DNA Sequencing and Data Analysis

Prof Noam Shomron
Amit Levon

Lecture 4, May 8, 2025

DNA Sequencing and Data Analysis

Sequence Mapping and Alignment
SAM & BAM File Formats

Thursday 18:30 to 21:00

C.L03

nshomron@gmail.com

amit.levon@post.runi.ac.il

Why Do We Need Sequence Mapping?

Determine the origin of an unknown sequence

Find homologous sequences

Determine genomic position of a sequence

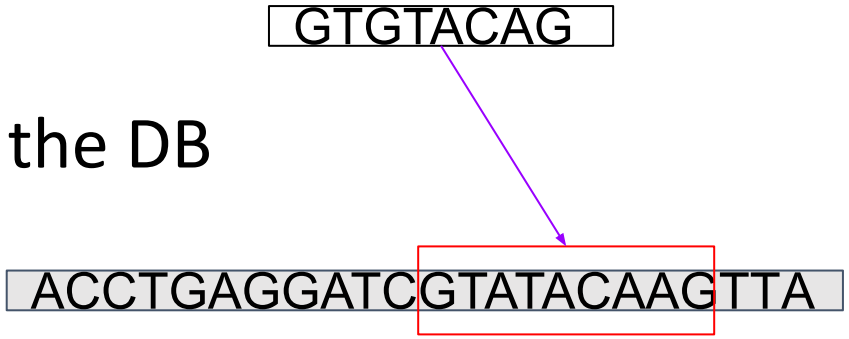
Identify genomic variants between samples (variant calling)

Determine the function of a sequence (annotation)

Two Stages of Sequence Mapping

1. **SEARCH -**

   Roughly find the position of the query in the DB



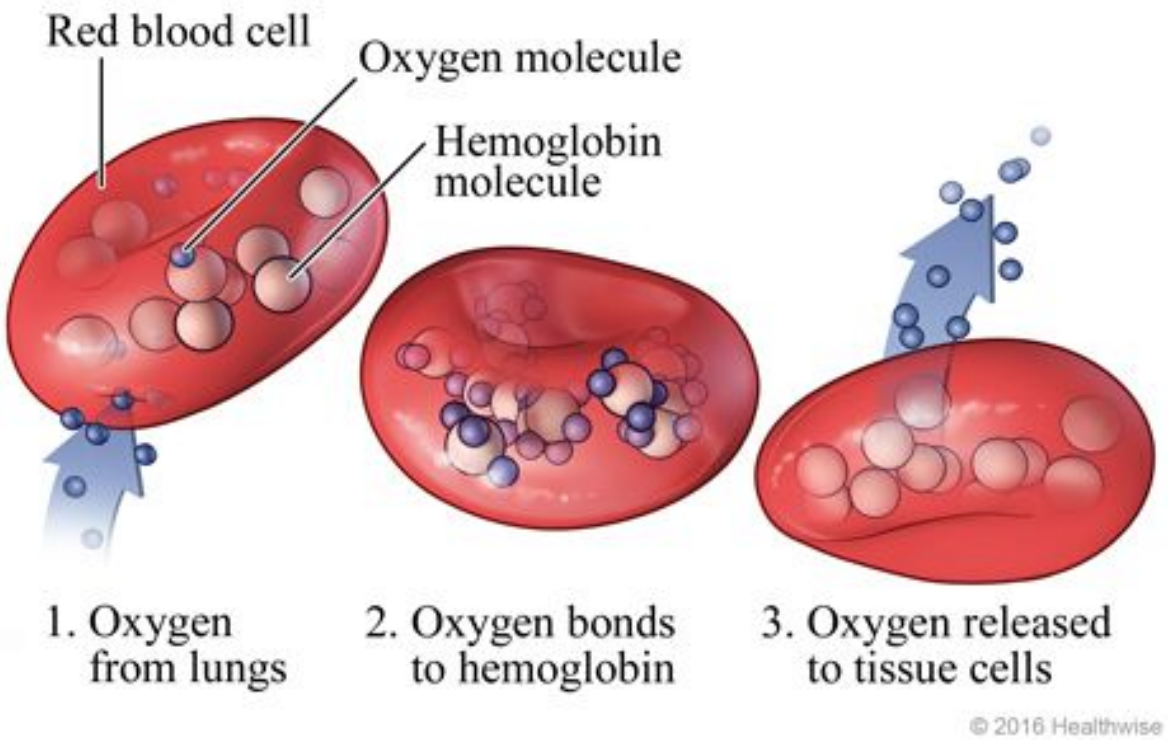
1. **ALIGN -**

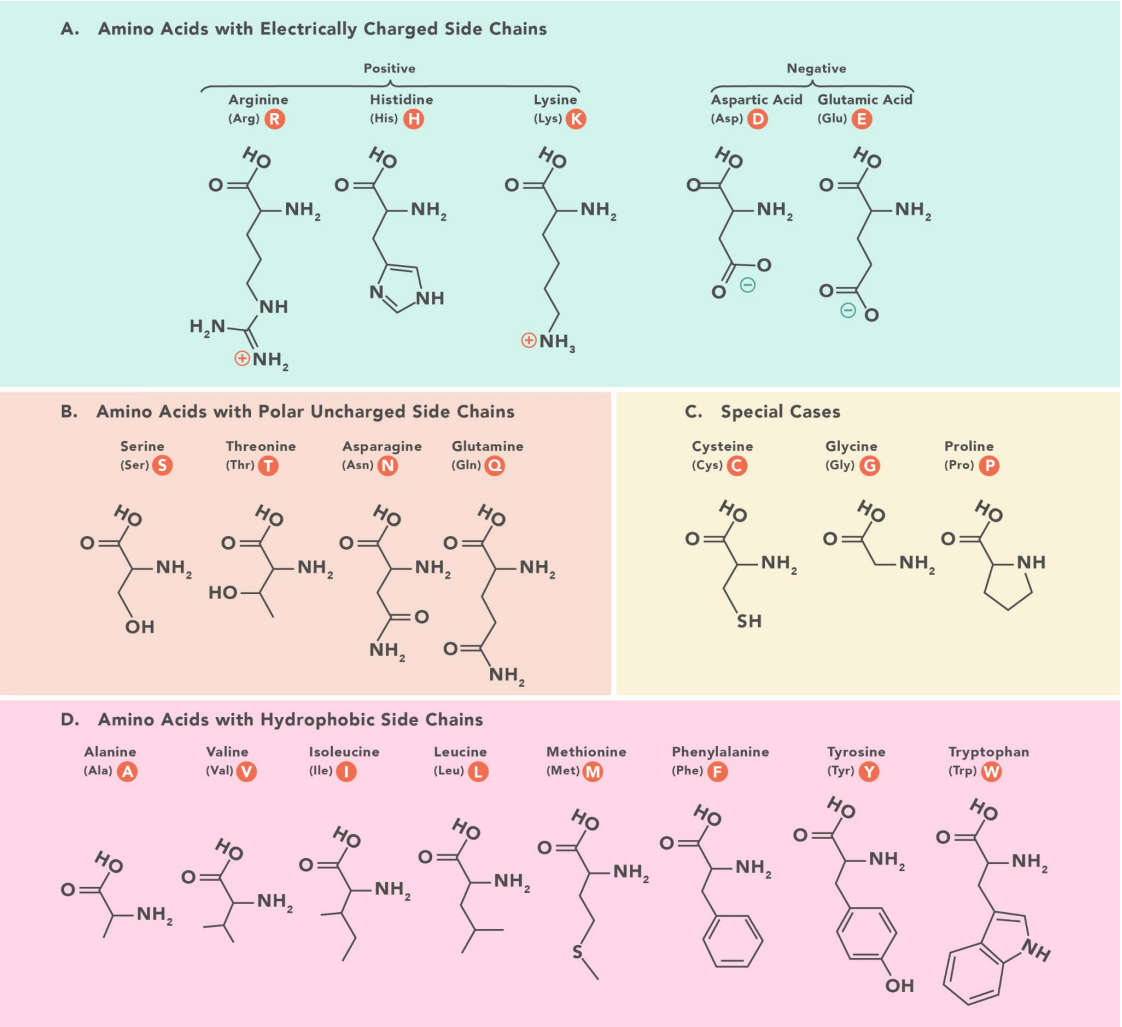
   Find the exact pairwise alignment of the query and the DB sequences

G	T	G	T	A	C	A	-	G
G	T	A	T	A	C	A	A	G

Pairwise Alignment

Hemoglobin Homologous





Pairwise Alignment

Hemoglobin Homologous

```
# NCBI Reference Sequence: NP_000508.1 (human hemoglobin subunit A)
r1 = skbio.Protein("MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADAL\
            TNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR")

# NCBI Reference Sequence: NP_001004376.1 (chicken hemoglobin subunit A)
r2 = skbio.Protein("MVLSAADKNNVKGIFTKIAGHAEEYGAETLERMFTTYPPTKTYFPHFDLSHGSAQIKGHGKKVVAAL\
            IEAANHIDDIAGTLSKLSDLHAHKLRVDPVNFKLLGQCFLVVVAIHHPAALTPEVHASLDKFLCAVGTVLTAKYR")

# GenBank: QFF91579.1 (sei whale hemoglobin subunit A)
q1 = skbio.Protein("MVLFPADKSNVKATWAKIGNHGAEYGAEALERMFMNFPSTKTYFPHFDLGHDSAQVKGHGKKVADAL\
            TKAAGHMDNLLDALSDLSDLHAHKLRVDPVNFKLLSHCLLVTLALHLPAEFTPSVHASLDKFLASVSTVLTSKYR")
```

Pairwise Alignment

The Hamming Distance

Hamming distance is the number of symbols or positions of two strings at which their corresponding characters are different

```
def hamming_distance(string1, string2):
    if (len(string1) != len(string2)):
        raise Exception('Strings must be of equal length.')
    dist_counter = 0
    for n in range(len(string1)):
        if string1[n] != string2[n]:
            dist_counter += 1
    return dist_counter / len(string1)
```

The Hamming distance between r1 and q1 is: 0.1690
The Hamming distance between r2 and q1 is: 0.3169

Pairwise Alignment

The Hamming Distance

```
# NCBI Reference Sequence: XP_028905054.1 (platypus hemoglobin subunit A);
q2 = skbio.Protein("MLTDAEKKEVTALWGKAAGHGEEYGAEALERLFQAFPTTKTYFSHFDLSHGSAQIKAHGKKVADA\
LSTAAGHFDDMDSALSALSDLHAHKLRVDPVNFKLLAHCILVVLARHCPGEFTPSAHAAMDKFLSKVATVLTSKYR")
q2
```

```
Protein
----------------------------------------------------------------------
Stats:
    length: 141
    has gaps: False
    has degenerates: False
    has definites: True
    has stops: False
----------------------------------------------------------------------
0   MLTDAEKKEV TALWGKAAGH GEEYGAEALE RLFQAFPTTK TYFSHFDLSH GSAQIKAHGK
60  KVADALSTAA GHFDDMDSAL SALSDLHAHK LRVDPVNFKL LAHCILVVLA RHCPGEFTPS
120 AHAAMDKFLS KVATVLTSKY R
```

```
q2 = skbio.Protein("MLTDAEKKEVTALWGKAAGHGEEYGAEALERLFQAFPTTKTYFSHFDLSHGSAQIKAHGKKVADA\
LSTAAGHFDDMDSALSALSDLHAHKLRVDPVNFKLLAHCILVVLARHCPGEFTPSAHAAMDKFLSKVATVLTSKYR-")
```

