

高通量测序数据处理学习记录（一）：比对软件STAR的使用



面面的徐爷 (/u/fe854ffa1f9e) +关注

0.4 2017.09.13 11:31* 字数 915 阅读 6091 评论 4 喜欢 12

(/u/fe854ffa1f9e)



high-throughput seq analysis急先锋——STAR的使用介绍

在所有物是人非的景色里，我最中意你。

正体

这次给大家带来的是ENCODE project的御用比对软件STAR，ENCODE项目是一个由美国国家人类基因组研究所(NHGRI)在2003年9月发起的一项公共联合研究项目，旨在找出人类基因组中所有功能组件[。这是既完成人类基因组计划后国家人类基因组研究所开始的最重要的项目之一。所有在该项目中产生的数据都会被迅速的在公共数据库中公开。

在我之前的那篇RNA-seq数据分析---方法学文章的实战练习

(<https://www.jianshu.com/p/1f5d13cc47f8>)文章里关于比对软件的比较中STAR也展现了不俗的表现。所以在处理比对时我也考虑了将HISAT2与STAR共同使用，查看它们的表现情况，选取适合的比对工具。

STAR的安装



```
cd biosoft && mkdir STAR && cd STAR
wget https://github.com/alexdobin/STAR/archive/2.5.3a.tar.gz
tar -xzf 2.5.3a.tar.gz
cd STAR-2.5.3a

# for easy use, add bin/ to your PATH
```

下载需要参考基因组并进行index构建

```
# downloading dna index fasta file
nohup wget -r -np -nH -nd -R index.html -L ftp://ftp.ensembl.org/pub/release-90/fasta

# download gtf annotation file
nohup wget ftp://ftp.ensembl.org/pub/release-90/gtf/homo_sapiens/Homo_sapiens.GRCh38

mkdir STAR_index && cd STAR_index
STAR --runMode genomeGenerate --genomeDir ~/reference/STAR_index/ --genomeFastaFiles

# --sjdbOverhang 数值为reads长度-1
# Mode 为generate
# --genomeFastaFiles --sjdbGTFfile 分别对应fasta文件和GTF文件
```

STAR的使用

```
# STAR的manual里面给了最基本的比对参数示例
STAR
--runThreadN NumberOfThreads
--genomeDir /path/to/genomeDir
--readFilesIn /path/to/read1 [/path/to/read2 ]

# 基本示例，针对fastq.gz文件增加--readFilesCommand gunzip -c 参数/--readFilesCommand zcat
STAR --runThreadN 20 --genomeDir ~/reference/STAR_index/ --readFilesCommand zcat --readFilesIn

# 输出unsorted or sorted bam file
--outSAMtype BAM Unsorted 实际上就是-name 的sort，下游可以直接接HTSeq
--outSAMtype BAM SortedByCoordinate
--outSAMtype BAM Unsorted SortedByCoordinate 两者都输出
```

额外参数说明



```
# 单独指定注释文件，而不用在构建的时候使用
--sjdbGTFfile /path/to/ann.gtf
--sjdbFileChrStartEnd /path/to/sj.tab

# ENCODE参数

# 减少伪junction的几率
--outFilterType BySJout

# 最多允许一个reads被匹配到多少个地方
--outFilterMultimapNmax 20

# 在未有注释的junction区域，最低允许突出多少个bp的单链序列
--alignSJoverhangMin 8

# 在有注释的junction区域，最低允许突出多少个bp的单链序列
--alignSJDBoverhangMin 1

# 过滤掉每个paired read mismatch数目超过N的数据，999代表着忽略这个过滤
--outFilterMismatchNmax 999

# 相对paired read长度可以允许的mismatch数目，如果read长度为100，数值设定为0.04，则会过滤掉
--outFilterMismatchNoverReadLmax 0.04

# 最小的intro长度
--alignIntronMin 20

# 最大的intro长度
--alignIntronMax 1000000

# maximum genomic distance between mates，翻译不出来，自行理解
--alignMatesGapMax 1000000
```

(/apps/
utm_sc
banner

STAR的输出

STAR可以根据你的参数设定输出多个结果文件，包含各种信息，下面对默认参数情况下的输出文件做了一个详细的展示，有些不好翻译的地方我选择使用原汁原味的manual text

