DNA测序数据比对软件BWA原理及应用

一、介绍

通过二代测序我们可以获得很多reads（一条条碱基序列），但是从只看这些reads，我们得不到任何信息。将高通量测序读取（reads）数据比对（映射）参考基因组是RNA-seq数据分析中的关键步骤。将序列读取数据映射到参考基因组有助于基因发现、基因定量、剪接变体（可变剪接）分析、变异调用以及识别嵌合（融合）基因。所以如果我们想知道reads是从基因组的哪个位置转录过来的，就需要将这些reads比对到参考基因组上，从而进行下一步分析。

STAR（Spliced Transcripts Alignment to a Reference）是用于将RNA-seq读取数据与参考基因组序列进行高度准确和超快速的剪接感知（*splice aware*）比对的工具。注意，STAR是一个专门针对RNA-seq数据映射的比对工具，这意味着不能用于比对DNA数据。与其它的RNA-seq比对工具相比，其具有较高的准确率，映射速度较其他比对软件高 50 多倍。STAR在识别经典和非经典剪接位点方面具有很高的精确性，还可以检测到嵌合（融合）转录本。除了映射短读取数据（例如≤ 200 bp），STAR还可以准确地映射长读取数据（例如来自PacBio或Ion Torrent的数Kbp读取数据）。STAR在变异检测（SNP和INDEL）方面具有更好的灵敏度，因此，STAR被用于GATK最佳实践工作流程，用于从RNA-seq数据中识别短变异。

STAR的缺点是它是一种对RAM（内存）要求较高的比对工具，所以平时我们在处理大量的数据样本时需要高性能计算平台的帮助，而且STAR的比对速度可能会根据可用内存而有所不同。

