WEEK 6

第3组

陈泳佐 2019012239

（p3-p5, p17-p20)

谢芸歆 2019012264

(p6-p16)

Content

I. Homework

II. Mapping

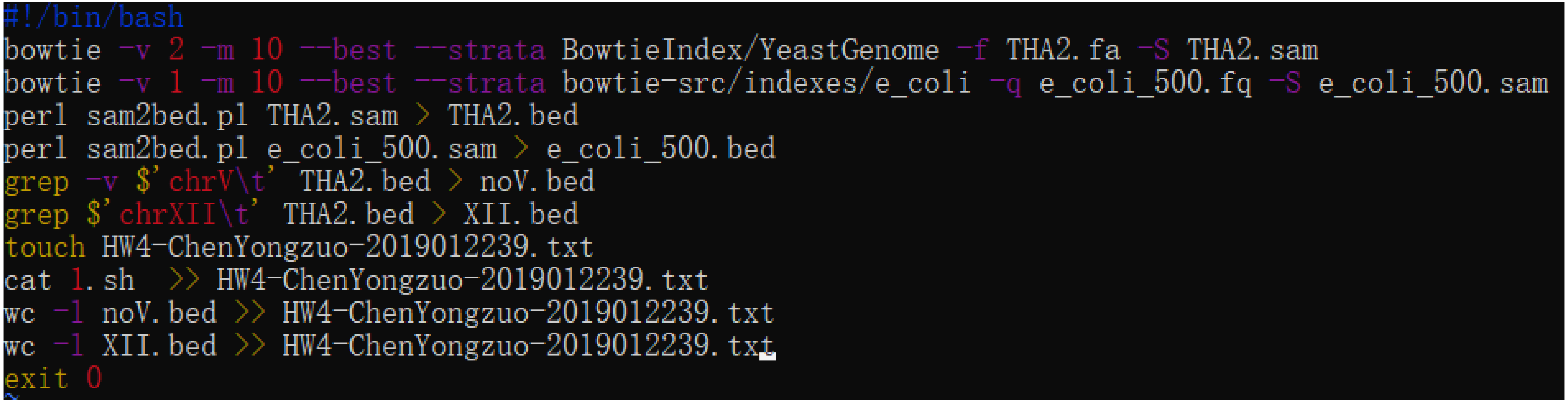
  I. DNA mapping

  II. RNA mapping

III. Genome browser resources

IV. NGS methods overview

1. Homework

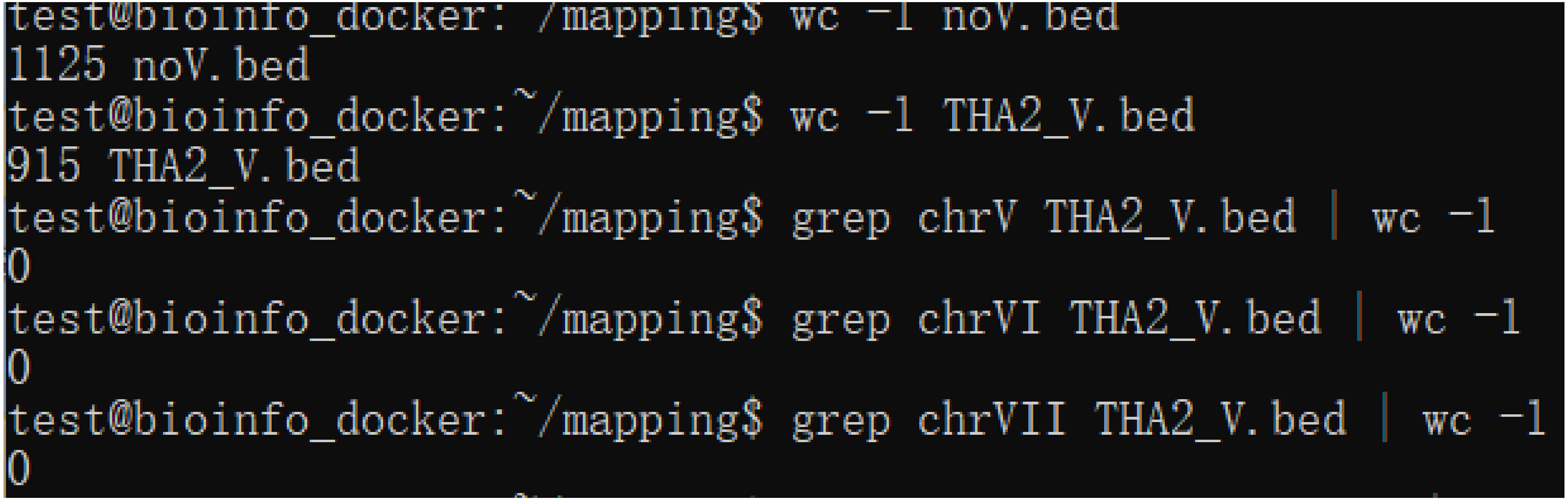


```
#!/bin/bash
bowtie -v 2 -m 10 --best --strata BowtieIndex/YeastGenome -f THA2.fa -S THA2.sam
bowtie -v 1 -m 10 --best --strata bowtie-src/indexes/e_coli -q e_coli_500.fq -S e_coli_500.sam
perl sam2bed.pl THA2.sam > THA2.bed
perl sam2bed.pl e_coli_500.sam > e_coli_500.bed
grep -v $'chrV\t' THA2.bed > noV.bed
grep $'chrXII\t' THA2.bed > XII.bed
touch HW4-ChenYongzuo-2019012239.txt
cat 1.sh  >> HW4-ChenYongzuo-2019012239.txt
wc -l noV.bed >> HW4-ChenYongzuo-2019012239.txt
wc -l XII.bed >> HW4-ChenYongzuo-2019012239.txt
exit 0
```

1.–v or –n ?

