生物信息学基本概念与工具

ZHANG Yang

2017 年 12 月 8 日

目录

Part I

生物信息学基本概念介绍

Part II

生物信息学常用文件格式

Part III

生物信息学常用工具

Chapter 1

Samtools

Samtools 包含了一系列操作高通量测序数据的工具，主要用于操作 SAM 格式、BAM 格式等。主要包括的模块包括：

Samtools	读写、编辑以及排序 SAM/BAM/CRAM 文件
BCFtools	读写 BCF2/VCF/gVCF 文件；SNP 和 InDel 的提取、过滤等
HTSlib	读写高通量数据格式的 C 语言库

通过 Samtools 可以很方便的对 SAM、BAM 文件进行排序、合并、索引，并可以很方便地提取任意区域的 Reads。Samtools 能够识别和打开 FTP（“ftp://”）或 HTTP（“http://”）服务器上的 BAM 文件（不能打开 SAM 文件）。

1.1 用法举例

```
samtools view -bt ref_list.txt -o aln.bam aln.sam.gz
```

```
samtools sort -T /tmp/aln.sorted -o aln.sorted.bam aln.bam
```

```
samtools index aln.sorted.bam
```

```
samtools idxstats aln.sorted.bam
```

```
samtools flagstat aln.sorted.bam
```

```
samtools stats aln.sorted.bam
```

```
samtools bedcov aln.sorted.bam
```

```
samtools depth aln.sorted.bam
```

```
samtools view aln.sorted.bam chr2:20,100,000-20,200,000
```

```
samtools merge out.bam in1.bam in2.bam in3.bam
```

```
samtools faidx ref.fasta
```

```
samtools tview aln.sorted.bam ref.fasta
```

```
samtools split merged.bam
```

```
samtools quickcheck in1.bam in2.cram
```

```
samtools dict -a GRCh38 -s "Homo sapiens" ref.fasta
```

```
samtools fixmate in.namesorted.sam out.bam
```

```
samtools mpileup -C50 -gf ref.fasta -r chr3:1,000-2,000 in1.bam in2.bam
```

```
samtools flags PAIRED,UNMAP,MUNMAP
```

```
samtools fastq input.bam > output.fastq
```

```
samtools fasta input.bam > output.fasta
```

```
samtools addreplacerg -r 'ID:fish' -r 'LB:1334' -r 'SM:alpha' -o output.bam input.bam
```

```
samtools collate aln.sorted.bam aln.name_collated.bam
```

```
samtools depad input.bam
```

```
samtools markdup in.algnsorted.bam out.bam
```

表 1.1: samtools view [region...] 设置举例

chr1	输出所有参考序列名称为 chr1 的比对结果
chr2:1000000	输出所有参考序列名称为 chr2 的，比对位点在 1000000 之后的比对结果
chr3:1000-2000	输出所有参考序列名称为 chr3 的，比对位点在 1000 之后、2000 之前的比对结果
'*'	输出文件末尾的、未能比对的 reads。
.	输出所有比对结果，等同于未指定 REGION。

1.2 Samtools 可用命令和参数

1.2.1 samtools view

samtools view 的基本用法如下：

```
samtools view [options] in.sam|in.bam|in.cram [region...]
```

改命令主要用于将比对文件（SAM、BAM 或 CRAM 文件）中满足给定条件的比对结果输出到标准输出（屏幕）。

可以在 [region...] 处按 RNAME[:STARTPOS[-ENDPOS]] 的格式指定满足给定条件的比对结果所在的染色体区域（当给定多个区域时，某些比对结果可能输出多次）。指定[region...] 的形式举例（表 ??）：

samtools view 可以指定一系列的参数（options，表 1.2）：

表 1.2: samtools view 可选参数

参数	值	说明
-b		Output in the BAM format
-C		Output in the CRAM format (requires -T)
-1		Enable fast BAM compression (implies -b)
-u		Output uncompressed BAM. This option saves time spent on compression/decompression and is thus preferred when the output is piped to another samtools command.
-h		Include the header in the output.
-H		Output the header only.
-c		Instead of printing the alignments, only count them and print the total number. All filter options, such as -f, -F, and -q, are taken into account.
-?		Output long help and exit immediately.
-o	FILE	Output to FILE [stdout].

