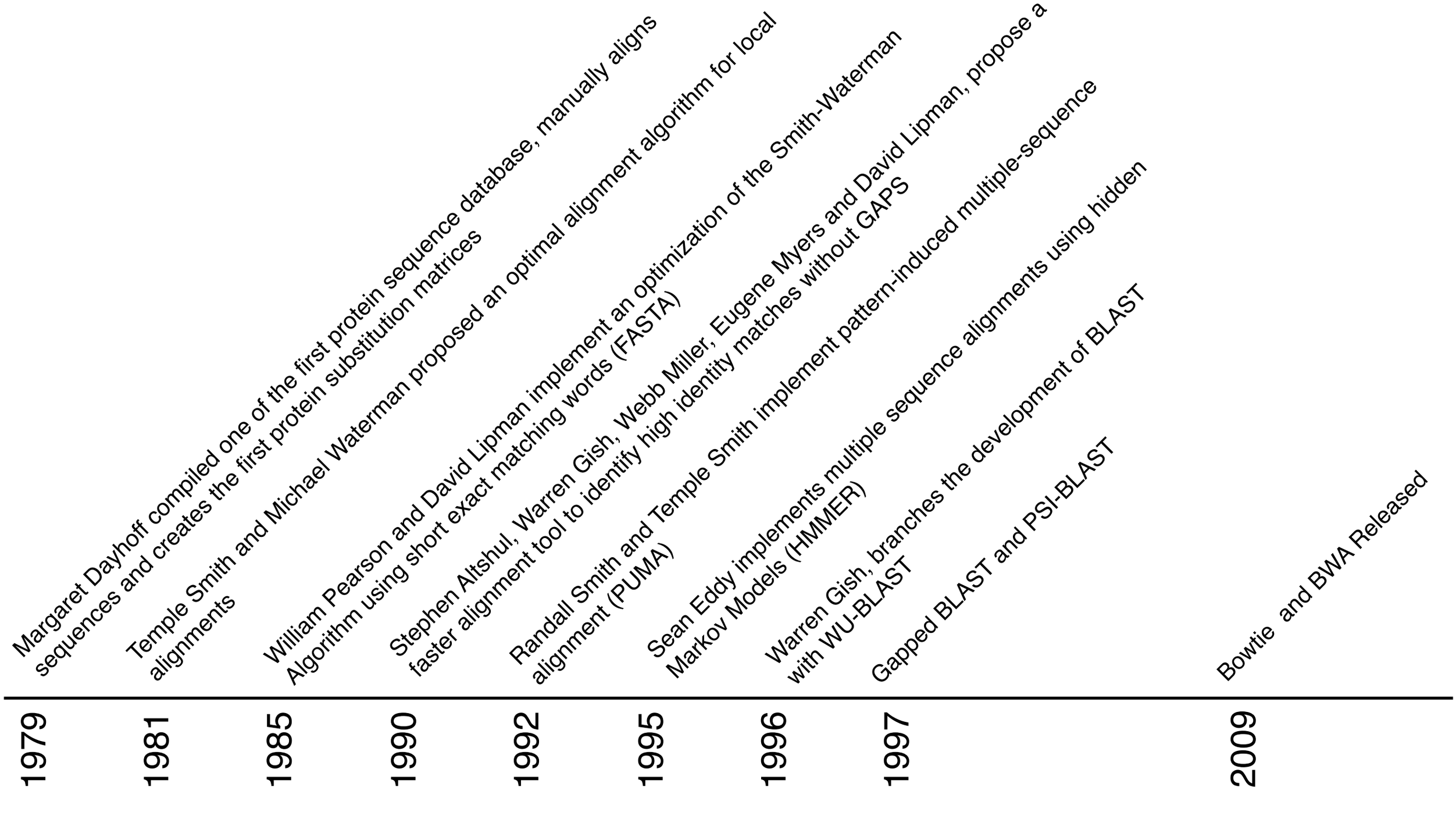
Short Read Alignment

Mapping Reads to a Reference

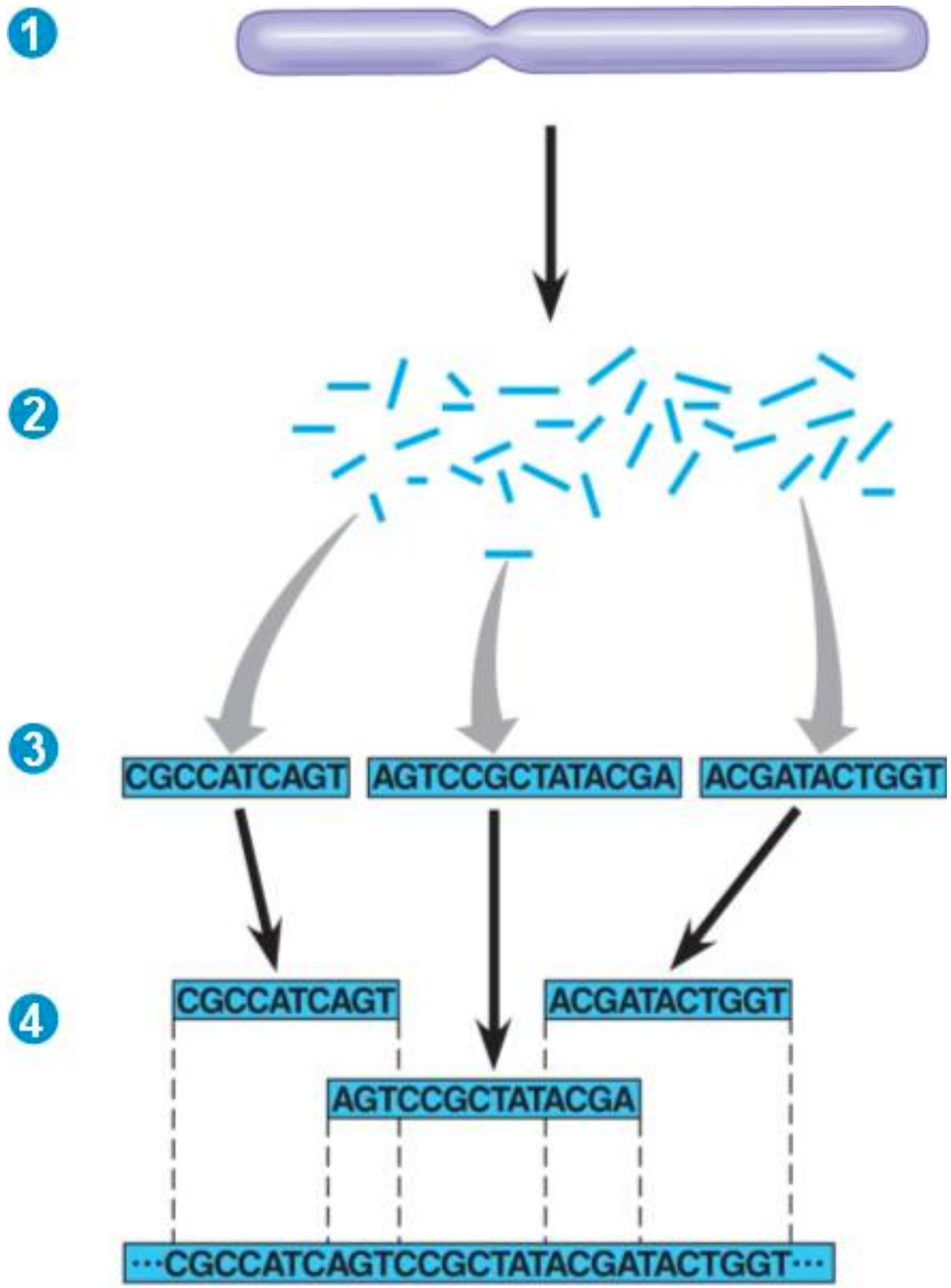
Brandi Cantarel, PhD
BICF
08/01/2016

- Introduction to Mapping
- Short Read Aligners
- DNA vs RNA
- Alignment Quality
- Pitfalls and Improvements

