Lesson 5

Sequence mapping part II

NGS reads

FASTA

Header
Sequence
Header
Sequence
Header
Sequence

```
>VIT_201s0011g03530.1
AATTAAGCATAAATACTCACTCTTACCCCCTTATTTTCTTATCTCTCATCACTTTTGGTGCGAAG
GACCATGAGAACAAGCTGCAATGGGTGTAGGGTTCTTCGCAAGGCATGCAGCCAAGACTGCATCA
>VIT_201s0011g03540.1
CAGGTAGCGTGAAGTTAAACCCTAGCGCTTTAGACAAACAGCTGTAGTCACCGCCCACAAACACC
AGCCTCTGAGACACCACCTCAAACCTTTCCACTTAAATACACATCCCTCACACCCTTTTCAATTC
>VIT_201s0011g03550.1
CATGCAAAGCTGAACGCGATGCTGTGATTGGTGGTAAGTGGTAGTTGAGTAAATTTGACAGTGAA
GCCGAAATGGTAAAAGACTAAGGCTAGAAGTAGAATACCACTGTTCTTCTCATCACGTGGGCCCA
```

FASTQ

Read header

Sequence

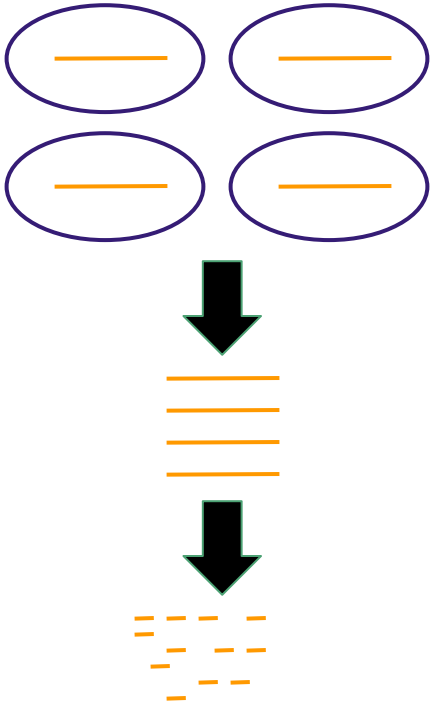
Q-score

Read 1

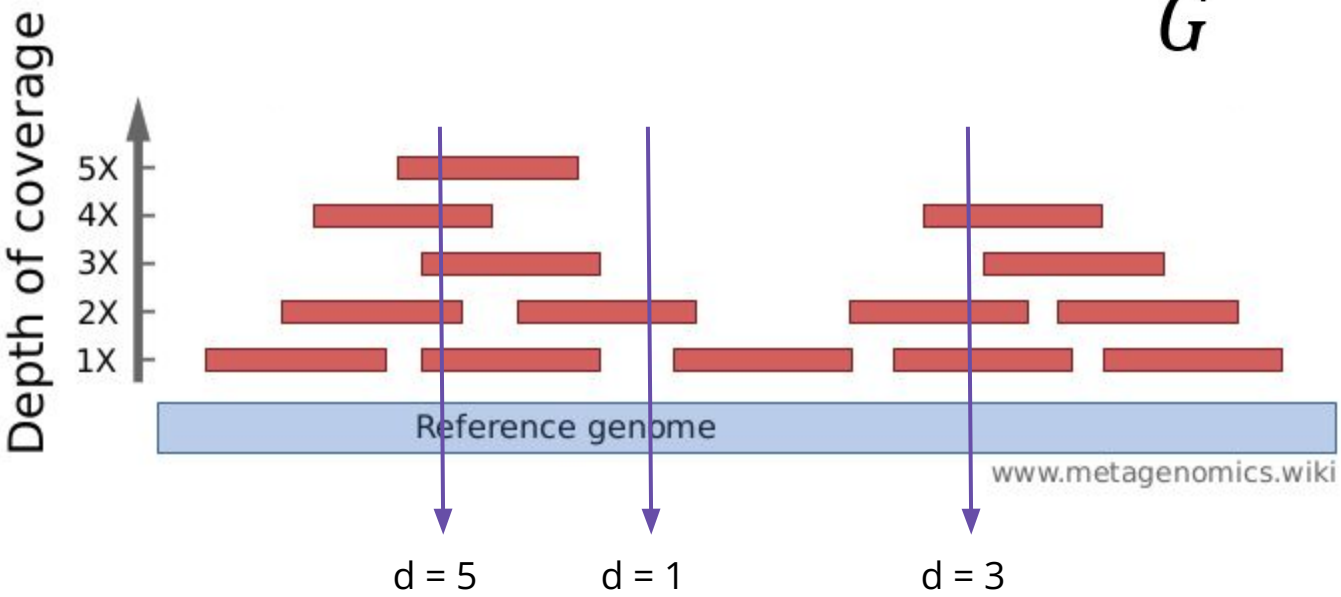
Read 2

```
@ERR2348785.250578 250578/1
CCCAAAAGGCGGACCATTGGGCTGGCAAGGTACCAAATCTAGTTCTGTCTCTTATACA
+
BBBFFFFFFFFFFIIFIFFFFFFI<FFIIFFFIIIFFFIFFIIFIB<FFFIFFFFFF
@ERR2348785.250579 250579/1
ATAACATACCATTTACATGTTGGAGGAAAAGATAAACGCTGTCTCTTATACACATCTC
+
BBBFFFFFFFFFIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIFIIIIIIIIIFIIIIII
```

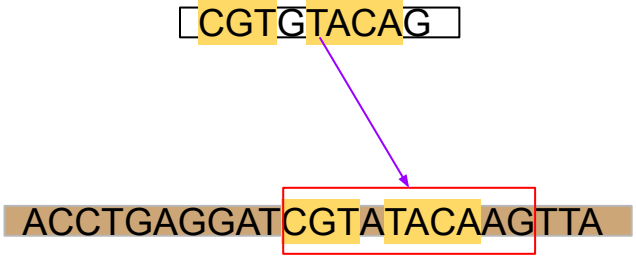
Sequencing depth

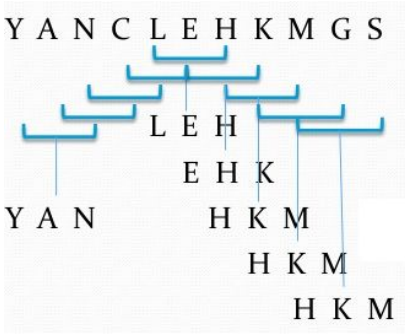


$$D = \frac{L * N}{G}$$



Sequence mapping - BLAST





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

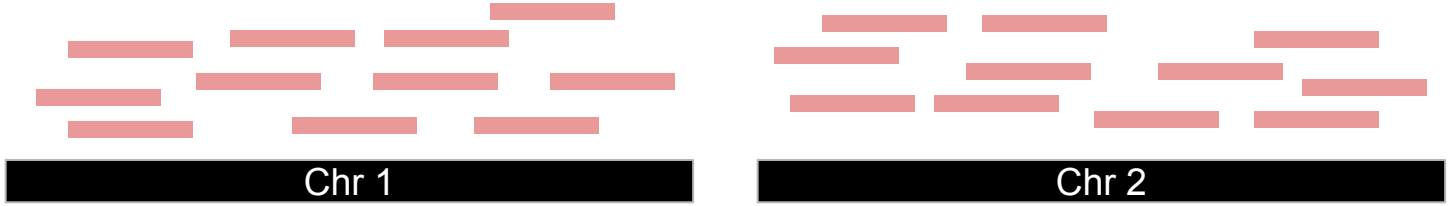
	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

By the end of this lesson you will...