续表 1.2 · · ·

参数	值	说明
-U	FILE	Write alignments that are not selected by the various filter options to FILE. When this option is used, all alignments (or all alignments intersecting the regions specified) are written to either the output file or this file, but never both.
-t	FILE	A tab-delimited FILE. Each line must contain the reference name in the first column and the length of the reference in the second column, with one line for each distinct reference. Any additional fields beyond the second column are ignored. This file also defines the order of the reference sequences in sorting. If you run: 'samtools faidx <ref.fa>', the resulting index file <ref.fa>.fai can be used as this FILE.
-T	FILE	A FASTA format reference FILE, optionally compressed by bgzip and ideally indexed by samtools faidx. If an index is not present, one will be generated for you.
-L	FILE	Only output alignments overlapping the input BED FILE [null].
-r	STR	Only output alignments in read group STR [null].
-R	FILE	Output alignments in read groups listed in FILE [null].
-q	INT	Skip alignments with MAPQ smaller than INT [0].
-l	STR	Only output alignments in library STR [null].
-m	INT	Only output alignments with number of CIGAR bases consuming query sequence INT [0]
-f	INT	Only output alignments with all bits set in INT present in the FLAG field. INT can be specified in hex by beginning with '0x' (i.e./^0x[0-9A-F]+/) or in octal by beginning with '0' (i.e. /^0[0-7]+/) [0].
-F	INT	Do not output alignments with any bits set in INT present in the FLAG field. INT can be specified in hex by beginning with '0x' (i.e. /^0x[0-9A-F]+/) or in octal by beginning with '0' (i.e. /^0[0-7]+/) [0].
-G	INT	Do not output alignments with all bits set in INT present in the FLAG field. This is the opposite of -f such that -f12 -G12 is the same as no filtering at all. INT can be specified in hex by beginning with '0x' (i.e. /^0x[0-9A-F]+/) or in octal by beginning with '0' (i.e./^0[0-7]+/) [0].
-x	STR	Read tag to exclude from output (repeatable) [null]
-B		Collapse the backward CIGAR operation.

续表 1.2 ···

参数	值	说明
-s	FLOAT	Output only a proportion of the input alignments. This subsampling acts in the same way on all of the alignment records in the same template or read pair, so it never keeps a read but not its mate. The integer and fractional parts of the -s INT.FRAC option are used separately: the part after the decimal point sets the fraction of templates/pairs to be kept, while the integer part is used as a seed that influences which subset of reads is kept. When subsampling data that has previously been subsampled, be sure to use a different seed value from those used previously; otherwise more reads will be retained than expected.
-@	INT	Number of BAM compression threads to use in addition to main thread [0].
-S		Ignored for compatibility with previous samtools versions. Previously this option was required if input was in SAM format, but now the correct format is automatically detected by examining the first few characters of input.

1.2.2 samtools sort

samtools sort 命令用于给定的条件对比对结果进行排序。默认情况下，按照按照个比对结果最左边的坐标。如果给出`-n`，则按比对结果的 read 名称进行排序。

排序规则为：

如果存在 “-t” 参数，比对记录首先根据跟定的比对标签（alignment tag）；然后如果存在 “-n” 参数，则按 read 名称排序，否则按坐标排序。

默认情况下，会将结果输出到标准输出（一般是屏幕）；如果给出`-o`，则会将结果写入`-o` 指定的文件。

samtools sort 的基本用法如下：

```
samtools sort [-l level] [-m maxMem] [-o out.bam] [-O format]
    [-n] [-t tag] [-T tmpprefix] [-@ threads]
    [in.sam|in.bam|in.cram]
```

samtools sort 的可选参数包括（表 1.3）：

表 1.3: samtools sort 可选参数

参数	值	说明
-l	INT	小写字母 “L”，level；设置期望的压缩级别。0：不压缩；1 速度最快，压缩程度较低；9：压缩程度最高；（与 gzip 的压缩级别类似）
-m	INT	小写字母 “m”，memory；每个线程最大的（大致）内存需求。可以通过 K、M 或 G 指定单位。为防止参数大量临时文件，最小值需大于 1M。

续表 1.3 · · ·

参数	值	说明
-n		按 read 名称（SAM 格式的 QNAME 字段）排序。
-t	TAG	Sort first by the value in the alignment tag TAG, then by position or name (if also using -n).
-o	FILE	将输出写入到 FILE 中。
-O	FORMAT	输出文件的格式（sam、bam 和 cram）；默认值 bam。
-T	PREFIX	如果 PREFIX 是文件夹，则命名临时文件为 “PREFIX/samtools.mmm.mmm.tmp.nnnn.bam”。否则，命名临时文件为 “PREFIX.nnnn.bam”。
-@	INT	设置线程数，默认为单线程。

1.2.3 samtools index

对 BAM 或 CRAM 创建索引，以便于快速随机访问。

samtools index 的基本用法如下：

```
samtools index [-bc] [-m INT] aln.bam|aln.cram [out.index]
```

samtools index 的可选参数包括（表 1.4）：

表 1.4: samtools sort 可选参数

参数	值	说明
-b		创建 BAI 索引。
-c		创建 CSI 索引。默认情况下，最小间隔大小为 2^{14}。
-m	INT	创建 CSI 索引，将最小间隔大小设置为 2^{INT}。