History of Sequence Similarity



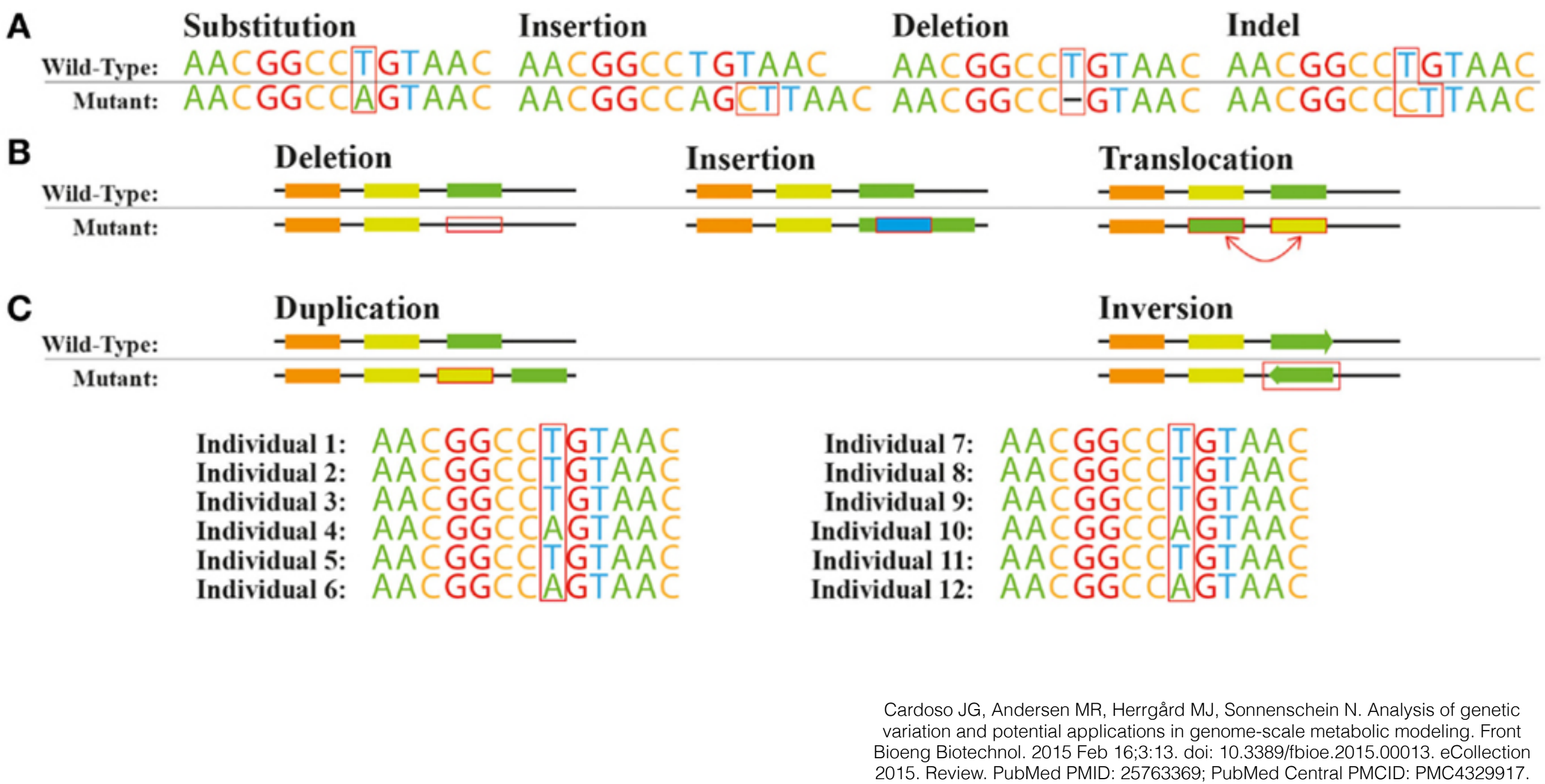
Whole Genome Shotgun



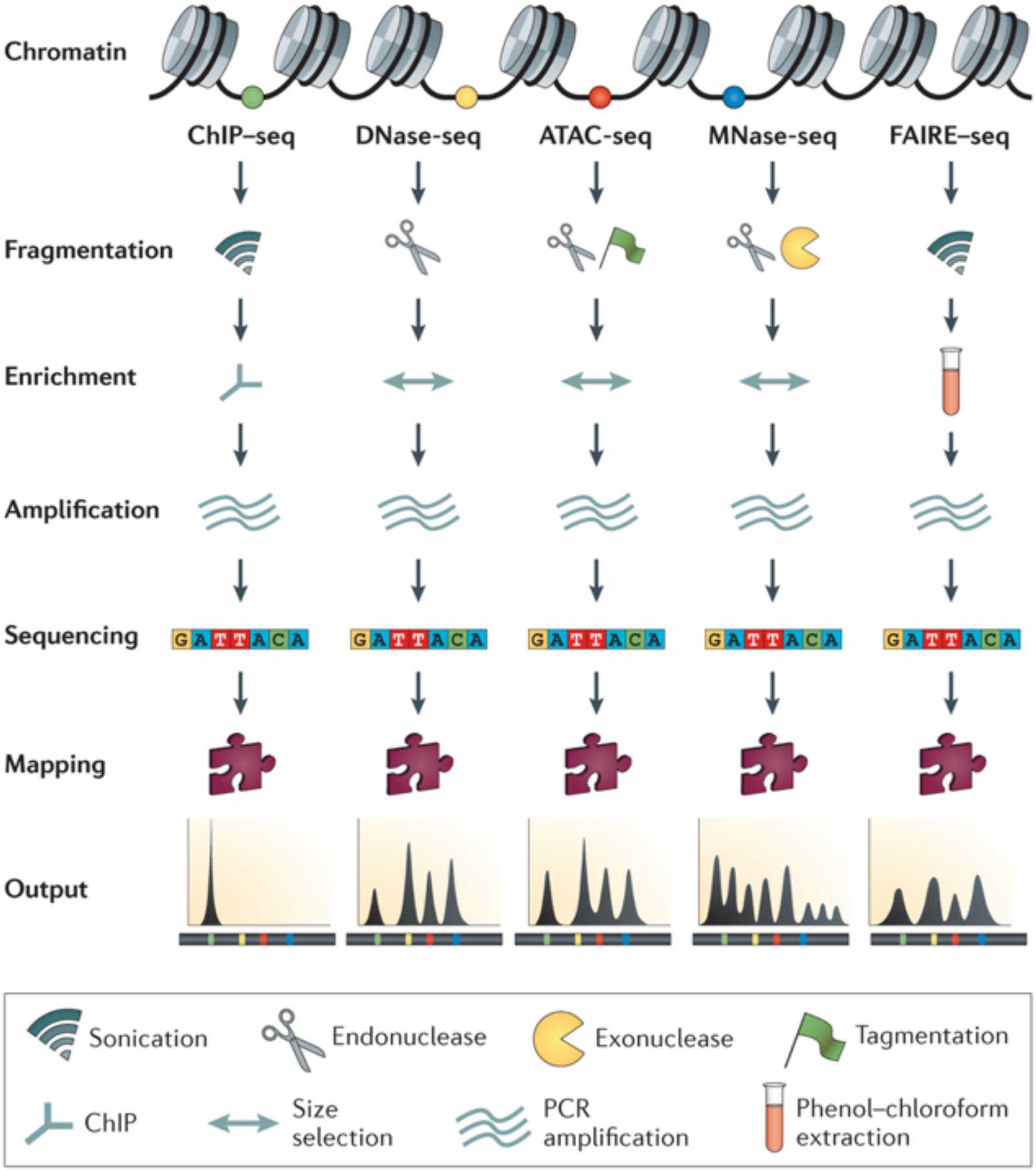
Utility of Mapping

- DNASeq
  - Identify Variation
- RNASeq
  - Estimate the abundance of transcripts and genes
- ChromatinSeq
  - Determine the structure of DNA (open, close, bound to proteins, etc)

Genetic Variation



Cardoso JG, Andersen MR, Herrgård MJ, Sonnenschein N. Analysis of genetic variation and potential applications in genome-scale metabolic modeling. Front Bioeng Biotechnol. 2015 Feb 16;3:13. doi: 10.3389/fbioe.2015.00013. eCollection 2015. Review. PubMed PMID: 25763369; PubMed Central PMCID: PMC4329917.