二、原理

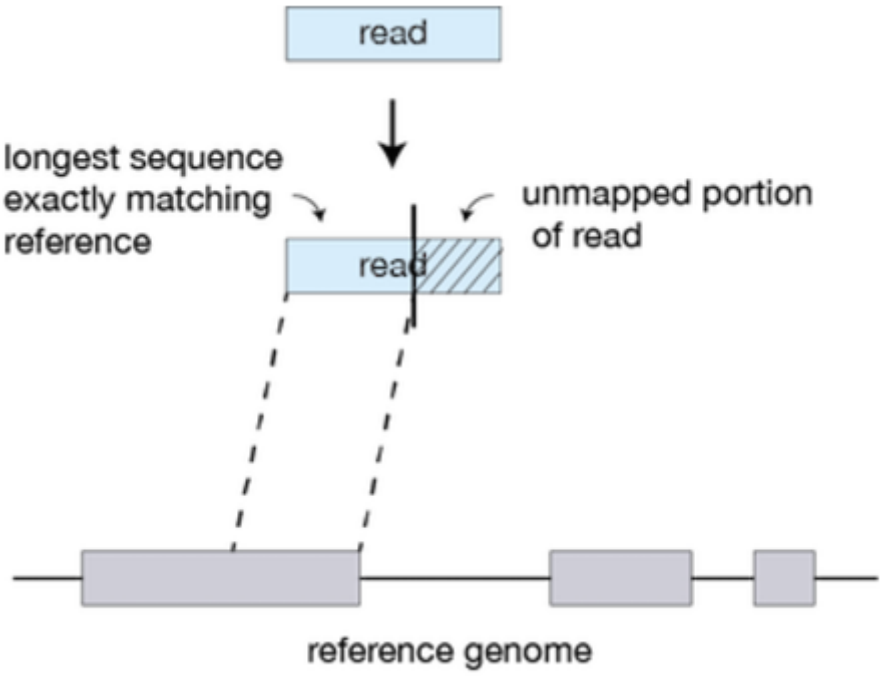
STAR在比对过程中，主要分为2步：

（1）Seed search

（2）Clustering, stitching and scoring

1、Seed search

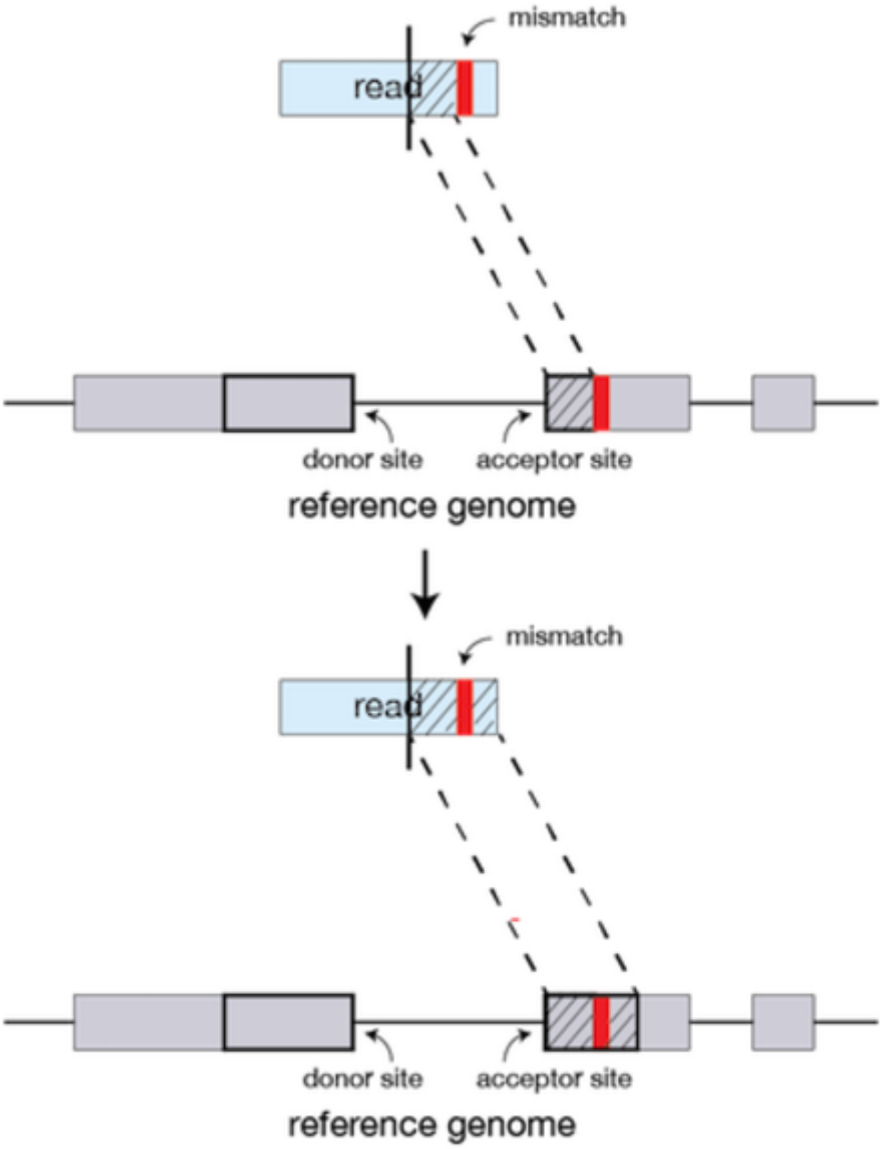
这部分可以算作是STAR工具的核心。STAR先对一个Read在参考基因组上去搜索，找到一个完全匹配的最长序列。这个最长的匹配序列称为最大可映射前缀（Maximal Mappable Prefix，MMP）。



