華中農業大學
HUAZHONG AGRICULTURAL UNIVERSITY

生物信息学进展

实验报告

题目：ChIP-Seq 数据分析（序列比对）- 中期报告

姓名：张子栋
学号：2020317210101
学院：信息学院

湖北 · 武汉
2023 年 3 月 2 日

目录

1 实验材料

1.1 线虫参考基因组

1.1.1 WormBase

WormBase 是一个专门收录线虫的基因组信息的数据库，它支持使用线虫作为模式生物的研究者，提供了线虫的基因组序列、注释、变异、表达、互作等数据。WormBase 还包括一个子项目 WormBase ParaSite，它收录了其他线虫和扁形动物寄生虫的基因组信息。参考基因组是一个完整的线虫基因序列，基因组注释文件是对参考基因组中的基因、转录本、外显子等特征的描述。

通过 WormBase : Nematode Information Resource 网站获取线虫参考基因组。

1.2 线虫 ChIP-seq 数据

来自于上次实验。

1.3 实验软件

1.3.1 BWA

BWA 是一种能够将差异度较小的序列比对到一个较大的参考基因组上的软件包。它由三个不同的算法组成：BWA-backtrack，BWA-SW 和 BWA-MEM。BWA-backtrack 适用于比对 Illumina 的序列，reads 长度最长能到 100 bp。BWA-SW 和 BWA-MEM 适用于比对长序列，支持的长度为 70 bp - 1 Mbp；同时支持剪接性比对。

BWA 的使用需要先对参考基因组建立索引，然后再进行比对。比对的结果是一个 SAM 格式的文件，可以用 samtools 进行后续的处理。

- 建立索引

```
bwa index ref.fa
```

- 比对

```
bwa mem ref.fa reads.fq > aln.sam
```

- 查看帮助

```
bwa -h
```

1.3.2 Bowtie

Bowtie 是一个超快的，存储高效的短序列片段比对程序。它能够将短的 DNA 序列片段（reads）比对到人类基因组或其他较大的参考基因组上。它有两个版本：Bowtie 和 Bowtie2。Bowtie2 是 Bowtie 的升级版，能够比对更长的 reads，支持局部比对和剪接性比对。

Bowtie 的使用也需要先对参考基因组建立索引，然后再进行比对。比对的结果是一个 SAM 或 BAM 格式的文件，可以用 samtools 进行后续的处理。

- 建立索引

```
bowtie-build ref.fa ref
```

- 比对

```
bowtie -x ref -U reads.fq -S aln.sam
```

- 查看帮助

```
bowtie --help
```

Bowtie 和 BWA 都是常用的短序列比对工具，它们有一些相似之处，也有一些不同之处。

相似之处：

- 都是基于 Burrows-Wheeler Transform（BWT）的算法，利用后缀数组和 FM-index 进行比对。
  - BWT 是 Burrows-Wheeler Transform 的缩写，它是一种数据转换算法，可以把一个文本转换成一个相似的文本，使得相同的字符更容易聚集在一起，从而方便后续的压缩。BWT 的原理是对文本的所有循环移位进行字典序排序，然后取最后一列作为转换后的文本。BWT 是可逆的，也就是说可以从转换后的文本恢复原文本。
- 都能够比对单端或双端的 reads，支持多种格式的输入和输出。
- 都能够处理比对错误和插入缺失，但是对于较长的插入缺失，效果不佳。

不同之处：

- BWA 有两种模式：BWA-MEM 和 BWA-ALN，前者适用于较长的 reads（$>$ 70 bp），后者适用于较短的 reads（$<$ 70 bp）。Bowtie 有两个版本：Bowtie 和 Bowtie2，前者只支持全局比对，后者支持局部比对和剪接性比对。
- BWA 比 Bowtie2 更准确，但是 Bowtie2 比 BWA 更快。BWA 比 Bowtie 更敏感，但是 Bowtie 比 BWA 更节省内存。
- BWA 和 Bowtie2 都能够比对较长的 reads，但是 BWA-MEM 对于较长的 reads 更优于 Bowtie2。Bowtie 只能够比对较短的 reads（$<$ 50 bp）。
- BWA 和 Bowtie2 都能够处理比对错误和插入缺失，但是 BWA-MEM 对于较短的插入缺失更优于 Bowtie2。Bowtie 只能够处理比对错误，不能处理插入缺失。

1.3.3 Samtools

samtools 是一个用于处理 SAM 或 BAM 格式的比对结果的软件包，它可以实现以下功能：

- 转换 SAM 和 BAM 格式，例如：

```
samtools view -bS in.sam > out.bam
```

- 排序和合并 BAM 文件，例如：

```
samtools sort in.bam -o out.sorted.bam
```

- 去除 PCR 重复，例如：

```
samtools rmdup in.bam out.rmdup.bam
```

- 统计比对结果，例如：

```
samtools flagstat in.bam
```

- 查看比对结果，例如：

```
samtools tview in.bam ref.fa
```

- 生成 BCF 文件，用于 SNP 和 Indel 的分析，例如：

```
samtools mpileup -uf ref.fa in.bam | bcftools call -c - > out.bcf
```

1.3.4 Bedtools

bedtools 是一个用于处理基因组数据的软件，它可以对 BED、BAM、GFF 等格式的文件进行各种操作，如求交集、并集、覆盖度、分组统计等。bedtools 的主要功能有：

