CESA2
>Os03t0837100-01_CESA5
>Os05t0176100-01_CESA1
>Os07t0208500-01_CESA8
>Os07t0252400-01_CESA6
>Os07t0424400-01_CESA3
>Os09t0422500-01_CESA9
>Os10t0467800-01_CESA7
```

实战：

示例三：原始fq数据的基本信息的统计

seqkit stat mySY137Run.fastq

file	format	type	num_seqs	sum_len	min_len	avg_len	max_len
FDES20H000008-1a_L1_1_clean.rd.fq.gz	FASTQ	DNA	352,721,005	52,908,150,750	150	150	150
FDES20H000008-1a_L1_2_clean.rd.fq.gz	FASTQ	DNA	352,721,005	52,908,150,750	150	150	150

实战：

示例四：合并多个文件中相同ID的行

less U9.AA.lifted.anchors | csvtk -tH join -f '1;1' - U9.BB.lifted.anchors | csvtk -tH cut -f 4,2,1|head

```
MtU9_01T0246600.1       Macma4_11_g12230.1      1690
MtU9_01T0246700.1       Macma4_11_g12140.1      1420
MtU9_01T0246900.1       Macma4_11_g12120.1      1460
MtU9_01T0247000.1       Macma4_11_g12110.1      215
MtU9_01T0247100.1       Macma4_11_g12100.1      283
MtU9_01T0247200.1       Macma4_11_g12090.1      762
MtU9_01T0247300.1       Macma4_11_g12080.1      1290
MtU9_01T0247700.1       Macma4_11_g12020.1      1170
MtU9_01T0247800.1       Macma4_11_g12030.1      623
MtU9_01T0248000.1       Macma4_11_g12050.1      1560
MtU9_01T0248100.1       Macma4_11_g12280.1      860
MtU9_01T0248200.1       Macma4_11_g12290.1      3280
```

```
MtU9_01T0246700.1       Mb_11_t10860.1  1390
MtU9_01T0246900.1       Mb_11_t10880.1  773
MtU9_01T0247000.1       Mb_11_t10890.1  220
MtU9_01T0247100.1       Mb_11_t10900.1  277
MtU9_01T0247300.1       Mb_11_t10910.1  1290
MtU9_01T0247700.1       Mb_11_t10940.1  941
MtU9_01T0248000.1       Mb_11_t10960.1  1510
MtU9_01T0248100.1       Mb_11_t10980.1  854
MtU9_01T0248200.1       Mb_11_t10990.1  2280
```

```
(jcvi) [rzhou@login01 MCScan]$ less U9.AA.lifted.anchors | csvtk -tH join -f '1;1' - U9.BB.lifted.anchors | head
MtU9_01T0000400.1       Macma4_01_g00040.1      5820    Mb_03_t22770.1  5810
MtU9_01T0000400.1       Macma4_01_g00040.1      5820    Mb_03_t22760.1  976L
MtU9_01T0000500.1       Macma4_01_g00050.1      2990    Mb_03_t22750.1  2910
MtU9_01T0000600.1       Macma4_01_g00060.1      3570    Mb_03_t22740.1  3580
MtU9_01T0000700.1       Macma4_01_g00070.1      738     Mb_03_t22730.1  723
MtU9_01T0000800.1       Macma4_01_g00080.1      1920    Mb_03_t22720.1  1920
MtU9_01T0000900.1       Macma4_01_g00090.1      1300    Mb_03_t22710.1  712L
MtU9_01T0001000.1       Macma4_01_g00110.1      465     Mb_03_t22700.1  465
MtU9_01T0001100.1       Macma4_01_g00120.1      927     Mb_03_t22680.1  920
MtU9_01T0001200.1       Macma4_01_g00130.1      1900    Mb_03_t22670.1  1880
(jcvi) [rzhou@login01 MCScan]$ less U9.AA.lifted.anchors | csvtk -tH join -f '1;1' - U9.BB.lifted.anchors | csvtk -tH cut -f
4,2,1| head
Mb_03_t22770.1  Macma4_01_g00040.1      MtU9_01T0000400.1
Mb_03_t22760.1  Macma4_01_g00040.1      MtU9_01T0000400.1
Mb_03_t22750.1  Macma4_01_g00050.1      MtU9_01T0000500.1
Mb_03_t22740.1  Macma4_01_g00060.1      MtU9_01T0000600.1
Mb_03_t22730.1  Macma4_01_g00070.1      MtU9_01T0000700.1
Mb_03_t22720.1  Macma4_01_g00080.1      MtU9_01T0000800.1
Mb_03_t22710.1  Macma4_01_g00090.1      MtU9_01T0000900.1
Mb_03_t22700.1  Macma4_01_g00110.1      MtU9_01T0001000.1
Mb_03_t22680.1  Macma4_01_g00120.1      MtU9_01T0001100.1
Mb_03_t22670.1  Macma4_01_g00130.1      MtU9_01T0001200.1
```

实战：

示例五：文件展示

less U9_gene_tpm_matrix.txt|cut -f 1-6 | csvtk -tH pretty| head