Pairwise Alignment

The Hamming Distance

The Hamming distance between r1 and q2 is: 0.90845

The Hamming distance between r2 and q2 is: 0.92254

```
q2_aligned = skbio.Protein("M-LTDAEKKEVTALWGKAAGHGEEYGAEALERLFQAFPTTKTYFSHFDLSHGSAQIKAHGKKVADA\
            LSTAAGHFDDMDSALSALSDLHAHKLRVDPVNFKLLAHCILVVLARHCPGEFTPSAHAAMDKFLSKVATVLTSKYR")
print(r1)
print(q2_aligned)
```

```
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPV
M-LTDAEKKEVTALWGKAAGHGEEYGAEALERLFQAFPTTKTYFSHFDLSHGSAQIKAHGKKVADALSTAAGHFDDMDSALSALSDLHAHKLRVDPV
```

The Hamming distance between r1 and q2_aligned is: 0.27465

The Hamming distance between r2 and q2_aligned is: 0.34507

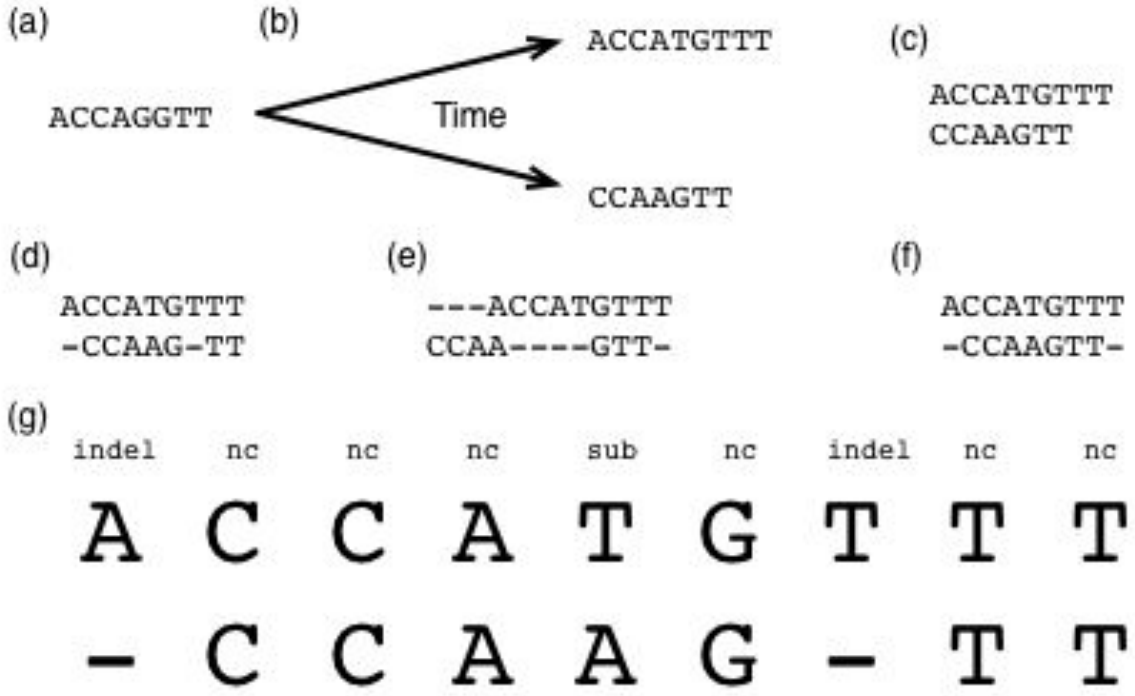
What Is Sequence Alignment?

Mutations:

Substitutions, where one DNA base is replaced with another

Insertions, where one or more contiguous DNA bases are inserted into a sequence

Deletions, where one or more contiguous DNA bases are deleted from a sequence.