Nature Reviews | Genetics

Chromatin Seq

How to Make an Alignment

```
PMILGYWNVRGL
:     .
PPYTIVYFPVRG
```

```
PMILGYWNVRGL
:  .  :. :::
PPYTIVYFPVRG
```

```
PM-ILGYWNVRGL
:  :. :. :::
PPYTIV-YFPVRG
```

	P	M	I	L	G	Y	W	N	V	R	G	L
P	:											
Y		.			:	.						
T							.	.				
I		:	.					.				.
V	.	.	.					:				.
Y	.				:	.						
F	.	.	.		.	.						.
P	:											
V	.	.	.					:				.
R									:			
G										:		

Local Alignment

```
AAPMILGYWNVRGLBB
  :  .  :. :::
DDPPYTIVYFPVRGCC
```

Global Alignments

Global Alignment

```
-PMILGYWNVRGL
 :  .  :. :::
PPYTIVYFPVRG-
```

Basis:

$F_{0j} = d * j$

$F_{i0} = d * i$

Recursion, based on the principle of optimality:

$$F_{ij} = \max(F_{i-1,j-1} + S(A_i, B_j),\ F_{i,j-1} + d,\ F_{i-1,j} + d)$$

The pseudo-code for the algorithm to compute the F matrix therefore looks like this:

```
for i=0 to length(A)
  F(i,0) ← d*i
for j=0 to length(B)
  F(0,j) ← d*j
for i=1 to length(A)
  for j=1 to length(B)
  {
    Match ← F(i-1,j-1) + S(A_i, B_j)
    Delete ← F(i-1, j) + d
    Insert ← F(i, j-1) + d
    F(i,j) ← max(Match, Insert, Delete)
  }
```

Local Alignments

Local Alignment

```
 AAPMILGYWNVRGLBB
   :  .  :. :::
DDPPYTIVYFPVRGCC
```

A matrix H is built as follows:

$$H(i,0) = 0,\ 0 \leq i \leq m$$

$$H(0,j) = 0,\ 0 \leq j \leq n$$

if $a_i = b_j$ then $w(a_i,b_j) = w(\text{match})$ or if $a_i! = b_j$ then $w(a_i,b_j) = w(\text{mismatch})$

$$H(i,j) = \max \left\{ \begin{array}{cc} 0 & \\ H(i-1,j-1) + \ w(a_i,b_j) & \text{Match/Mismatch} \\ H(i-1,j) + \ w(a_i,-) & \text{Deletion} \\ H(i,j-1) + \ w(-,b_j) & \text{Insertion} \end{array} \right\},\ 1 \leq i \leq m, 1 \leq j \leq n$$

Where:

- a,b = Strings over the Alphabet Σ
- m = length(a)
- n = length(b)
- $H(i,j)$ - is the maximum Similarity-Score between a suffix of a[1...i] and a suffix of b[1...j]
- $w(c,d),\ c,d \in \Sigma \cup \{'-'\}$, '-' is the gap-scoring scheme

- Introduction to Mapping
- Short Read Aligners
- DNA vs RNA
- Alignment Quality
- Pitfalls and Improvements

Short Read Aligners

- Short read aligners assume that the read came from “intact” from the reference

- So the alignment is “global” from the read perspective and “local” from the reference perspective

Short Read Aligners

```
Read   GATCGCAGAGCTCGGGCATAGCTAGCGC
```

```
AGCTATGATCGCAGAGCTCGGGCATAGCTAGCGCTAGAGCTCGCGATCGCAGAGTCGA
Genome
```