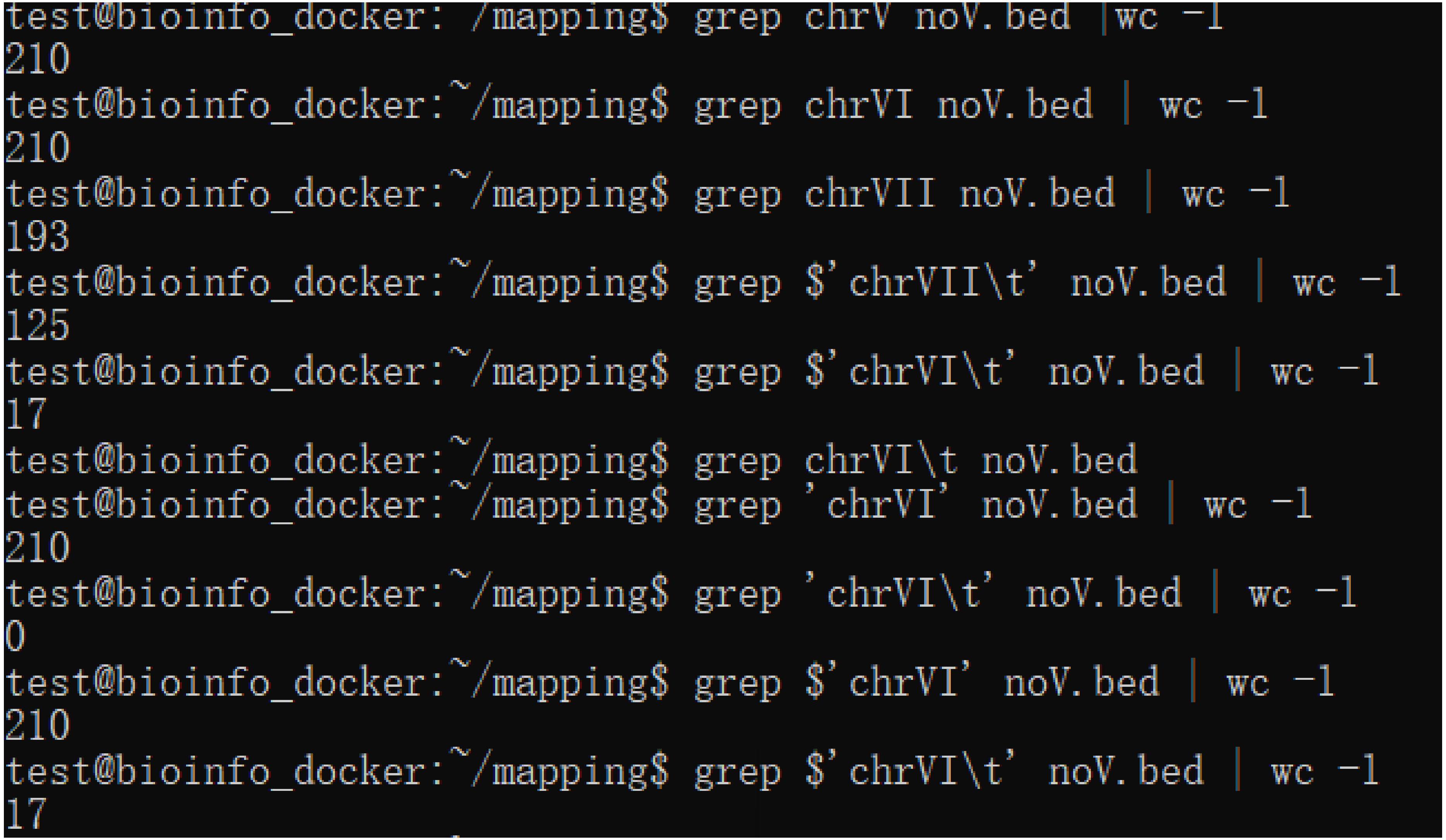
2.What is -m and what is –m --best --strata?

1. Homework



```
test@bioinfo_docker:~/mapping$ wc -l noV.bed
1125 noV.bed
test@bioinfo_docker:~/mapping$ wc -l THA2_V.bed
915 THA2_V.bed
test@bioinfo_docker:~/mapping$ grep chrV THA2_V.bed | wc -l
0
test@bioinfo_docker:~/mapping$ grep chrVI THA2_V.bed | wc -l
0
test@bioinfo_docker:~/mapping$ grep chrVII THA2_V.bed | wc -l
0
```

1. Homework



```
test@bioinfo_docker:/mapping$ grep chrV noV.bed |wc -l
210
test@bioinfo_docker:~/mapping$ grep chrVI noV.bed | wc -l
210
test@bioinfo_docker:~/mapping$ grep chrVII noV.bed | wc -l
193
test@bioinfo_docker:~/mapping$ grep $'chrVII\t' noV.bed | wc -l
125
test@bioinfo_docker:~/mapping$ grep $'chrVI\t' noV.bed | wc -l
17
test@bioinfo_docker:~/mapping$ grep chrVI\t noV.bed
test@bioinfo_docker:~/mapping$ grep 'chrVI' noV.bed | wc -l
210
test@bioinfo_docker:~/mapping$ grep 'chrVI\t' noV.bed | wc -l
0
test@bioinfo_docker:~/mapping$ grep $'chrVI' noV.bed | wc -l
210
test@bioinfo_docker:~/mapping$ grep $'chrVI\t' noV.bed | wc -l
17
```

2. Mapping

- **DNA** mapping — short reads map to genome: **bowtie, BWT, bowtie2** etc.
- **RNA** mapping
  - Map to **transcriptome** — similar to DNA mapping
  - Map to **genome** — have to consider splicing: **HISAT2, STAR** (TopHat2, Salmon)
  - **Reference-free** assembly — De novo assembly: **Trinity**

2.1 DNA Mapping

Common tools

- SOAP (2008): target single-end reads, suitable for small memory (4G)
- MAQ/BWA (2008): invented by Heng Li (BGI, 华大); BWA suitable for SNP analysis
- Bowtie (2009): strong in speed, not consider indels
- Bowtie2 (2012): different tool c.f. Bowtie, allow longer reads

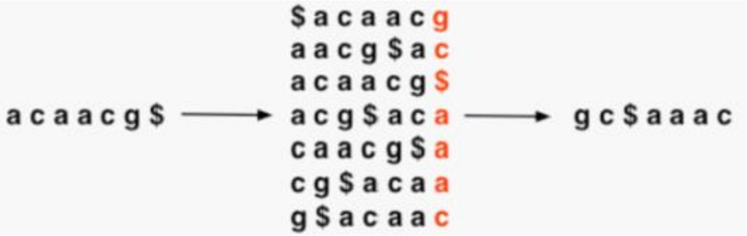
- Bowtie, Bowtie2 and BWA are all based on BWT, while MAQ is based on hash

https://www.plob.org/article/7181.html

2.1 DNA Mapping-BWT recall

Components of FM index

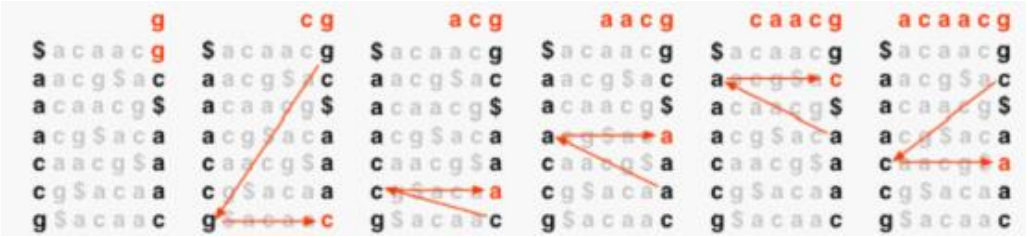
- BWT(T)