Simple Align

ACCATGTTT

CCAAGTT

ACCATGTTT

-CCAAGTT -

S = -1+1+1-1+1+1+1-1=4

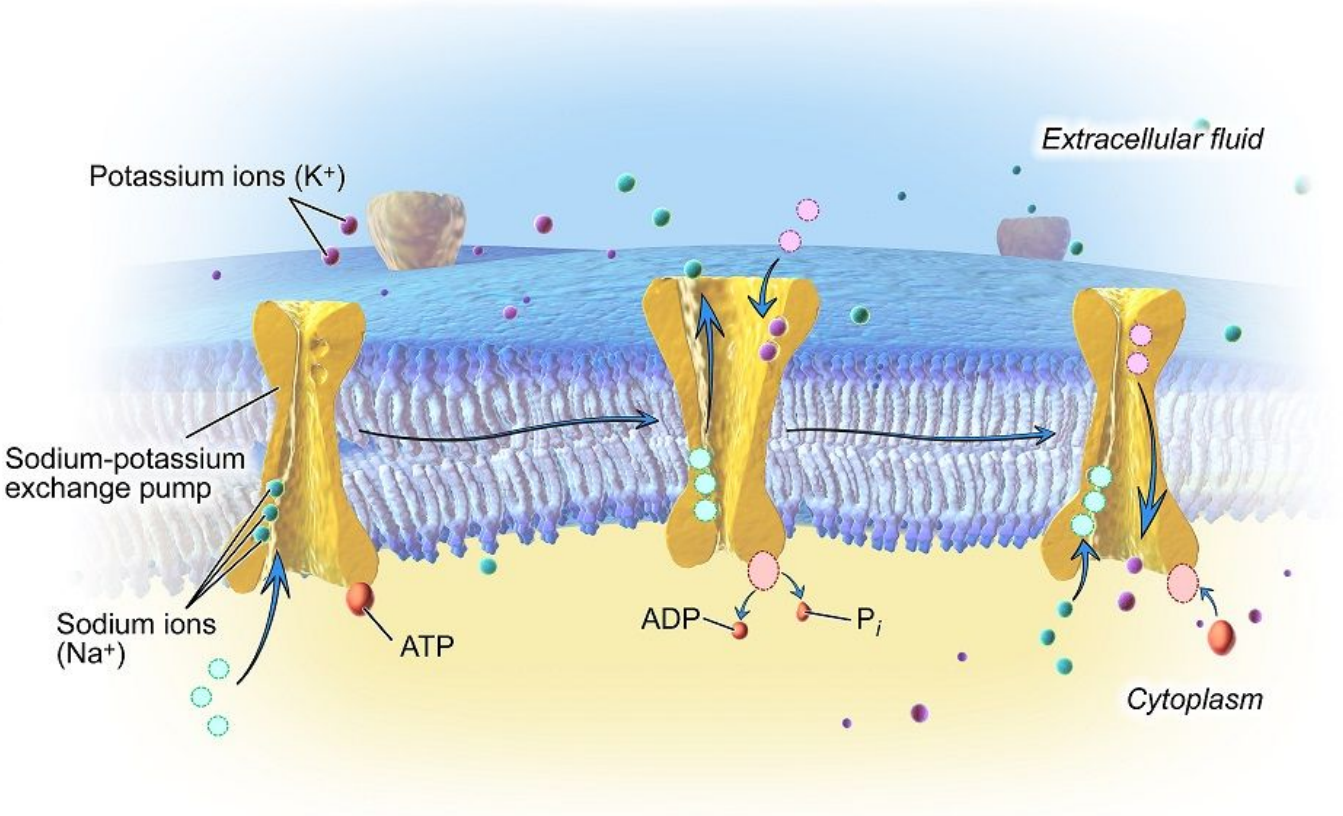
ACCA- -TGTTT

-CCAAG- - -TT

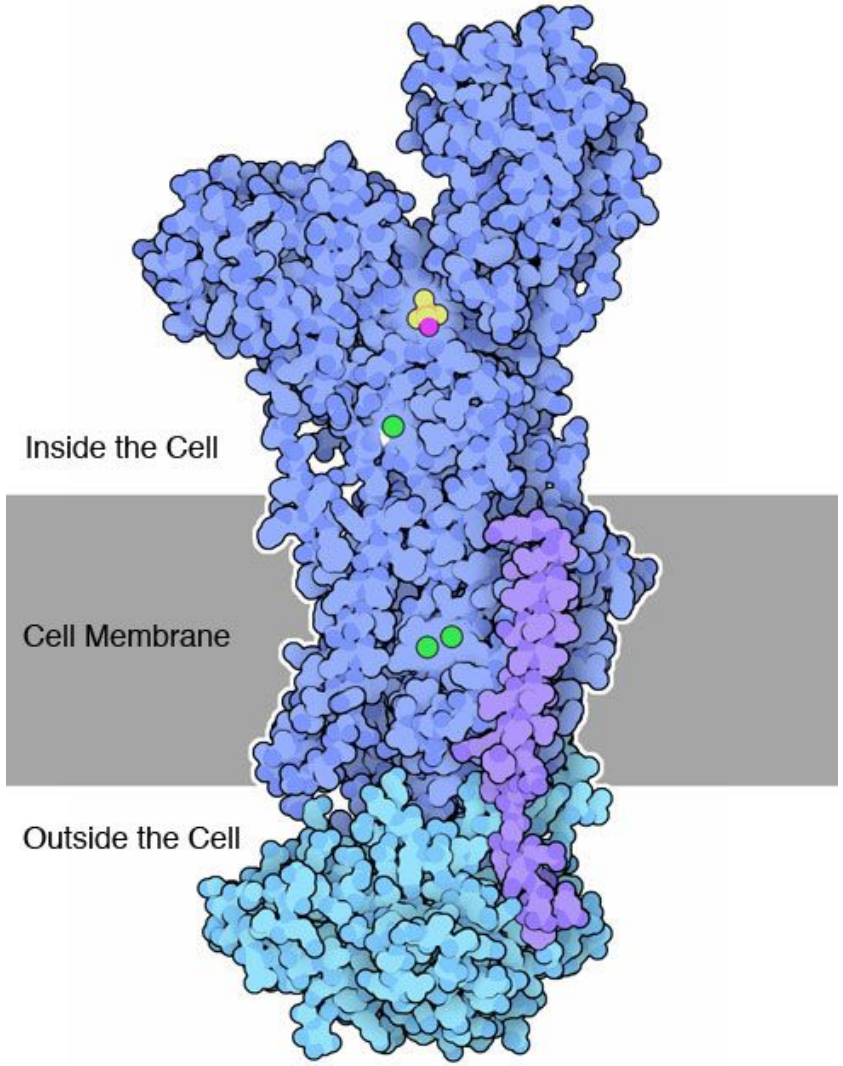
S = -1+1+1+1-1-1-1-1-1+1+1=-1

Simple Align

Too Simplistic



The Sodium-Potassium Exchange Pump



Simple Align

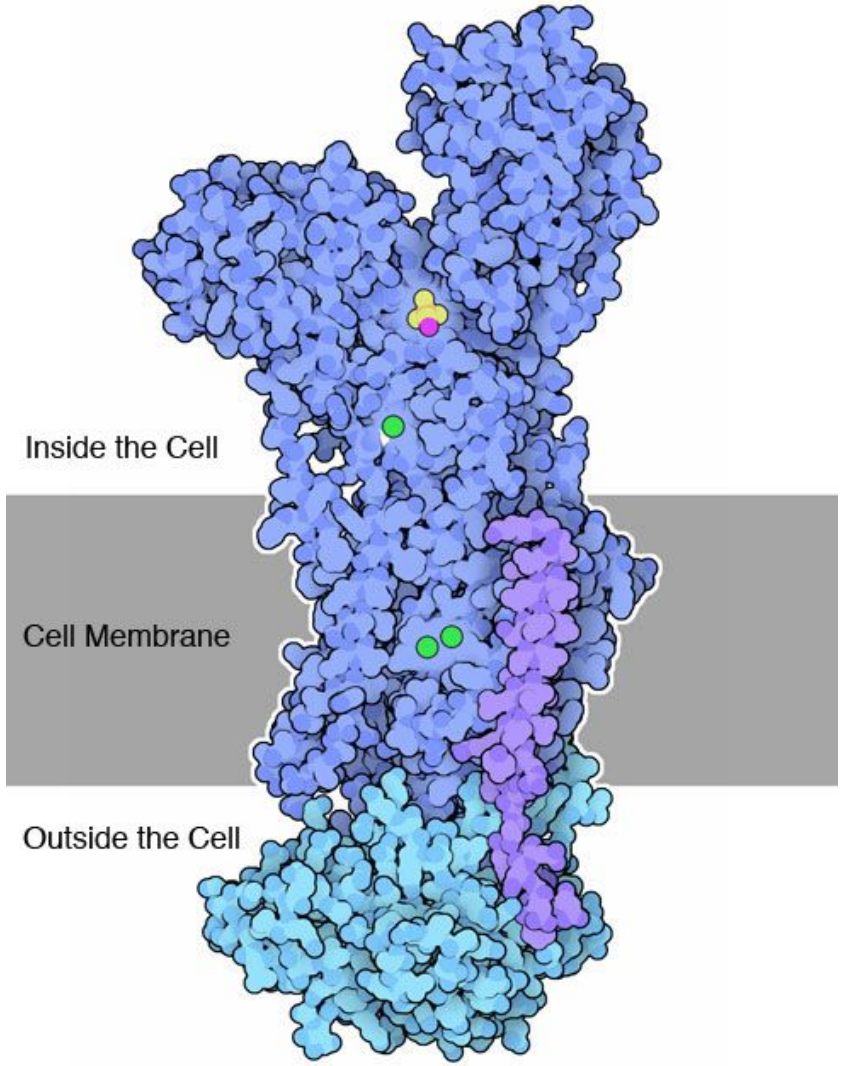
Too Simplistic

BLOSUM 62 scoring matrix

(positive values are shaded)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**4**																			
R	-1	**5**																		
N	-2	0	**6**																	
D	-2	-2	**1**	**6**																
C	0	-3	-3	-3	**9**															
Q	-1	**1**	0	0	-3	**5**														
E	-1	0	0	**2**	-4	**2**	**5**													
G	0	-2	0	-1	-3	-2	-2	**6**												
H	-2	0	**1**	-1	-3	0	0	-2	**8**											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	**4**										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	**2**	**4**									
K	-1	**2**	0	-1	-3	**1**	**1**	-2	-1	-3	-2	**5**								
M	-1	-1	-2	-3	-1	0	-2	-3	-2	**1**	**2**	-1	**5**							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	**6**						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	**7**					
S	**1**	-1	**1**	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	**4**				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	**1**	**5**			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	**1**	-4	-3	-2	**11**		
Y	-2	-2	-2	-3	-2	-1	-2	-3	**2**	-1	-1	-2	-1	**3**	-3	-2	-2	**2**	**7**	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	**3**	**1**	-2	**1**	-1	-2	-2	0	-3	-1	**4**

The values for amino acid substitutions were obtained from Henikoff S & Henikoff JG (1992) Amino acid substitutions matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89**: 10915-10919.

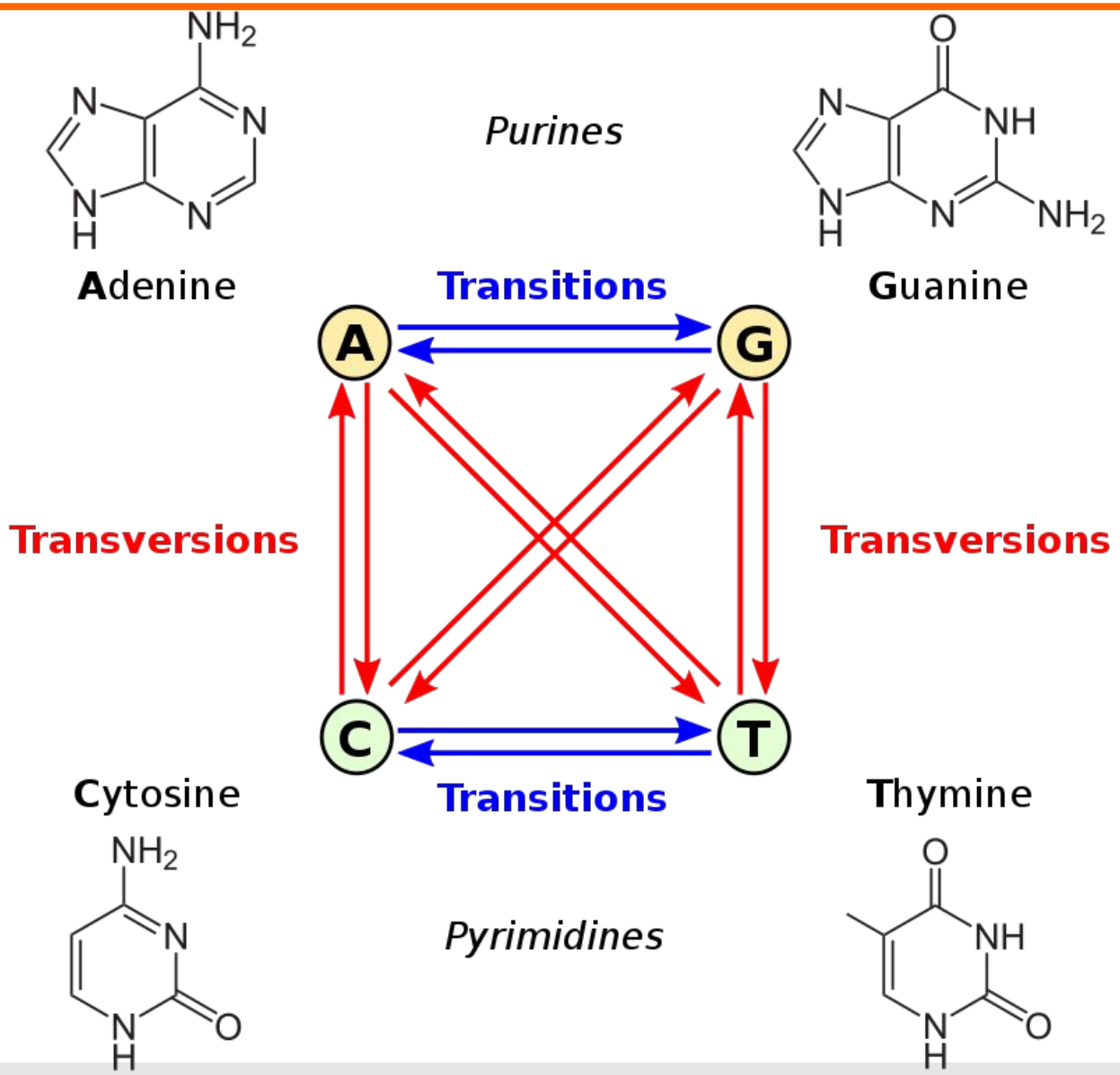


Simple Align

Seond letter

First letter	U	C	A	G	Third letter
U	UUU, UUC Phe UUA, UUG Leu	UCU, UCC, UCA, UCG Ser	UAU, UAC Tyr UAA Stop UAG Stop	UGU, UGC Cys UGA Stop UGG Trp	U C A G
C	CUU, CUC, CUA, CUG Leu	CCU, CCC, CCA, CCG Pro	CAU, CAC His CAA, CAG Gin	CGU, CGC, CGA, CGG Arg	U C A G
A	AUU, AUC, AUA Ile AUG Met	ACU, ACC, ACA, ACG Thr	AAU, AAC Asn AAA, AAG Lys	AGU, AGC Ser AGA, AGG Arg	U C A G
G	GUU, GUC, GUA, GUG Val	GCU, GCC, GCA, GCG Ala	GAU, GAC Asp GAA, GAG Glu	GGU, GGC, GGA, GGG Gly	U C A G

Simple Align



Local vs. Global Alignment

Global alignment - try to match entire sequences

Useful for closely-related sequences of similar size

Local alignment - allow partial matching

Useful for sequences expected to contain some similarity regions

Global Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Query Sequence

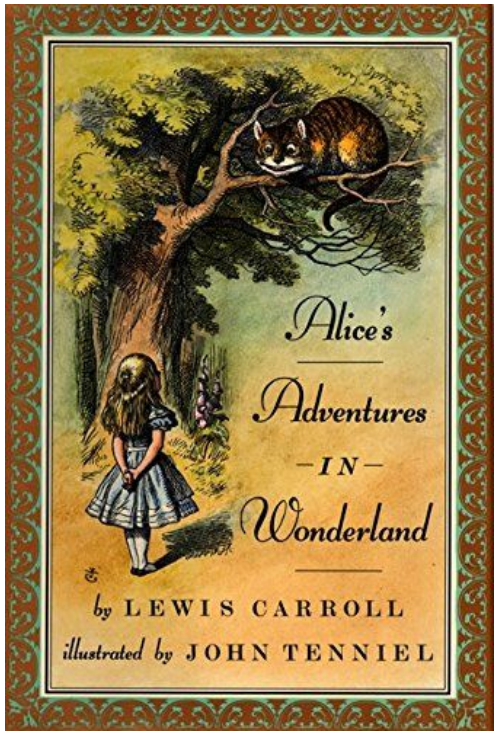
Local Alignment

Target Sequence

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
              |||| |||||| |||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Search

Imagine we have a big book...