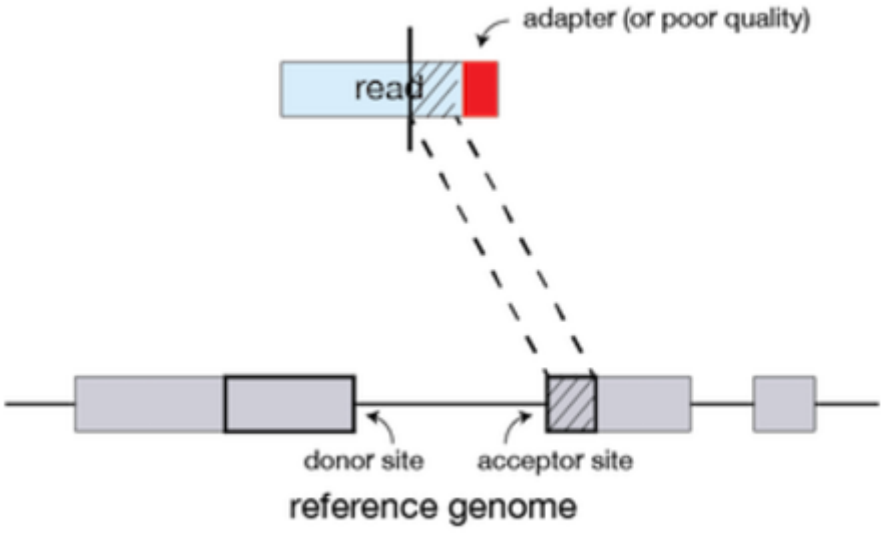
匹配到的 Read 的不同部分称为“seed”。所以对齐到基因组的第一个 MMP 称为*seed1*。然后，STAR 将再次搜索Read的未映射部分，以找到与参考基因组完全匹配的下一个最长序列 MMP，即 *seed2*。

其他意外情况：

如果由于存在一个或多个错配而导致 *MMP* 搜索未到达Read的末尾，*则MMP*将作为基因组中可以扩展的锚点，从而允许与错配进行比对。

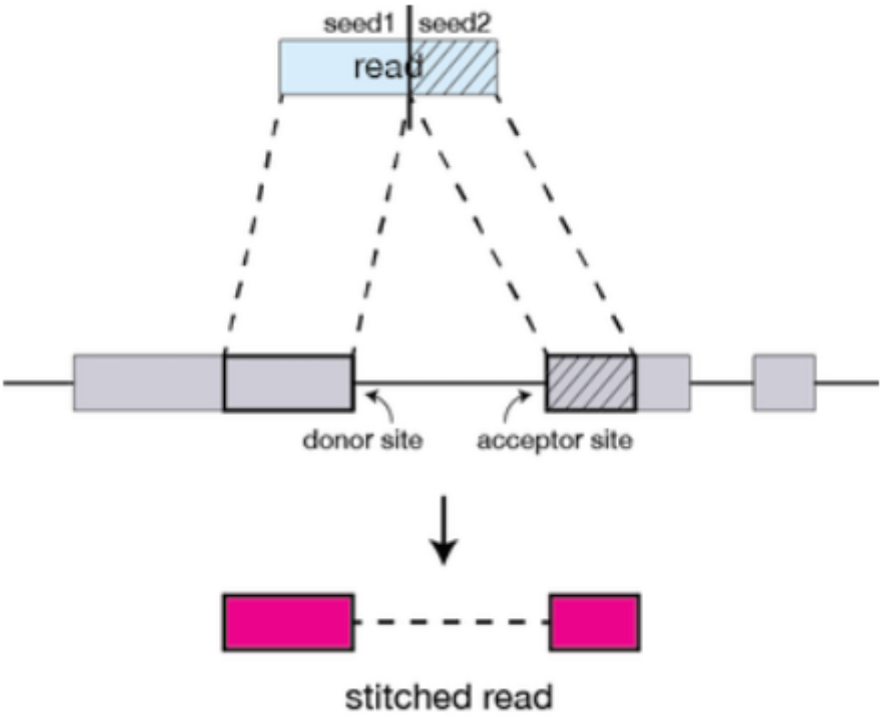


在某些情况下，扩展程序不会产生良好的基因组比对，从而可以识别 poly-A 尾、文库接头序列或测序质量差的尾部，这些将会被软件剪切。



2、Clustering, stitching and scoring

基于与一组‘anchor’种子或非多重映射种子的接近程度将种子聚集在一起，将单独的种子聚集在一起以创建完整的读取。然后根据读取的最佳对齐方式将种子拼接在一起（基于不匹配、插入缺失、间隙等进行评分）



