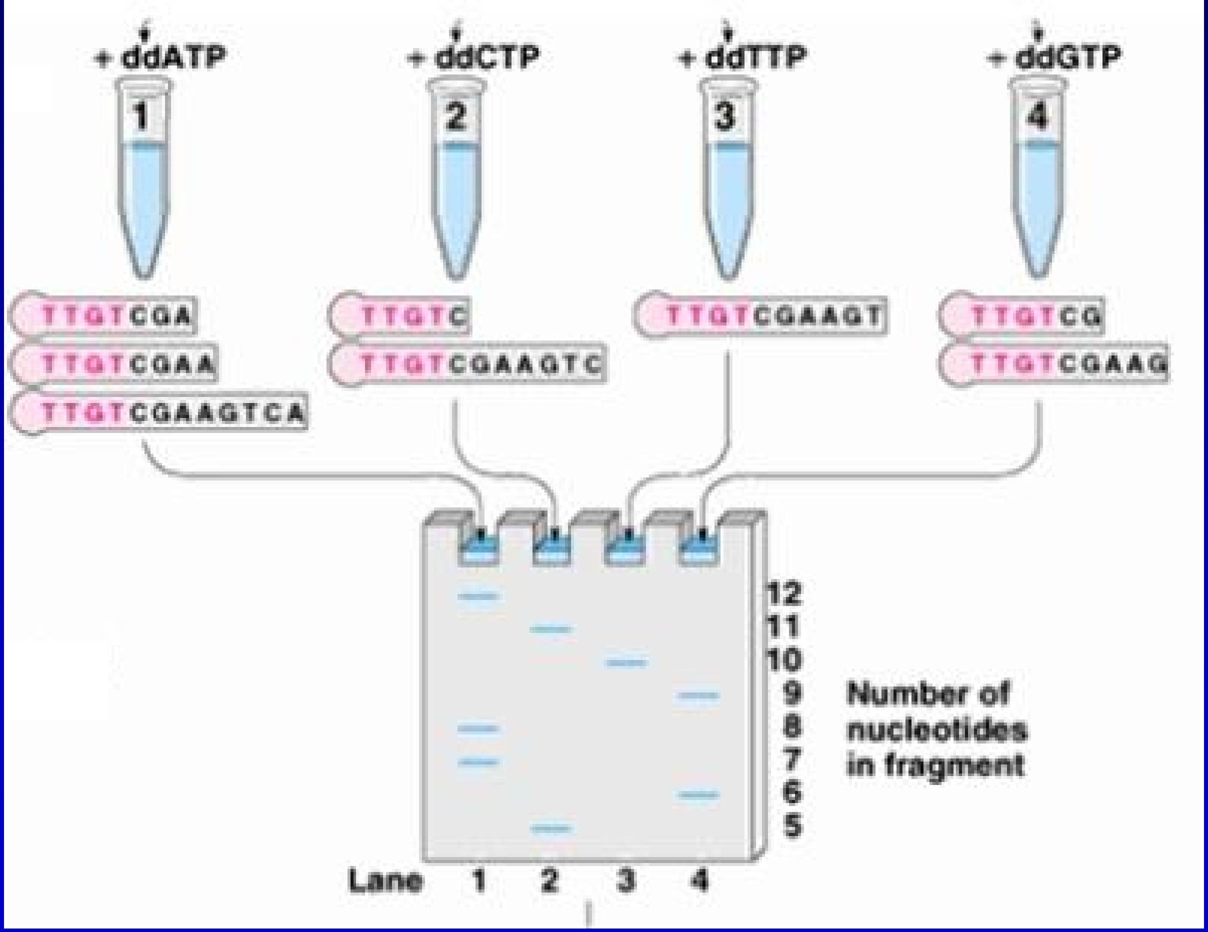
高通量测序数据分析

提纲

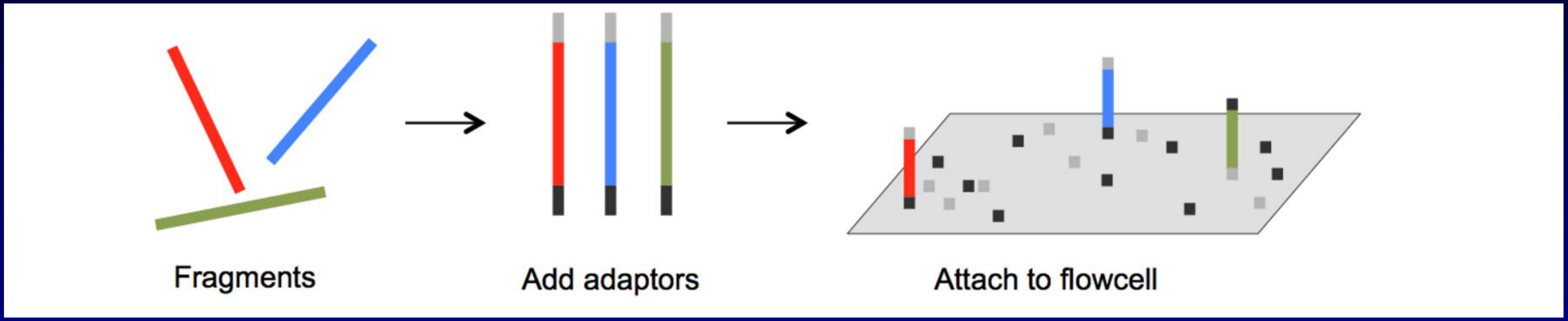
- 测序技术.
- 测序结果和质量控制Fastq 和 FASTQC.
- 序列匹配算法:
  - Seed.
  - Borrows-Wheeler transformation & LF mapping.
- 文件格式: SAM and BED.

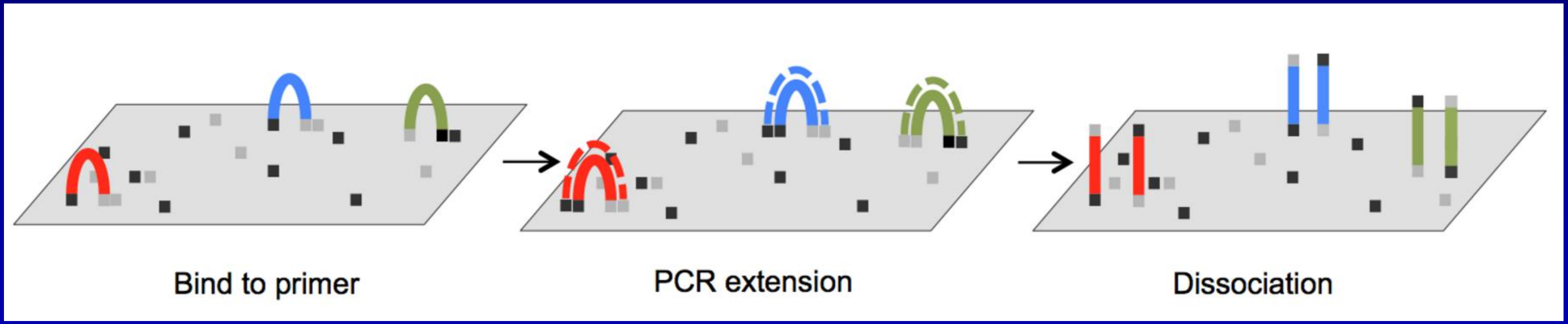
第一代: Sanger测序法

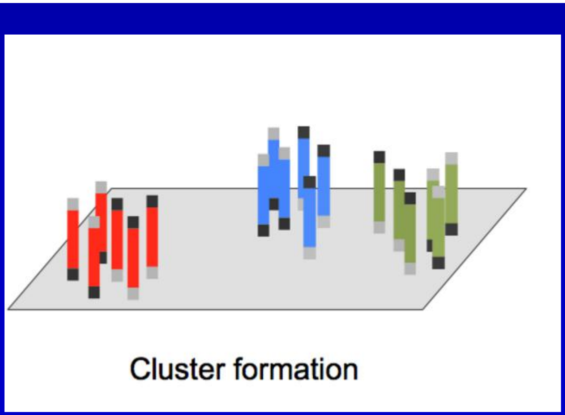
- Add one-stranded DNA sequence to four test tubes.
- Each tube contain all d**N**TPs + one dd**N**TP.