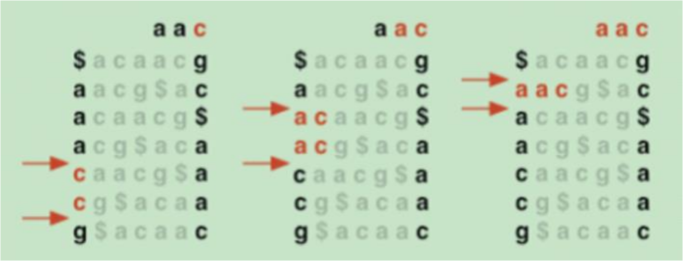


- Checkpoints: for every 448 characters, keep track of the positions

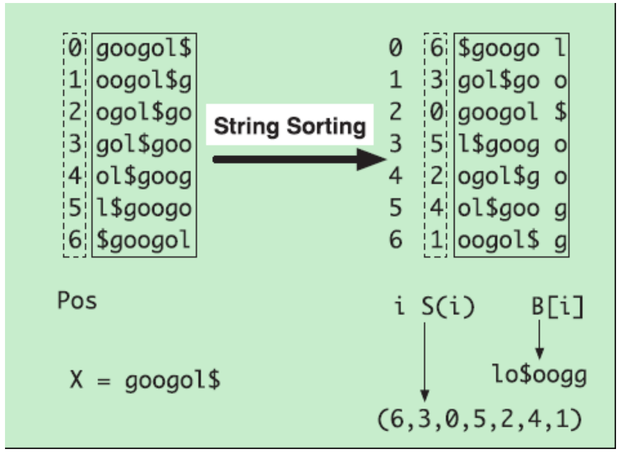
- Suffix array: SA[]

LF mapping: get unpermuted matrix



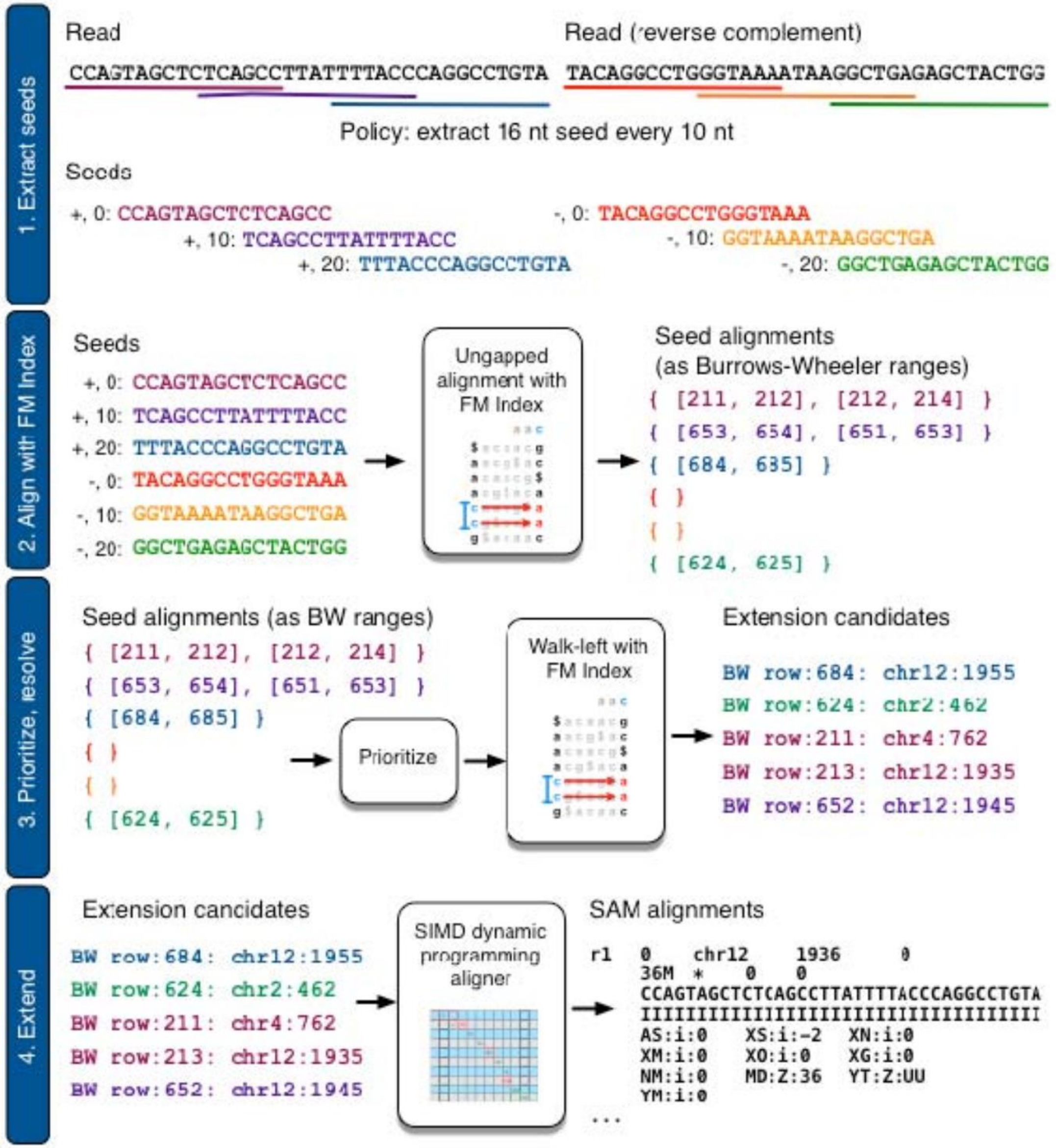


Exact match with BWT(T)