- Aligned.out.sam

Aligned.out.sam当然就是我们的比对结果啦！

```
E00516:168:H37WKCCXY:8:1101:6400:59130 99 1 92836373 255 20M1063N129M = S
```

我截取了一条比对信息
我们来看一下最后面的 NH:i:1 HI:i:1 AS:i:289 nM:i:0
NH:i:后面的数值代表着此条read比对到几个loci，1代表着unique map，数值大于1代表着multi-mapped
HI:i:后面的数值attributes enumerates multiple alignments of a read starting with 1，下游分析接cufflinks or stringtie的时候需要使用参数
AS:i:的数值代表着local alignment score (paired for paired-edn reads)
nM:i:的数值代表着the number of mismatches per (paired) alignment, not to be confused with
关于下游处理工具的兼容性还需要使用者自己仔细参考manual



- Log.out文件

Log.out文件记录了程序运行时的信息，可以用来回溯错误信息。

```
tail Log.out
Joined thread # 12
Completed: thread #13
Joined thread # 13
Joined thread # 14
Joined thread # 15
Joined thread # 16
Joined thread # 17
Joined thread # 18
Joined thread # 19
ALL DONE!
```

(/apps/
utm_sc
banner

- Log.progress.out文件

Log.progress.out报告比对进程情况，1分钟记录一次

```
tail Log.progress.out
Sep 08 17:57:52      33.1    23115987      285    94.1%    284.0    0.2%    4.0%
Sep 08 17:58:53      34.0    24349711      285    94.1%    284.0    0.2%    4.0%
Sep 08 18:00:23      33.5    24789186      285    94.1%    284.1    0.2%    4.0%
Sep 08 18:01:51      33.3    25493588      285    94.1%    284.0    0.2%    4.0%
Sep 08 18:02:58      33.5    26284824      285    94.1%    284.1    0.2%    4.0%
Sep 08 18:04:23      33.7    27163519      285    94.1%    284.1    0.2%    4.0%
Sep 08 18:05:36      33.1    27428080      285    94.1%    284.1    0.2%    4.0%
Sep 08 18:06:54      33.8    28659661      285    94.1%    284.1    0.2%    4.0%
Sep 08 18:08:00      34.3    29741743      285    94.1%    283.9    0.2%    4.0%
ALL DONE!
```

- Log.final.out文件

Log.final.out包含了比对结束后比对统计的信息

```
head Log.progress.out
      Time      Speed      Read      Read      Mapped      Mapped      Mapped      Mapped      Unr
      M/hr      number      length      unique      length      MMrate      multi      n
Sep 08 17:17:47      2.9      88583      288      94.2%      287.4      0.1%      4.0%
Sep 08 17:18:53     14.5     711158      282      94.1%      281.9      0.2%      4.0%
Sep 08 17:20:06     19.2    1329197      284      94.1%      283.8      0.2%      4.0%
Sep 08 17:21:19     32.7    2923414      284      94.1%      283.7      0.2%      4.0%
Sep 08 17:22:39     32.5    3629649      285      94.1%      283.9      0.2%      4.0%
Sep 08 17:23:49     32.4    4248206      285      94.1%      284.0      0.2%      4.0%
Sep 08 17:24:57     36.6    5483555      285      94.1%      284.3      0.2%      4.0%
Sep 08 17:26:03     35.7    6012744      285      94.1%      284.4      0.2%      4.0%
tail Log.progress.out
Sep 08 17:57:52      33.1    23115987      285    94.1%    284.0    0.2%    4.0%
Sep 08 17:58:53      34.0    24349711      285    94.1%    284.0    0.2%    4.0%
Sep 08 18:00:23      33.5    24789186      285    94.1%    284.1    0.2%    4.0%
Sep 08 18:01:51      33.3    25493588      285    94.1%    284.0    0.2%    4.0%
Sep 08 18:02:58      33.5    26284824      285    94.1%    284.1    0.2%    4.0%
Sep 08 18:04:23      33.7    27163519      285    94.1%    284.1    0.2%    4.0%
Sep 08 18:05:36      33.1    27428080      285    94.1%    284.1    0.2%    4.0%
Sep 08 18:06:54      33.8    28659661      285    94.1%    284.1    0.2%    4.0%
Sep 08 18:08:00      34.3    29741743      285    94.1%    283.9    0.2%    4.0%
ALL DONE!
```