… and we want to search it for a specific sentence

It would be
“ so nice if
something
made sense
for a
change.

Lewis Carroll
Alice in Wonderland

Search

- How can we do it in a timely manner?
  - Brute force
  - Indexing
- Do we allow slight changes?

  e.g. : *"it **c**ould be so nice if something made sense"*
- Do we allow insertions and deletions?

  e.g. : *"it would be ~~**so**~~ nice if something made **a little** sense"*
- What if the sentence is repeated in several places in the book?

It would be
" so nice if
something
made sense
for a
change.

Lewis Carroll
Alice in Wonderland

Sequence Mapping Challenges

Large DBs - millions to billions of nucleotides/AAs

Repetition - biological sequences tend to repeat

Noisy - sequencing errors and real biological variants

BLAST - Basic Local Alignment Search Tool

The most popular alignment tool

BLAST finds regions of similarity between biological sequences.

Compares nucleotide or protein sequences to sequence databases

Calculates the statistical significance of DB hits

Allows searching for **imperfect** sequence matches

Uses a **heuristic** algorithm to improve efficiency

bioljerk.tumblr.com

BLAST - Algorithm

1. Index the DB
2. Generate query words
3. compute neighbour words
4. Search the DB for exact word matches - seeds
5. Elongate and combine seeds to get final alignment
6. Score alignment

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

Mon, 17 Mar 2025

Improvements include upgrading to GCP Artifact Registry and better handling of job completion status in kubernetes version 1.30+.

ElasticBLAST 1.4.0 is now available!

More BLAST news...

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

BLAST Genomes

Enter organism common name, scientific name, or tax id

Search

Human Mouse Rat Microbes

Scale and Speed

We need to map millions to hundreds of millions of reads

Can we use Blast?

Blastn - ~100 reads / sec

Human genome - ~ 3Gb

Assume 100bp reads

How long to map x10 data to the human genome?

Hint: how many reads do we need?

Can We Use Blast in NGS?

Blastn - ~100 reads / sec

Human genome - ~ 3Gb

Assume 100bp reads

How long to map x10 data to the human genome?

Data required:
3 Gb x 10 = 30 Gb

Reads required:
30 Gb / 100 = 300 M reads

Time to map:
300 M reads / (100 reads/sec) = 3M sec = ~ **35 days**

BWA - Burrows-Wheeler Aligner

Specifically designed for mapping of short reads

Maps ~2,200 reads / sec (one CPU)

Allows parallel computing

Contains three algorithms - the most useful is **BWA-MEM**

BWA - Limitations

Only works for nucleotides (usually DNA, not RNA)

Less effective when:

- Queries are very long
- Reads are highly diverged from the reference
- Reads contain lots of sequencing errors

Usually offers a good accuracy-speed balance

BWA - Algorithm Overview

Step 1: Index the reference genome

Step 2: Search for reads

Indexing is based on the **Burrows-Wheeler's transformation**

Index allows easy searching:

- Quick
- Memory efficient

BWA - The Burrows Wheeler Transform

1. Create index of reference genome:

   Input: reference in fasta format

```
$ bwa index genome.fasta
```

1. Map reads to reference:

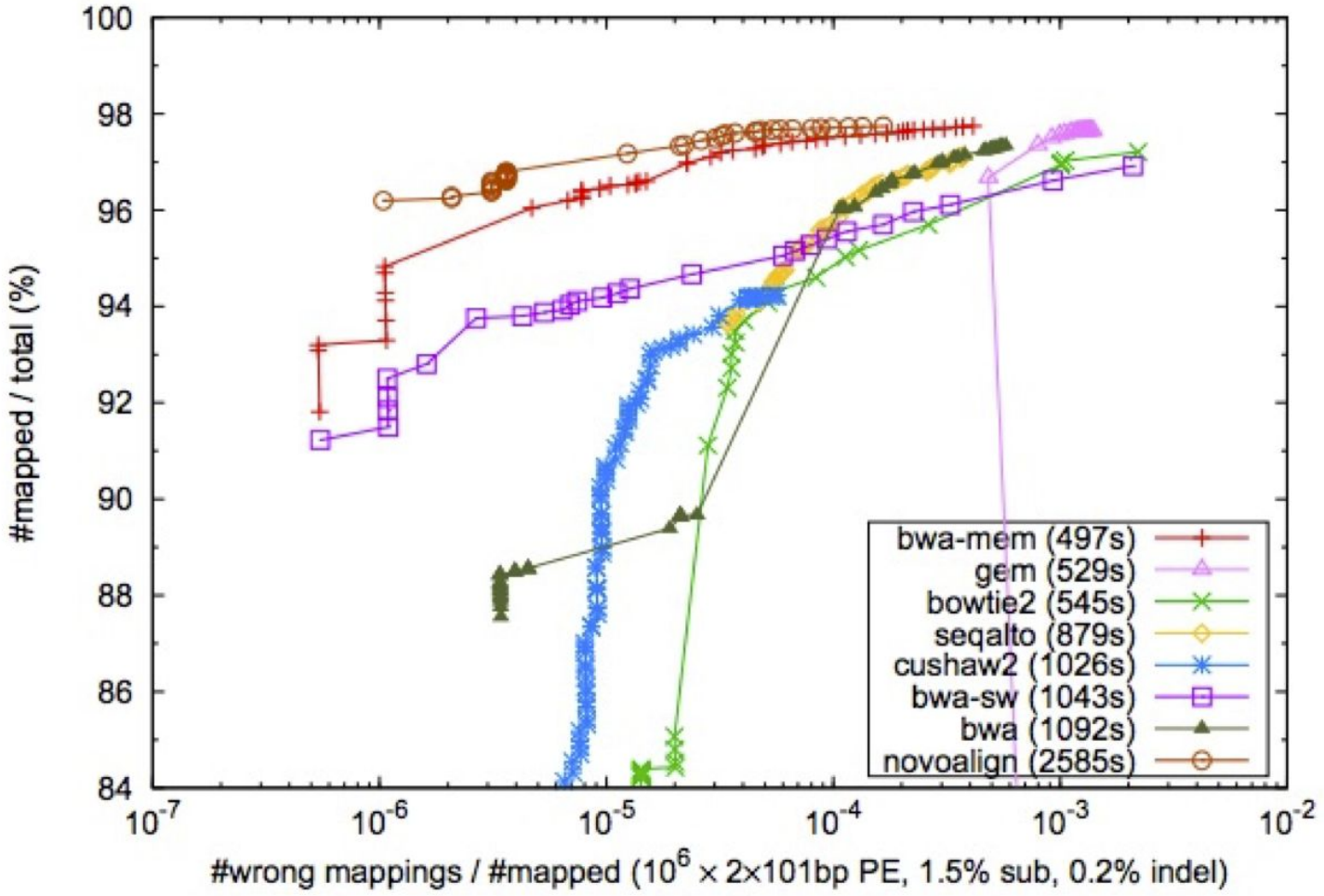
   Input: reads file or pair (for PE data) in fastq format

```
$ bwa mem genome.fasta reads_R1.fq reads_R2.fq -o aln.sam
```

Aligners Comparison

Aligner	Index	Applications	Availability
BWA-mem	Burrows-Wheeler	DNA, SE, PE	open-source
Bowtie2	Burrows-Wheeler	DNA, SE, PE	open-source
Novoalign	Hash-Based	DNA, SE, PE	propriety
TopHat	Burrows-Wheeler	RNA-seq	open-source
STAR	Hash-Based (reads)	RNA-seq	open-source
GSNAP	Hash-Based (reads)	RNA-seq	open-source

Aligners Comparison



https://arxiv.org/pdf/1303.3997v2.pdf

BWA-MEM Workflow

This takes a long time, but you do it <u>once</u>

Create BWT of reference genome.

```
$ bwa index grch38.fa
```

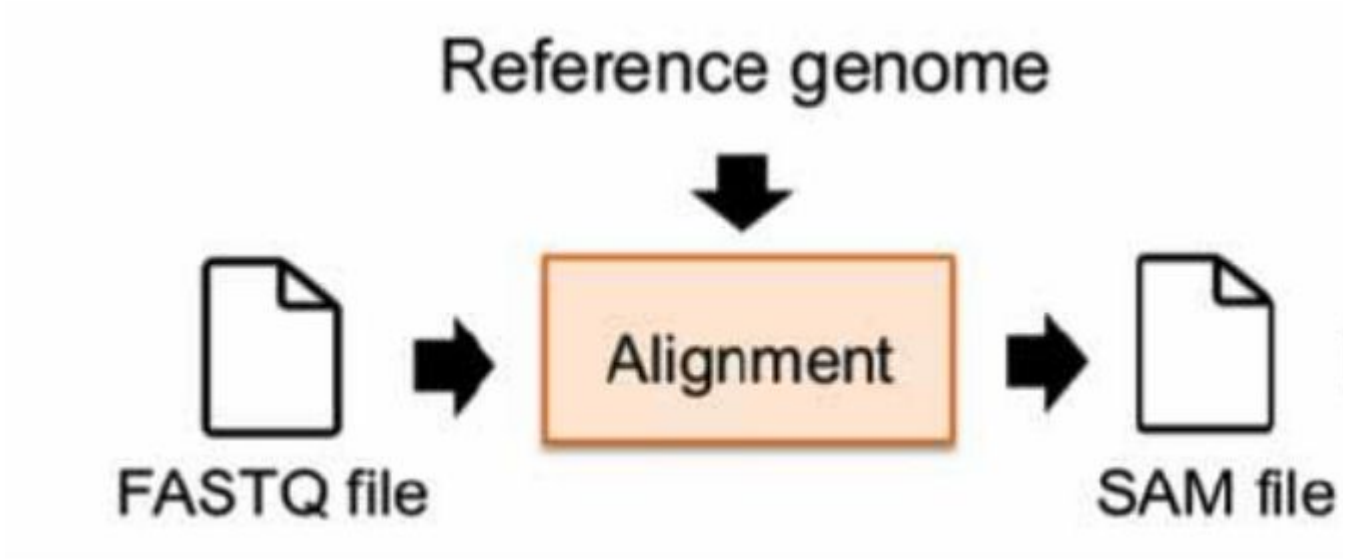
↓

Output is in SAM format.
Use multiple threads if you have a computer with multiple CPUs.