第二代: Illumina Sequencing

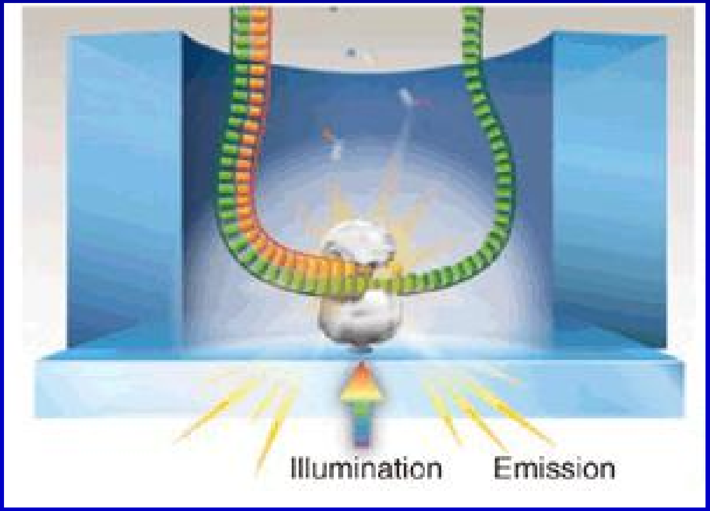


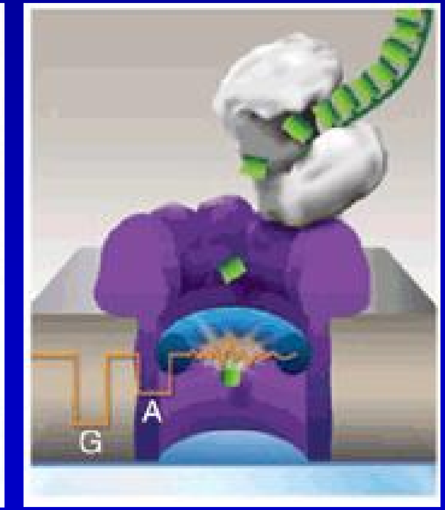




第三代： 纳米孔（Nanopores）

- Single molecule sequencing: no amplification.
- Fewer but much longer reads.
- Good for sequencing long reads, but not for read count applications.
- Technology still under active development.





FASTQ File

- Format:
  1. Sequence ID.
  2. Sequence.
  3. Quality ID.
  4. Quality score.

```
@HWI-EAS305:1:1:1:991#0/1
GCTGGAGGTTCAGGCTGGCCGGATTTAAACGTAT
+HWI-EAS305:1:1:1:991#0/1
MVXUWVRKTWWULRQQMMWWBBBBBBBBBBBBB
B
@HWI-EAS305:1:1:1:201#0/1
AAGACAAAGATGTGCTTTCTAAATCTGCACTAAT
+HWI-EAS305:1:1:1:201#0/1
PXX[[[[XTXYXTTWYYY[XXWWW[TMTVXWBBB
```

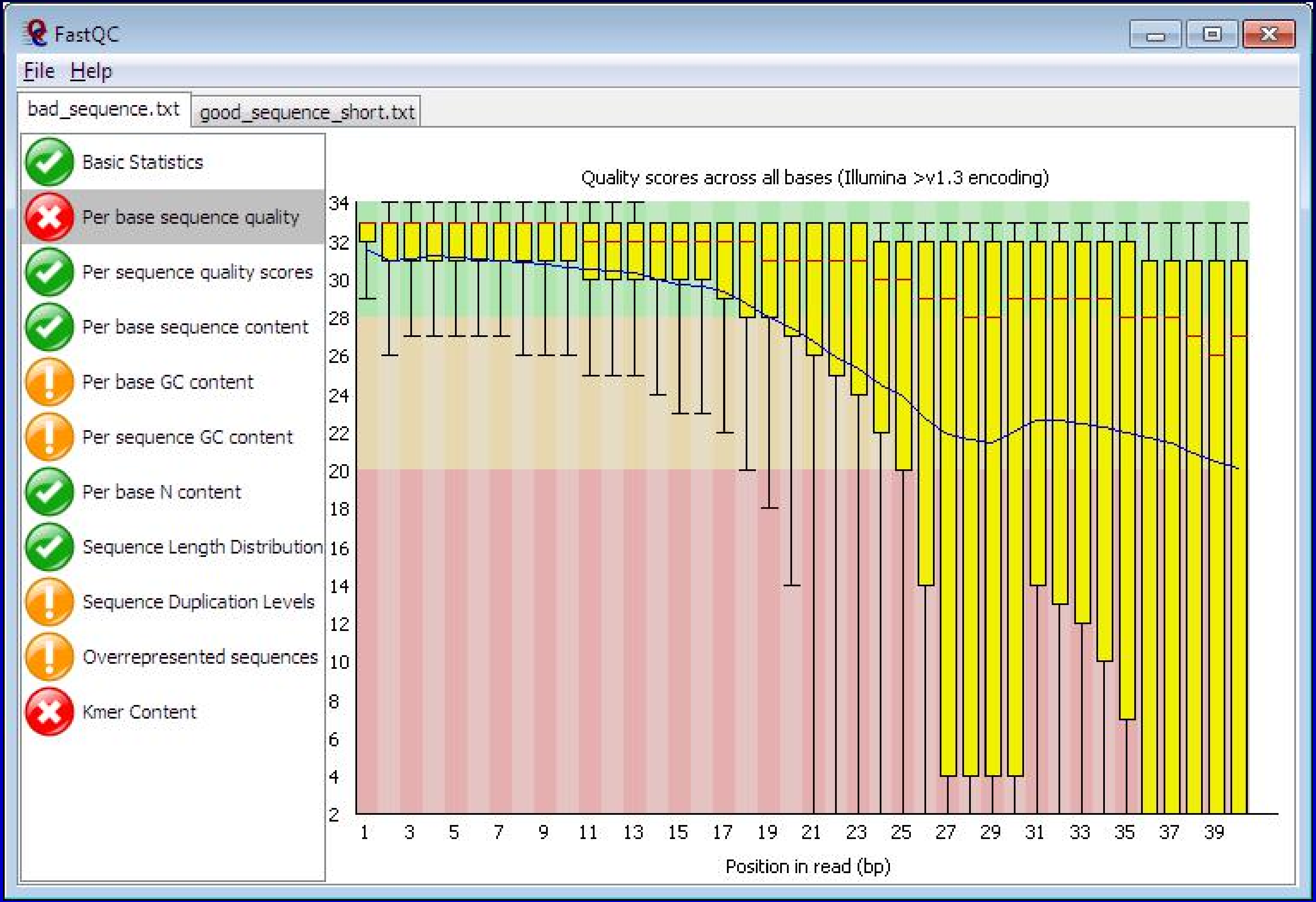
- Phred quality:
  - ASCII of: sequence quality + 33.
  - -10 $\log_{10}$ Pr(bp is wrongly sequenced).