Short Read Aligners

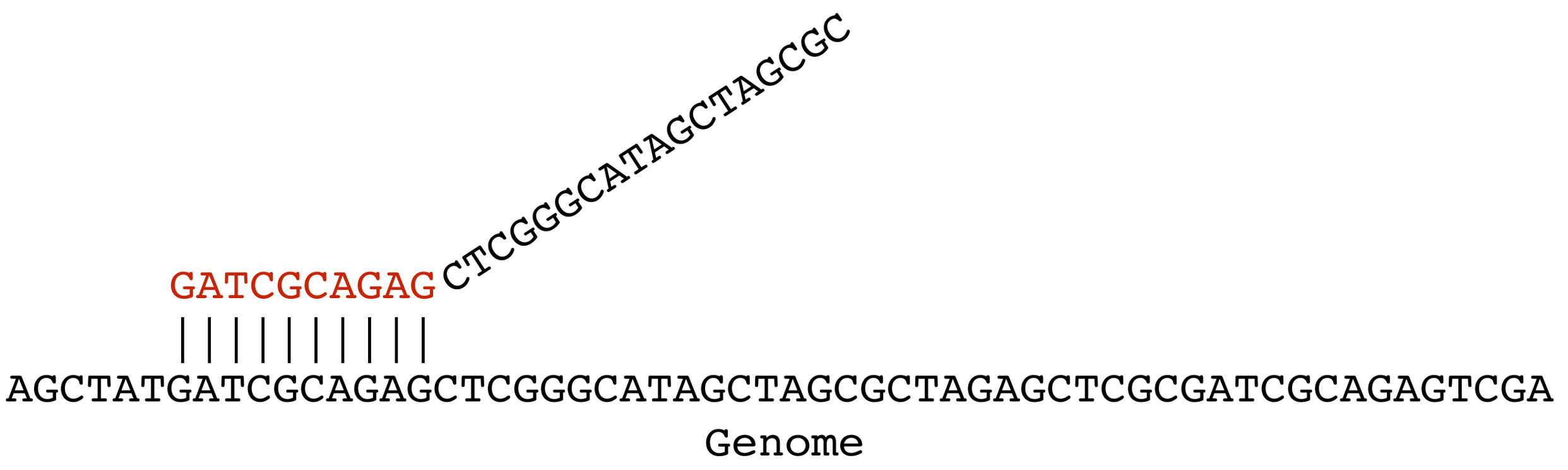
Read GATCGCAGAGCTCGGGCATAGCTAGCGC

Seed

AGCTATGATCGCAGAGCTCGGGCATAGCTAGCGCTAGAGCTCGCGATCGCAGAGTCGA

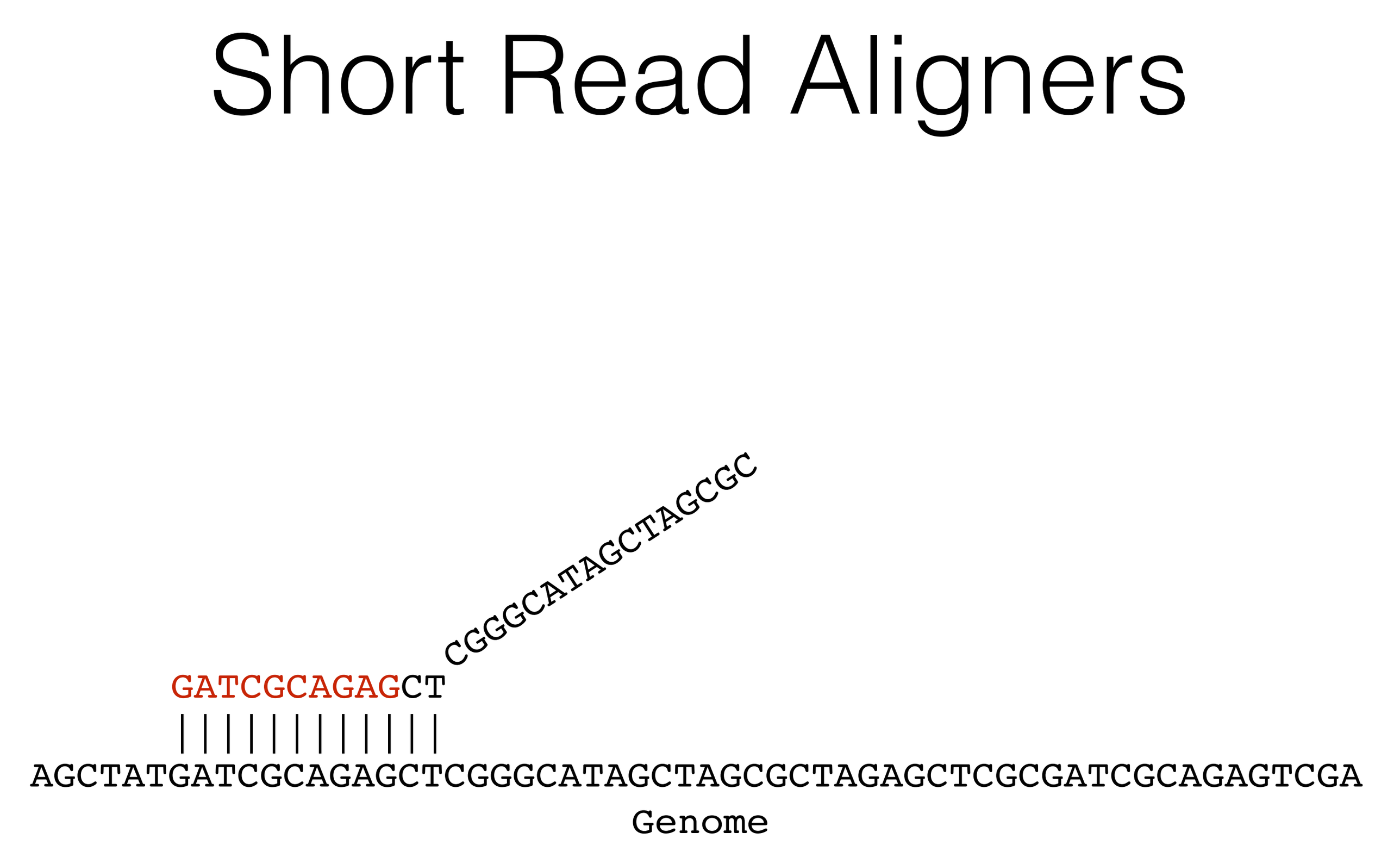
Genome

Short Read Aligners



Short Read Aligners

Short Read Aligners

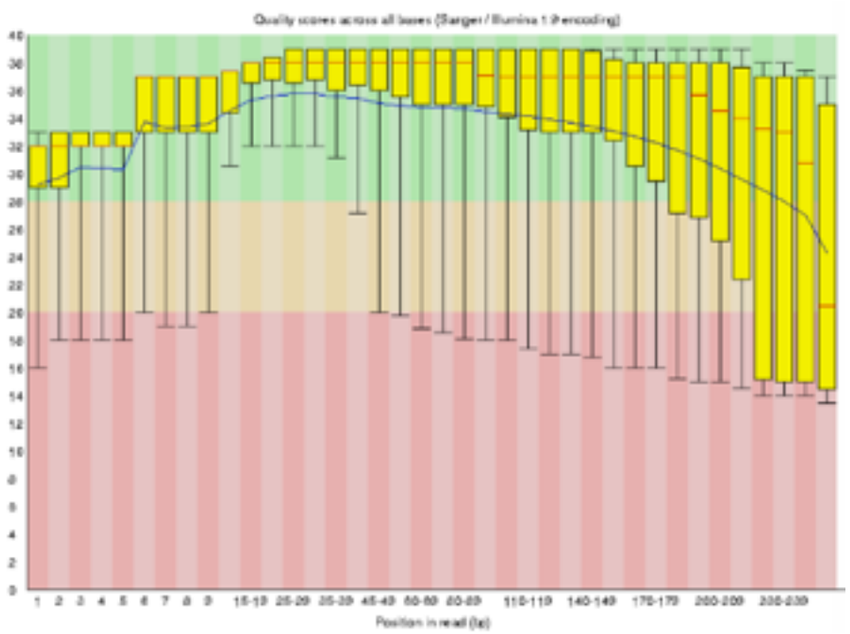


SRA Features: Seeding

- Seeding represents the first few tens of base pairs of a read. The seed part of a read is expected to contain less erroneous characters due to the specifics of the NGS technologies. Therefore, the seeding property is mostly used to maximize performance and accuracy. The alignments are then extended from the seed.

SRA Features: Base Quality

- Base quality scores provide a measure on correctness of each base in the read. The base quality score is assigned by a phred-like algorithm. The score Q is equal to -10 log10(e), where e is the probability that the base is wrong. Some tools use the quality scores to decide mismatch locations. Others accept or reject the read based on the sum of the quality scores at mismatch positions.

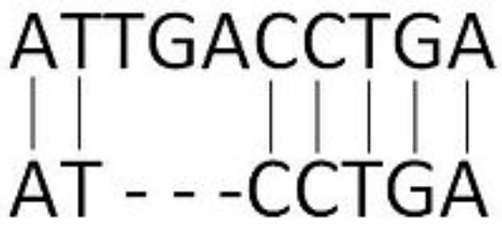


Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

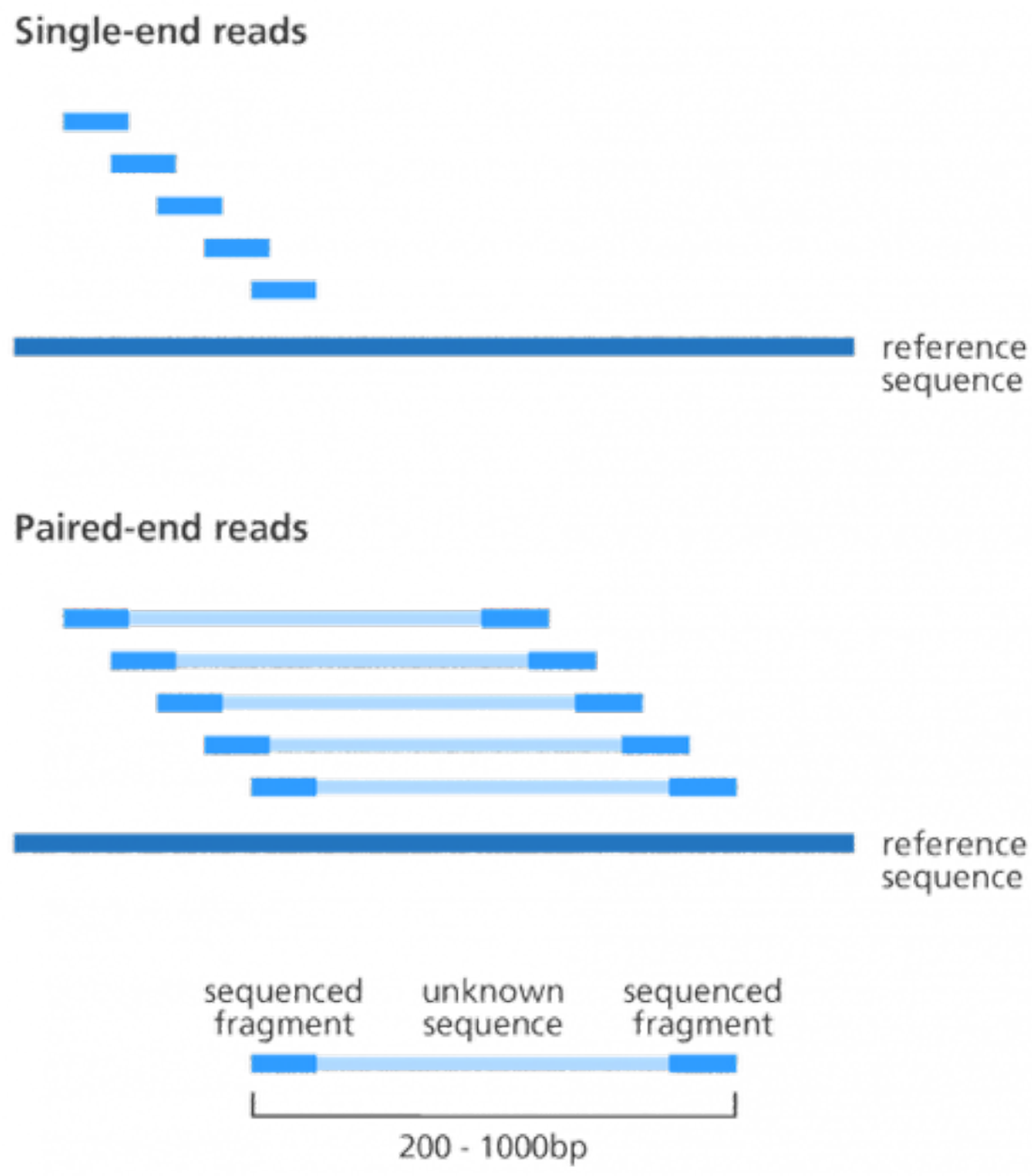
SRA Features: Gaps

- Existence of indels necessitates inserting or deleting nucleotides while mapping a sequence to a reference genome (gaps). The complexity of choosing a gap location increases with the read length. Therefore, some tools do not allow any gaps while others limit their locations and numbers.



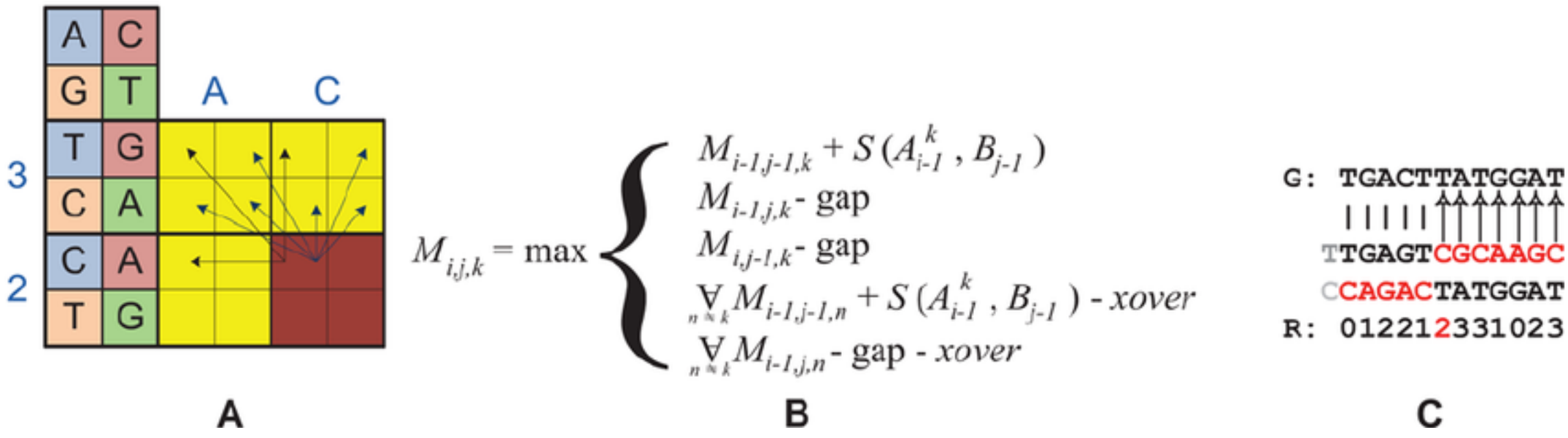
SRA Features: Paired-End

- Paired-end reads result from sequencing both ends of a DNA molecule. Mapping paired-end reads increases the confidence in the mapping locations due to having an estimation of the distance between the two ends.