Align paired-end FASTQ to BWT index.

```
$ bwa mem -t 16 grch38.fa 1.fq 2.fq > sample.sam
```

FASTQ to SAM



Sequence Alignment and Mapping (SAM)

BIOINFORMATICS **APPLICATIONS NOTE** *Vol. 25 no. 16 2009, pages 2078–2079*
doi:10.1093/bioinformatics/btp352

Sequence analysis

The Sequence Alignment/Map format and SAMtools

Heng Li[1,†], Bob Handsaker[2,†], Alec Wysoker[2], Tim Fennell[2], Jue Ruan[3], Nils Homer[4], Gabor Marth[5], Goncalo Abecasis[6], Richard Durbin[1,*] and 1000 Genome Project Data Processing Subgroup[7]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, [2]Broad Institute of MIT and Harvard, Cambridge, MA 02141, USA, [3]Beijing Institute of Genomics, Chinese Academy of Science, Beijing 100029, China, [4]Department of Computer Science, University of California Los Angeles, Los Angeles, CA 90095, [5]Department of Biology, Boston College, Chestnut Hill, MA 02467, [6]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109, USA and [7]http://1000genomes.org

Received on April 28, 2009; revised on May 26, 2009; accepted on May 30, 2009

Advance Access publication June 8, 2009

Associate Editor: Alfonso Valencia

Table 1. Mandatory fields in the SAM format

No.	Name	Description
1	QNAME	Query NAME of the read or the read pair
2	FLAG	Bitwise FLAG (pairing, strand, mate strand, etc.)
3	RNAME	Reference sequence NAME
4	POS	1-Based leftmost POSition of clipped alignment
5	MAPQ	MAPping Quality (Phred-scaled)
6	CIGAR	Extended CIGAR string (operations: MIDNSHP)
7	MRNM	Mate Reference NaMe ('=' if same as RNAME)
8	MPOS	1-Based leftmost Mate POSition
9	ISIZE	Inferred Insert SIZE
10	SEQ	Query SEQuence on the same strand as the reference
11	QUAL	Query QUALity (ASCII-33=Phred base quality)

Sequence Alignment and Mapping (SAM)

What critical information do we need for sequence alignments?

SAM Format

Col #	Name	Meaning	Example
1	QNAME	Read or Pair name	HWI:ST156_1:278:1:1058:4544:0
2	FLAG	Bitwise FLAG	***soon!***
3	RNAME	Reference sequence name	chr1
4	POS	1-based alignment start coordinate	8,724,005
5	MAPQ	Mapping quality	***soon!***
6	CIGAR	Extended CIGAR string	***soon!***
7	MRNM	If paired, the mate's reference seq.	chr1
8	MPOS	If paired, the mate's alignment start	8,724,505
9	ISIZE	If paired, the insert size	562
10	SEQ	The sequence of the query/mate	ACAAATTCAG...
11	QUAL	The quality string for the query/mate	HHH$^^%$$$...
12	OPT	Optional Tags	XA:i:2, MD:Z:0T34G15

http://samtools.sourceforge.net/samtools.shtml

SAM Format

Read name | SAM flag | Reference - Chr 5 | Position on Chr 5 | MAPQ | CIGAR | Location of paired read | Distance between paired reads | Read sequence & Phred

```
SRR1572628.1    83      SL4.0ch05       25938772        60      100M    =       25938399        -473    ATGTGAAGGCGTTTGACTGG
GCATGGAGTTTAAGAGTGAAACAAACACTTTTTGAAACTTAAGGTCTAAAATAAGTCAGAAATTGTGTGGCTGTAAATCA    CCCCDBDDDDEEFFFFFHGFGFFHJIGIIEIJIJJJJJJI
HIHFBDBJJJJIIIJIGIJJJIJJJJJJJIJJJJJJJJJJJJIJJJIHHHHHFFFFFCCC    NM:i:0  MD:Z:100        AS:i:100        XS:i:21
SRR1572628.1    163     SL4.0ch05       25938399        60      100M    =       25938772        473     CACAGATGCTGGGAGATCTA
TTTCATATCCATGCAGTAAAATCAAAACATGGGCAGAAACCAAATCAATCTTGCCTCGACCCAAAGTTCTCATCAGTATC    CCCFFFFFHHHHHJJIJJJJJJJIJJIJJJIJJIJJGIIJ
JJJJJJJJJJHIJJIJJIJJJJJJJJJJIIJJIHHHHFFFDDDBBCCDEEDEEDDD@CDD    NM:i:0  MD:Z:100        AS:i:100        XS:i:0
SRR1572628.2    83      SL4.0ch11       27210743        60      100M    =       27210394        -449    TCACAAGCATGACGGACCAT
CACAAGCTCCGTAACCCCACACTTAGTCAGACATCCCTATCTTCAATCAGCAACTGCACTAAGCTGCAACCTACGGACCG    3A<A@<CC?895@>3@CC@;CBCC@?=(DDC?:5GGGC=C
=GFAGFDHHD?D3DIIHFB?DEFDEG?GGFHF@CA?CBIHIHFGGEC+CHFCADDDD@@@    NM:i:0  MD:Z:100        AS:i:100        XS:i:65
SRR1572628.2    163     SL4.0ch11       27210394        60      100M    =       27210743        449     CTCCAAACTTCATGACGAAC
CTACAGGACGGACTGTCGTAGACTTTACGGACCATCGTGGACTCCGTCACCCCACACTTAGACTTACCCATCTTCCTTCA    @@@FFFFFHHHHGIIIII6FGGGIIIG>BGHFH:FBFIB=
CG>FGIGHGGAHFECDF?BCBECDCB?ABD@BB7<8?<CCDCC@@A>39CCC@CDD<CCC    NM:i:0  MD:Z:100        AS:i:100        XS:i:25
SRR1572628.3    83      SL4.0ch01       55877048        60      100M    =       55876674        -474    TCTATCATCTTCGTGTTTAG
GATTATTGGTGAAATAGTACTACCACATAGAACTTTTTGGCAACACGCCTACTGTTCTTTTTCACCGCCTATTTGATCCT    CDEDDDDDDDDDDDDDCCDDEDDCBDEDDEEDEFC@A@F
CAHE?HEEIJIHJJIJJJJJIGFIIJIIIHEJJJJIJJJJIHF@IHIHGHHHFFFFFCCC    NM:i:0  MD:Z:100        AS:i:100        XS:i:0
SRR1572628.3    163     SL4.0ch01       55876674        60      100M    =       55877048        474     TAGCTAGACGTAAACAAAAG
GAAGAAAATGATCCCAAATAGACATAACATCAACCGAGTATGATAATTTTTCTCCATTCAAGTCTAGGATGAAGAATGTT    BCCFFDFFGHFHHJJJJIJIJJJHJJIIIIIJJIJJJIIJ
JBFHCHIGHGECHIJJIICEHIJGIHFEGHHFFFFFFFEEECCCCCDDDDDDDDDCCCAC    NM:i:0  MD:Z:100        AS:i:100        XS:i:0
```

MAPQ

MAPQ - mapping quality

Definition: $-10 \log_{10} \Pr\{\text{mapping position is wrong}\}$

The higher - the better

Usually between 0 and 60

Calculation of MAPQ is differ between aligners

It considers alignment score, Phred score and alternative mappings

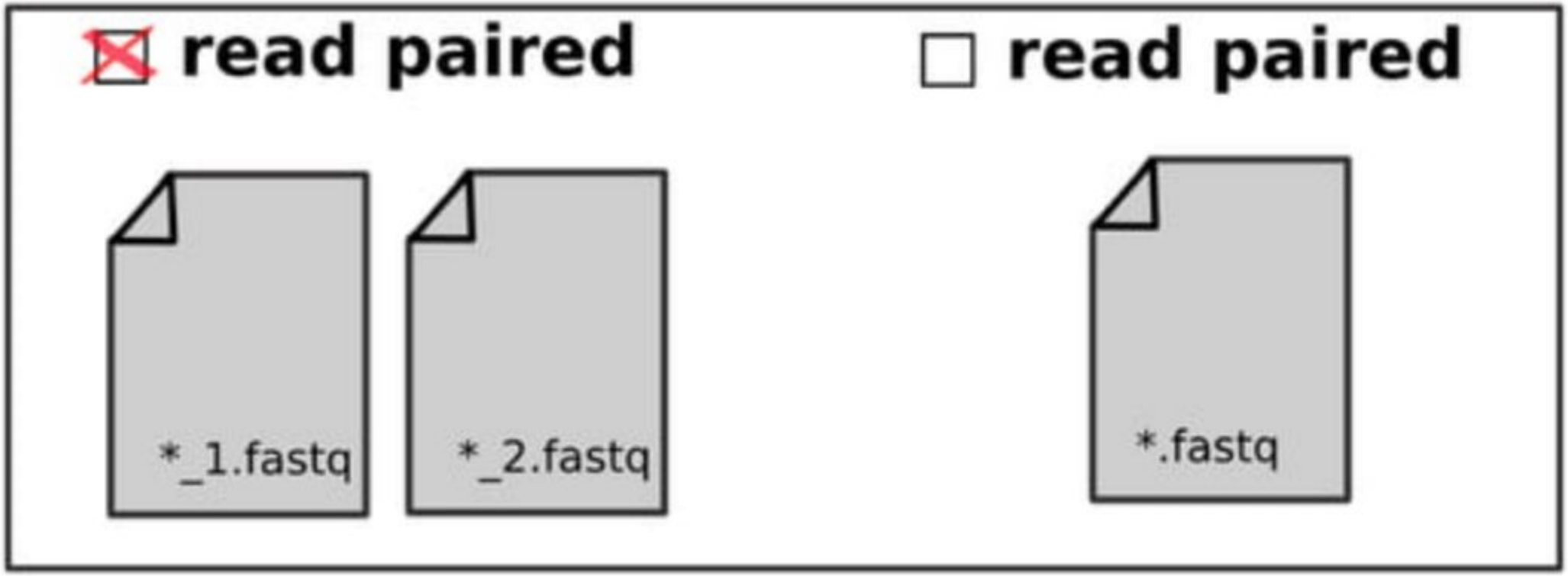
As a rule of thumb:

- MAPQ > 30 is considered a good mapping
- MAPQ 0 usually means ambiguous mapping