- Know how to use BWA for short read mapping
- Understand the Sequence Alignment Map (SAM) and BAM file formats
- Be able to view and manipulate SAM/BAM files using samtools
- Be familiar with the IGV genome browser and how to use it for viewing BAM files

Short read mapping

- Map (search and align) Illumina reads to a reference genome
- Find the most likely position of a read in the genome
- Probably the most common task in genomics



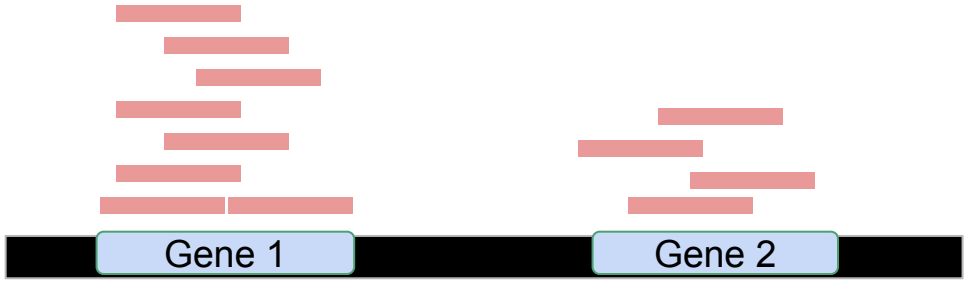
Short read mapping to reference genome - why?

- Variant calling / genotyping

  DNA

- Gene expression profiling

  RNA



The challenge - 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?

- Blastn - ~100 reads / sec
- Human genome - ~ 3Gb
- Assume 100bp reads

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**

How long to map x10 data to the human genome?

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)

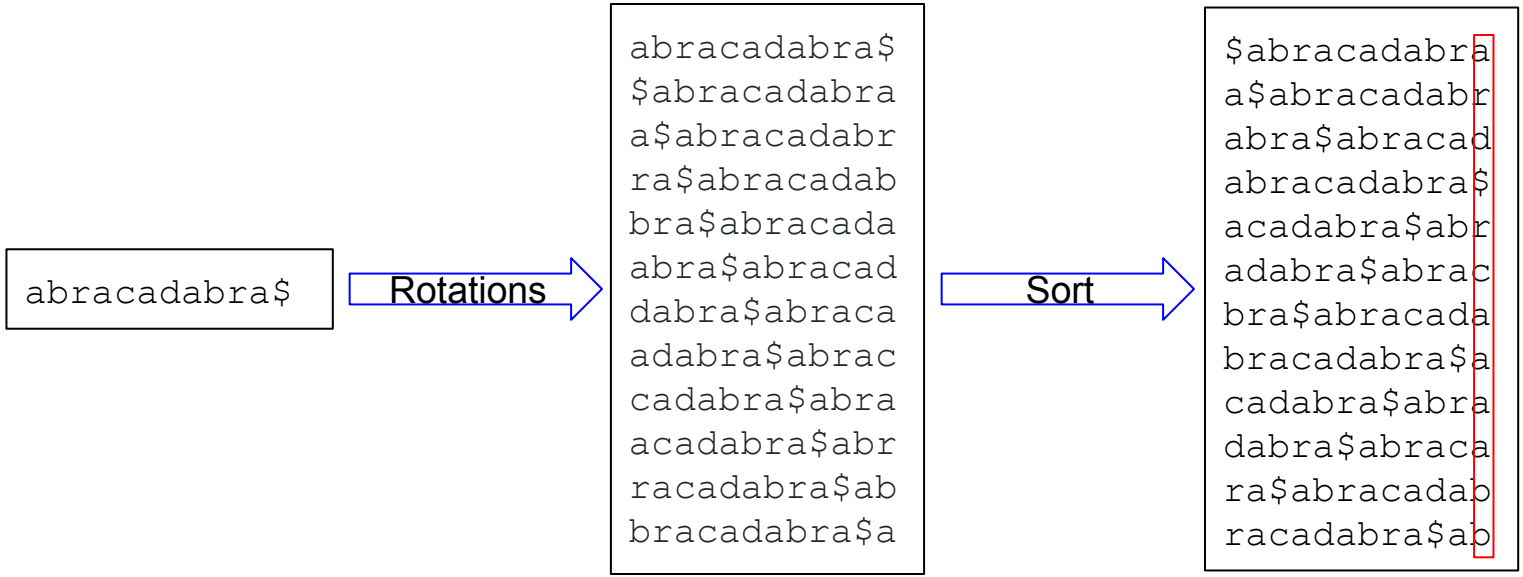
- Less effective when:
  - Queries are very long
  - Reads are highly diverged from the reference
  - Reads contain lots of sequencing errors

- 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

The Burrows-Wheeler's transformation



BWT(abracadabra$) = ard$rcaaaabb

The Burrows-Wheeler's transformation

- BWT is **reversible** - we can get back from BWT(G) to G

- BWT(G) tends to cluster the same characters together - easy to compress

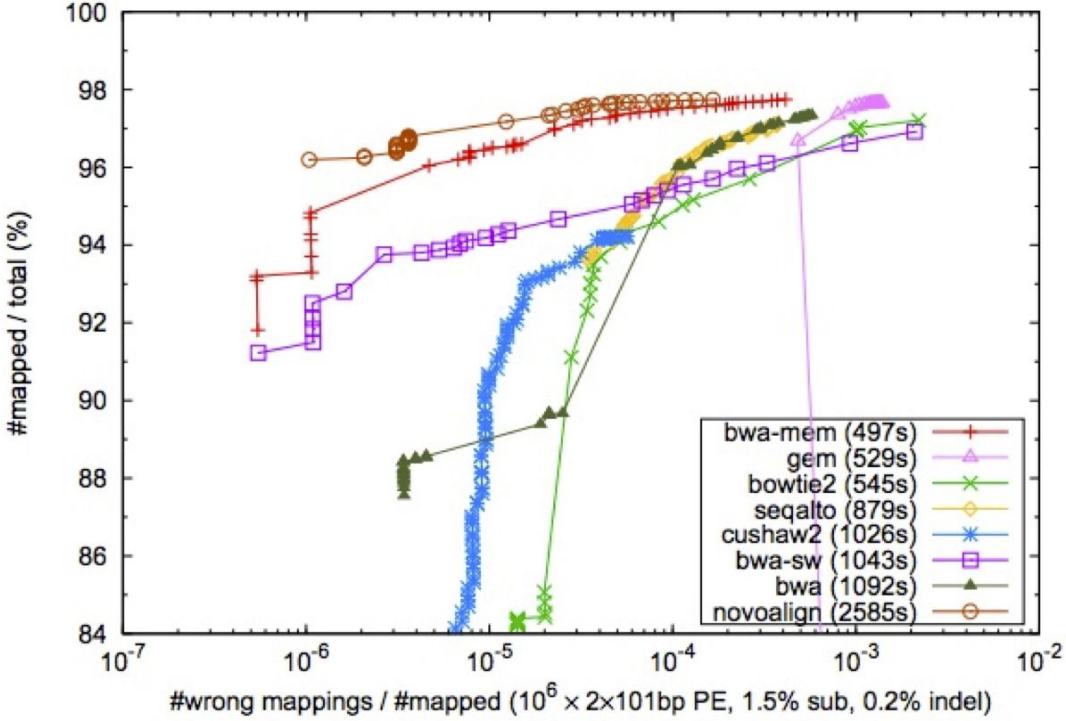
- Using some additional data structures, BWT(G) can be searched efficiently