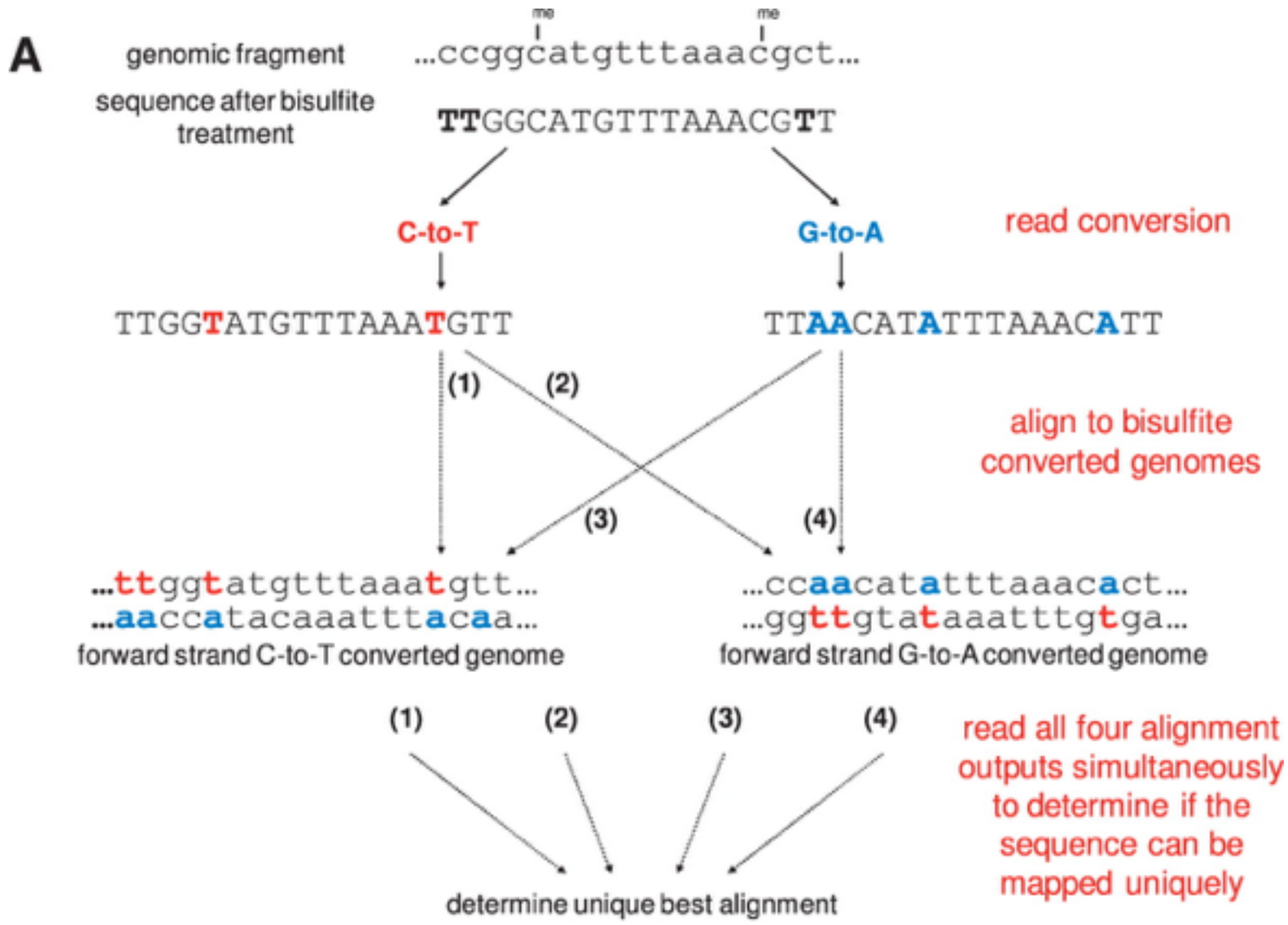
SRA Features: Color-Space

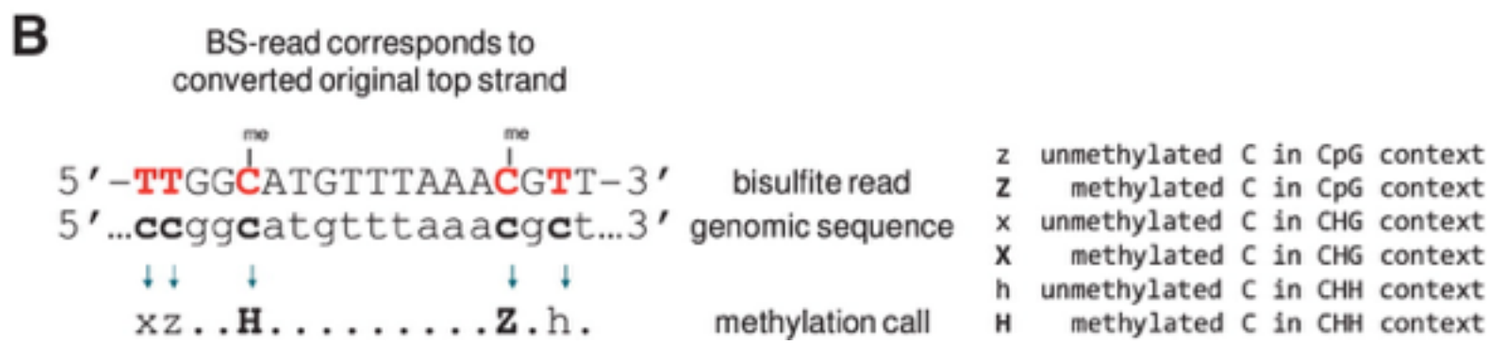
- Color space read is a read type generated by SOLiD sequencers. In this technology, overlapping pairs of letters are read and given a number (color) out of four numbers [17]. The reads can be converted into bases, however, performing the mapping in the color space has advantages in terms of error detection.



SRA Features: Bisulphite

- Bisulphite treatment is a method used for the study of the methylation state of the DNA [3]. In bisulphite treated reads, each unmethylated cytosine is converted to uracil. Therefore, they require special handling in order not to misalign the reads.