Worst quality

Best quality

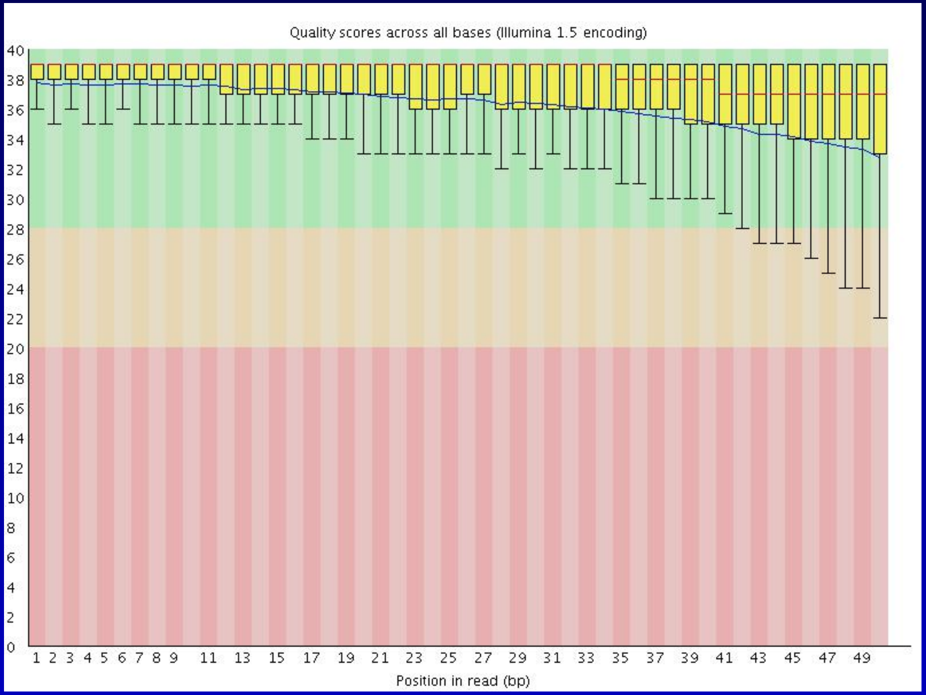
```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

FASTQC


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
1
3
5
7
9
11
13
15
17
19
21
23
25
27
29
31
33
35
37
39
Position in read (bp)

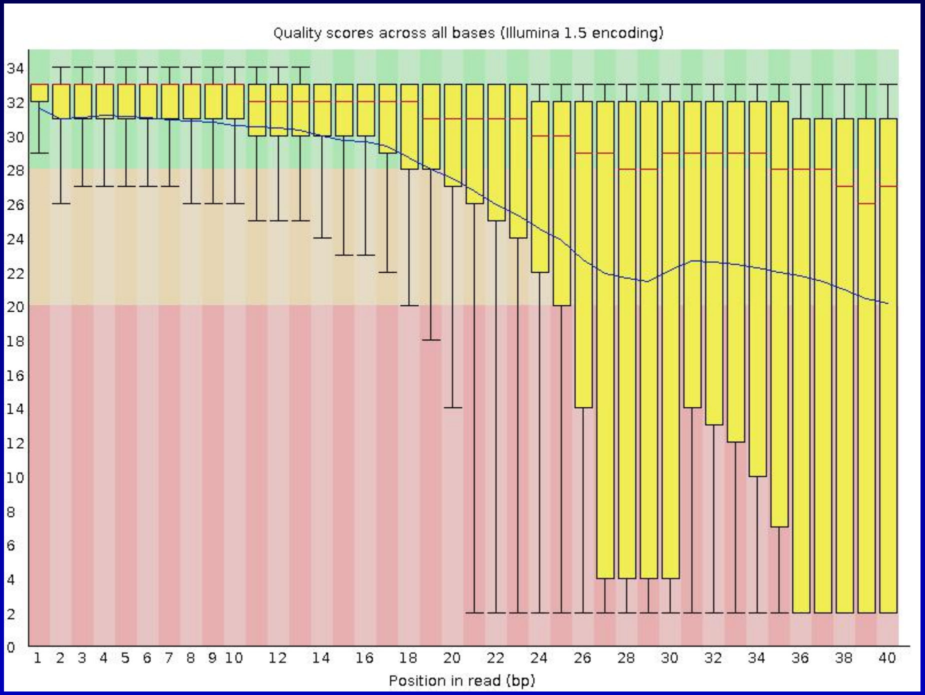
FASTQC: Per Base Sequence Quality

Good quality!



- Consistent.
- High-quality along the read.

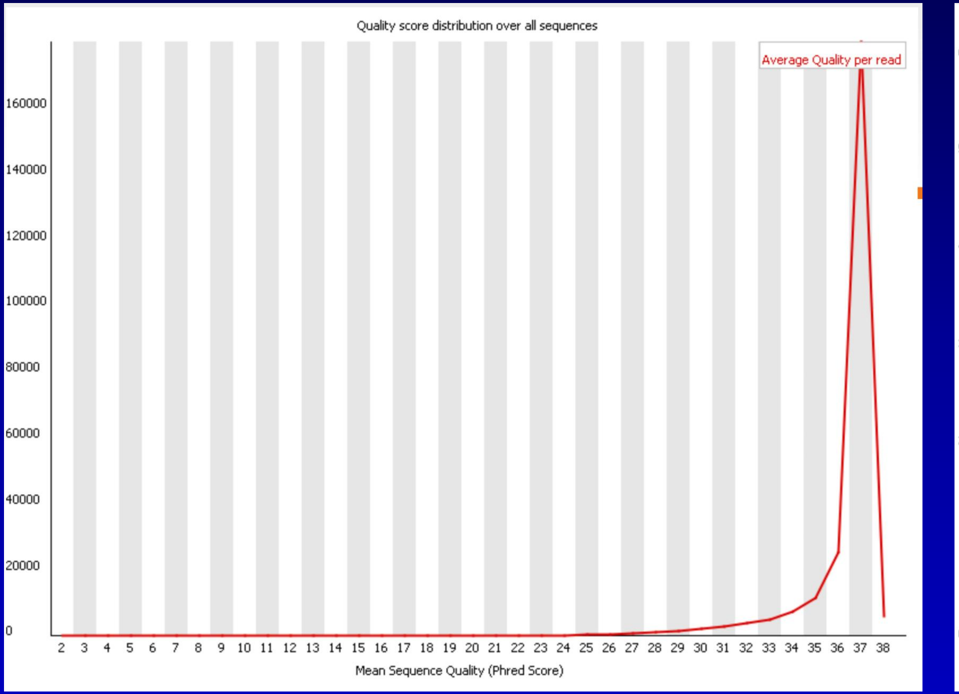
Poor quality!



- High Variance.
- Quality decreases with length.

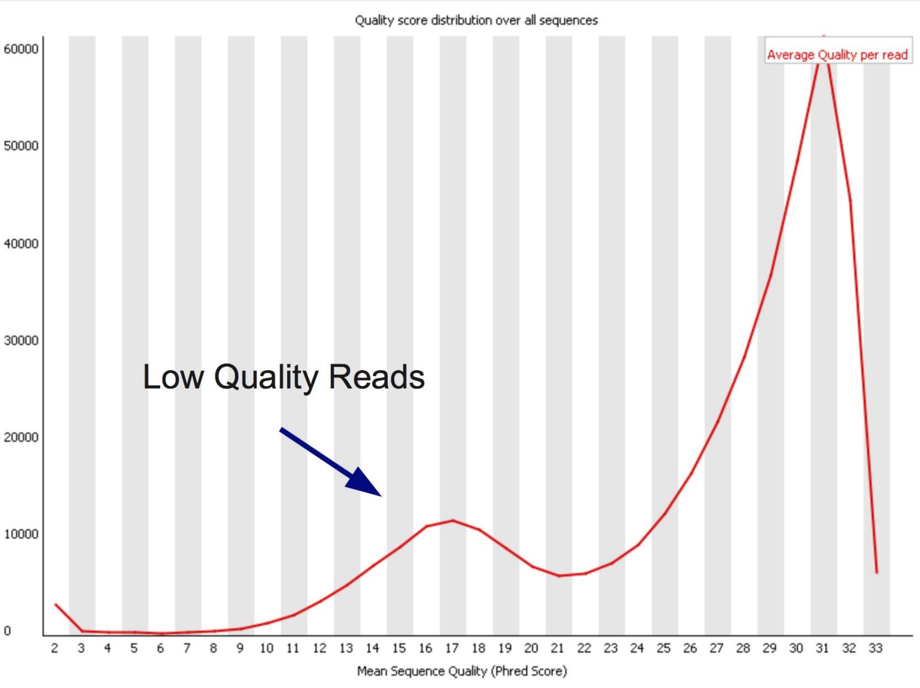
FASTQC: Per Sequence Quality Distribution

Good quality!



- Most are high-quality sequences.

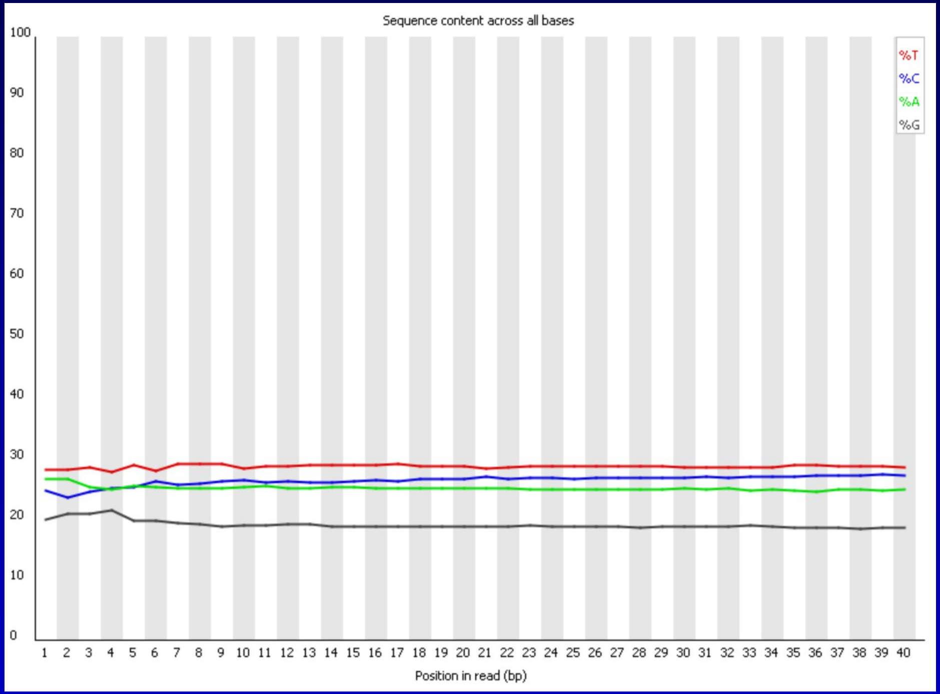
Poor quality!



- Distribution is not uniform.
- Presence of low quality reads.

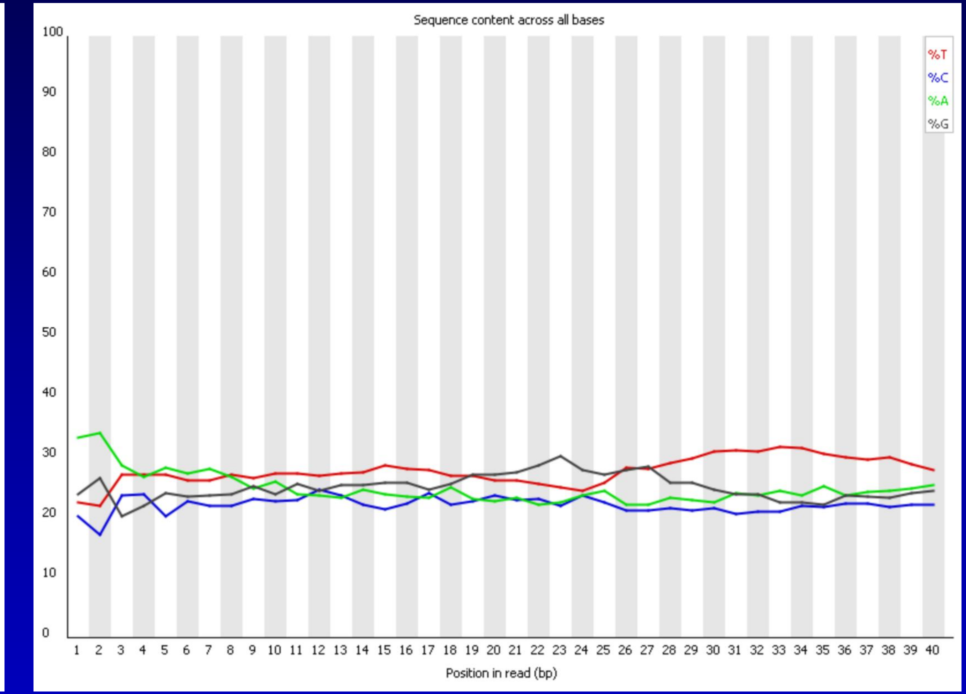
FASTQC: Nucleotide Content Per Position

Good quality!



- Smooth over length.

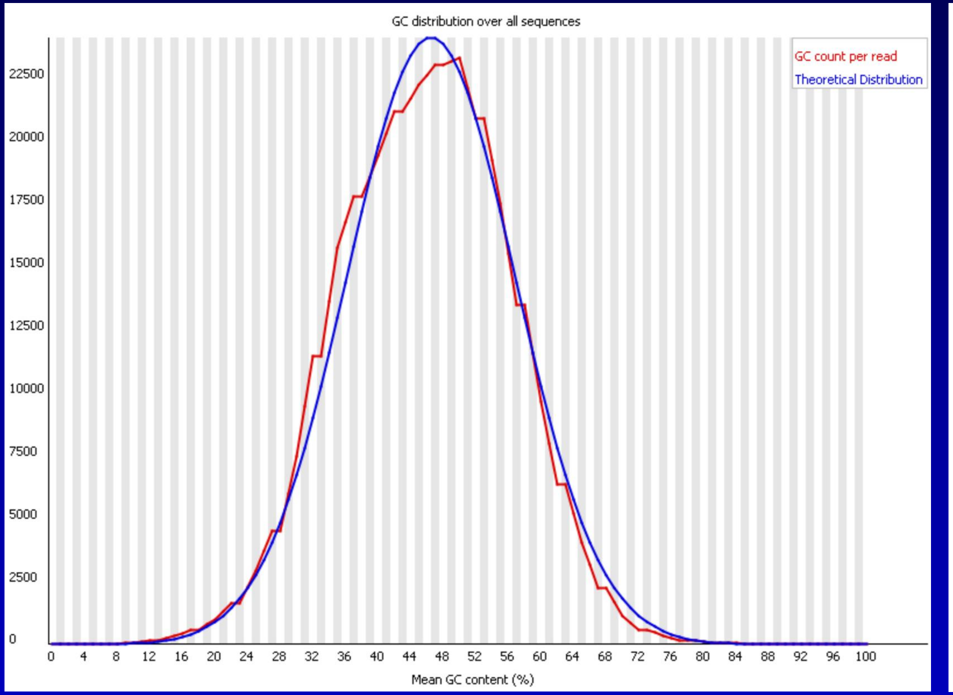
Poor quality!



- Sequence-position bias.

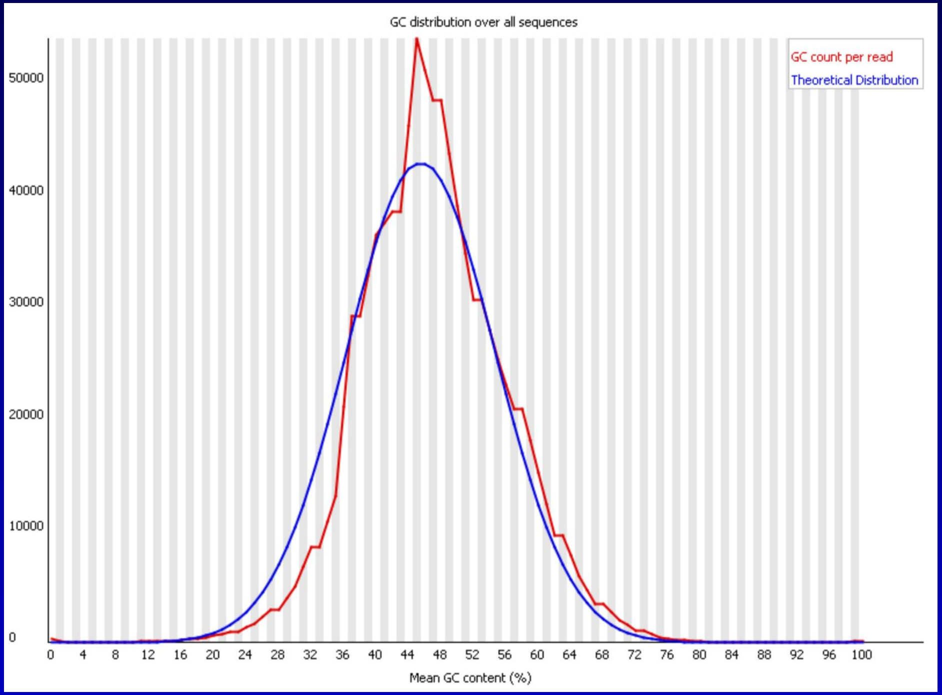
FASTQC: Per Sequence GC Content

Good quality!



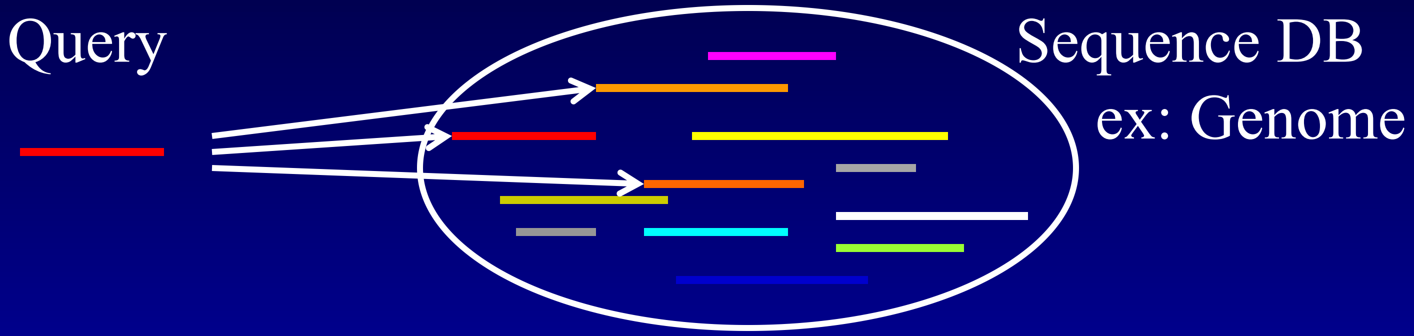
- Fits with expectation.

Poor quality!



- Does not fit with expectation.

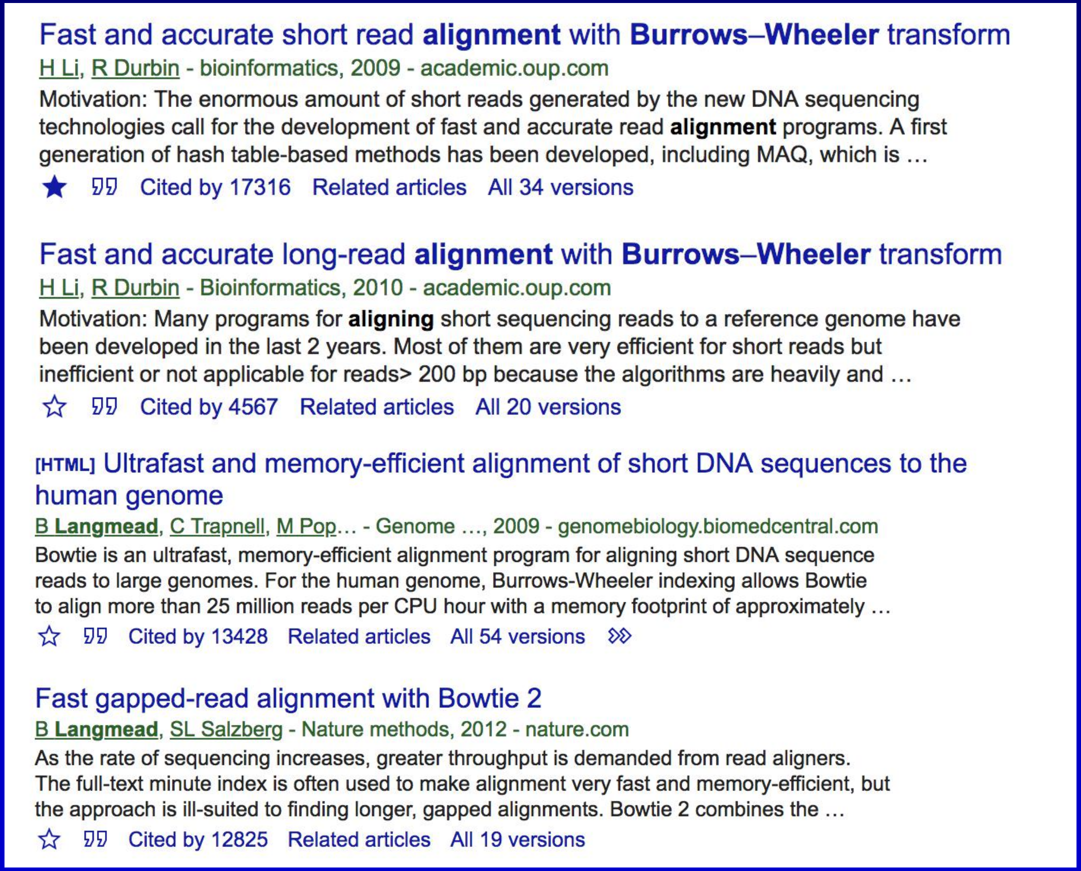
Read Mapping



- 如何把测得的序列片段匹配到基因组上？

Burrows-Wheeler Alignment

- Most widely used tools:
  - bwa (http://bio-bwa.sourceforge.net/).