BWT(abracadabra$) = ard$rcaaaabb

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.

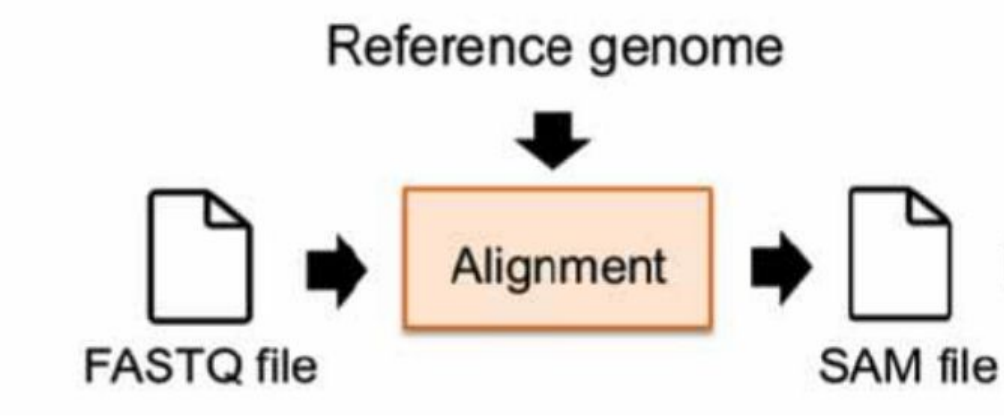
```
$ bwa index grch38.fa
```

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 BAM



Sequence Alignment and Mapping

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 28, 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)

The SAM format sections

- Header
  - Lines start with ‘@’
  - Meta-data - General information about the file
- Alignments
  - Contains the actual read mapping information
  - Each line has 11 mandatory fields (columns)
  - Additional fields may be included
  - Fields are separated by tabs

```
@SQ     SN:SL4.0ch00    LN:9643250
@SQ     SN:SL4.0ch01    LN:90863682
@SQ     SN:SL4.0ch02    LN:53473368
@SQ     SN:SL4.0ch03    LN:65298490
@SQ     SN:SL4.0ch04    LN:64459972
@SQ     SN:SL4.0ch05    LN:65269487
@SQ     SN:SL4.0ch06    LN:47258699
@SQ     SN:SL4.0ch07    LN:67883646
@SQ     SN:SL4.0ch08    LN:63995357
@SQ     SN:SL4.0ch09    LN:68513564
@SQ     SN:SL4.0ch10    LN:64792705
@SQ     SN:SL4.0ch11    LN:54379777
@SQ     SN:SL4.0ch12    LN:66688036
@PG     ID:bwa  PN:bwa  VN:0.7.12-r1039 CL:bwa mem -t 10 /groups/itay_mayrose/nosnap/liorglic/Projects/GPAD/data/S_lycopersi
cum_chromosomes.4.00.fa /groups/itay_mayrose/nosnap/liorglic/Projects/GPAD/data/SRR1572628_1.fastq /groups/itay_mayrose/nosn
ap/liorglic/Projects/GPAD/data/SRR1572628_2.fastq
SRR1572628.1    83      SL4.0ch05       25938772        60      100M    =       25938399        -473    GTGAAGGCGTTTGACTGG
GCATGGAGTTTAAGAGTGAAACAAACACTTTTTGAAACTTAAGGTCTAAAATAAGTCAGAAATTGTGTGGCTGTAAATCA    CCCCDBDDDDEEFFFFFHG[illegible]FHJIGIIEIJIJJJJJJI
HIHFBDBJJJJIIIJIGIJJJIJJJJJJJIJJJJJJJJJJJJJIJJJIHHHHHFFFFCCC    NM:i:0  MD:Z:100        AS:i:100        XS:i:21
SRR1572628.1    163     SL4.0ch05       25938399        60      100M    =       25938772        473     CACAGATGCTGGGAGATCTA
TTTCATATCCATGCAGTAAAATCAAAACATGGGCAGAAACCAAATCAATCTTGCCTCGACCCAAAGTTCTCATCAGTATC    CCCFFFFFHHHHHJJIJJJJJJJJI[illegible]JJJIJJIJJGIIJ
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    CDEDDDDDDDDDDDDDDCCDDEDDCBDEDDEEDEFC@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
```

What command will fetch only header lines?

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
HIHFBDBJJJJIIIJIGIJJJIJJJJJJJIJJJJJJJJJJJJJIJJJIHHHHHFFFFFCCC    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    CDEDDDDDDDDDDDDDDCCDDEDDCBDEDDEEDEFC@A@F
CAHE?HEEIJIHJJIJJJJJJIGFIIJIIIHEJJJJIJJJJIHF@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