- SJ.out.tab文件

SJ.out.tab包含了剪切信息，其实目前我还没怎么看懂，等以后再来补坑。

```
head SJ.out.tab
1 14830 14969 2 2 0 1 9 69
1 14844 14969 2 2 0 0 2 30
1 15039 15795 2 2 1 2 7 53
1 15948 16606 2 2 1 1 1 41
1 16028 16606 2 2 0 0 1 57
1 16311 16606 2 2 0 2 0 67
1 16766 16853 2 2 0 2 0 43
1 16766 16857 2 2 1 17 108 73
1 16766 16875 2 2 0 0 1 61
1 16789 16875 2 2 0 0 1 53

# 参数释义
column 1: chromosome
column 2: first base of the intron (1-based)
column 3: last base of the intron (1-based)
column 4: strand (0: undened, 1: +, 2: -)
column 5: intron motif: 0: non-canonical; 1: GT/AG, 2: CT/AC, 3: GC/AG, 4: CT/GC, 5: A
column 6: 0: unannotated, 1: annotated (only if splice junctions database is used)
column 7: number of uniquely mapping reads crossing the junction
column 8: number of multi-mapping reads crossing the junction
column 9: maximum spliced alignment overhang
```

(/apps/
utm_sc
banner

写在最后

其实我探究STAR的最终目的实现利用STAR的Chimeric and circular alignments.
我自己处理的数据里面存在着fusion-protein，而其余的比对软件暂时还没发现有这个功能的

当使用--chimSegmentMin参数的时候，STAR可以把read拆分为两部分，分别进行比对

STAR-Fusion是一个package，可以承接STAR的chimeric output，点我看代码
(<https://links.jianshu.com/go?to=https%3A%2F%2Fgithub.com%2FSTAR-Fusion%2FSTAR-Fusion>.)

当然STAR还可以做2-pass mapping，可以detect more splicesreads mapping to novel junctions



使用--quantMode GeneCounts参数还可以达到HTSeq的效果哦，可以帮你生成count matrix，省去你HTSeq的功夫，有空回来做一个比对，看HTSeq和GeneCounts的效率。

(/apps/
utm_sc
banner

参考文献：

<https://github.com/alexdobin/STAR/blob/master/doc/STARmanual.pdf>

([https://links.jianshu.com/go?](https://links.jianshu.com/go?to=https%3A%2F%2Fgithub.com%2Falexdobin%2FSTAR%2Fblob%2Fmaster%2Fdoc%2FSTARmanual.pdf)

[to=https%3A%2F%2Fgithub.com%2Falexdobin%2FSTAR%2Fblob%2Fmaster%2Fdoc%2FSTARmanual.pdf](https%3A%2F%2Fgithub.com%2Falexdobin%2FSTAR%2Fblob%2Fmaster%2Fdoc%2FSTARmanual.pdf))



日常Bob镇楼

以下为高通量测序数据处理系列快速通道：

高通量测序数据处理学习记录（零）：NGS分析如何选择合适的参考基因组和注释文件
(<https://www.jianshu.com/p/58decf8fb6d6>)

高通量测序数据处理学习记录（一）：比对软件STAR的使用
(<https://www.jianshu.com/p/eca16bf2824e>)



小礼物走一走，来简书关注我

赞赏支持

高通量数据处理学习记录 (/nb/38671008)

举报文章 © 著作权归作者所有



面面的徐爷 (/u/fe854ffa1f9e) ♂

写了 53844 字，被 674 人关注，获得了 360 个喜欢
(/u/fe854ffa1f9e)

+ 关注

中科院上海生科院硕博连读生。

喜欢 | 12



更多分享



下载简书 App ▶

随时随地发现和创作内容



/apps/redirect?utm_source=note-bottom-click

被以下专题收入，发现更多相似内容



生物信息学与算法 (/c/826a1e944a7d?

utm_source=desktop&utm_medium=notes-included-collection)



RNASeq ... (/c/a6473c639dd8?utm_source=desktop&utm_medium=notes-

included-collection)

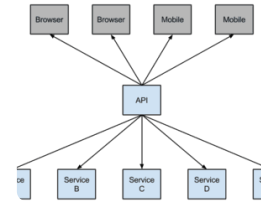


高通量测序数据处理 (/c/14f3bc38ad71?

utm_source=desktop&utm_medium=notes-included-collection)

(/apps/
utm_sc
banner

(/p/46fd0faecac1?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommend

Spring Cloud (/p/46fd0faecac1?utm_campaign=maleskine&utm_conte...