  - bowtie (http://bowtiebio.sourceforge.net/bowtie2/index.shtml).



Fast and accurate short read **alignment** with **Burrows–Wheeler** transform

H Li, R Durbin - bioinformatics, 2009 - academic.oup.com

Motivation: The enormous amount of short reads generated by the new DNA sequencing technologies call for the development of fast and accurate read **alignment** programs. A first generation of hash table-based methods has been developed, including MAQ, which is …

★ 99 Cited by 17316 Related articles All 34 versions

Fast and accurate long-read **alignment** with **Burrows–Wheeler** transform

H Li, R Durbin - Bioinformatics, 2010 - academic.oup.com

Motivation: Many programs for **aligning** short sequencing reads to a reference genome have been developed in the last 2 years. Most of them are very efficient for short reads but inefficient or not applicable for reads> 200 bp because the algorithms are heavily and …

☆ 99 Cited by 4567 Related articles All 20 versions

[HTML] Ultrafast and memory-efficient alignment of short DNA sequences to the human genome

B **Langmead**, C Trapnell, M Pop… - Genome …, 2009 - genomebiology.biomedcentral.com

Bowtie is an ultrafast, memory-efficient alignment program for aligning short DNA sequence reads to large genomes. For the human genome, Burrows-Wheeler indexing allows Bowtie to align more than 25 million reads per CPU hour with a memory footprint of approximately …

