第二次实验作业

29 MARCH 2021

> 实验背景
>
> > 在50多年的生物信息学发展中，尤其是随着DNA测序技术的出现与不断发展，生物信息学所处理的问题中很大一部分为核酸序列[1]。在下一代测序技术（Next Generation Sequencing, NGS）出现之后，使用计算机来辅助核酸序列的大数据处理更显得尤为重要与基础。而生物信息学主流工作平台仍然是开源、相对自由、易部署的多用户操作系统Linux，因此，掌握Linux系统的基本操作、及其上一些处理核酸序列数据的工具，是生物信息学学习与实践中重要一环
>
> 实验方法
>
> > 使用Linux系统内置的函数与命令对文本进行截取、替换、计数等基本操作，并使用管道符 `|` 来进行命令的分割与结果的流水线处理。
> >
> > 使用 `GFFread` [2]命令行工具，结合序列注释数据，对GFF3/GTF格式的人21号染色体序列提取特征与信息
>
> 运行环境（见下）
>
> ```
> platform          x86_64-pc-linux-gnu
> arch              x86_64
> os                linux-gnu
> system            x86_64, linux-gnu
> R version         4.0.5 (2021-03-31)
> nickname          Lost Library Book
> radian version    0.5.10
> python version    3.7.3
> ```

基因组序列与注释数据——以人21号染色体为例

数据速览

基因组序列

`chr21.fa`文件开头如下

```
>chr21
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

其中，`>`开头的行为起始行，后跟下文序列的描述或元数据；其后为若干行序列数据，忽略所有空白字符，直至下一个`>`开头的起始行或文件结尾

注释数据[3]

`annotation.gtf`文件开头如下（此处由于服务器上下载时中断，因此采用最新的全基因组注释文件重新构建`annotation.gtf`）

```
##description: evidence-based annotation of the human genome (GRCh38), version 37 (
Ensembl 103)
##provider: GENCODE
##contact: gencode-help@ebi.ac.uk
##format: gtf
##date: 2020-12-07
chr21   HAVANA  gene    5011799 5017145 .   +   .   gene_id "ENSG00000279493.1"; g
ene_type "protein_coding"; gene_name "FP565260.4"; level 2; havana_gene "OTTHUMG00
000189354.1";
chr21   HAVANA  transcript  5011799 5017145 .   +   .   gene_id "ENSG00000279493.1
"; transcript_id "ENST00000624081.1"; gene_type "protein_coding"; gene_name "FP565
260.4"; transcript_type "protein_coding"; transcript_name "FP565260.4-201"; level
2; protein_id "ENSP00000485664.1"; transcript_support_level "5"; tag "mRNA_start_NF
"; tag "cds_start_NF"; tag "basic"; tag "appris_principal_1"; havana_gene "OTTHUMG0
0000189354.1"; havana_transcript "OTTHUMT00000479422.1";
```

其中`##`开头的为注释行，此文件中前五行均为注释，分别为：

1. 描述，此文件是基于证据建立的人基因组序列注释信息
2. 提供者，GeneCode
3. 联系方式
4. 格式，此处为`gtf`
5. 日期

此行之后为正式的注释信息，其中有多个`tab`分隔的字段，从左至右依次为：

1. 序列名称：`chr21`
2. 数据来源：`HAVANA`，ENSEMBL的手动注释项目
3. 特征名：`gene`基因，`transcript`转录本
4. 起始位点
5. 终止为点
6. 得分
7. 链方向：`+`正向，`-`反向
8. 读码框：`0`表示首位密码子与特征起始位置重合，`1`首位密码子在特征起始位置后一位，`2`后两位
9. 属性：包含了序列相关的额外注释信息，可选的见下

- `gene_id`：基因ID
- `gene_type`：基因类型，`protein_coding`等
- `gene_name`：基因名
- `level`：？（未查到对应文档记录）
- `havana_gene`：HAVANA项目中的基因
- `transcript_id`：转录本ID
- `transcript_name`：转录本名
- `transcript_type`：转录本类型
- `protein_id`：蛋白ID
- `transcript_support_level`：基于模型的转录本支持等级
- `havana_transcript`：HAVANA项目中的转录本
- `tag`：额外的注释标签

统计长度

总长