https://blog.csdn.net/stormlovetao/article/details/7048481

2.1 DNA Mapping-Bowtie2

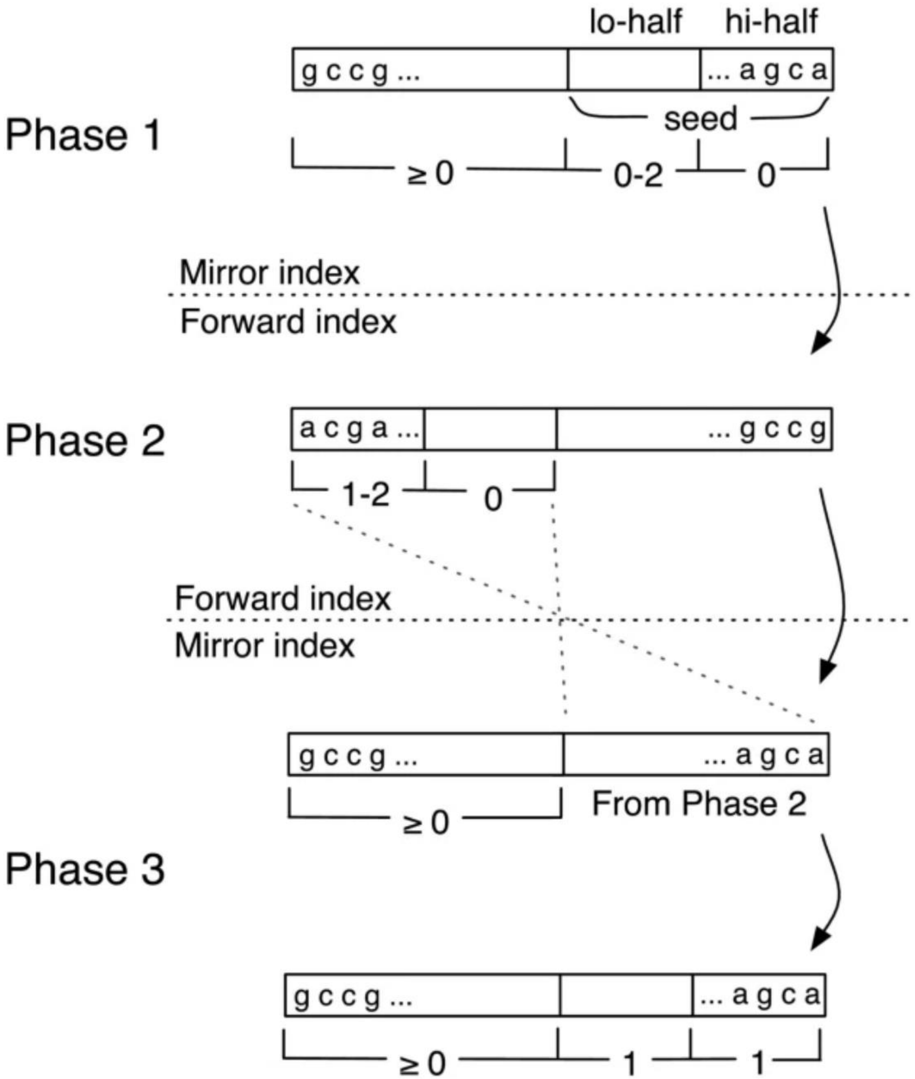


```
r1   0    chr12    1936     0
     36M  *    0    0
     CCAGTAGCTCTCAGCCTTATTTTACCCAGGCCTGTA
     IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
     AS:i:0    XS:i:-2   XN:i:0
     XM:i:0    XO:i:0    XG:i:0
     NM:i:0    MD:Z:36   YT:Z:UU
     YM:i:0
...
```

- Step1**: extract seed substrings** from the reads and the complements
- Step2: **Align** the substrings to ref genome ungapped using FM index, yielding Burrows-Wheeler (BW) ranges
- Step3: Prioritize BW ranges by their sizes, and randomly select the rows to **get the offset** to the ref genome by FM index “walk-left” procedure
- Step4: **Extension**, performing SIMD (Single Instruction Multiple Data)-accelerated **dynamic programming** in the vicinity of each alignments, until (1) all seeds examined (2) sufficient alignments examined (3) limit reached
  (gap allowed here)

https://www.nature.com/articles/nmeth.1923

2.1 DNA Mapping



Bowtie--Phased maq-like search

Seed: first 28 bases on the high-quality end of the read (14+14, hi/lo-half)

Example: 2 allowed mismatches

Case1: no mismatch in seed

Case2: no mismatch in hi-half, 1/2 mismatches in lo-half

Case3: no mismatch in lo-half, 1/2 mismatches in hi-half

Case4: 1 mismatch each in lo/hi-half

A three-phase approach minimizing backtracking

- **Phase 1**: use the mirror index to find alignments for cases 1 and 2.
- Phases 2 and 3: cooperate to find alignments for case 3, **phase 2** finds partial alignments with mismatches only in the hi-half, and **phase 3** extends those partial alignments into full alignments
- Finally, **phase 3** invokes the aligner to find alignments for case 4

https://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25

2.2 RNA Mapping-map to genome

Common tools

- HISAT2
- STAR
- TopHat2, …

- RNA mapping (to genome) requires additional consideration of splicing while also caring about indels, mismatches and multiple copies.

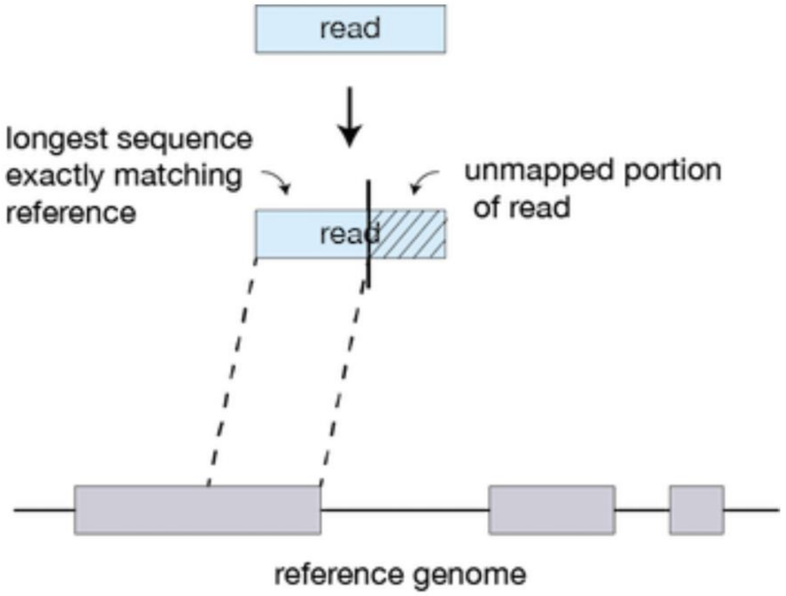
2.2 RNA Mapping-STAR aligner

2 main process:

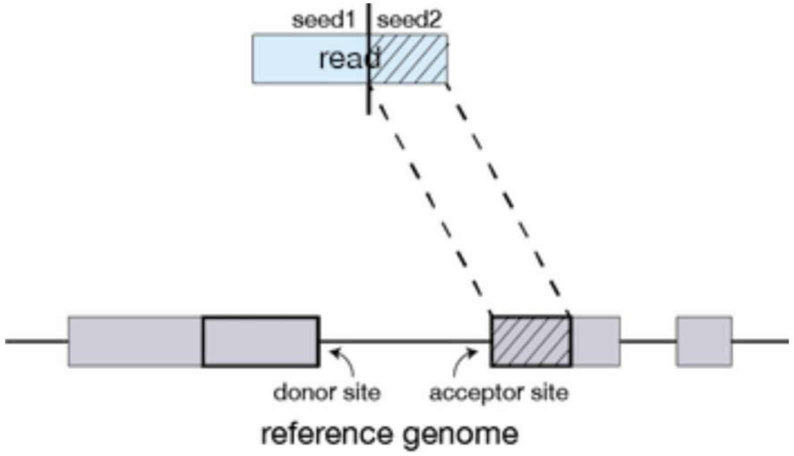
- Seed searching
- Clustering, stitching and scoring

Utilize SA[], split before iteration
-> high efficiency

Process 1:



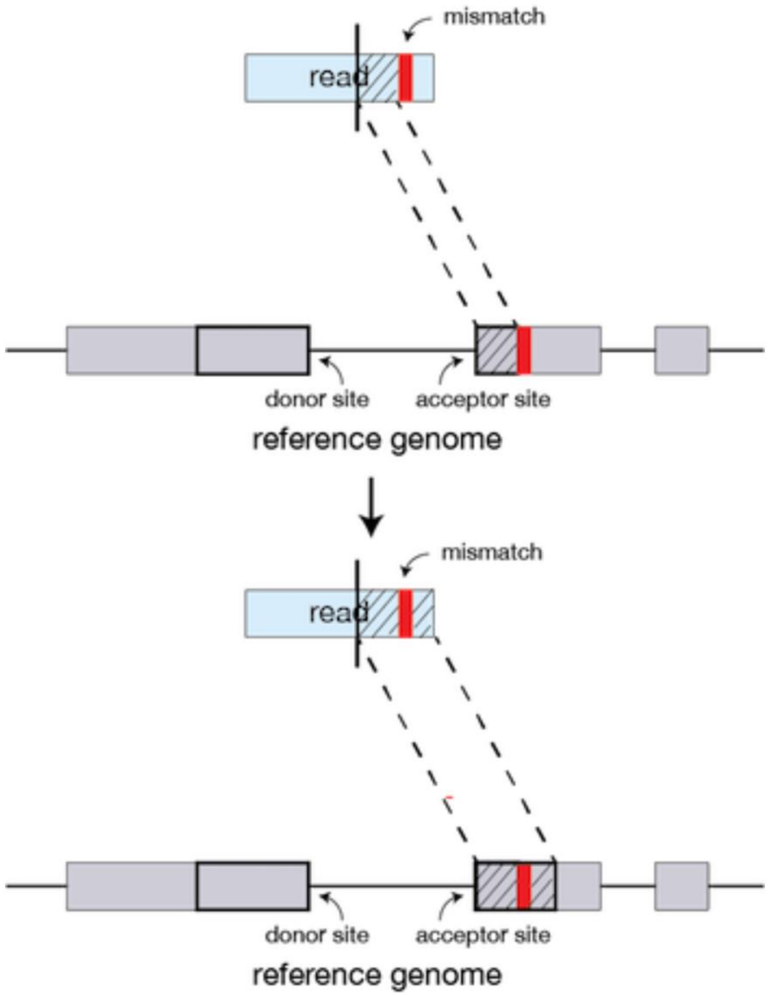
Maximal Mappable Prefixes (MMPs)



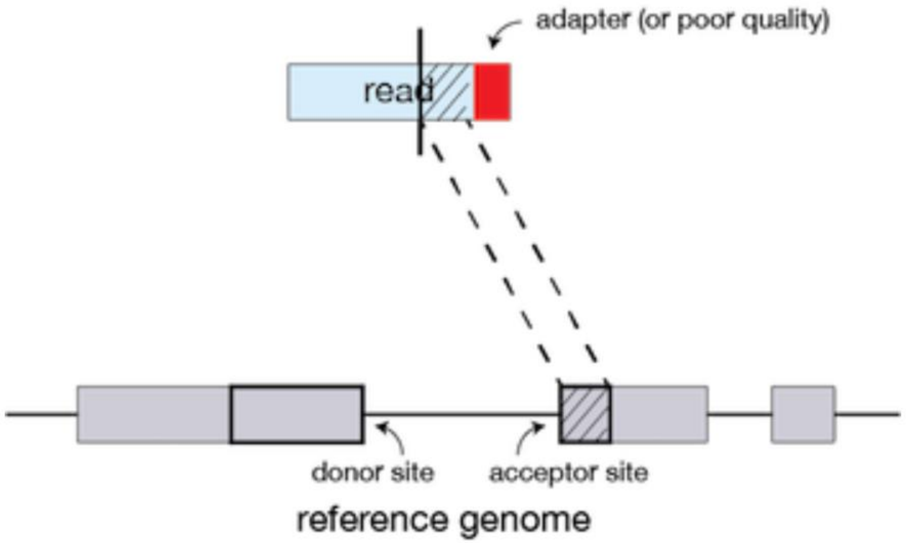
Search unmapped portion

https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

2.2 RNA Mapping-STAR aligner



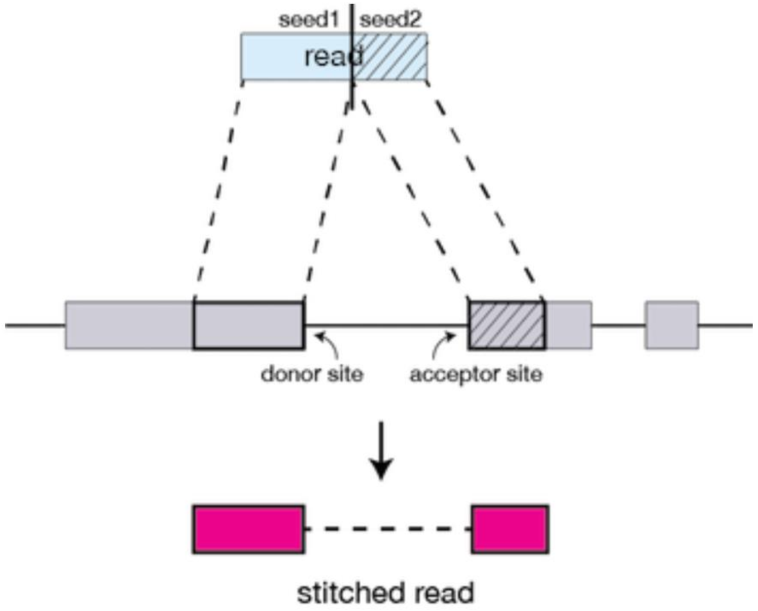
Mismatches and soft-clipping



https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

2.2 RNA Mapping-STAR aligner

Process 2: clustering, stitching and scoring



Anchor seeds: least multi-mapping

- **Clustering** is based on anchors (alignment windows)
- **Stitching** is based on **scoring** (mismatches, indels, gaps etc.)

Deletions that are longer than a user defined minimum intron size are considered splice junction (**gaps**)

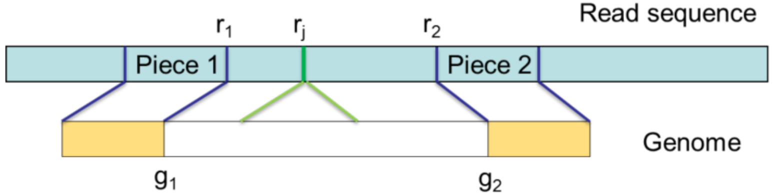
⇒ Gap opening + log(gap length)
GT/AG>GC/AG..