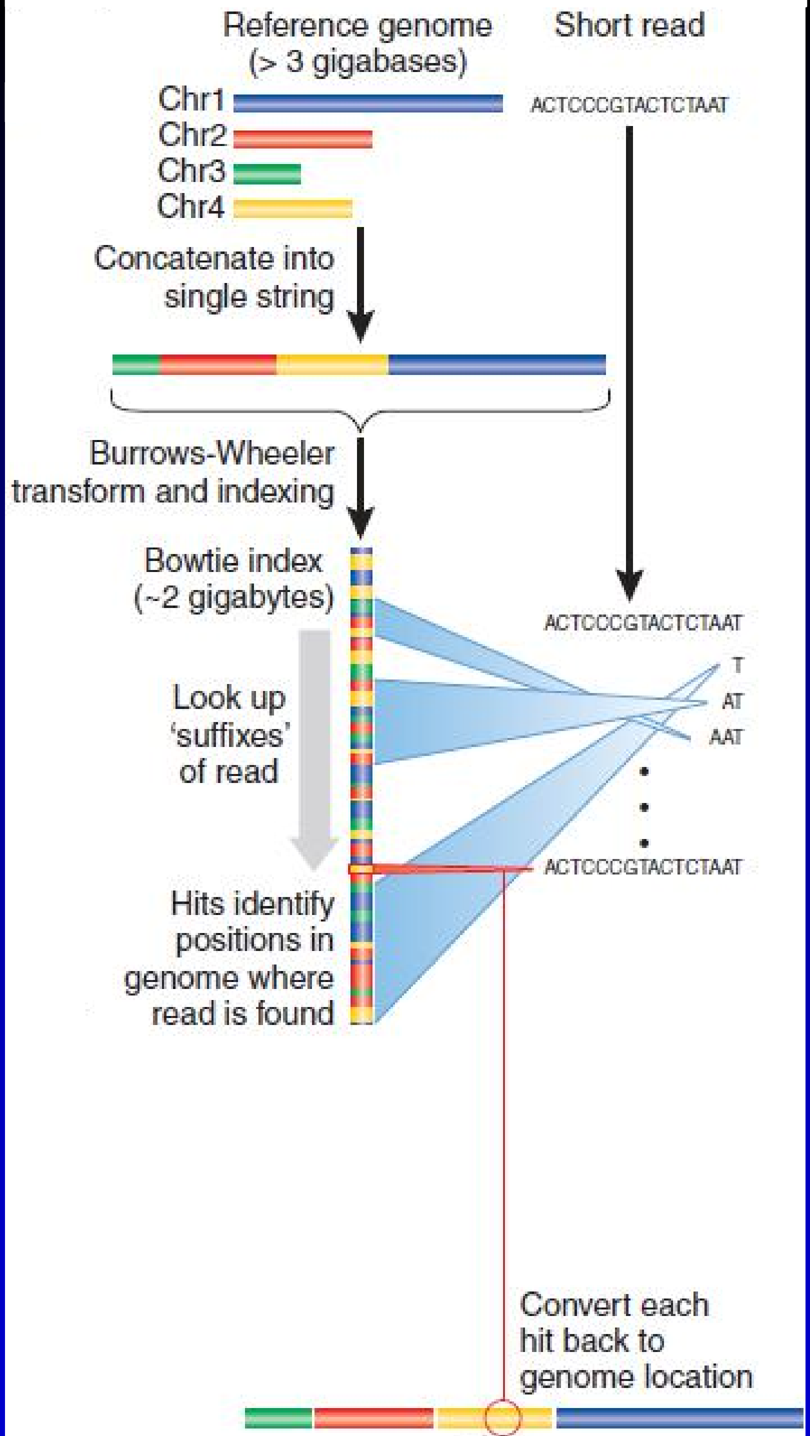
☆ 99 Cited by 13428 Related articles All 54 versions »

Fast gapped-read alignment with Bowtie 2

B **Langmead**, SL Salzberg - Nature methods, 2012 - nature.com

As the rate of sequencing increases, greater throughput is demanded from read aligners. The full-text minute index is often used to make alignment very fast and memory-efficient, but the approach is ill-suited to finding longer, gapped alignments. Bowtie 2 combines the …

☆ 99 Cited by 12825 Related articles All 19 versions


Reference genome
(> 3 gigabases)
Short read
Chr1
Chr2
Chr3
Chr4
ACTCCCGTACTCTAAT
Concatenate into single string
Burrows-Wheeler transform and indexing
Bowtie index (~2 gigabytes)
ACTCCCGTACTCTAAT
T
AT
AAT
Look up 'suffixes' of read
ACTCCCGTACTCTAAT
Hits identify positions in genome where read is found
Convert each hit back to genome location

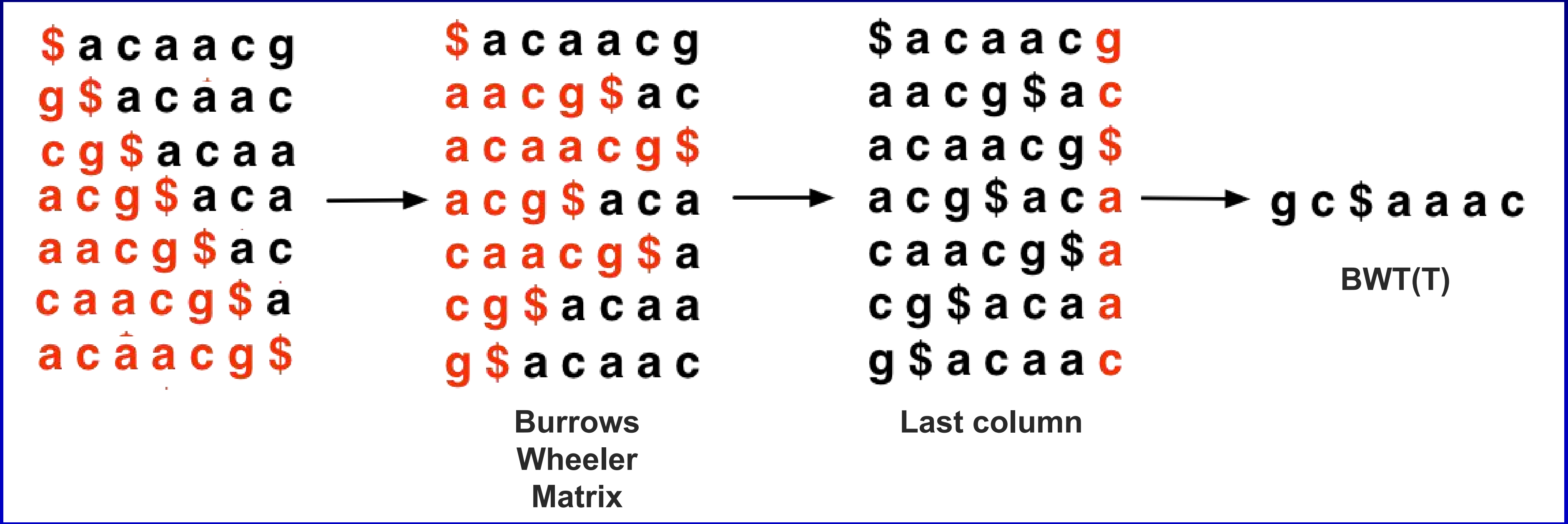
Burrows-Wheeler

- Use Burrows-Wheeler transform to store entire reference genome as a lookup index.
- Align tag base by base from the end.
- All active locations are reported.
- If no match is found, then back up and try a substitution.

Trapnell & Salzberg, Nat Biotech 2009.

Burrows-Wheeler Transform

- 原始序列 T = acaacg$
- 编码序列BWT(T)=gc$aaac