1.2.4 samtools idxstats

输出与给定文件相关的索引文件的状态。故在使用此命令前，应该对输入文件使用 index 命令处理。

samtools idxstats 的基本用法如下：

```
samtools idxstats in.sam|in.bam|in.cram
```

输出内容包括：

- 参考序列名称
- 参考序列长度
- 比对上的 reads 数目
- 未必对上的 reads 数目

1.2.5 samtools flagstat

计算并输出统计信息到标准输出（一般指屏幕）。

samtools flagstat 的基本用法如下：

```
samtools flagstat in.sam|in.bam|in.cram
```

1.2.6 samtools stats

samtools stats 收集 BAM 文件的统计信息，并输入到文本文件中。输出文件可以通过 ***plot-bamstats*** 。

samtools stats 的基本用法如下：

```
samtools stats [options] in.sam|in.bam|in.cram [region...]
```

表 1.5: samtools stats 可选参数

参数	长参数	值	说明
-c	–coverage	MIN,MAX,STEP	设置指定区域的覆盖分布情况（MIN, MAX, STEP）[1,1000,1]
-d	–remove-dups		Exclude from statistics reads marked as duplicates
-f	–required-flag	STR|INT	Required flag, 0 for unset. See also ‘samtools flags‘ [0]
-F	–filtering-flag	STR|INT	Filtering flag, 0 for unset. See also ‘samtools flags‘ [0]
	–GC-depth	FLOAT	the size of GC-depth bins (decreasing bin size increases memory requirement) [2e4]
-h	–help		输出帮助信息
-i	–insert-size	INT	Maximum insert size [8000]
-I	–id	STR	Include only listed read group or sample name []
-l	–read-length	INT	Include in the statistics only reads with the given read length []
-m	–most-inserts	FLOAT	Report only the main part of inserts [0.99]
-P	–split-prefix	STR	A path or string prefix to prepend to filenames output when creating categorised statistics files with -S/–split. [input filename]
-q	–trim-quality	INT	The BWA trimming parameter [0]
-r	–ref-seq	FILE	Reference sequence (required for GC-depth and mismatches-per-cycle calculation). []

续表 1.5 · · ·

参数	长参数	值	说明
-S	–split	TAG	In addition to the complete statistics, also output categorised statistics based on the tagged field TAG (e.g., use –split RG to split into read groups). Categorised statistics are written to files named `<prefix>_<value>`.bamstat, where prefix is as given by –split-prefix (or the input filename by default) and value has been encountered as the specified tagged field's value in one or more alignment records.
-t	–target-regions	FILE	Do stats in these regions only. Tab-delimited file chr,from,to, 1-based, inclusive. []
-x	–sparse		Suppress outputting IS rows where there are no insertions. 。

1.2.7 samtools bedcov

1.2.8 samtools depth

1.2.9 samtools merge

1.2.10 samtools faidx

1.2.11 samtools tview

1.2.12 samtools split

1.2.13 samtools quickcheck

1.2.14 samtools dict

1.2.15 samtools fixmate

1.2.16 samtools mpileup

1.2.17 samtools flags

1.2.18 samtools fastq/a

1.2.19 samtools collate

1.2.20 samtools reheader

1.2.21 samtools cat

1.2.22 samtools rmdup

1.2.23 samtools addreplacerg

1.2.24 samtools calmd

1.2.25 samtools targetcut

1.2.26 samtools phase

1.2.27 samtools depad

1.2.28 samtools markdup

1.2.29 samtools help

显然，该命令用于输出 samtools 的帮助信息。另外，如果给出了 samtools 的命令名，该命令将输出有关 samtools 特定命令的帮助信息。

samtools help 的基本用法如下：

```
1 samtools help view
```

Part IV

R 程序设计语言

Chapter 2

R 语言帮助系统

Chapter 3

R 语言控制语句

Chapter 4

R 语言内置数据结构

Chapter 5

R 语言内置函数

Chapter 6

R 语言面向对象编程

Chapter 7

R 语言输入和输出

Chapter 8

R 语言异常与错误处理

Chapter 9

R 语言标准库简介

Part V

Python 程序设计语言

Chapter 10

Python 帮助系统

Chapter 11

Python 控制语句

11.1 if 语句

```
if_stmt ::=  "if" expression ":" suite
( "elif" expression ":" suite )*
["else" ":" suite]
```

```
x = int(input("Please enter an integer: "))

if x < 0:
    x = 0
    print('Negative changed to zero')
elif x == 0:
    print('Zero')
elif x == 1:
    print('Single')
else:
    print('More')
```

11.2 for 语句

```
for_stmt ::=  "for" target_list "in" expression_list ":" suite
["else" ":" suite]
```

```
animals = ['cat', 'dog']

```

```
3 for animal in animals:
4     print(animal, len(animal))
```

11.3 while 语句

```
while_stmt ::=  "while" expression ":" suite
["else" ":" suite]
```

11.4 break 语句

11.5 continue 语句

11.6 pass 语句

Chapter 12

Python 内置数据结构

Chapter 13

Python 内置函数

Chapter 14

Python 面向对象编程

Chapter 15

Python 输入和输出

Chapter 16

Python 异常与错误处理

Chapter 17

Python 标准库简介

Part VI

C++ 程序设计语言

Chapter 18

C++ 帮助系统

Chapter 19

C++ 控制语句

Chapter 20

C++ 数据结构

附录 A

附录

参考文献