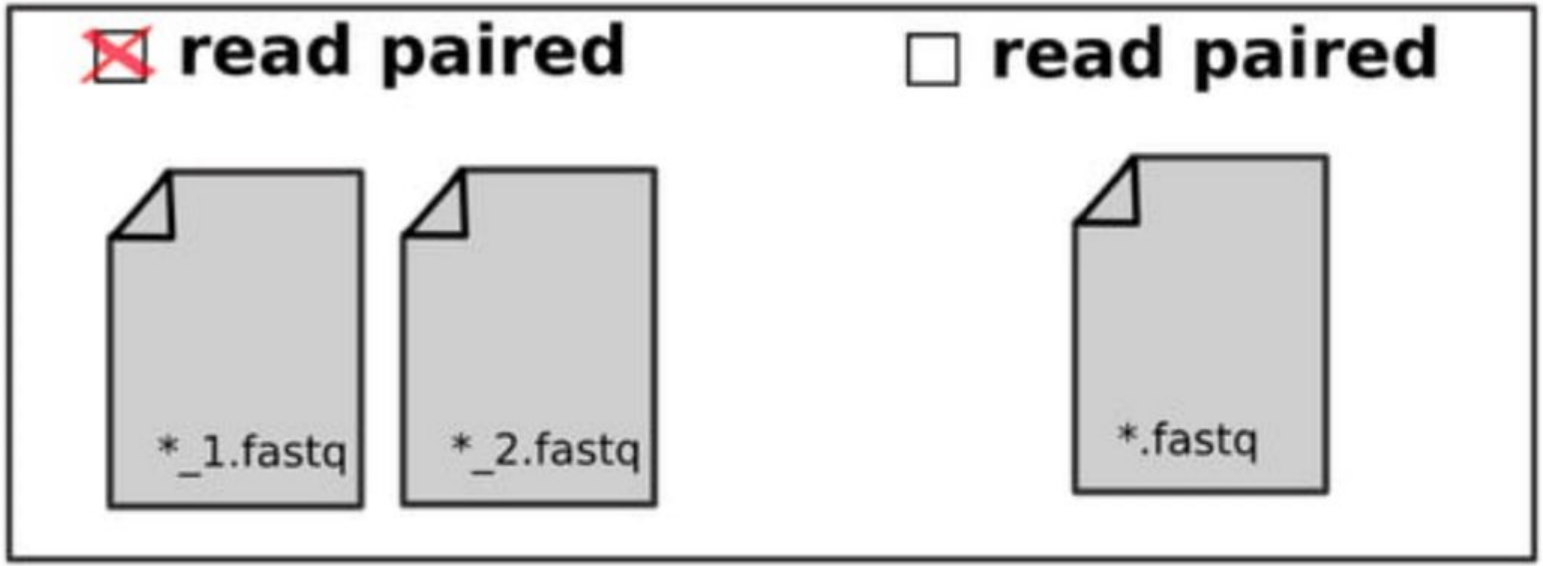
As a rule of thumb:

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

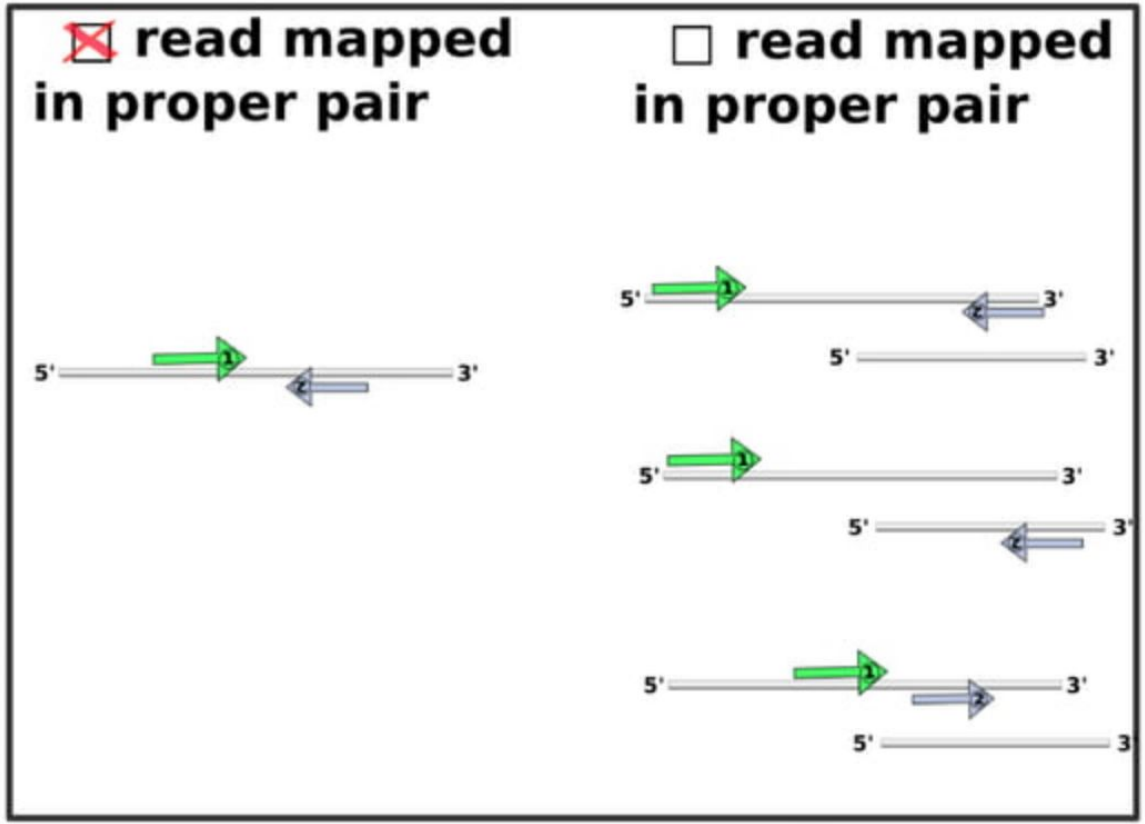
SAM Flag

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

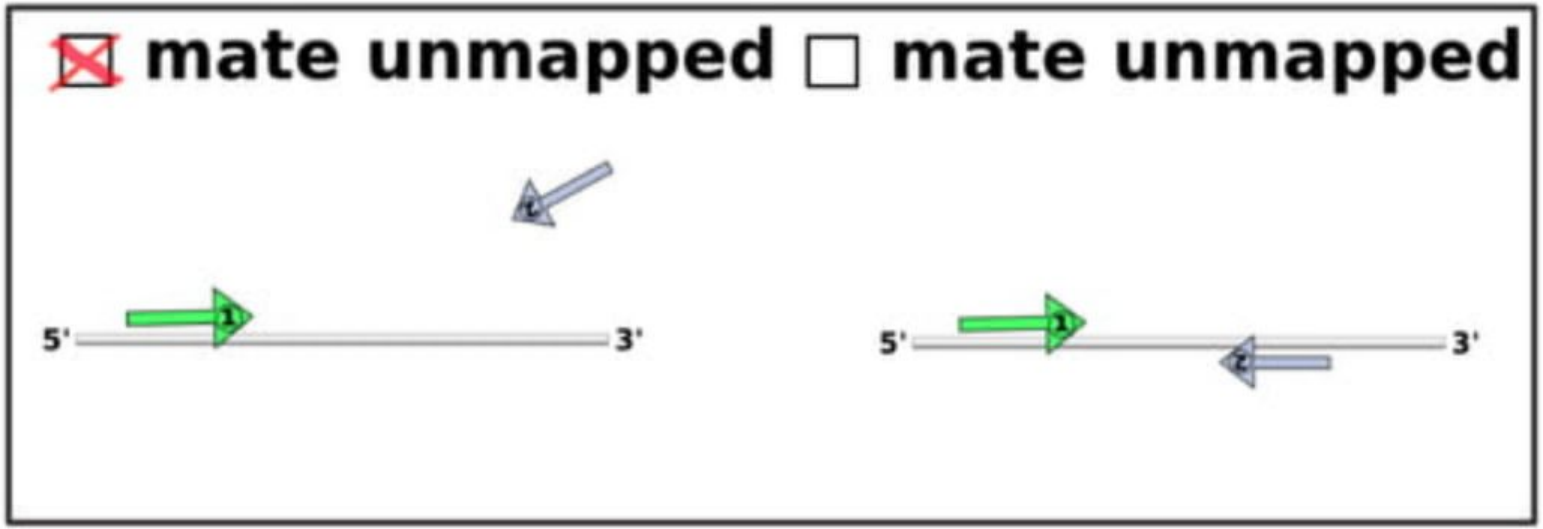


SAM Flag

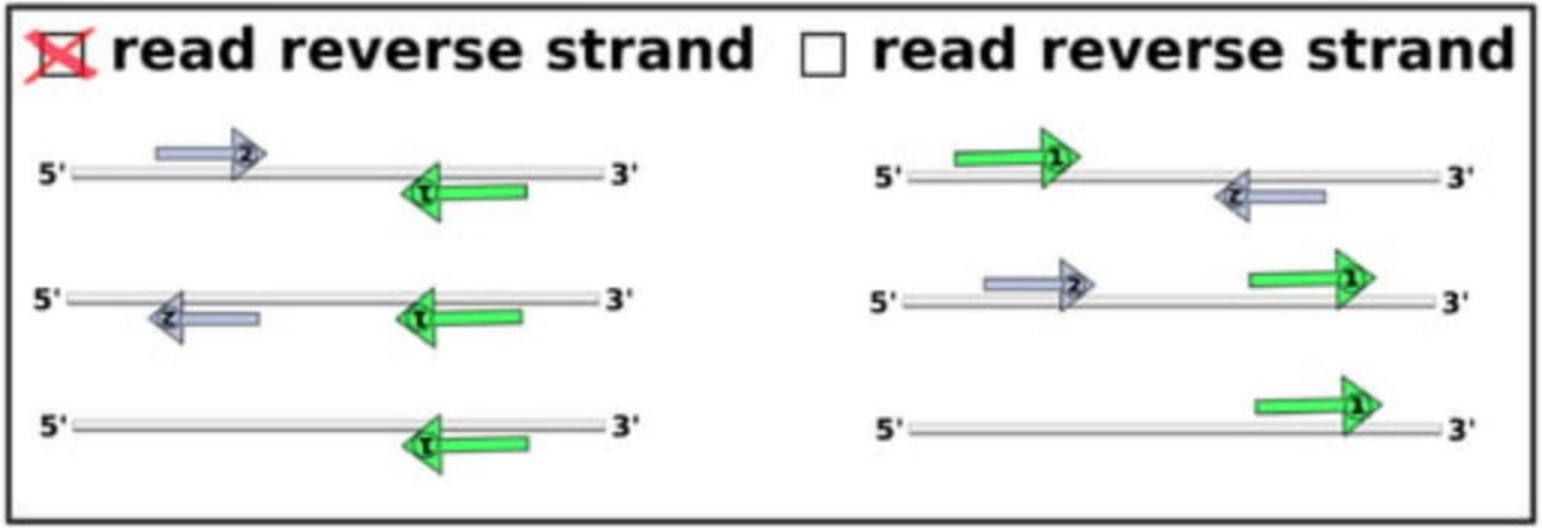


SAM Flag

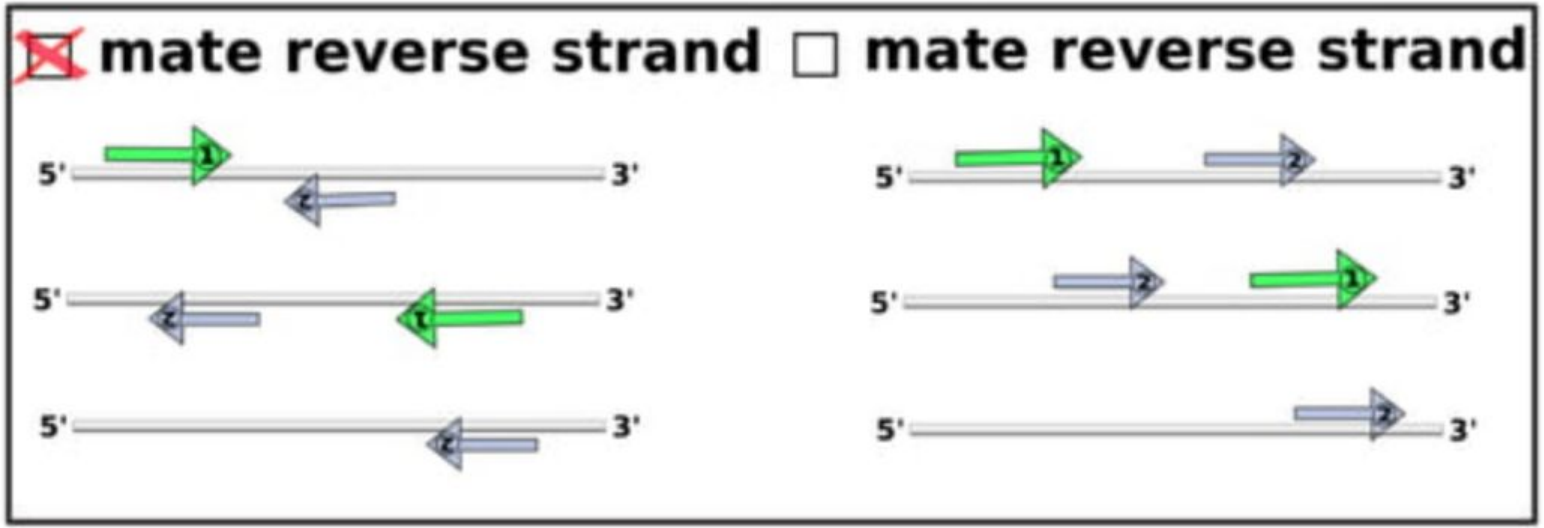
SAM Flag



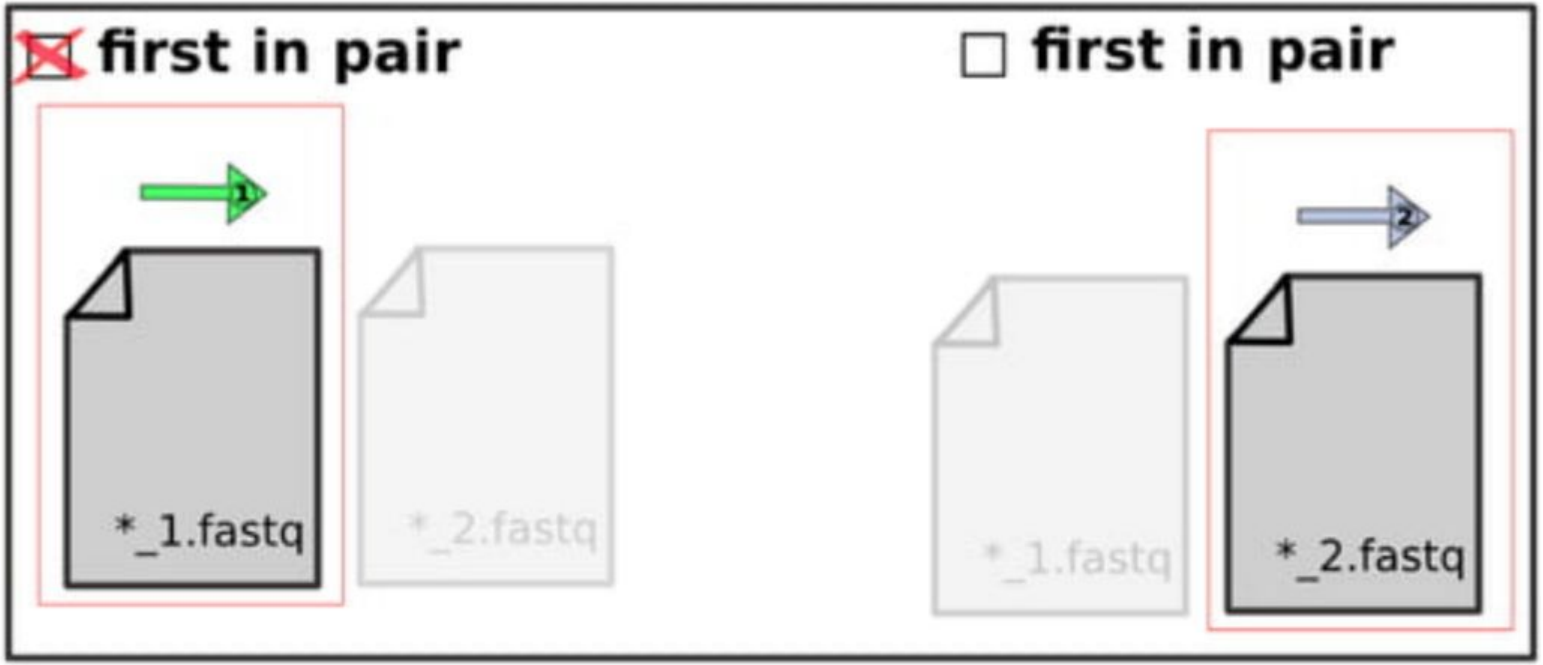
SAM Flag



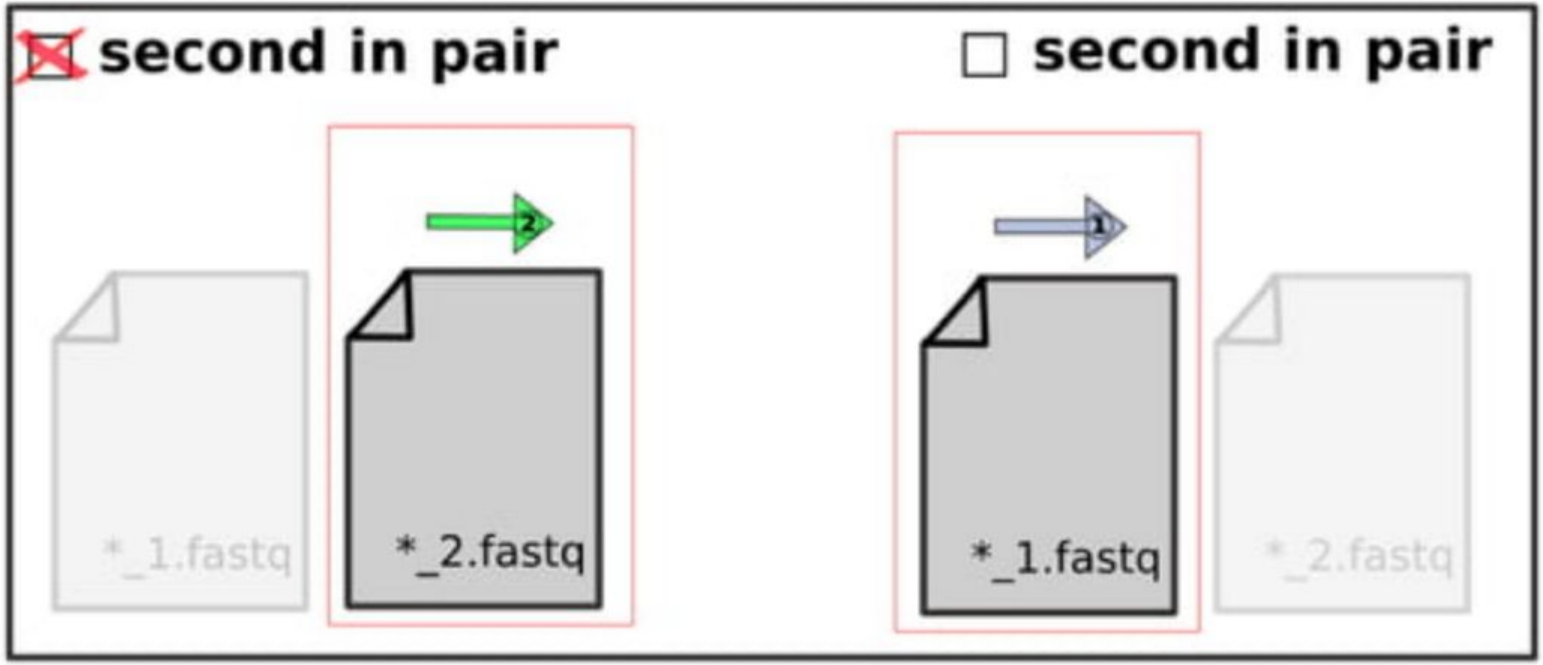
SAM Flag



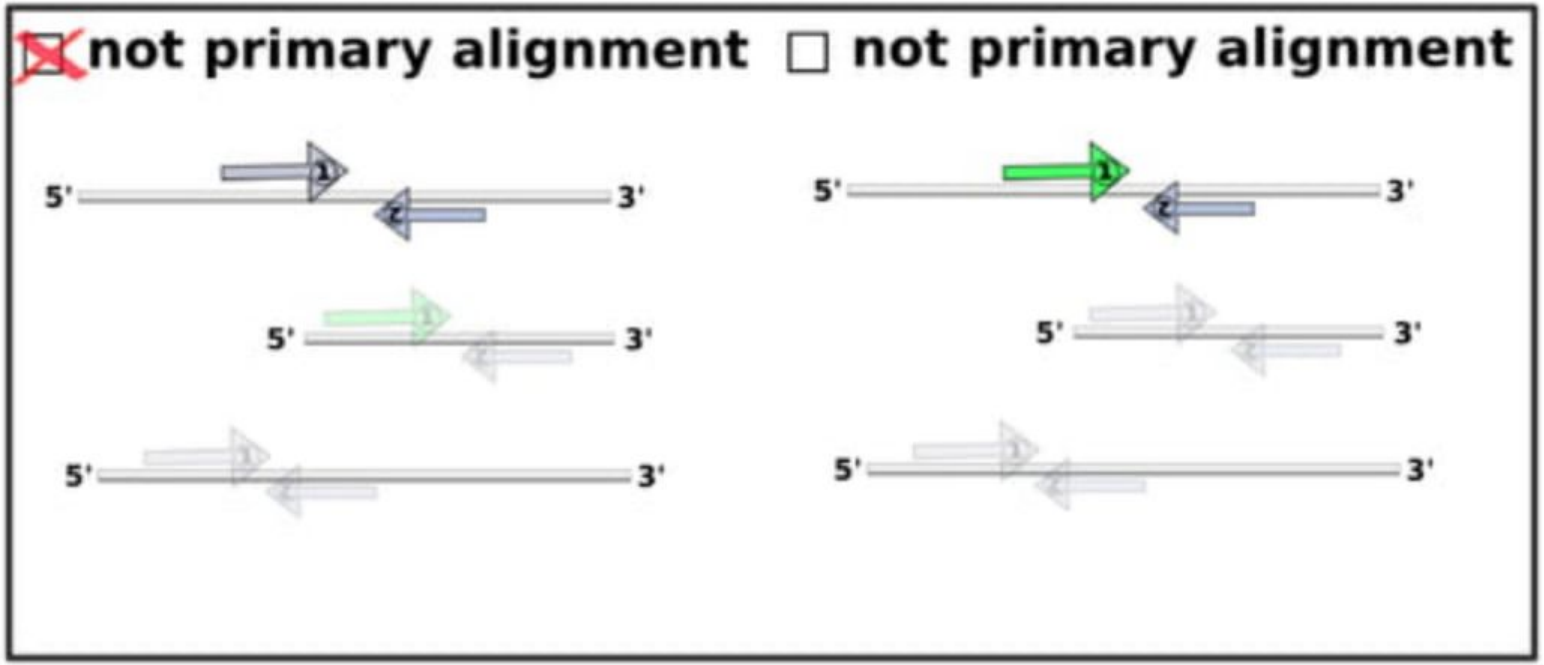
SAM Flag



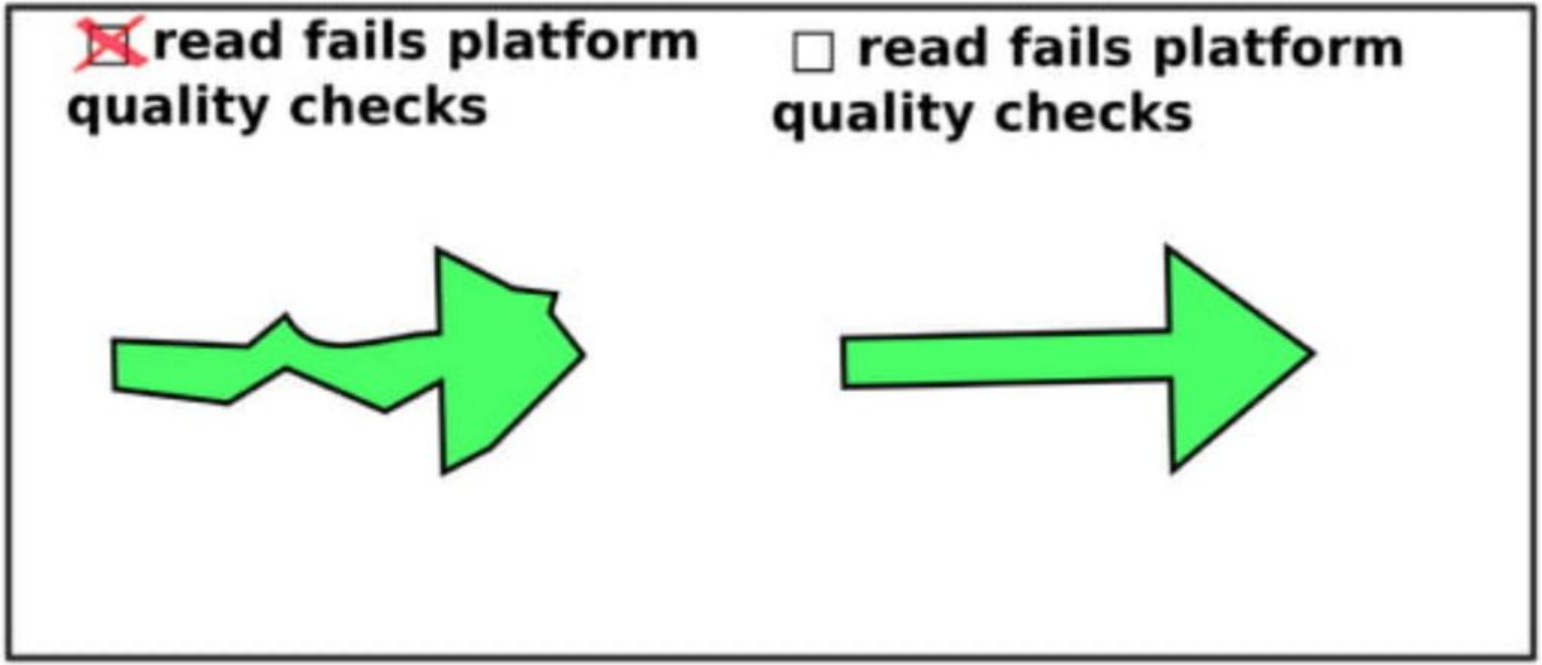
SAM Flag


second in pair
second in pair
2
1
*_1.fastq
*_2.fastq
*_1.fastq
*_2.fastq

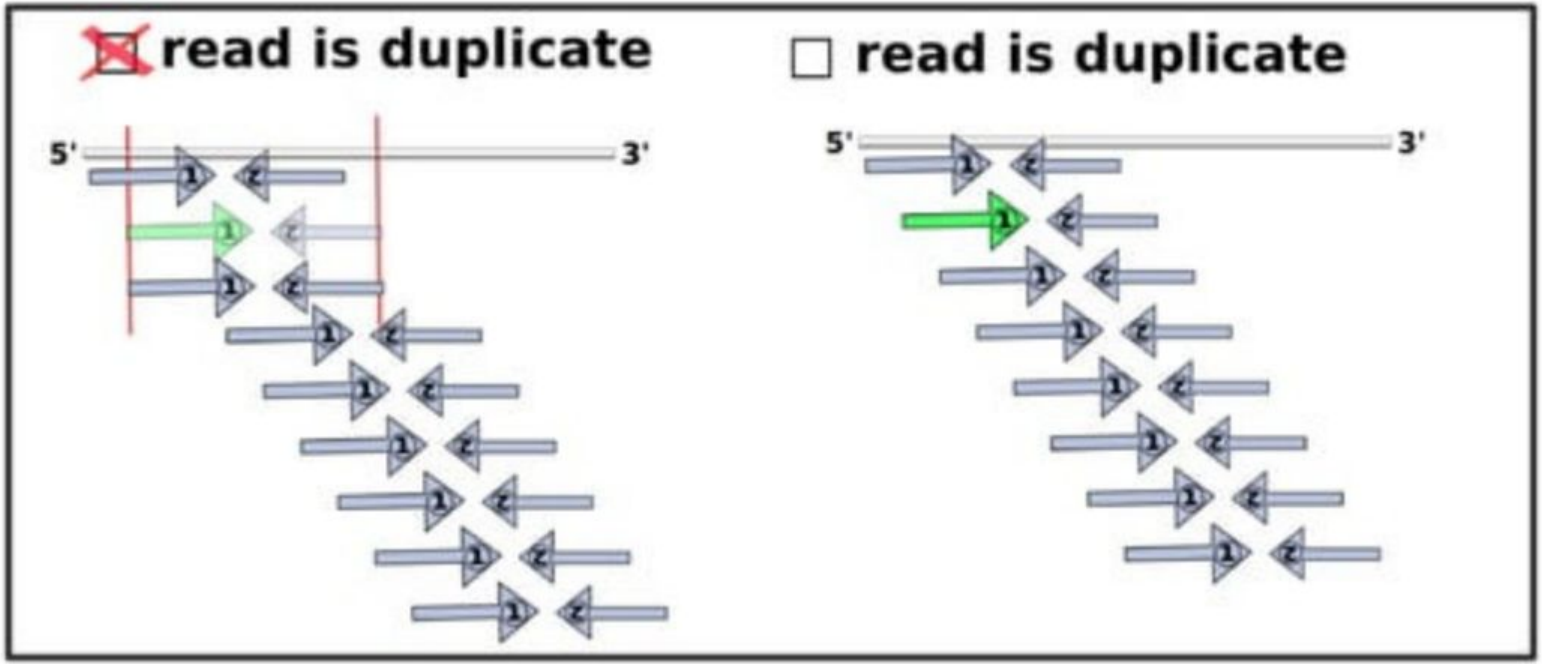
SAM Flag


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

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

Decoding SAM flags

https://broadinstitute.github.io/picard/explain-flags.html

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 Additional fields

```
SRR1572628.1    83      SL4.0ch05       25938772        60      100M    =       25938399        -473    ATGTGAAGGCGTTTGACTGG