```
$ a c a a c g        $ a c a a c g        $ a c a a c g
g $ a c a a c        a a c g $ a c        a a c g $ a c
c g $ a c a a        a c a a c g $        a c a a c g $
a c g $ a c a   →    a c g $ a c a   →    a c g $ a c a   →   g c $ a a a c
a a c g $ a c        c a a c g $ a        c a a c g $ a
c a a c g $ a        c g $ a c a a        c g $ a c a a
a c a a c g $        g $ a c a a c        g $ a c a a c
```

Burrows Wheeler Matrix

Last column

BWT(T)

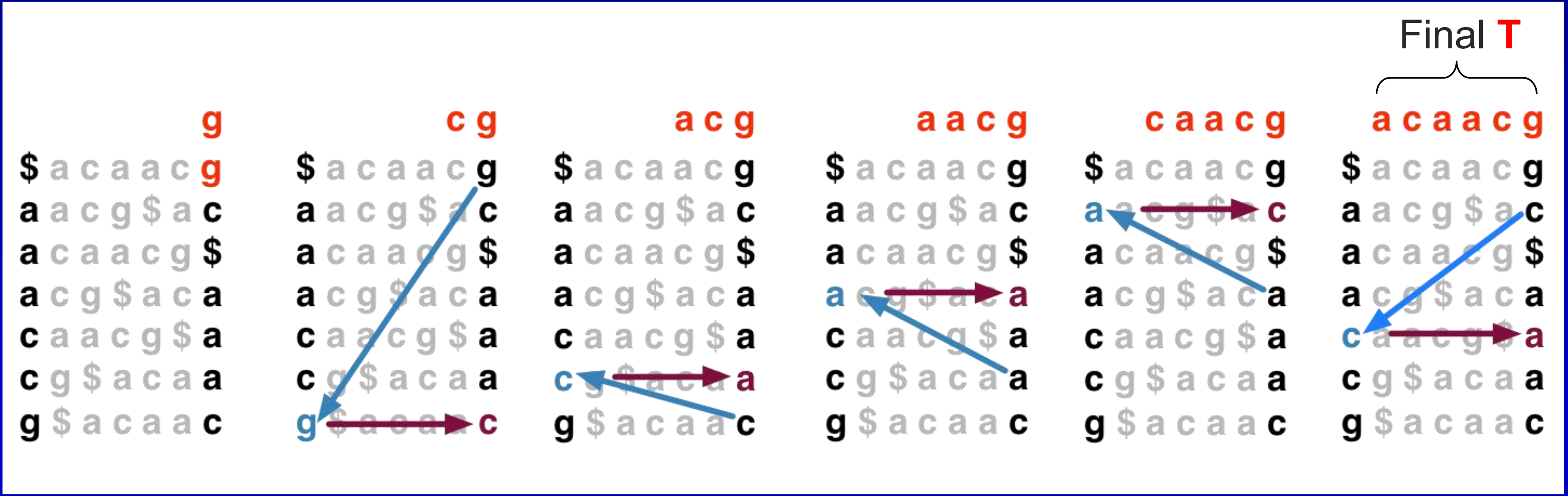
Burrows M, Wheeler DJ: **A block sorting lossless data compression algorithm.** *Digital Equipment Corporation, Palo Alto, CA* 1994, Technical Report 124; 1994

BWT: LF Mapping

- To recreate T from BWT(T), repeatedly apply rule:

  T = BWT[LF(i)] + T; i = LF(i)

  – Where LF(i) maps row i to row whose first character corresponds to i's last per LF Mapping.



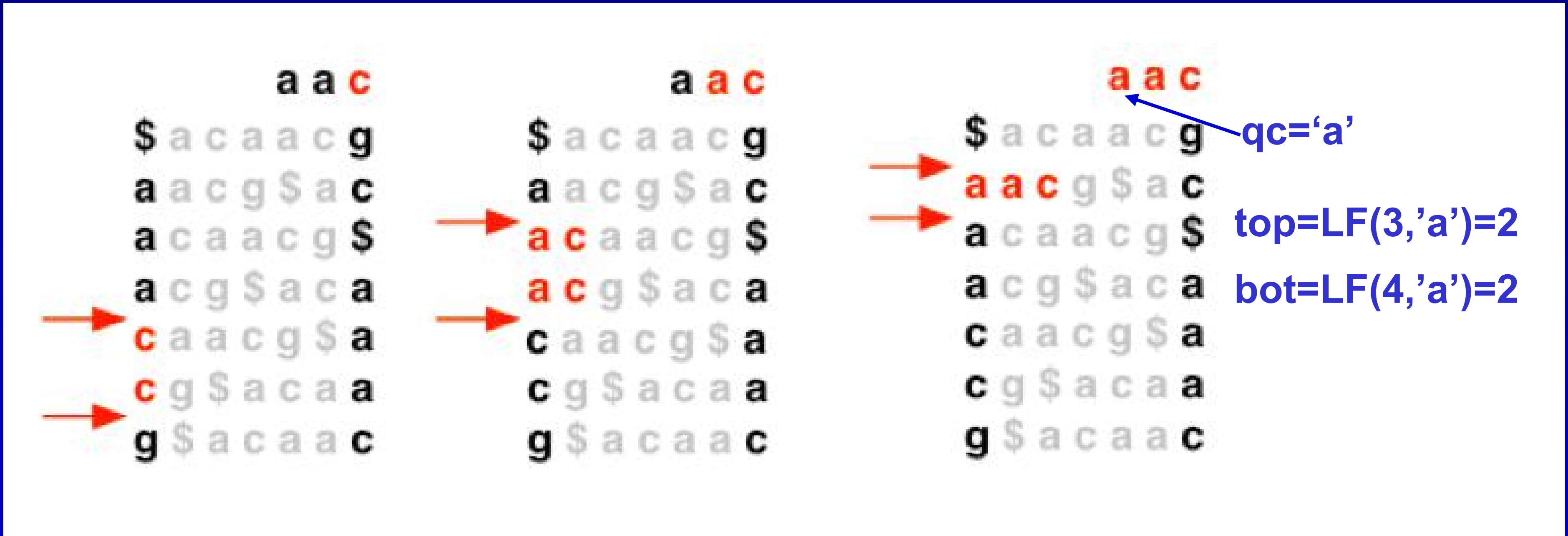
Final T

g	c g	a c g	a a c g	c a a c g	a c a a c g
