The Biostar Handbook: A Beginner's Guide to Bioinformatics[5]

#course

short read aligners

HTS mappers

> “Short read aligners are commonly used software tools in bioinformatics, designed to align a very large number of short reads (billions). ”

> “Most short read aligners will find only alignments that are reasonably similar to the target.
> This means that the algorithm “gives up” searching beyond a certain threshold.
> Short read aligners are designed to find regions of high similarity.
> Most short read aligners typically cannot handle long reads or become inefficient when doing so.
> There is also a limit to how short a read can be.
> ”

limitation of short read aligners

短序列比对工具则是要快速从众多潜在可选联配中找到最优的位置；而blast是找到最多的联配

> “The word “mapper” is often used to emphasize that the optimal alignment of a read is not guaranteed. The purpose of a mapper tool is locating a region in a genome, not producing an optimal alignment to that region.”
>
> “**Mapping**
>
> A mapping is a region where a read sequence is placed.
> A mapping is regarded to be correct if it overlaps the true region.
>
> **Alignment**

> An alignment is the detailed placement of each base in a read.
> An alignment is regarded to be correct if each base is placed correctly."
>
> "studies examining SNPs and variations in a genome would be primarily alignment-oriented."
>
> "studies focusing on RNA-Seq would be essentially mapping-oriented."

也就是说mapping侧重于把序列放到正确的位置，而不管这个序列的一致性，而alignment则是主要让序列和参考序列尽可能的配对，而不管位置。比如说变异检测就要优先保证alignment，而RNA-Seq则要尽可能保证把reads放到正确的位置。

> **"mental checklist"**
>
> **+ Can the aligner handle the volume of data relative to computational power**
> **+ Can the aligner be customized to report the type of data you need (all alignments or just best alignment, can you filter the output)**
> **+ Will the aligner produce attributes of the alignment that you need (alternative alignment locations, secondary alignment...)**

联配算法： 全局，局部还是半全局
需要报道非线性重排(non-linear arrangements)
比对工具如何处理InDels
比对工具支持可变剪切
比对工具能够过滤出符合需要的联配
比对工具能找到嵌合联配(chimeric alignments)

bwa

> "The new algorithm is the so-called bwa mem algorithm (where mem stands for Maximally Exact Matches)."

aln和mem分别处理低于100bp和大于70bp的短读

> "They first need to build an index from a known reference (this only needs to be done once).
> The reads in FASTA/FASTQ files are then aligned against this index."

建立索引和比对索引

> “The resulting file is in a so-called SAM (Sequence Alignment Map) format, one that you will undoubtedly love and hate (all at the same time). It is one of the most recent bioinformatics data formats, one that by today has become the standard method to store and represent all high-throughput sequencing results.”

SAM is one of the most recent bioinfomatics data formats

bowtie/bowtie2

> “There are two versions of bowtie. The latest version, bowtie2”

bowtie也有1和2两代，处理50bp以下和50bp以上的短读

> “They first need to build an index from a known reference (this only needs to be done once).
> The reads in FASTA/FASTQ format are then aligned against this index.”

建立索引和比对索引

bwa or bowtie2

proj > results > 2019-4-2

The performance of bwa is better than bowtie2, but if we modify the bowtie2 parameters, the alignment rate is improved:

```
bowtie2 --very-sensitive-local -x $REF -1 $R1 -2 $R2 > bowtie.sam
# 10s, 63.21%
time bowtie2 -D 20 -R 3 -N 1 -L 20 -x $REF -1 $R1 -2 $R2 > bowtie.sam
# 11s, 87.11%
```

> “Bowtie2 fills in a lot more information, it also it offers more ways to format the output and filter the alignments. So it has its applications and may be indispensable in some situations.”

> “The help for bowtie2 lists the parameters that are set when
> **—-very-sensitive-local** is set, so students started with those and kept tweaking and re-running the alignments”

也就是说bwa的默认参数是经过很好的优化来保证在默认参数下的结果，是不是我们都要选择bwa呢？也不能如此绝对，毕竟bowtie2的SAM结果保留了更多的信息。