Short Read Aligners

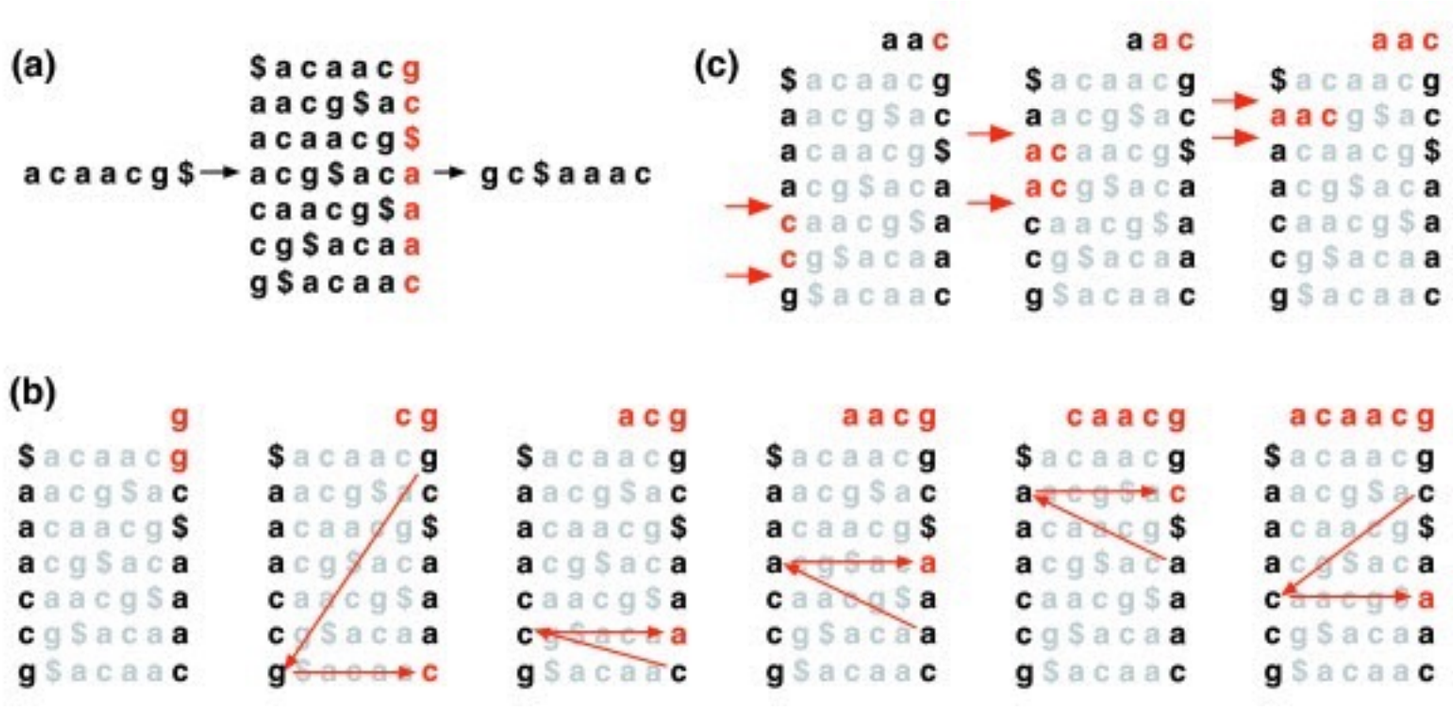
Table 1 Features supported by the tools

	Bowtie	Bowtie2	BWA	SOAP2	MAQ	RMAP	GSNAP	FANGS	Novoalign	mrFAST	mrsFAST
Seed mm.	Up to 3		Any	Up to 2	Any	Any					
Non-seed mm.	QS	AS	Count	Count	QS	Count	Count	Count	QS	Count	Count
Var. seed len.	> 5		Any	> 28							
Mapping qual.		Yes	Yes		Yes				Yes		
Gapped align.		Yes	Yes	PE	PE		Yes	Yes	Yes	Yes	
Colorspace	Yes		Yes		Yes				Yes		
Splicing							Yes				
SNP tolerance							Yes				
Bisulphite reads						Yes	Yes		Yes	Yes	

PE: paired-end only, mm.: mismatches, QS: base quality score, count: total count of mismatches in the read, AS: alignment score, and empty cells mean not supported.

SRA: Burrows-Wheeler Transform

- BWT is an efficient data indexing technique that maintains a relatively small memory footprint when searching through a given data block. BWT was extended by Ferragina and Manzini to a newer data structure, named FM-index, to support exact matching. By transforming the genome into an FM-index, the lookup performance of the algorithm improves for the cases where a single read matches multiple locations in the genome. However, the improved performance comes with a significantly large index build up time compared to hash tables.



SRA: BWA

- BWA is a BWT based tool. The BWA tool uses the Ferragina and Manzini matching algorithm to find exact matches. To find inexact matches, the authors provided a new backtracking algorithm that searches for matches between substring of the reference genome and the query within a certain defined distance.

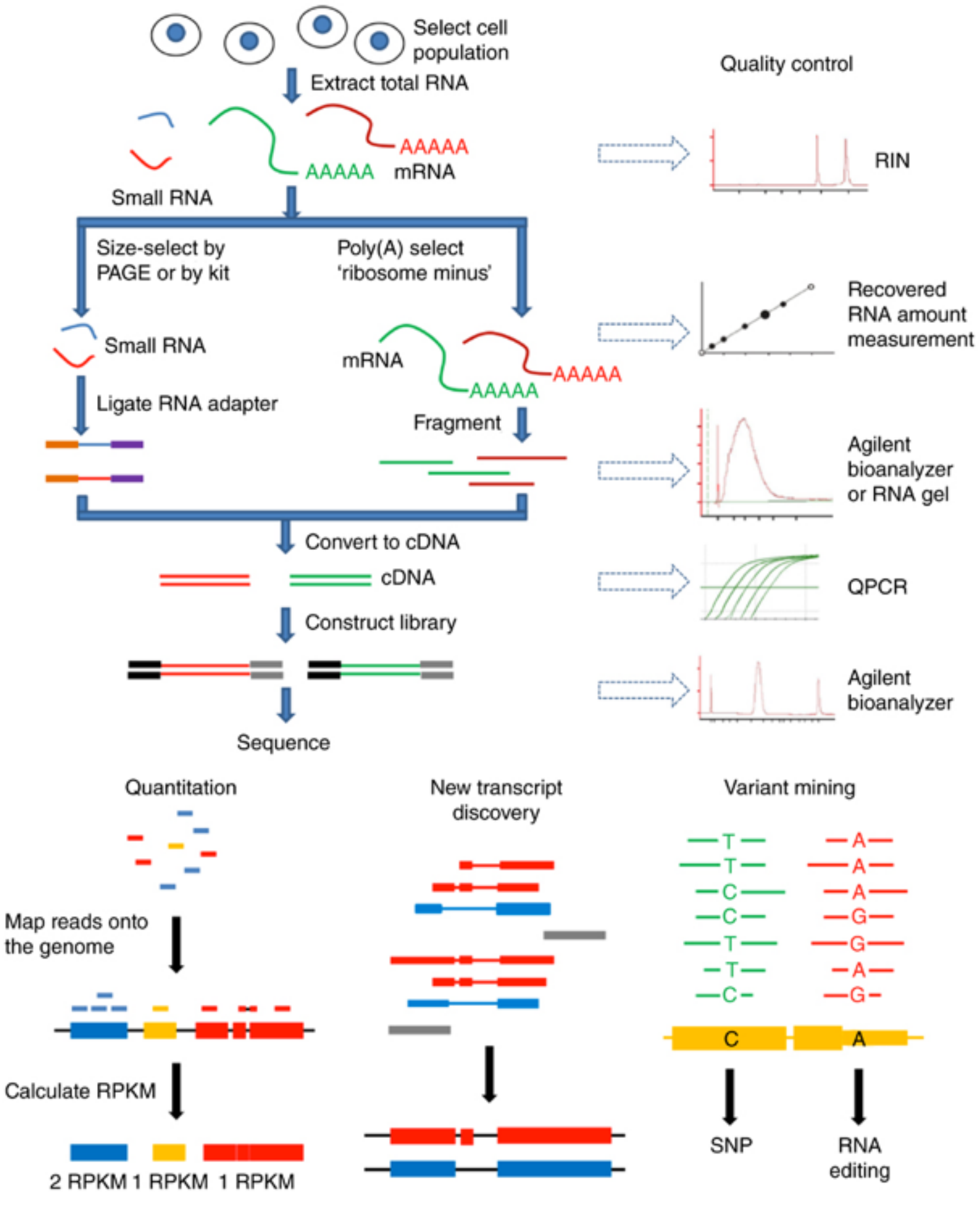
- BWA is fast, and can do gapped alignments. When run without seeding, it will find all hits within a given edit distance. Long read aligner is also fast, and can perform well for 454, Ion Torrent, Sanger, and PacBio reads. BWA is actively maintained and has a strong user community.

SRA: Bowtie

- Bowtie starts by building an FM-index for the reference genome and then uses the modified Ferragina and Manzini [39] matching algorithm to find the mapping location. There are two main versions of Bowtie namely Bowtie and Bowtie 2. Bowtie 2 is mainly designed to handle reads longer than 50 bps. Additionally, Bowtie 2 supports features not handled by Bowtie.

- Bowtie2 is faster than BWA for some types of alignment, but it takes a hit in sensitivity and specificity in some applications.