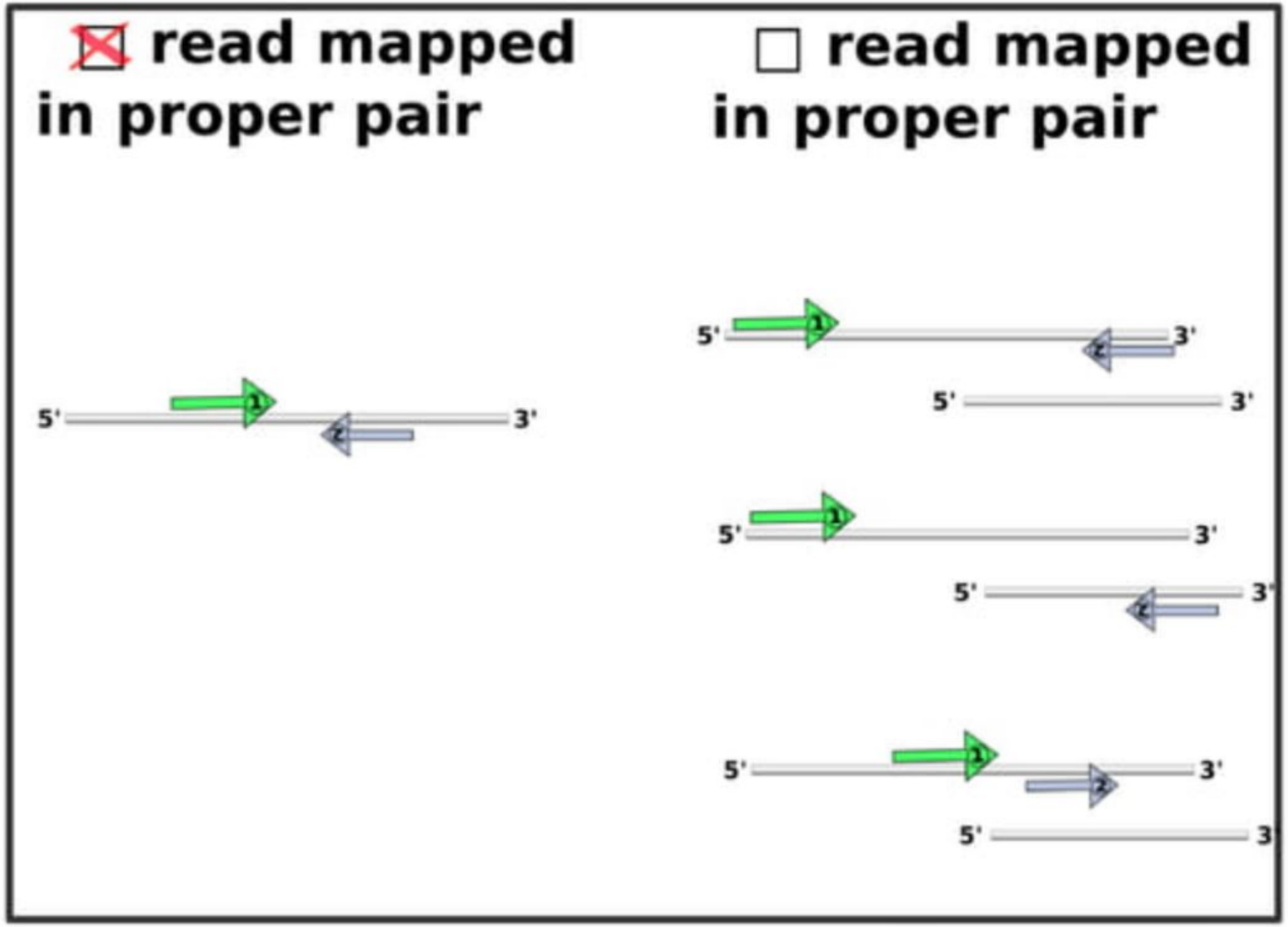
SAM Flag

base2	base10	base16	Meaning	Applies to:
00000000001	1	0x0001	The read originated from a paired sequencing molecule	Both
00000000010	2	0x0002	The read is mapped in a **proper** pair	Pairs only
00000000100	4	0x0004	The query sequence itself is unmapped	Both
00000001000	8	0x0008	The query's mate is unmapped	Pairs only
00000010000	16	0x0010	Strand of the query (0 for forward; 1 for reverse strand)	Both
00000100000	32	0x0020	Strand of the query's mate	Pairs only
00001000000	64	0x0040	The query is the first read in the pair	Pairs only
00010000000	128	0x0080	The read is the second read in the pair	Pairs only
00100000000	256	0x0100	The alignment is not primary	Both
01000000000	512	0x0200	The read fails platform/vendor quality checks	Both
10000000000	1024	0x0400	The read is either a PCR duplicate or an optical duplicate	Both