Scoring scheme

$$S = +\sum_{match} \overset{+1}{P_m} - \sum_{mismatch} \overset{-1}{P_{mm}} - \sum_{inserion} P_{ins} - \sum_{deletion} P_{del} - \sum_{gap} P_{gap}$$

$$P_{ins/del} = P_{ins/del}^{open} + P_{ins/del}^{extend} \cdot L_{ins/del}$$

Junction point selection



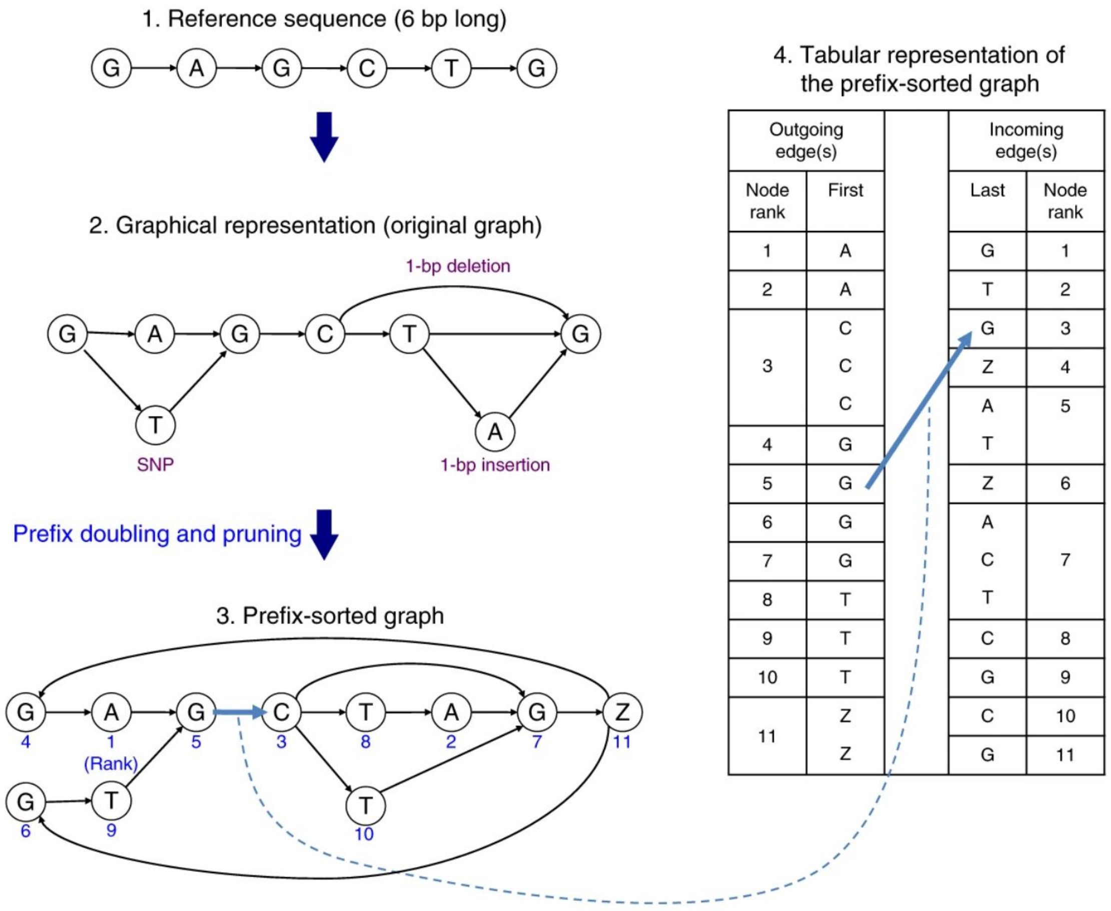
$$\max_{r_1<r_j<r_2}\left\{\sum_{r=1}^{r_j-r_1}\left[\begin{array}{ll}1 & if\ R(r_1+r)=G(g_1+r)\ \&\ R(r_1+r)\neq G(g_1+r+\Delta)\\ -1 & if\ R(r_1+r)\neq G(g_1+r)\ \&\ R(r_1+r)=G(g_1+r+\Delta)\\ 0 & otherwise\end{array}\right] - P_{gap}(r_j)\right\}$$

Select best alignment based on scores

https://academic.oup.com/bioinformatics/article/29/1/15/272537?login=true

https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/03_alignment.html

2.2 RNA Mapping-HISAT2


1. Reference sequence (6 bp long)
G A G C T G
2. Graphical representation (original graph)
1-bp deletion
G A G C T G
T
SNP
A
1-bp insertion
Prefix doubling and pruning
3. Prefix-sorted graph
G 4
A 1
(Rank)
G 5
C 3
T 8
A 2
G 7
Z 11
G 6
T 9
T 10
4. Tabular representation of the prefix-sorted graph

Outgoing edge(s)		Incoming edge(s)	
Node rank	First	Last	Node rank
1	A	G	1
2	A	T	2
3	C	G	3
	C	Z	4
	C	A	5
4	G	T	
5	G	Z	6
6	G	A	7
7	G	C	
8	T	T	
9	T	C	8
10	T	G	9
11	Z	C	10
	Z	G	11

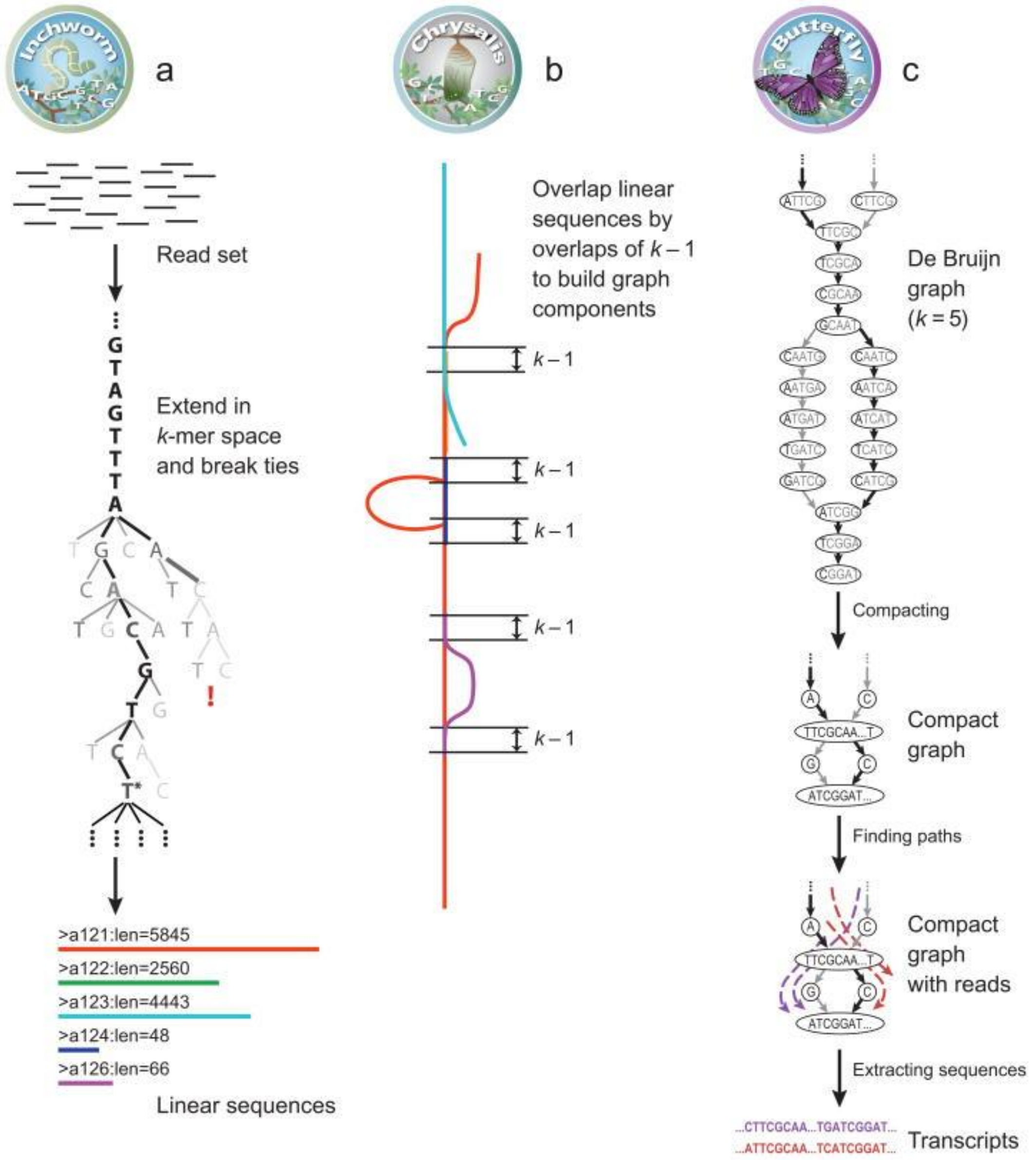
HISAT2 (hierarchical indexing for spliced alignment of transcripts 2)