```
cat chr21.fa | \
grep -v ">" | \
tr [:lower:] [:upper:] | \
grep -o [ATCGN] | \
wc -l > \
length.txt && \
cat length.txt
```

- line 1: `cat chr21.fa`打印`chr21.fa`文本
- line 2: `grep -v ">"`去掉首行注释
- line 3: `tr [:lower:] [:upper:]`将所有小写转化为大写
- line 4: `grep -o [ATCGN]`抓取字符，每个字符一行
- line 5: `wc -l`统计字符数（实际为行数）
- line 6: `> length.txt`结果写入`length.txt`
- line 7: `cat length.txt`打印结果

得到所读入的21号染色体长度为`46709919`。

N个数

在`fasta`等格式的记录核酸信息的文本语言中，除了`ATCG`代表DNA的四种碱基外，还有其他的拓展符号[4]来表示简并的碱基对，其中`N`指此位置可以为`ATCG`中任一碱基；大片连续出现在`fasta`等格式文件中时通常意味着此段序列未知。

```
cat chr21.fa | \
grep -v ">" | \
tr [:lower:] [:upper:] | \
grep -o [N] | \
wc -l > \
lengthN.txt && \
cat lengthN.txt
```

与统计全长类似，我们只需修改命令行，将`grep`匹配的字符由`[ATCGN]`修改为`[N]`即可，得到结果为`6621300`。

统计各feature频次

```
cat annotation.gtf | \
grep -v "##*" | \
cut -f 3 | \
sort | \
uniq -c > \
features.txt && \
cat features.txt
```

- line 1: `cat annotation.gtf`打印`annotation.gtf`文件中文本
- line 2: `grep -v "##*"`除去以`##`开头的注释行
- line 3: `cut -f 3`获取feature字段的内容
- line 4: `sort`排序结果
- line 5: `uniq -c`对结果计数
- line 6: `> features.txt`将结果写入`features.txt`
- line 7: `cat features.txt`打印结果

得到结果

```
   8519 CDS
  17880 exon
    875 gene
    923 start_codon
    897 stop_codon
   3021 transcript
   3280 UTR
```

统计所有gene ID与频次

```
cat annotation.gtf | \
grep -v "##" | \
cut -f 9 | \
grep "transcript_id" | \
cut -d ";" -f 1,2 | \
sed "s/;//g" | \
cut -d ' ' -f 2,4 | \
sort | \
uniq -c | \
sort -b -g > \
gene_id.txt && \
cat gene_id.txt
```

- line 1: `cat annotation.gtf` 打印 `annotation.gtf` 文件中文本
- line 2: `grep -v "##*"` 除去以 `##` 开头的注释行
- line 3: `cut -f 9` 获取注释字段的内容
- line 4: `grep "transcript_id"` 获取含有转录ID的注释信息行
- line 5: `cut -d ';' -f 1,2` 获取字段中第一、二项 `gene_id "ENSG*"; transcript_id "ENST*"`
- line 6: `sed "s/;//g"` 去除多余分号
- line 7: `cut -d ' ' -f 2,4` 获取 `gene_id` 和 `transcript_id` 中编号
- line 8: `sort` 排序结果
- line 9: `uniq -c` 对结果计数
- line 10: `sort -b -g` 根据统计数排序结果
- line 11: `> gene_id.txt` 将结果写入 `gene_id.txt`
- line 12: `cat gene_id.txt` 打印结果

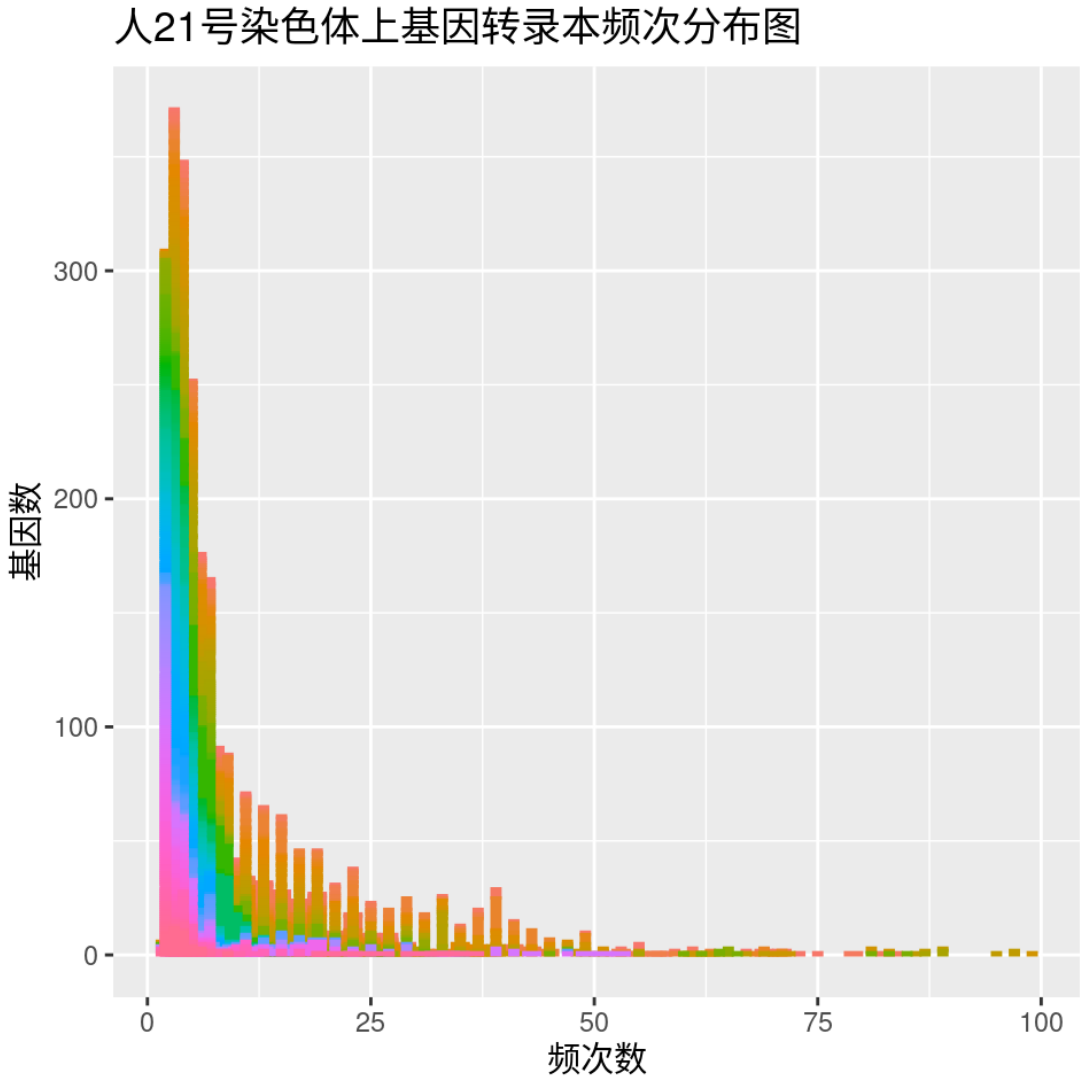
得到结果为