$ a c a a c g	$ a c a a c g	$ a c a a c g	$ a c a a c g	$ a c a a c g	$ a c a a c g
a a c g $ a c	a a c g $ a c	a a c g $ a c	a a c g $ a c	a a c g $ a c	a a c g $ a c
a c a a c g $	a c a a c g $	a c a a c g $	a c a a c g $	a c a a c g $	a c a a c g $
a c g $ a c a	a c g $ a c a	a c g $ a c a	a c g $ a c a	a c g $ a c a	a c g $ a c a
c a a c g $ a	c a a c g $ a	c a a c g $ a	c a a c g $ a	c a a c g $ a	c a a c g $ a
c g $ a c a a	c g $ a c a a	c g $ a c a a	c g $ a c a a	c g $ a c a a	c g $ a c a a
g $ a c a a c	g $ a c a a c	g $ a c a a c	g $ a c a a c	g $ a c a a c	g $ a c a a c

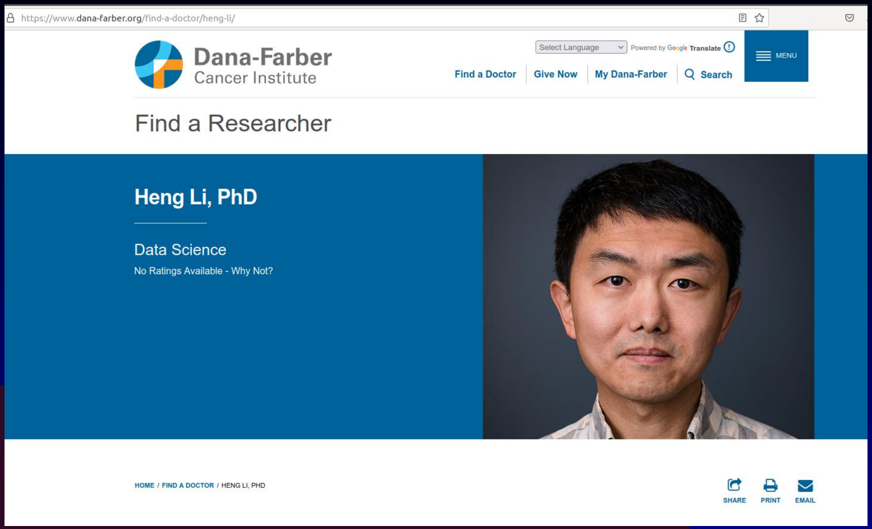
BWT(T) to retrieve alignments

T = acaacg

Q = aac



Bowtie2软件



https://www.dana-farber.org/find-a-doctor/heng-li/

Dana-Farber Cancer Institute

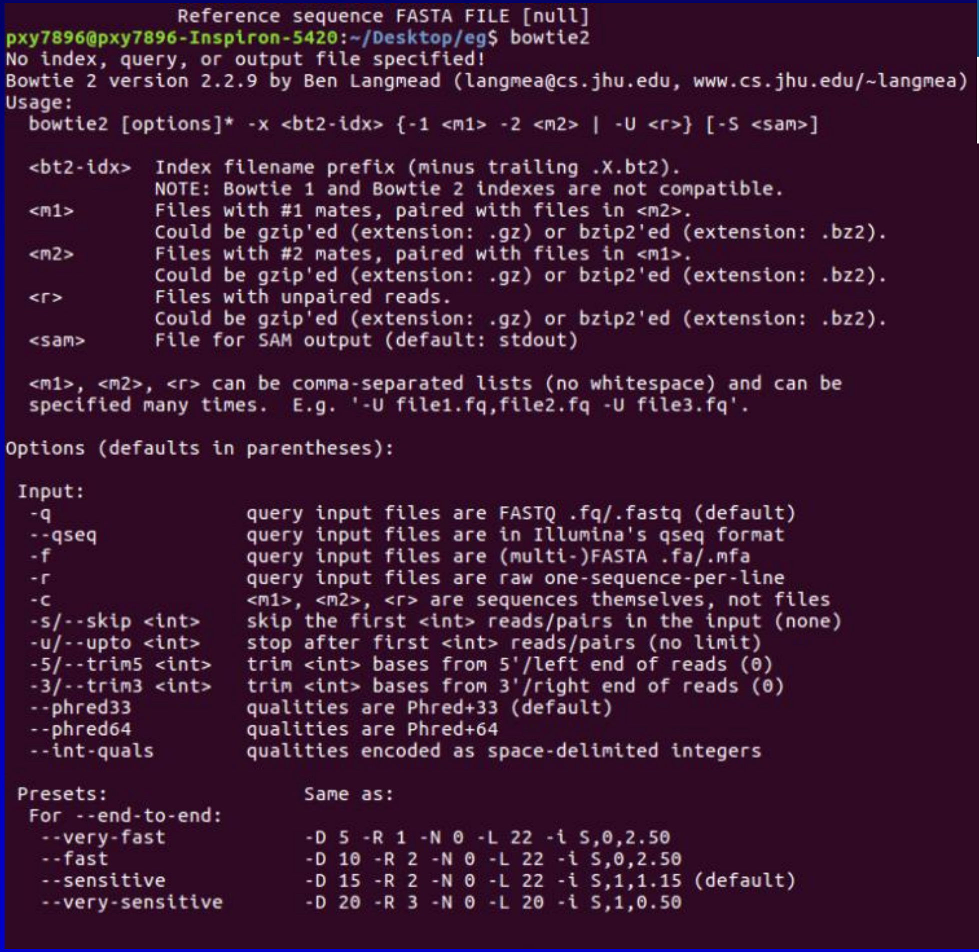
Find a Researcher

Heng Li, PhD

Data Science

No Ratings Available - Why Not?

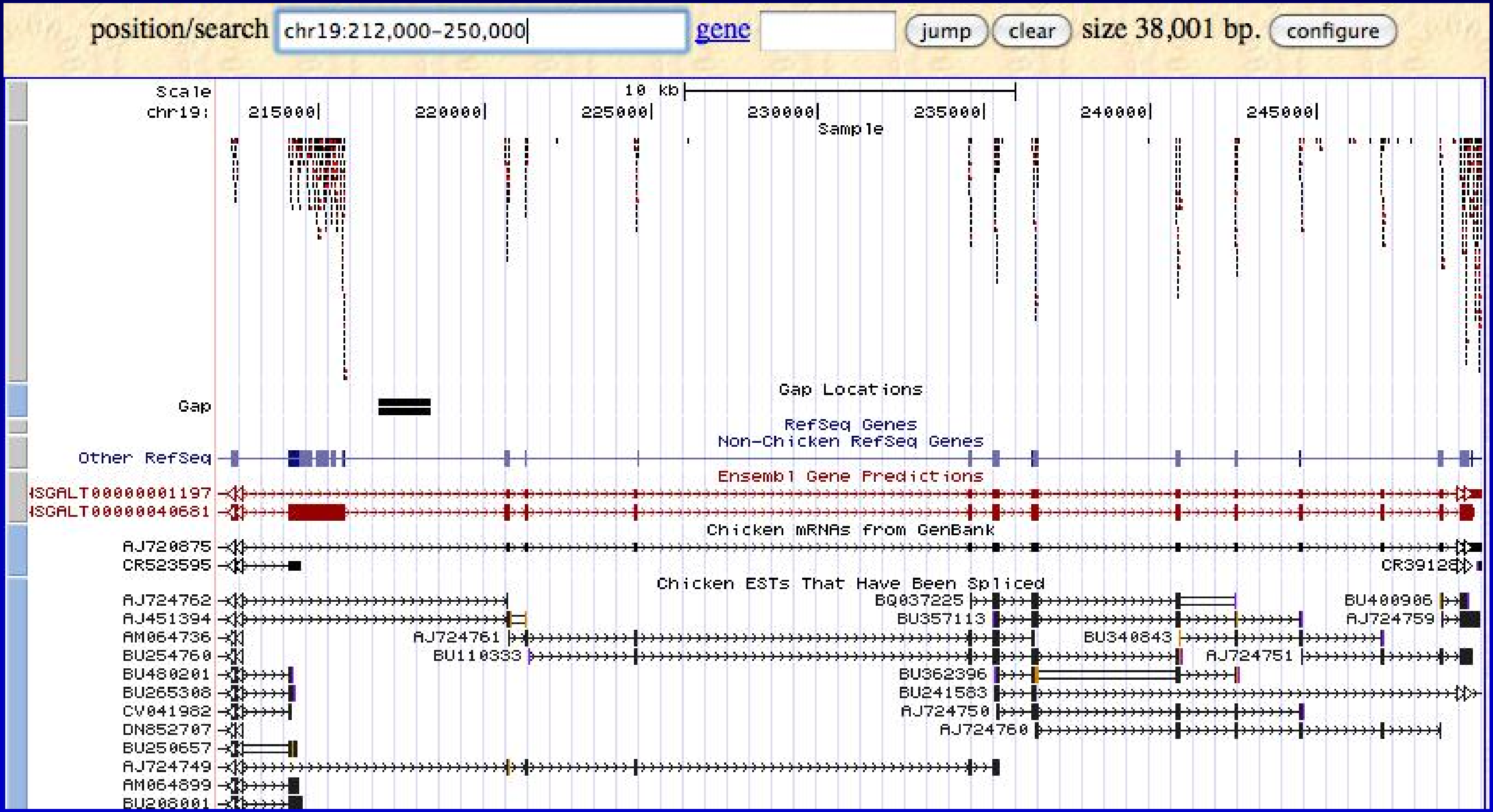
HOME / FIND A DOCTOR / HENG LI, PHD