- align both DNA and RNA sequences using a graph Ferragina Manzini index
- A conservative method, more precision but less recall
- TopHat2 announced preceded

https://www.nature.com/articles/s41587-019-0201-4

https://www.jianshu.com/p/681e02e7f9af

2.2 RNA Mapping-de novo assembly



Trinity

- Consist of 3 components, Inchworm, Chrysalis and Butterfly.
- Rely on partitioning the sequence data into individual disconnected graphs
- Process independently to extract full-length isoforms and gain paralogous gene transcripts

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3571712/#!po=14.0000

Mapping总结

存在的问题：

1.序列比对 or 全局比对？

2.对全基因组遍历？

3.如何界定容忍度？

实现流程：

1.建立索引：Hashing/BWT-FM

2.确定mapping的位置（单一seed/hybrid seed）

3.序列比对（动态规划/非动态规划)

3. Genome Browser

IGV:

官网下载：https://software.broadinstitute.org/software/igv/download

视频：https://youtu.be/6_1ZcVw7ptU

https://www.bilibili.com/video/av30448472/

https://cloud.tsinghua.edu.cn/d/ad22768345664924b202/files/?p=%2FVideo%2FNGS%20Data%20Analysis%2FGenome%20Browser%20-%20IGV%20-%20Zhiyu%20Xu.mp4

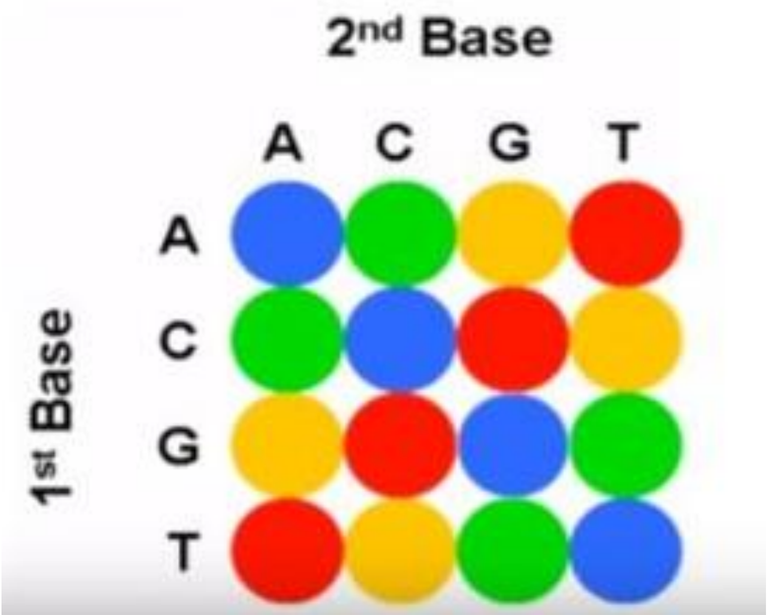
UCSC：

网站：https://genome.ucsc.edu/

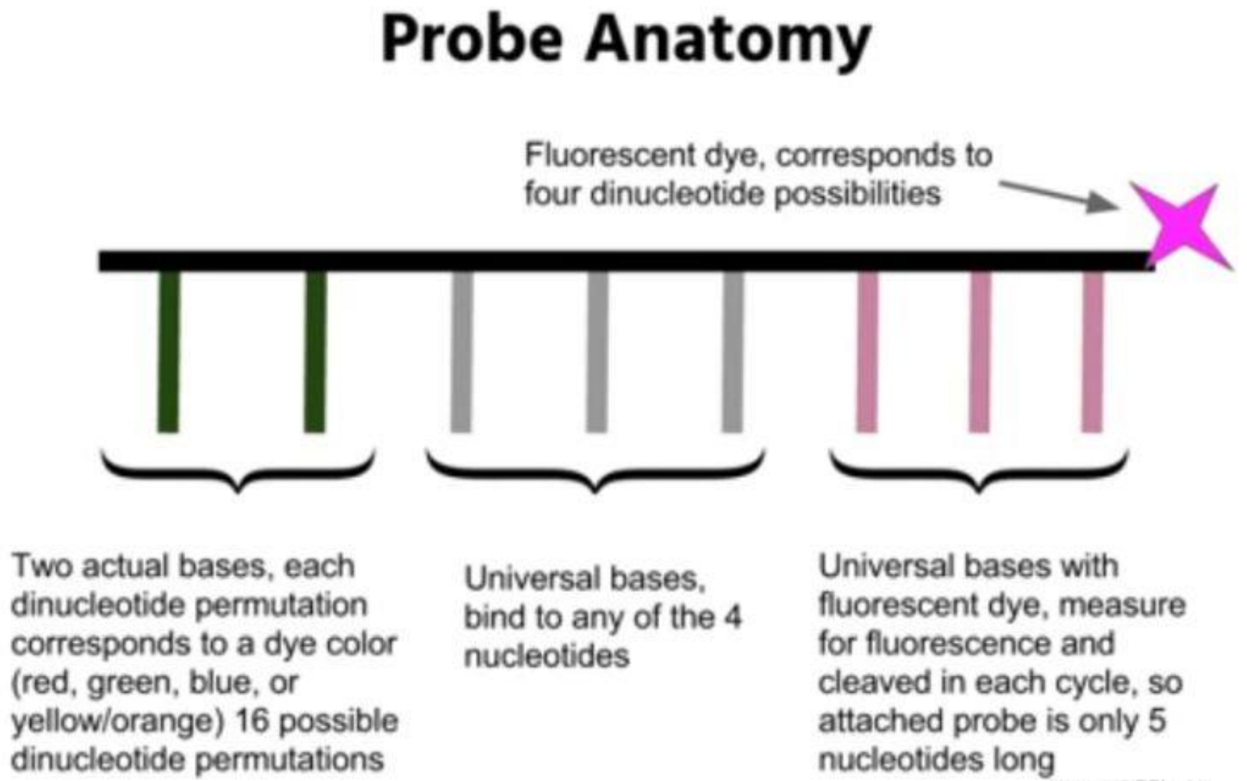
视频：https://cloud.tsinghua.edu.cn/d/ad22768345664924b202/files/?p=%2FVideo%2FNGS%20Data%20Analysis%2FGenome%20Browser%20-%20UCSC%20-%20Zhiyu%20Xu.mp4