```
  3 "ENSG00000279687.1" "ENST00000623188.1"
 11 "ENSG00000279493.1" "ENST00000624081.1"
 11 "ENSG00000280071.4" "ENST00000625036.3"
 12 "ENSG00000280071.4" "ENST00000620015.4"
 13 "ENSG00000280071.4" "ENST00000624810.3"
 17 "ENSG00000277117.5" "ENST00000620481.4"
 17 "ENSG00000277117.5" "ENST00000623795.1"
 19 "ENSG00000277117.5" "ENST00000612610.4"
 19 "ENSG00000277117.5" "ENST00000623903.3"
 19 "ENSG00000277117.5" "ENST00000623960.4"
...
 86 "ENSG00000182871.16" "ENST00000359759.8"
 87 "ENSG00000182871.16" "ENST00000355480.10"
 87 "ENSG00000185658.14" "ENST00000342449.8"
 89 "ENSG00000182871.16" "ENST00000651438.1"
 89 "ENSG00000185658.14" "ENST00000333229.6"
 89 "ENSG00000185658.14" "ENST00000380800.7"
 95 "ENSG00000182670.13" "ENST00000354749.6"
 97 "ENSG00000182670.13" "ENST00000355666.5"
 97 "ENSG00000182670.13" "ENST00000399017.6"
 99 "ENSG00000160299.17" "ENST00000359568.10"
```

绘制频次分布图

```
library(ggplot2)
library(tidyverse)
gene_id <- read.table("./gene_id.txt")
colnames(gene_id) <- c("Freq", "geneID", "transID")
DT::datatable(gene_id)
```

```
gene_id %>%
    ggplot(aes(x = Freq)) +
    geom_bar(
        aes(color = geneID, fill = geneID),
        position = "stack"
    ) +
    labs(
        x = "频次数",
        y = "基因数",
        title = "人21号染色体上基因转录本频次分布图"
    ) +
    theme(
        legend.position = "none",
        legend.text = element_text(
            size = 6
        )
    )
```



gffread使用

gffread提取注释信息

gffread[2]命令提供了从GTF注释文件和参考基因组文件中提取出特定信息的功能，如下所示：