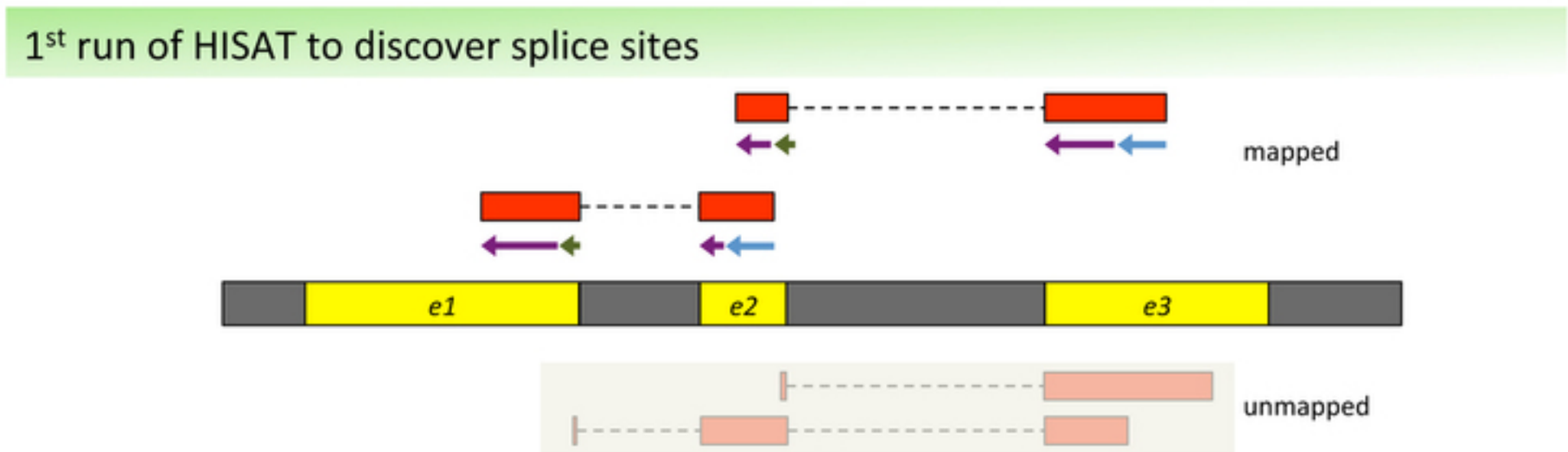
- Introduction to Mapping
- Short Read Aligners
- DNA vs RNA
- Alignment Quality
- Pitfalls and Improvements


Select cell population
Extract total RNA
AAAAA
AAAAA mRNA
Small RNA
Size-select by PAGE or by kit
Poly(A) select 'ribosome minus'
Small RNA
Ligate RNA adapter
mRNA
AAAAA
Fragment
Convert to cDNA
cDNA
Construct library
Sequence
Quality control
RIN
Recovered RNA amount measurement
Agilent bioanalyzer or RNA gel
QPCR
Agilent bioanalyzer
Quantitation
Map reads onto the genome
Calculate RPKM
2 RPKM 1 RPKM 1 RPKM
New transcript discovery
Variant mining
T T C C T T C
A A A G G A G
C A
SNP
RNA editing

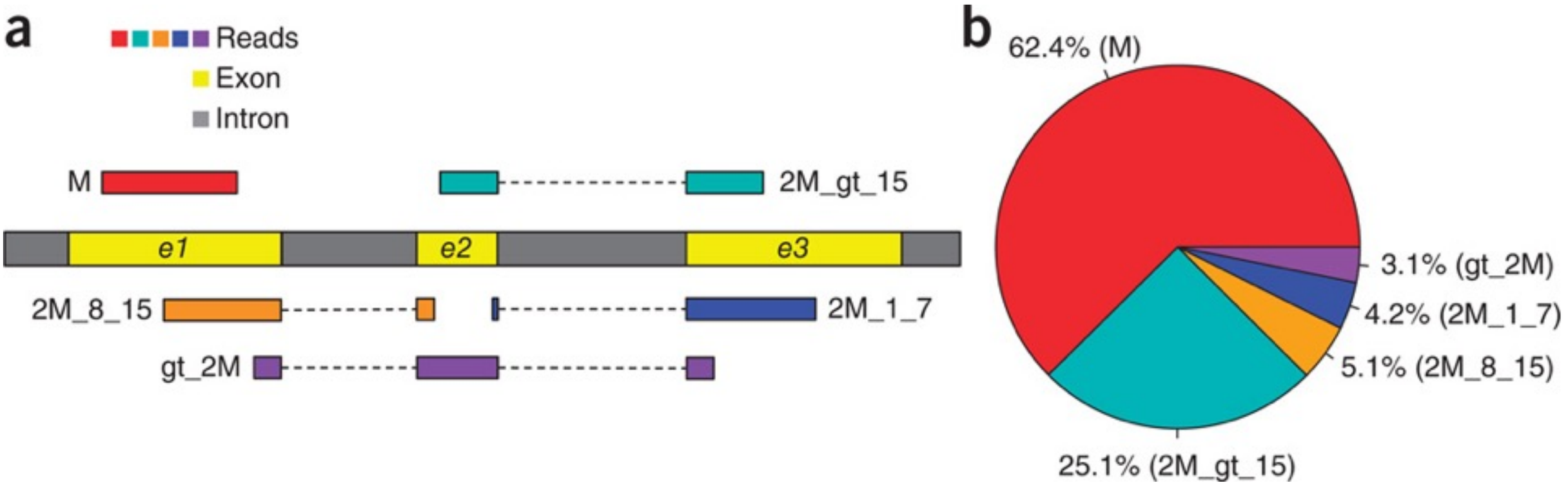
RNASeq

SRA Features: Splice-Aware

- Splicing refers to the process of cutting the RNA to remove the non-coding part (introns) and keeping only the coding part (exons) and joining them together. Therefore, when sequencing the RNA, a read might be located across exon-exon junctions. The process of mapping such reads back to the genome is hard due to the variability of the intron length. For instance, the intron length ranges between 250 and 65,130 nt in eukaryotic model organisms [37].



HiSAT2



RNA-seq read types and their relative proportions from 20 million simulated 100-bp reads. (a) Five types of RNA-seq reads: (i) M, exonic read; (ii) 2M_gt_15, junction reads with long, >15-bp anchors in both exons; (iii) 2M_8_15, junction reads with intermediate, 8- to 15-bp anchors; (iv) 2M_1_7, junction reads with short, 1- to 7-bp, anchors; and (v) gt_2M, junction reads spanning more than two exons. (b) Relative proportions of different types of reads in the 20 million 100-bp simulated read data.

Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015 Apr;12(4):357-60. doi: 10.1038/nmeth.3317. Epub 2015 Mar 9. PubMed PMID: 25751142; PubMed Central PMCID: PMC4655817.

HiSAT2

Sensitivity and precision of leading spliced aligners

Program	no. of splice sites reported	no. of true splice sites reported	sensitivity (%)	Precision (%)
HISATx1	91,904	85,546	97.3	93.1
HISATx2	90,331	85,603	97.3	94.8
HISAT	90,300	85,587	97.3	94.8
STAR	95,892	84,678	96.3	88.3
STARx2	92,254	84,734	96.3	91.8
GSNAP	92,547	85,598	97.3	92.5
OLego	86,779	82,879	94.2	95.5
TopHat2	96,474	79,705	90.6	82.6

Sensitivity and precision of leading spliced aligners for 87,944 true splice sites contained in 20 million simulated reads from the human genome, with a mismatch rate of 0.5%. Sensitivity is the percentage of true splice sites found out of the total that were present. Precision (or positive predictive value) is the percentage of reported splice sites that are correct.

Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015 Apr;12(4):357-60. doi: 10.1038/nmeth.3317. Epub 2015 Mar 9. PubMed PMID: 25751142; PubMed Central PMCID: PMC4655817.

- Introduction to Mapping
- Short Read Aligners
- DNA vs RNA
- Alignment Quality
- Pitfalls and Improvements