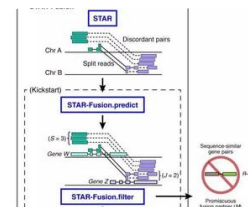
Spring Cloud为开发人员提供了快速构建分布式系统中一些常见模式的工具（例如配置管理，服务发现，断路器，智能路由，微代理，控制总线）。分布式系统的协调导致了样板模式，使用Spring Cloud开发人员可...



卡卡罗2017 (/u/d90908cb0d85?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommend

(/p/7092a2eb5727?



utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommend

我是如何学习Gene Fusion分析的 (/p/7092a2eb5727?utm_campaign=ma...

你看到的不仅仅是一个教程，更是一个自学经验！一、Fusion原理 基因融合（Gene fusion）是指将两个或多个基因的编码区首尾相连，置于同一套调控序列（包括启动子、增强子和终止子等）的控制之下，构成...



生信技能树 (/u/d645f768d2d5?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommend

Android - 收藏集 (/p/dad51f6c9c4d?utm_campaign=maleskine&utm_c...

Android 自定义View的各种姿势1 Activity的显示之ViewRootImpl详解 Activity的显示之ViewRootImpl初探 Activity的显示之Window和View Android系统的创世之初以及Activity的生命周期 图解Andro...



passiontim (/u/e946d18f163c?

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommend


(/p/1f5d13cc47f8?

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommend

RNA-seq数据分析---方法学文章的实战练习 (/p/1f5d13cc47f8?utm_camp...



前言 这次给大家带来的是16年发表在NATURE PROTOCOLS上面的一篇处理RNA-seq数据的文章：Transcript-level expression analysis of RNA-seq...

 面面的徐爷 (/u/fe854ffa1f9e?)




(/apps/
utm_sc
utm_s
commen
banner

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommenc

军训评优感悟 (/p/62e1b40cf60a?utm_campaign=maleskine&utm_conte...


流水，时光匆匆，半个月的军训生活即将结束，因为经历所以感动，因为感动所以伤感。从一开始的陌生，到现在的熟悉；从一开始的叫苦连天，到现在的乐趣盎然；从一开始的白皙，到现在的黑黝。我们原本分别...

 贪狼噬魂 (/u/f3f886290745?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommenc

只有九首歌曲的音乐文件夹 (/p/a287f357f95a?utm_campaign=maleskine...


在我读初中的时候，关于电子产品的最大奢望是一个MP3，管他是不是苹果或者索尼，只要能够出身，看着比那该死的复读机酷就行！我不是一个音乐发烧友，只是耳闻HiFi、森海萨尔、独立芯片，记得最多的时...

 akingm1949 (/u/de489139826d?)

utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommenc

冬 (/p/5d846b268ee5?utm_campaign=maleskine&utm_content=note&u...

文/陌宇轩 我无法 在萧瑟的冬天 勾画炫彩的画面 如果 享用更超大的奢华 我只能想象 异乡的壁炉

 小哲小诗 (/u/75844d4eee37?)


utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommenc

(/p/fee999b2e36d?)

utm_campaign=maleskine&utm_content=note&utm_medium=seo_notes&utm_source=recommenc

读书的三种境界 (/p/fee999b2e36d?utm_campaign=maleskine&utm_con...

这是 清水一点通 日更的第 186篇，希望能帮助你。现在的读书人越来越多了，这是件好事，读书使人进步。只有读书人才知道除了读书本身外，还能...

 清水一点通 (/u/083de874314b?)



utm_campaign=maleskine&utm_content=user&utm_medium=seo_notes&utm_source=recommenc