```
gffread -W -w exons.fasta -x cds.fasta -y protein.fasta -g chr21.fa annotation.gtf
```

其中，`gffread`的`-W`参数指定向文件中写入额外的序列定位信息，`-w`参数指定提取外显子信息，`-x`参数指定提取编码序列信息，`-y`参数指定提取翻译后蛋白序列信息，`-g`参数指定参考序列，最后输入的为`GTF`格式的注释文件。

蛋白序列计数

在获得了各种特征的序列信息之后，可以借助于`EMBOSS`软件包中的`seqcount`工具，一个简单的命令行完成计数

```
seqcount protein.fasta -auto -stdout
```

得到结果为共有`1049`个编码蛋白的基因

长非编码RNA

长非编码RNA（long non-coding RNA, lncRNA）为长度在200bp以上、不编码氨基酸的RNA转录物，文献中常报道其在转录调控、转录后修饰、表观遗传等方面有生物学意义，但也有文献中实验表明一些lncRNA实际上也会翻译为多肽[5]。lncRNA的数据目前主要来源于手动标记，因此在提取21号染色体上的lncRNA转录本时也依照注释文件中的信息来筛选。

我们可以通过筛选`gtf`注释信息中的关键词来提取一份仅包含lncRNA的注释文件

```
cat annotation.gtf | \
grep "lncRNA" > \
lncRNA.gtf
```

- line 1: `cat annotation.gtf`打印文件内容
- line 2: `grep "lncRNA"`仅保留含有`lncRNA`的行
- line 3: `> lncRNA.gtf`将筛选后结果写入文件

然后我们可以通过指定`gffread`工具中`--force-exons`参数来将lncRNA识别为外显子，然后将结果写入文件

```
gffread -W -w lncRNA.fasta -g chr21.fa lncRNA.gtf
```

最终得到结果记录于`lncRNA.fasta`中，同样可以使用`seqcount`工具完成计数

```
seqcount lncRNA.fasta -auto -stdout
```

可以得到，所选注释文件中共含有此1047个lncRNA

注释格式转换

`gffread`提供了简单的`gff <--> gtf`格式转换的命令行工具

```
gffread annotation.gtf -o annotation.gff
```

反之亦然，只需要加参数`-T`指定输出格式为gtf

可以查看转换后得到的gff文件的头几行

```
# gffread annotation.gtf -o annotation.gff
##gff-version 3
chr21   HAVANA  mRNA    5011799 5017145 .   +   .   ID=ENST00000624081.1;geneID=EN
SG00000279493.1;gene_name=FP565260.4
chr21   HAVANA  exon    5011799 5011874 .   +   .   Parent=ENST00000624081.1
chr21   HAVANA  exon    5012548 5012687 .   +   .   Parent=ENST00000624081.1
chr21   HAVANA  exon    5014386 5014471 .   +   .   Parent=ENST00000624081.1
chr21   HAVANA  exon    5016935 5017145 .   +   .   Parent=ENST00000624081.1
chr21   HAVANA  CDS 5011799 5011874 .   +   0   Parent=ENST00000624081.1
chr21   HAVANA  CDS 5012548 5012687 .   +   2   Parent=ENST00000624081.1
chr21   HAVANA  CDS 5014386 5014471 .   +   0   Parent=ENST00000624081.1
chr21   HAVANA  CDS 5016935 5017098 .   +   1   Parent=ENST00000624081.1
```

其中`##`开头的行依然为注释行，在转换过程中原文件头中的注释信息被忽略，取而代之的是输入的命令与gff格式版本。其后为注释文件主体，包含了若干个字段：

- 字段1~8：与gtf文件注释格式的含义基本一致
- 字段9：包含转录本/mRNA的ID`ID`，基因ID`geneID`，基因名`gene_name`，以及一个额外的属性`Parent`

不同于gtf格式，gff格式中多包含的`Parent`属性指向了一个转录本ID，也就是说，gff格式中注释信息为一种层级结构，以转录本为最大的注释，下面从属了若干个外显子或CDS的字段

全基因组注释信息简单统计

使用`rsync`下载最新的人类全基因组新数据，并且下载全基因组注释数据

```
mkdir hg38 && rsync -avzP rsync://hgdownload.cse.ucsc.edu/goldenPath/hg38/chromosomes/

mkdir hg38/affi && mv hg38/*_*.fa hg38/affi && cat hg38/*.fa > hg38/wholeHumanGenome.f

wget ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_37/gencode.v37.an
```

其中全基因组数据中包含了一些未配对的序列与其他杂项，并不为我们所用，故将其移入一个子文件夹中，并将剩下所需要的基因组序列整合为一个`fasta`文件`wholeHumanGenome.fa`，此后我们依照上文所述的`gffread`使用方法来对注释信息做一定提取