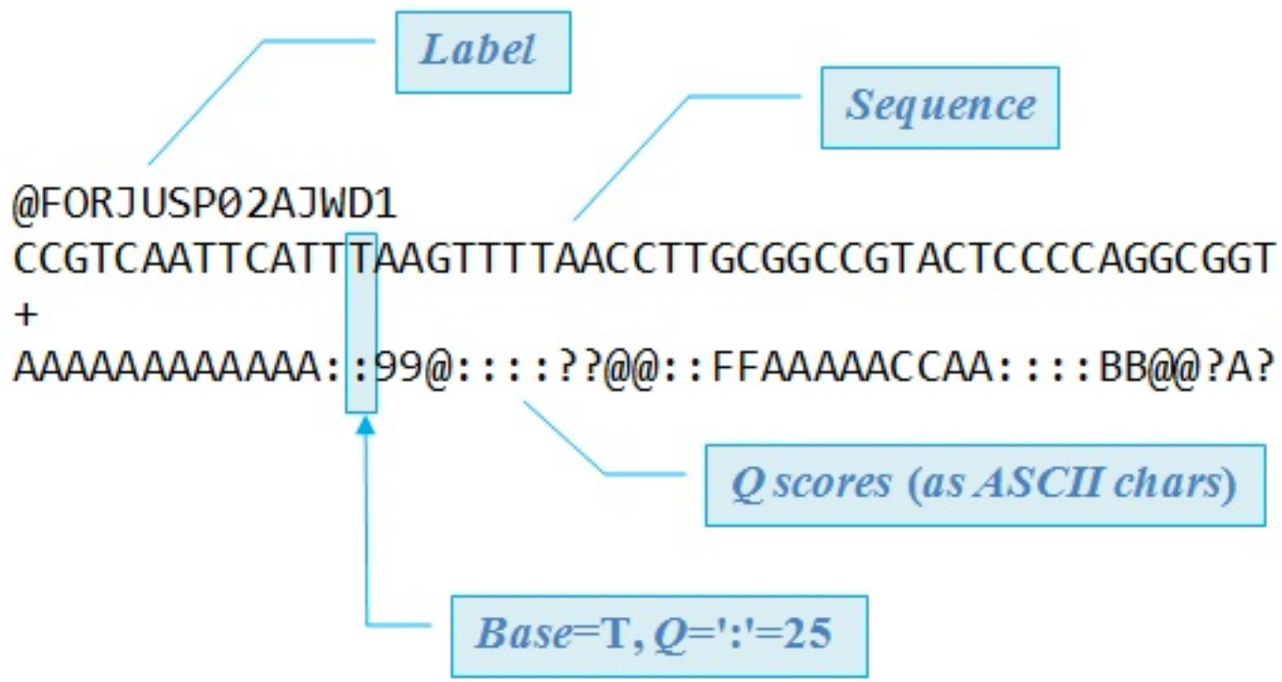
Mapping Quality

- Probability that a read is mapped incorrectly
- Factors include:
  - uniqueness (one best scoring alignment)
  - number of mismatches
  - number of gaps
  - quality of the bases (Phred)

Alignment Metrics

- Alignment Rate (Mapping Rate)
- Paired Alignments
  - Properly Paired Mapping
    - Comparing the rate of read pairs mapped within a certain proximity
  - Average Insert Size
    - Distance between adapter sequences
- Duplication Rate
  - Same fragment size can be duplicated during library preparation (PCR) or during sequencing (colony formation)

File Formats: FastQ



File Formats: SAM

Field	Regular expression	Range	Description
QNAME	[^ \t\n\r]+		Query pair NAME if paired; or Query NAME if unpaired [2]
FLAG	[0-9]+	[0,2^{16}-1]	bitwise FLAG (Section 2.2.2)
RNAME	[^ \t\n\r@=]+		Reference sequence NAME [3]
POS	[0-9]+	[0,2^{29}-1]	1-based leftmost POSition/coordinate of the clipped sequence
MAPQ	[0-9]+	[0,2^{8}-1]	MAPping Quality (phred-scaled posterior probability that the mapping position of this read is incorrect) [4]
CIGAR	([0-9]+[MIDNSHP])+|*		extended CIGAR string
MRNM	[^ \t\n\r@]+		Mate Reference sequence NaMe; "=" if the same as <RNAME> [3]
MPOS	[0-9]+	[0,2^{29}-1]	1-based leftmost Mate POSition of the clipped sequence
ISIZE	-?[0-9]+	[-2^{29},2^{29}]	inferred Insert SIZE [5]
SEQ	[acgtnACGTN.=]+|*		query SEQuence; "=" for a match to the reference; n/N/. for ambiguity; cases are not maintained [6,7]
QUAL	[!-~]+|*	[0,93]	query QUALity; ASCII-33 gives the Phred base quality [6,7]
TAG	[A-Z][A-Z0-9]		TAG
VTYPE	[AifZH]		Value TYPE
VALUE	[^\t\n\r]+		match <VTYPE> (space allowed)

File Formats: SAM

```
@HD     VN:1.4
@SQ     SN:insert       LN:599
@SQ     SN:ref1 LN:45
@SQ     SN:ref2 LN:40
@SQ     SN:ref3 LN:4
@RG     ID:fish PG:donkey
@RG     ID:cow  PU:13_&^&&*(:332
@RG     PU:*9u8jkjjkjd: ID:colt
@PG     ID:bull PP:donkey
@PG     ID:donkey
@PG     ID:moose
@PG     PP:moose        ID:cow
@CO
r000    99      insert  50      30      10M     =       80      30      ATTTAGCTAC      AAAAAAAAAA      RG:Z:cow        PG:Z:bull
r000    211     insert  80      30      10M     =       50      -30     CCCAATCATT      AAAAAAAAAA      RG:Z:cow        PG:Z:bull
r001    163     ref1    7       30      8M4I4M1D3M      =       37      39      TTAGATAAAGAGGATACTG     *       XX:B:S,12561,2,20,112   YY:i:100        RG:Z:fish       PG:Z:
r002    0       ref1    9       30      1S2I6M1P1I1P1I4M2I      *       0       0       AAAAGATAAGGGATAAA       *       XA:Z:abc        XB:i:-10        PG:Z:colt
r003    0       ref1    9       30      5H6M    *       0       0       AGCTAA  *       RG:Z:cow
r004    0       ref1    16      30      6M14N1I5M       *       0       0       ATAGCTCTCAGC    *       RG:Z:colt       PG:Z:colt
r003    16      ref1    29      30      6H5M    *       0       0       TAGGC   *       RG:Z:cow        PG:Z:colt
r001    83      ref1    37      30      9M      =       7       -39     CAGCGCCAT       *       RG:Z:fish       PG:Z:colt
x1      0       ref2    1       30      20M     *       0       0       AGGTTTTATAAAACAAATAA     *       RG:Z:colt       PG:Z:bull
x2      0       ref2    2       30      21M     *       0       0       GGTTTTATAAAACAAATAATT    ??????????????????????  RG:Z:colt       PG:Z:bull
x3      0       ref2    6       30      9M4I13M *       0       0       TTATAAAACAAATAATTAAGTCTACA       ???????????????????????????     RG:Z:fish       PG:Z:bull
x4      0       ref2    10      30      25M     *       0       0       CAAATAATTAAGTCTACAGAGCAAC        ????????????????????????        RG:Z:fish       PG:Z:bull
x5      0       ref2    12      30      24M     *       0       0       AATAATTAAGTCTACAGAGCAACT         ????????????????????????        RG:Z:fish       PG:Z:bull
x6      0       ref2    14      30      23M     *       0       0       TAATTAAGTCTACAGAGCAACTA ??????????????????????? RG:Z:cow
u1      4       *       0       30      23M     *       0       0       TAATTAAGTCTACAGAAAAAAAA ???????????????????????
```

@SQ = Contigs/Chromosomes
@RG = Read Group
@PG = Program Info

File Formats: SAM Flags

Flag	Chr	Description
0x0001	p	the read is paired in sequencing
0x0002	P	the read is mapped in a proper pair
0x0004	u	the query sequence itself is unmapped
0x0008	U	the mate is unmapped
0x0010	r	strand of the query (1 for reverse)
0x0020	R	strand of the mate
0x0040	1	the read is the first read in a pair
0x0080	2	the read is the second read in a pair
0x0100	s	the alignment is not primary
0x0200	f	the read fails platform/vendor quality checks
0x0400	d	the read is either a PCR or an optical duplicate

Bitwise:

00000000001 => 0x0001 => 2^0 = 1 => PAIRED
00000000010 => 0x0002 => 2^1 = 2 => PAIR MAPPED
00000000100 => 0x0004 => 2^2 = 4 => READ UNMAPPED
00000001000 => 0x0008 => 2^3 = 8 => MATE UNMAPPED
00000010000 => 0x0001 => 2^4 = 16 => READ REVERSE
00000100000 => 0x0002 => 2^5 = 32 => MATE REVERSE
00001000000 => 0x0004 => 2^6 = 64 => FIRST IN PAIR
00010000000 => 0x0008 => 2^7 = 128 => SECOND IN PAIR
00100000000 => 0x0001 => 2^8 = 256 => ALIGN NOT PRIM.
01000000000 => 0x0002 => 2^9 = 512 => QUALITY FAILS
10000000000 => 0x0004 => 2^10 = 1024 => PCR DUPLICATE

- Introduction to Mapping
- Short Read Aligners
- DNA vs RNA
- Alignment Quality
- Pitfalls and Improvements

Pitfalls

- All of the down stream analysis is based on the alignment
- Improper alignment can lead to false positive variate calls
- Inaccurate abundance calculations

Misalignment

- The Human Genome has many duplications
- Misalignment can result from sequence repetition
- Misalignment can be improved with:
  - Increased read lengths
    - Finding multiple 35bp matches is more likely than finding multiple 100bp matches
  - Paired Sequencing
    - A mate pair can “anchor” another when there is multiple mapping of the other mate pair.