4. Next Generation Sequencing (NGS)

1. 454焦磷酸测序法
2. Illumina 的 Solexa DNA簇测序法
3. ABI的SOLiD平台测序法



	A	C	G	T
A	🔵	🟢	🟡	🔴
C	🟢	🔵	🔴	🟡
G	🟡	🔴	🔵	🟢
T	🔴	🟡	🟢	🔵



Probe Anatomy

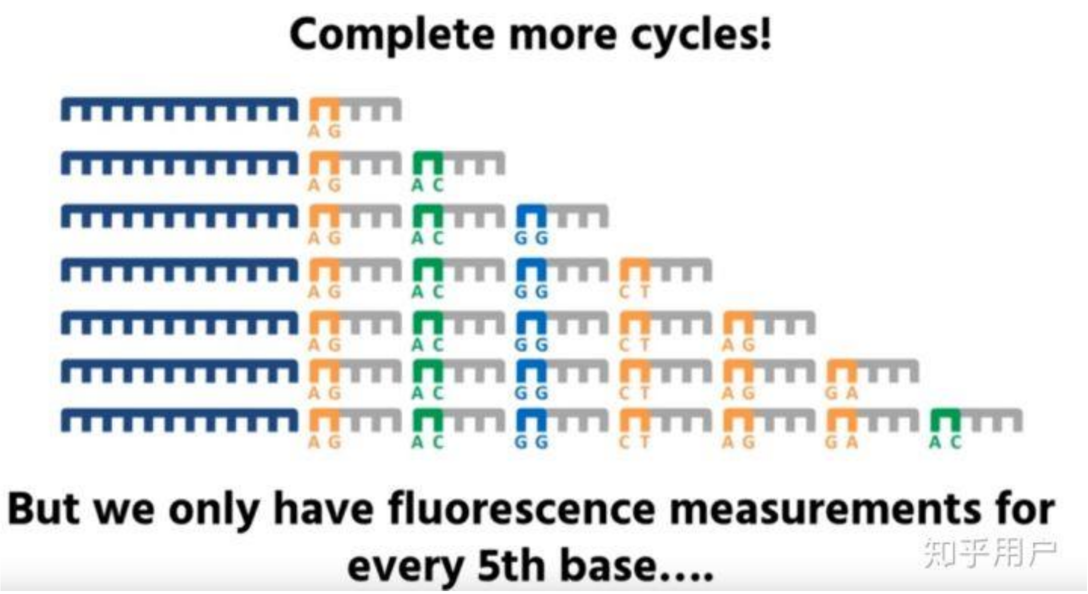
Fluorescent dye, corresponds to four dinucleotide possibilities

Two actual bases, each dinucleotide permutation corresponds to a dye color (red, green, blue, or yellow/orange) 16 possible dinucleotide permutations

Universal bases, bind to any of the 4 nucleotides

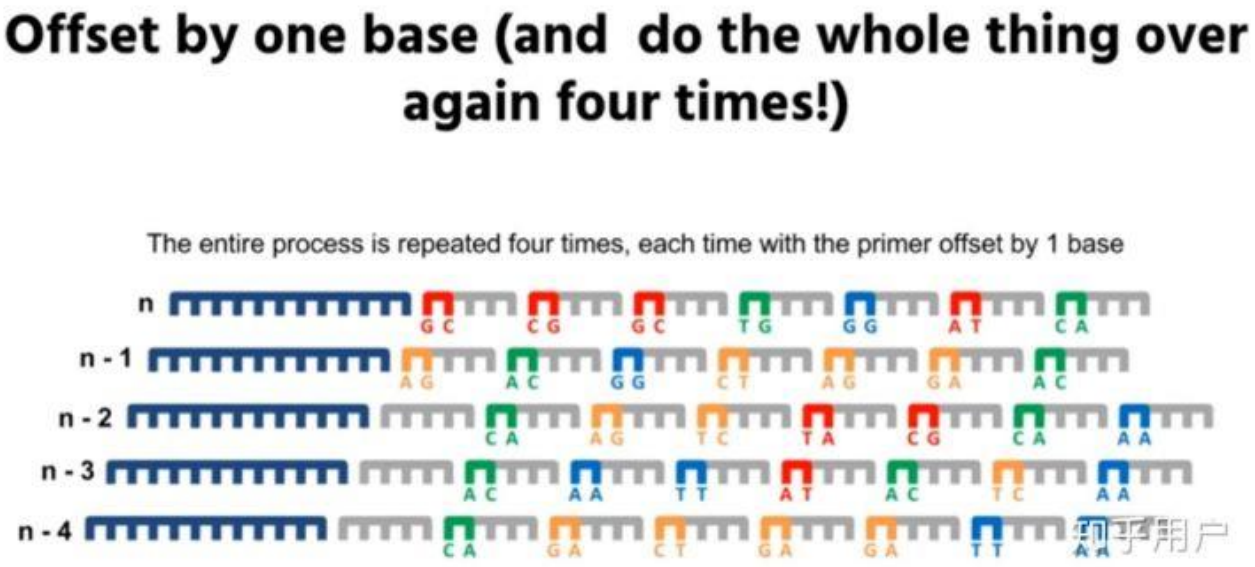
Universal bases with fluorescent dye, measure for fluorescence and cleaved in each cycle, so attached probe is only 5 nucleotides long

SOLiD测序法



Complete more cycles!

But we only have fluorescence measurements for every 5th base....



Offset by one base (and do the whole thing over again four times!)

The entire process is repeated four times, each time with the primer offset by 1 base

n G C C G G C T G G G A T C A
n - 1 A G A C G G C T A G G A A C
n - 2 C A A G T C T A C G C A A A
n - 3 A C A A T T A T A C T C A A
n - 4 C A G A C T G A G A T T A A
知乎用户

https://www.zhihu.com/question/29781261