```
(base) [rzhou@login01 hisat]$ less U9_gene_tpm_matrix.filter.txt|cut -f 1-6 | head
gene_id TT_1    TT_2    TT_4    U9Y_1   U9Y_2
MtU9_08G0314600 4.04223 1.543398        5.942754        1.092574        2.321096
MtU9_05G0286000 36.665344       56.11039        63.713169       69.5075 35.27919
MtU9_03G0176000 0.0     0.10721 0.059769        0.0     0.0
MtU9_06G0120600 3.59566 4.334239        3.182409        24.21722        11.098662
MtU9_08G0186900 0.488709        0.743878        0.0     1.134533        6.225074
MtU9_05G0005000 53.706512       34.929195       63.138439       30.444492       5.501335
MtU9_01G0319600 17.565212       14.35022        15.341734       2.973236        4.564662
MtU9_03G0150400 0.0     0.0     0.0     0.0     0.0
MtU9_04G0408100 71.025589       46.286537       51.218918       172.635727      70.262276
(base) [rzhou@login01 hisat]$ less U9_gene_tpm_matrix.filter.txt|cut -f 1-6 | csvtk -tH pretty| head
gene_id                 TT_1            TT_2            TT_4            U9Y_1           U9Y_2
MtU9_08G0314600         4.04223         1.543398        5.942754        1.092574        2.321096
MtU9_05G0286000         36.665344       56.11039        63.713169       69.5075         35.27919
MtU9_03G0176000         0.0             0.10721         0.059769        0.0             0.0
MtU9_06G0120600         3.59566         4.334239        3.182409        24.21722        11.098662
MtU9_08G0186900         0.488709        0.743878        0.0             1.134533        6.225074
MtU9_05G0005000         53.706512       34.929195       63.138439       30.444492       5.501335
MtU9_01G0319600         17.565212       14.35022        15.341734       2.973236        4.564662
MtU9_03G0150400         0.0             0.0             0.0             0.0             0.0
MtU9_04G0408100         71.025589       46.286537       51.218918       172.635727      70.262276
```

其他常用小工具：Bedtools （bed、vcf、gff）

一款对genomic features进行比较、相关操作和注释的工具

bedtools主要使用bed格式的前三列，bed最多可以有12列

Chrom	start(0)	end	geneID	score	strand	thickstart	thickend	itemRGB	blockCount	blockSiz	blockStarts
Chr01	105687	107957	MtU9_01T0001000.1	0	-	105687	107957	0	2	1444,240,	0,2030,
Chr01	109083	112409	MtU9_01T0001100.1	0	+	109083	112409	0	3	572,153,348,	0,2038,2978,
Chr01	115932	117228	MtU9_01T0001200.1	0	+	115932	117228	0	1	1296,	0,
Chr01	121163	121532	MtU9_01T0001300.1	0	-	121163	121532	0	1	369,	0,

bedtools 能干什么呢

1 Intersect 取两个区间坐标的overlap
2 getfasta 提取任意给定区间的序列
3 makewindows 生成固定窗口或者滑动窗口的区间文件
4 closest 看A和B之间距离
5 subtract 去除两个区间中有overlap的部分，保留uniq片段
6 merge 取两个区间的并集
7 coverage 计算特定区间的覆盖度
8 maskFastaFromBed 对特定区间的序列进行mask
9 flank 提取区间上下游序列（常用于取基因的promoter区域）

-
-
-

使用小工具，很爽，一直用一直爽

好处：
检查自己程序
为您节省大量（不）编写 Python/R 脚本的时间,提高工作效率

坏处：
久久不写代码，编程能力容易退化

Thanks