```
               Reference sequence FASTA FILE [null]
pxy7896@pxy7896-Inspiron-5420:~/Desktop/eg$ bowtie2
No index, query, or output file specified!
Bowtie 2 version 2.2.9 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jhu.edu/~langmea)
Usage:
  bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>} [-S <sam>]

  <bt2-idx>  Index filename prefix (minus trailing .X.bt2).
             NOTE: Bowtie 1 and Bowtie 2 indexes are not compatible.
  <m1>       Files with #1 mates, paired with files in <m2>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <m2>       Files with #2 mates, paired with files in <m1>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <r>        Files with unpaired reads.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <sam>      File for SAM output (default: stdout)

  <m1>, <m2>, <r> can be comma-separated lists (no whitespace) and can be
  specified many times.  E.g. '-U file1.fq,file2.fq -U file3.fq'.

Options (defaults in parentheses):

 Input:
  -q                 query input files are FASTQ .fq/.fastq (default)
  --qseq             query input files are in Illumina's qseq format
  -f                 query input files are (multi-)FASTA .fa/.mfa
  -r                 query input files are raw one-sequence-per-line
  -c                 <m1>, <m2>, <r> are sequences themselves, not files
  -s/--skip <int>    skip the first <int> reads/pairs in the input (none)
  -u/--upto <int>    stop after first <int> reads/pairs (no limit)
  -5/--trim5 <int>   trim <int> bases from 5'/left end of reads (0)
  -3/--trim3 <int>   trim <int> bases from 3'/right end of reads (0)
  --phred33          qualities are Phred+33 (default)
  --phred64          qualities are Phred+64
  --int-quals        qualities encoded as space-delimited integers

 Presets:                 Same as:
  For --end-to-end:
   --very-fast            -D 5 -R 1 -N 0 -L 22 -i S,0,2.50
   --fast                 -D 10 -R 2 -N 0 -L 22 -i S,0,2.50
   --sensitive            -D 15 -R 2 -N 0 -L 22 -i S,1,1.15 (default)
   --very-sensitive       -D 20 -R 3 -N 0 -L 20 -i S,1,0.50
```

Visualization

- Visualize BAM / BED files in genome browsers (UCSC or IGV)



position/search chr19:212,000-250,000

size 38,001 bp.

生物信息学工作的层次

- 0级 (Level 0)：为建模、而建模
- 1级（Level 1）：给数据、能分析
- 2级（Level 2）：想新招、玩数据
- 3级（Level 3）：玩数据、作发现
- X级（Level X）：玩科学、讲政治