三、使用

1、安装

推荐使用conda安装

```
conda install bioconda::star
```

2、建立索引

可在GENCODE等数据库下载人的基因组文件（FASTA格式）和基因组注释文件（GTF或GFF3格式）用来创建基因组索引

```
#PBS -N STAR_index
#PBS -l nodes=1:ppn=10
#PBS -l walltime=50:00:00
#PBS -S /bin/bash
```

```
#PBS -q pub_fat
cd $PBS_O_WORKDIR

STAR --runThreadN 10 \
     --runMode genomeGenerate \
     --genomeDir STAR_index \
     --genomeFastaFiles GRCh38.p14.genome.fa.gz \
     --sjdbGTFfile gencode.v47.annotation.gtf.gz \
     --sjdbOverhang 149
```

--runThreadN 线程数 :设置线程数
--runMode genomeGenerate : 设置模式为构建索引
--genomedDir 索引文件存放路径 : 必须先创建文件夹
--genomeFastaFiles 基因组fasta文件路径
--sjdbGTFfile gtf文件路径 : 可选项，高度推荐,用于提高比对精确性
--sjdbOverhang 读段长度: 后续回帖读段的长度。一般为设置为最大读长-1，目前一般测序读长为150bp。默认值为100

3、比对

```
#PBS -N STAR_run
#PBS -l nodes=1:ppn=10
#PBS -l walltime=50:00:00
#PBS -S /bin/bash
#PBS -q pub_fat
cd $PBS_O_WORKDIR

STAR --runThreadN 10 \
     --runMode alignReads \
     --genomeDir STARindex \
     --readFilesIn Read1.fastq.gz Read2.fastq.gz \
     --readFilesCommand zcat \
     --outSAMtype BAM SortedByCoordinate \
     --outFileNamePrefix outdir/out_index \
     --twopassMode Basic
```

--runThreadN 设置线程数
--runMode alignReads : 默认就是比对模式，可以不填写
--genomeDir: 索引文件夹
--readFilesIn FASTA/Q文件路径

--readFilesCommand zcat: 如果输入格式是gz结尾，那么需要加上zcat，否则会报错。如果输入格式不是压缩格式则不需要设置此命令

--outSAMtype: 输出SAM文件的格式，是否排序

--outFileNamePrefix：指定文件夹和前缀

--twopassMode：启用Twopass模式，先按索引进行第一次比对，而后把第一次比对发现的新剪切位点信息加入到索引中进行第二次比对。这个参数可以保证更精准的比对情况，但是费时也费内存。如果不需要Twopass模式则可以不指定此参数

4、比对结果统计

比对结束后，需要了解比对结果的情况，可以采用samtools flagstat进行统计

```
samtools flagstat test.bam >  flagstat.tax
```

samtools flagstat统计bam文件比对后每一个参数的解释如下：

```
14608455 + 0 in total (QC-passed reads + QC-failed reads) ## reads 总数
37967 + 0 secondary ##出现比对到参考基因组多个位置的reads数
0 + 0 supplementary ##可能存在嵌合的reads数
0 + 0 duplicates ##重复的reads数
14590894 + 0 mapped (99.88% : N/A) ##比对到参考基因组上的reads数
14570488 + 0 paired in sequencing ##属于PE read的reads总数。
7285244 + 0 read1 ##PE read中Read 1 的reads 总数。
7285244 + 0 read2 ##PE read中Read 2 的reads 总数。
14507068 + 0 properly paired (99.56% : N/A) ##完美比对的reads总数。PE 两端reads比对到同一条序列，且根据比对结果推断的插入片段大小符合设置的阈值。
14551500 + 0 with itself and mate mapped ##PE两端reads都比对上参考序列的reads总数。
1427 + 0 singletons (0.01% : N/A) ##PE两端reads，其中一端比上，另一端没比上的reads总数。
26260 + 0 with mate mapped to a different chr ##PE read中，两端分别比对到两条不同的序列的reads总数。
17346 + 0 with mate mapped to a different chr (mapQ>=5) ##PE read 中，两端分别比对到两条不同的序列，且mapQ>=5的reads总数。
```

主要查看比对率是否正常，一般都在99%左右