GCATGGAGTTTAAGAGTGAAACAAACACTTTTTGAAACTTAAGGTCTAAAATAAGTCAGAAATTGTGTGGCTGTAAATCA    CCCCDBDDDDEEFFFFFHGFGFFHJIGIIEIJIJJJJJJJI
HIHFBDBJJJJIIIJIGIJJJIJJJJJJJIJJJJJJJJJJJJJIJJJIHHHHHFFFFFCCC    NM:i:0  MD:Z:100        AS:i:100        XS:i:21
```

- Alignment software may output additional fields containing more information
- Additional fields will always look like:

  <Tag>:<type>:<value>

- Should be specified in software documentation
- Some examples:
  - NM - number of mismatches
  - AS - raw alignment score

SAM to BAM

Do it <u>once</u>

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
```

SAM - unmapped reads

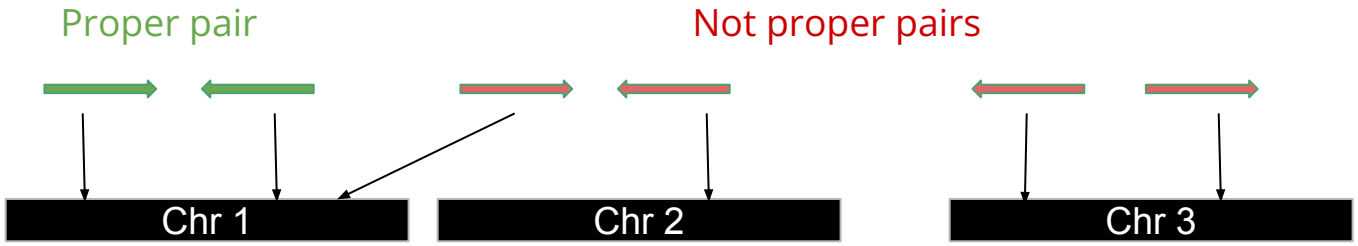
What could cause a read to be unmapped?

- A read appears even if it is unmapped!
- Unmapped reads have:
  - flag 4
  - MAPQ 0
  - Missing info for other fields (* or 0)

```
SRR1569760.7733303      77      *       0       0       *       *       0       0       AAAAAAAGAGTTTCATTACGGGGAAATAAAGTTTTCCCCTTTTTC
CGATATTATAGCGAAAAGGTTATCTCTCTTATTGTGCATTGTTGTGCCACTACCATA   @@@DDDD:DDHHBGHGIBDFGIIHBGEHIGBDFFEGGCHGHGGEEH>FGIIIIIG@EEE>CCC>>ACCCCDC