```
mkdir features && gffread gencode.v37.annotation.gtf -g chromosomes/wholeHumanGenome.f
```

由于其文件过大，故不在附录中展示，但可以通过命令行工具简单统计各个基因组上的序列信息，以统计各基因组上的外显子序列数量为例：

```
cat features/exons.fasta | \
grep ">" | \
cut -d ' ' -f 3 | \
cut -d '(' -f 1 | \
cut -d ':' -f 2 | \
sort | \
uniq -c | \
sort -b -g
```

得到结果为

```
  13 chrM
 216 chrY
1049 chr21
1370 chr13
1785 chr18
2268 chr22
2395 chr20
3352 chr9
3518 chrX
3548 chr15
3552 chr10
3923 chr8
4074 chr14
4109 chr4
4519 chr5
4750 chr6
4875 chr7
5201 chr16
6245 chr12
6375 chr3
6957 chr17
6966 chr2
7025 chr11
7595 chr19
9161 chr1
```

其中最少的为线粒体基因组上的外显子，其次为Y染色体

还可以通过简单的`bash`脚本统计各个基因组的基本信息

```
function calcChrKnownBasePct() {
  if [ ! -f $1 ]; then
    exit -1
  fi
  local seqName=$(cat $1 | grep ">" | sed "s/>//g")
  local unknownBase=$(cat $1 | grep -v ">" | tr [:lower:] [:upper:] | grep -o [N] | w

  local knownBase=$(cat $1 | grep -v ">" | tr [:lower:] [:upper:] | grep -o [ATCG] | 

  local totalBase=$(echo "$(($knownBase + $unknownBase))")
  local percentage=$(printf %.5f "$((10 ** 7 * $knownBase / $totalBase))e-7")
  echo -e "$seqName\t$percentage\t$totalBase\t$knownBase\t$unknownBase"
  return
}

function statChrInfo() {
  if [ ! -d $1 ]; then
    exit -1
  fi
  touch $1/stats.txt
  echo -e "SeqInfo\tPercentage\tlength\tknown_length\tunknown_length" > $1/stats.txt

  for fileName in $1/chr*.fa; do
    echo -e "Working on file $fileName"
    echo -e "$(calcChrKnownBasePct $1/$fileName)" >> $1/stats.txt
    echo -e "$fileName processed"
  done
}
```

得到的结果如下，不难看出Y染色体上注释信息较少确实与配对的序列长度有关

```
SeqInfo Percentage  length  known_length    unknown_length
chr10   0.99601 133797422   133262962   534460
chr11   0.99591 135086622   134533742   552880
chr12   0.99897 133275309   133137816   137493
chr13   0.85676 114364328   97983125    16381203
chr14   0.84609 107043718   90568149    16475569
chr15   0.82989 101991189   84641325    17349864
chr16   0.90555 90338345    81805943    8532402
chr17   0.99595 83257441    82920204    337237
chr18   0.99647 80373285    80089605    283680
chr19   0.99698 58617616    58440758    176858
chr1    0.92579 248956422   230481012   18475410
chr20   0.99224 64444167    63944257    499910
chr21   0.85825 46709983    40088619    6621364
chr22   0.77058 50818468    39159777    11658691
chr2    0.99321 242193529   240548228   1645301
chr3    0.99901 198295559   198100135   195424
chr4    0.99757 190214555   189752667   461888
chr5    0.99850 181538259   181265378   272881
chr6    0.99574 170805979   170078522   727457
chr7    0.99764 159345973   158970131   375842
chr8    0.99745 145138636   144768136   370500
chr9    0.88002 138394717   121790550   16604167
chrM    0.99994 16569   16568   1
chrX    0.99264 156040895   154893029   1147866
chrY    0.46158 57227415    26415043    30812372
```

参考资料

[1] GAUTHIER J, VINCENT A T, CHARETTE S J, 等. A brief history of bioinformatics[J]. Brief Bioinform, 2019, 20(6): 1981–1996. DOI:10.1093/bib/bby063.

[2] PERTEA G, PERTEA M. GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved][J]. F1000Research, 2020, 9(304). DOI:10.12688/f1000research.23297.2.

[3] RUFFIER M, KÄHÄRI A, KOMOROWSKA M, 等. Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation[J]. Database (Oxford), 2017, 2017(1). DOI:10.1093/database/bax020.

[4] 百度百科. 简并碱基[Z/OL]([日期不详]). https://baike.baidu.com/item/%E7%AE%80%E5%B9%B6%E7%A2%B1%E5%9F%BA.

[5] WIKIPEDIA. Long non-coding RNA[Z/OL](2021). https://en.wikipedia.org/wiki/Long_non-coding_RNA.