SAM Flag

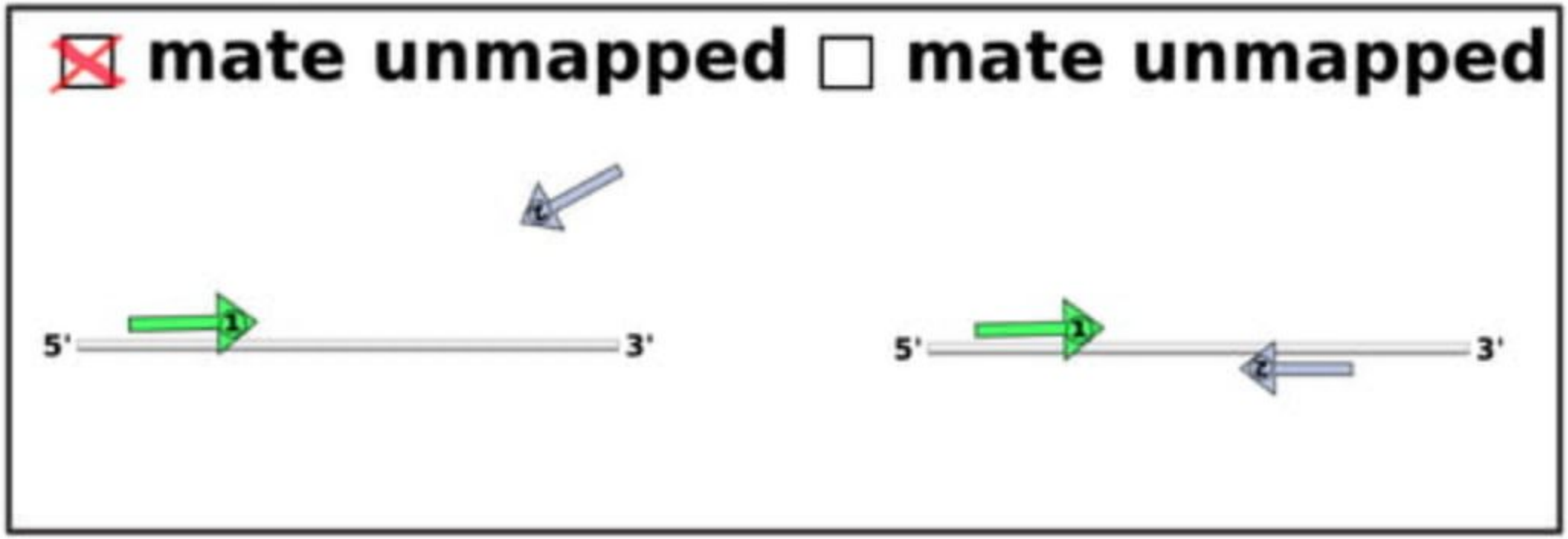


SAM Flag

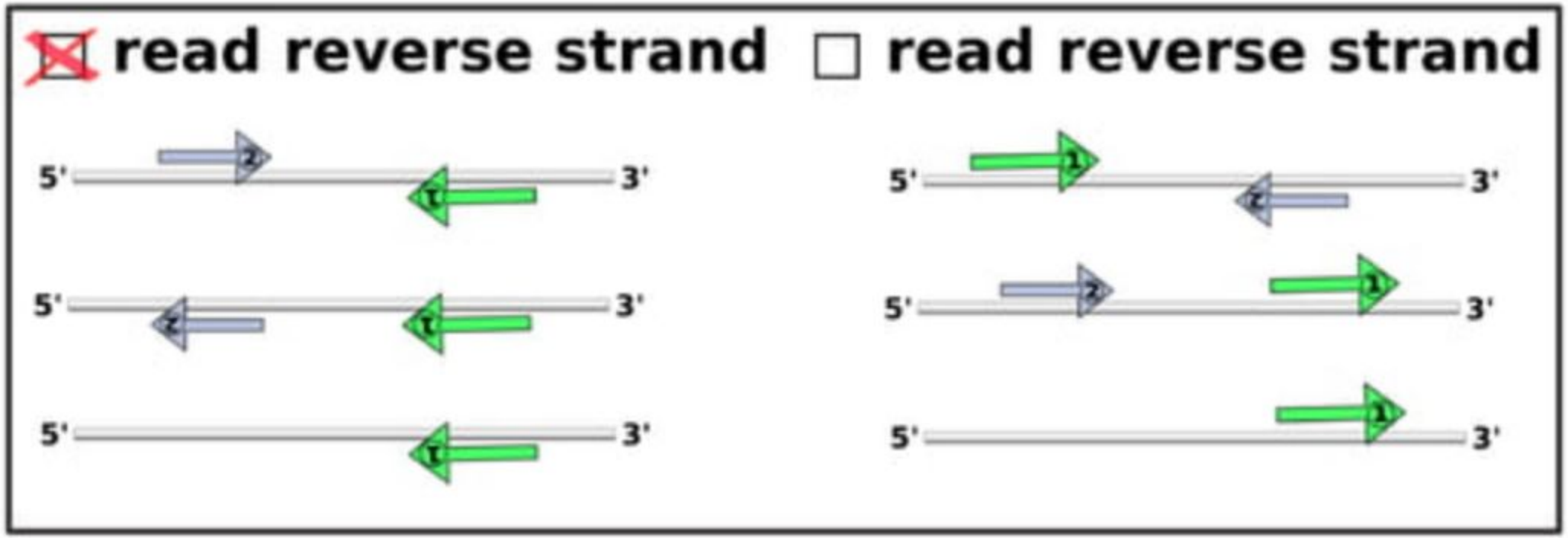


SAM Flag

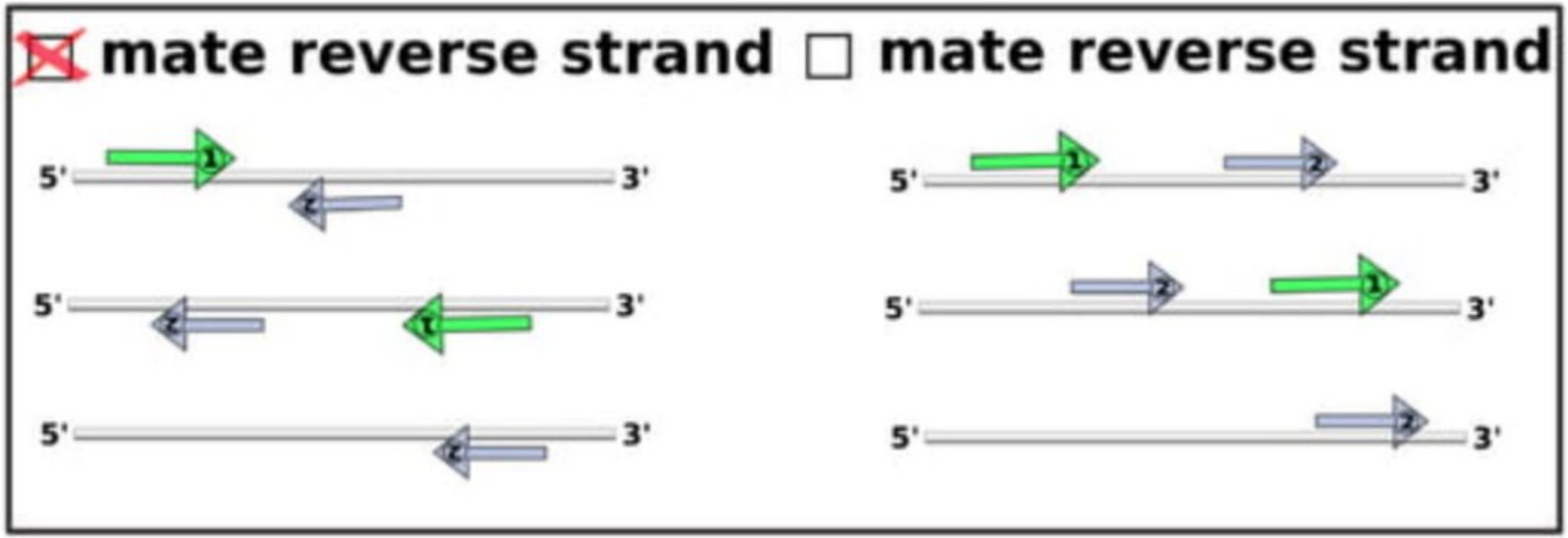
SAM Flag



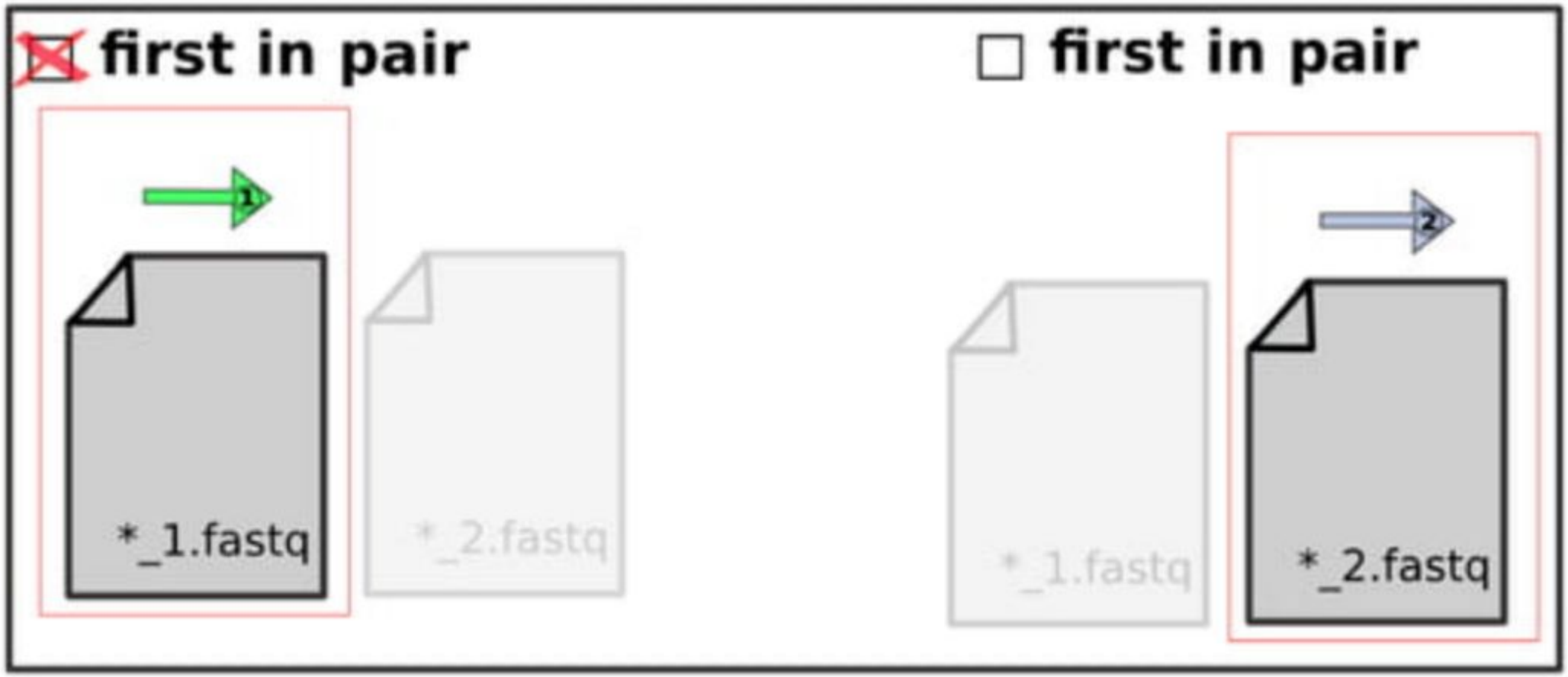
SAM Flag



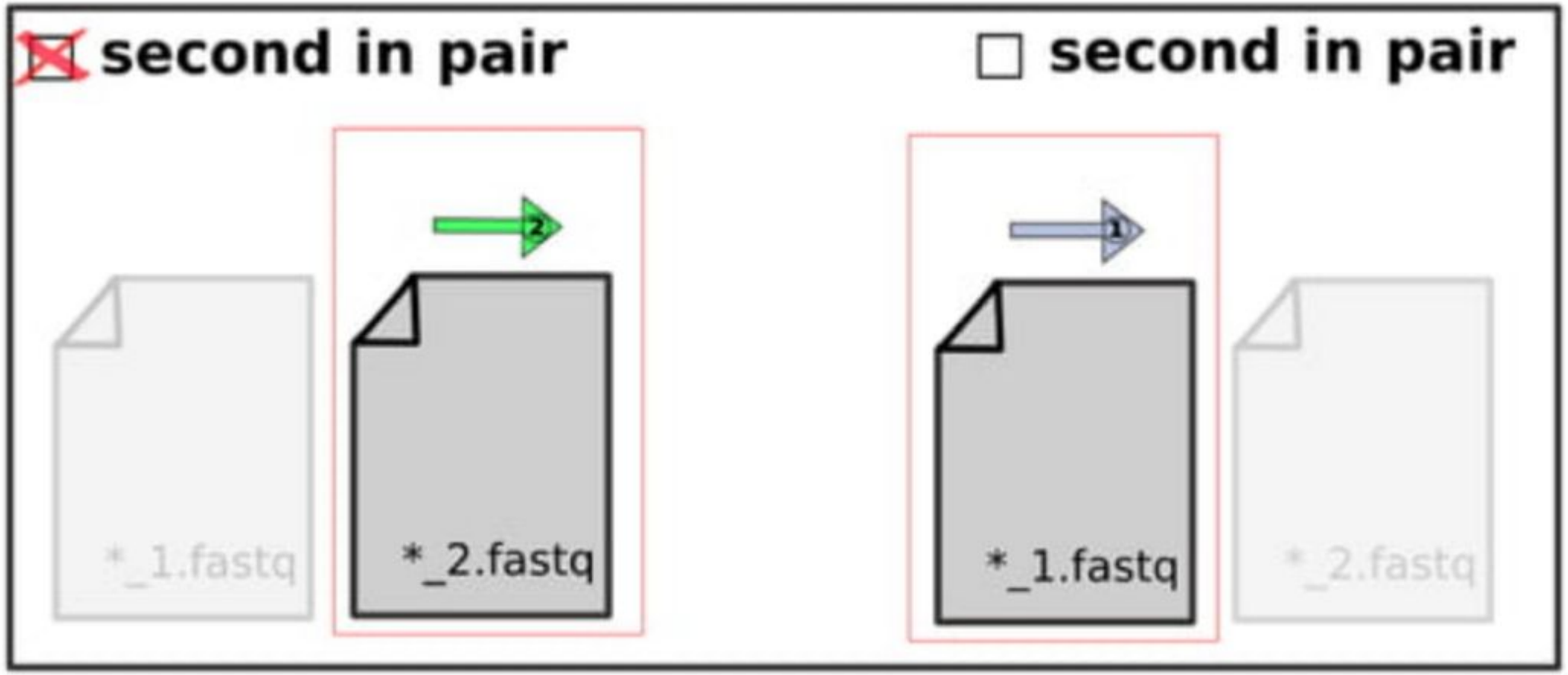
SAM Flag



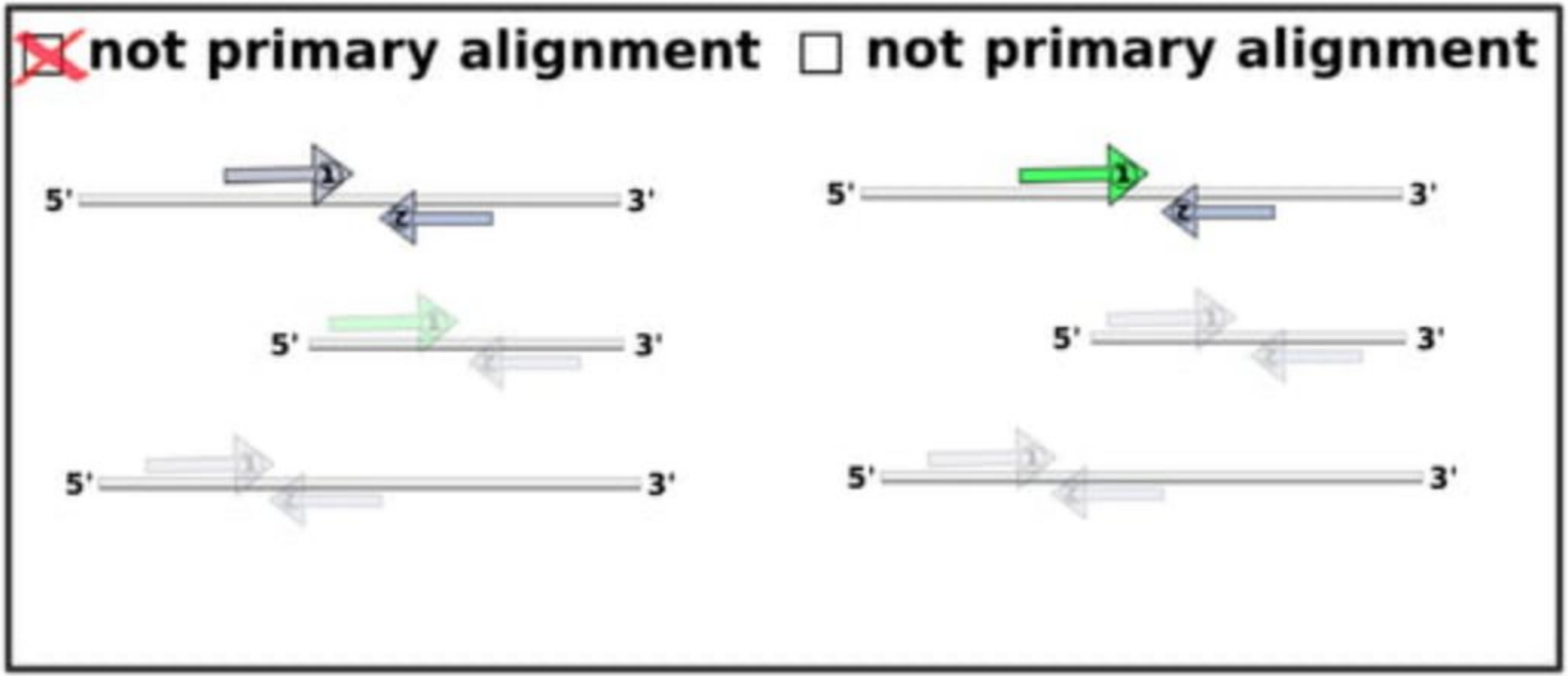
SAM Flag



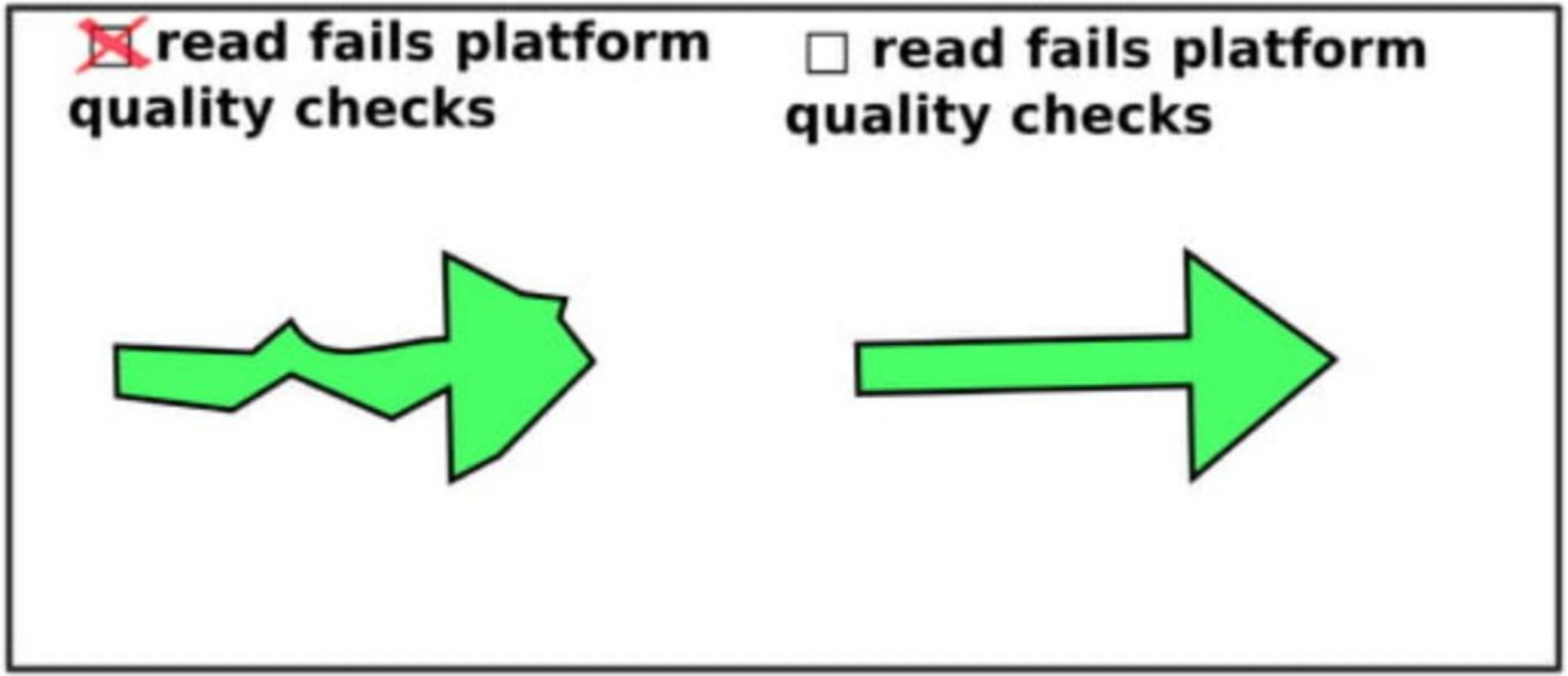
SAM Flag



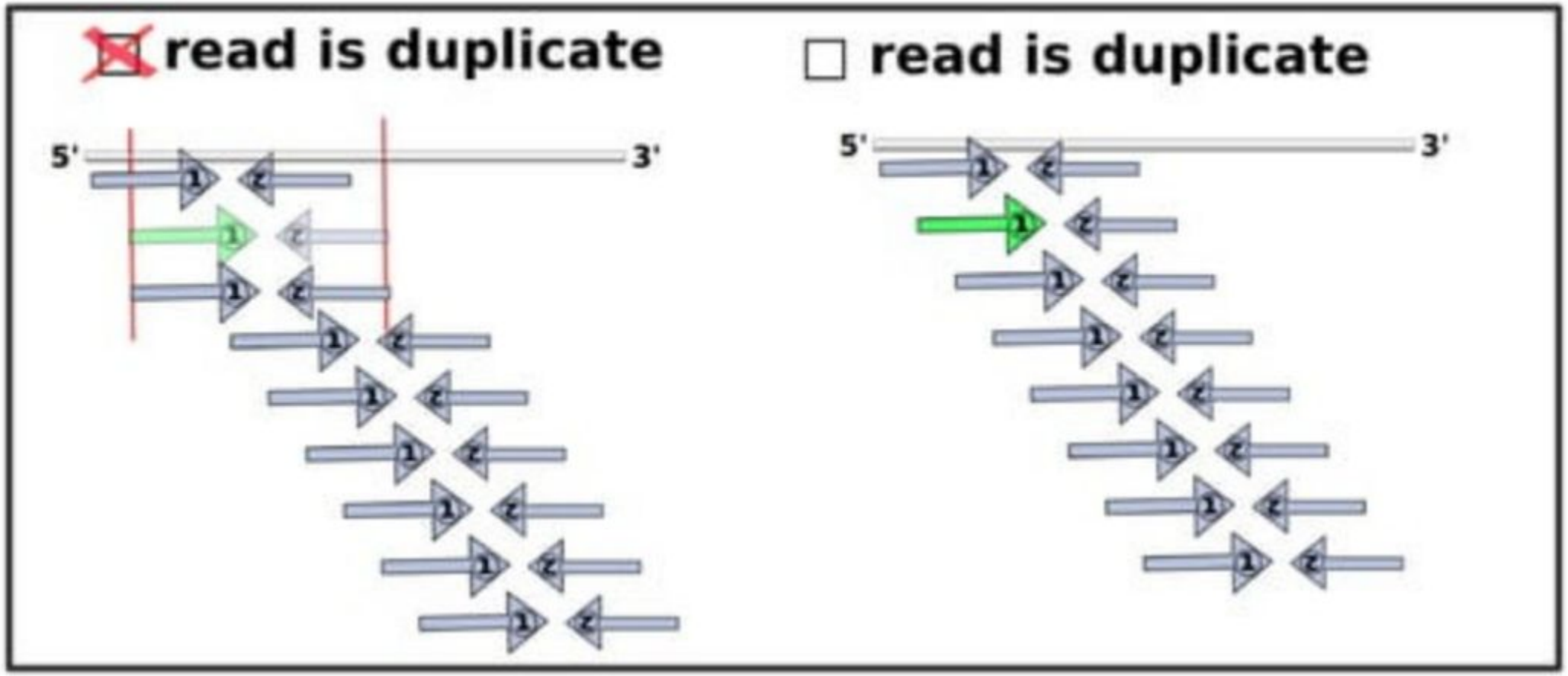
SAM Flag


not primary alignment
not primary alignment
5'
3'
5'
3'
5'
3'
5'
3'
5'
3'
5'
3'

SAM Flag


☒ read fails platform quality checks
☐ read fails platform quality checks

SAM Flag



ST-E00223:32:H5J57CCXX:4:1220:14651:8868 99 1 10086

base2	base10	base16	Meaning	Applies to:
00000000001	1	0x0001	The read originated from a paired sequencing molecule	Both
00000000010	2	0x0002	The read is mapped in a **proper** pair	Pairs only
00000000100	4	0x0004	The query sequence itself is unmapped	Both
00000001000	8	0x0008	The query's mate is unmapped	Pairs only
00000010000	16	0x0010	Strand of the query (0 for forward; 1 for reverse strand)	Both
00000100000	32	0x0020	Strand of the query's mate	Pairs only
00001000000	64	0x0040	The query is the first read in the pair	Pairs only
00010000000	128	0x0080	The read is the second read in the pair	Pairs only
00100000000	256	0x0100	The alignment is not primary	Both