- genomecov：计算基因组的覆盖度，即某些特征覆盖了基因组的哪些部分。
- groupby：对文件或流按照指定的列进行分组，并对另一列进行统计，类似于数据库的 “group by” 语句。
- intersect：计算两个文件或流中的特征的交集，即哪些特征在两个文件或流中都存在。
- merge：合并两个文件或流中的特征，即将有重叠的特征合并为一个特征。
- sort：对文件或流按照某些列进行排序，以便于其他操作。

2 实验过程及结果

2.1 下载数据

2.1.1 获取线虫参考基因组

下载线虫参考基因组到服务器。

```
wget --no-check-certificate https://downloads.wormbase.org/releases/current-production-release/species/c_elegans/PRJNA275000/c_elegans.PRJNA275000.WS286.genomic.fa.gz
```

2.2 BWA 比对

2.2.1 建立参考基因组索引

使用 BWA 软件进行比对之前，需要先建立索引，索引是一种数据结构，可以加快比对的速度，减少内存的占用。建立索引的原理是将参考基因组分割成多个子序列，然后对每个子序列建立一个哈希表，记录每个 k-mer 的出现位置。比对的时候，BWA 会将 reads 也分割成多个子序列，然后在哈希表中查找匹配的位置，从而找到最佳的比对位置。

具体命令：

```
bwa index c_elegans.PRJNA275000.WS286.genomic.fa.gz
```

得到文件：

```
bowtie_index
|-- c_elegans.PRJNA275000.WS286.genomic.fa
```

2.2.2 BWA 比对

具体命令：

```
bwa mem ~/bwa_index/c_elegans.PRJNA275000.WS286.genomic.fa
~/SRR14325856.processed.fastq > ~/bwa_result/c_elegans_ChIP-Seq.sam
```

得到文件：

```
bwa_result
|-- c_elegans_ChIP-Seq.sam
```

2.3 Bowtie 比对

2.3.1 建立参考基因组索引

使用 Bowtie2 软件进行比对之前，也需要先建立索引，索引的作用和原理与 BWA 软件类似，都是为了加快比对的速度，减少内存的占用。建立索引的方法是将参考基因组分割成多个子序列，然后对每个子序列建立一个 Burrows-Wheeler 变换（BWT）和后缀数组（SA），记录每个 k-mer 的出现位置。比对的时候，Bowtie2 会将 reads 也分割成多个子序列，然后在 BWT 和 SA 中查找匹配的位置，从而找到最佳的比对位置。

具体命令：

```
bowtie2-build bowtie_index/c_elegans.PRJNA275000.WS286.genomic.fa
bowtie_index/c_elegans.PRJNA275000.WS286.genomic.bowtie
```

得到文件：

```
bowtie_index/
|-- c_elegans.PRJNA275000.WS286.genomic.bowtie.1.bt2
|-- c_elegans.PRJNA275000.WS286.genomic.bowtie.2.bt2
|-- c_elegans.PRJNA275000.WS286.genomic.bowtie.3.bt2
|-- c_elegans.PRJNA275000.WS286.genomic.bowtie.4.bt2
|-- c_elegans.PRJNA275000.WS286.genomic.bowtie.rev.1.bt2
|-- c_elegans.PRJNA275000.WS286.genomic.bowtie.rev.2.bt2
|-- c_elegans.PRJNA275000.WS286.genomic.fa
```

2.3.2 Bowtie 比对

具体命令：

```
bowtie2 -p 10 -x ~/bowtie_index/c_elegans.PRJNA275000.WS286.genomic.bowtie -U
~/SRR14325856.fastq -S ~/bowtie_result/c_elegans_ChIP-Seq.sam
```

生成文件：

```
bowtie_result/
|-- c_elegans_ChIP-Seq.sam
```

2.4 Samtools 统计比对结果

2.4.1 BWA 比对结果

```
samtools view -S -b ~/bwa_result/c_elegans_ChIP-Seq.sam | samtools flagstat - >
~/bwa_result/flagstat.txt
```

```
6479093 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
9 + 0 supplementary
0 + 0 duplicates
3428962 + 0 mapped (52.92% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

2.4.2 Bowtie 比对结果

```
samtools view -S -b ~/bowtie_result/c_elegans_ChIP-Seq.sam | samtools flagstat - > ~/bowtie_result/flagstat.txt
```

```
6492282 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 secondary
0 + 0 supplementary
0 + 0 duplicates
3405738 + 0 mapped (52.46% : N/A)
0 + 0 paired in sequencing
0 + 0 read1
0 + 0 read2
0 + 0 properly paired (N/A : N/A)
0 + 0 with itself and mate mapped
0 + 0 singletons (N/A : N/A)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)
```

3 实验总结

- 在进行比对之前，需要检查测序数据的质量，去除低质量的序列，去除接头序列，去除重复序列等，以提高比对的准确性和效率。
- 在进行比对之后，需要检查比对结果的质量，去除低质量的比对，去除重复的比对，去除错误的比对等，以提高后续分析的准确性和效率。
- 在进行比对和统计的过程中，需要注意数据的格式，文件的路径，参数的设置，命令的输出等，以避免出现错误或者异常 。

本文同步至 GitHub: MarkdownNotes/生信进展实验2.pdf at main · Bluuur/MarkdownNotes (github.com)