01000000000	512	0x0200	The read fails platform/vendor quality checks	Both
10000000000	1024	0x0400	The read is either a PCR duplicate or an optical duplicate	Both

00001100011

$2^6+2^5+2^1+2^0 = 64+32+2+1 = 99$

Concise Idiosyncratic Gapped Alignment Report (CIGAR)

Encoding the details of the alignment

Operation	Meaning
M	Match*
D	Deletion w.r.t. reference
I	Insertion w.r.t. reference
N	Split or spliced alignment
S	Soft-clipping
H	Hard-clipping
P	Padding

Reference: ACCTGTC--TACCTTACG

Experimental: ACCT-TCCATACTTTATC

4M 1D 2M 2I 7M 2S

CIGAR string: 4M1D2M2I7M2S

LENGTH/OPERATION

CIGAR Extended

Operation	Meaning
=	Exact match
X	Mismatch
D	Deletion w.r.t. reference
I	Insertion w.r.t. reference
N	Split or spliced alignment
S	Soft-clipping
H	Hard-clipping
P	Padding

Reference: ACCTGTC--TACCTTACG

Experimental: ACCT-TCCATACTTTATC

4= 1D 2= 2I 3= 1X 3= 2S

CIGAR string: 4=1D2=2I3=1X3=2S

SAM to BAM

Do it once

Create BWT of reference genome.

```
$ bwa index grch38.fa
```

↓

Output is in SAM format

Align paired-end FASTQ to BWT index.

```
$ bwa mem -t 16 grch38.fa 1.fq 2.fq > sample.sam
```

↓

Output is in BAM format.
Unsorted!
random genomic order as reads are randomly placed in FASTQ by sequencer.

Convert SAM to BAM

```
$ samtools view -b sample.sam > sample.bam
```