CCCCC@D3@>@A@4>8<?@ACCCCCCCCC   AS:i:0  XS:i:0
```

SAM - paired-end data

- Both reads of a pair appear (as separate records)
- Records contain information about the paired read
- Several flags relate to paired information
- Especially flag 2 - “Read mapped in proper pair”



BAM & CRAM - binary SAM

- Compressed - smaller size
- Faster to read by a computer
- Impossible to read by humans - must use some conversion tool
- Required by some bioinformatic tools
- CRAM might contain only varitional changes from reference

Working with SAM files

- SAM files are text files
- You can view them - use `less`
- You can use Linux commands to manipulate the file:
  E.g.: get first 5 alignments:
  ```
  grep -v '^@' aln.sam | head -5
  ```

- Or you can use a dedicated command line tool - **samtools**

Samtools allows you to...

- View SAM and BAM files
- Select records that satisfy some criteria
- Convert between formats
- Manipulate SAM/BAM files
  - Sorting
  - Indexing
- Extract statistics

Running samtools

```
$ samtools
Program: samtools (Tools for alignments in the SAM format)
Version: 1.9 (using htslib 1.9)
Usage:   samtools <command> [options]
```

- samtools features many commands
- Each command has its own function and options
- Today we'll look at three commands:
  - samtools view
  - samtools sort
  - samtools index

samtools view - viewing files

View sam/bam

```
$ samtools view aln.bam | less
```

View sam/bam including header

```
$ samtools view -h aln.bam | less
```

View just header of sam/bam

```
$ samtools view -H aln.bam | less
```

Samtools sort

Sort a bam file by location

```
$ samtools sort aln.bam > aln.sort.bam
```

Sort a bam file by read name

```
$ samtools sort -n aln.bam > aln.sort_name.bam
```

Samtools index

- Useful for quick handling of a bam file
- Takes a while for large bam files
- Required by some software tools
- Only works on **sorted BAM** files
- Creates a new .bai file

```
$ samtools index aln.sort.bam
```

samtools view - filter records

Only print records with MAPQ >= 20

```
$ samtools view -q 20 aln.bam | less
```

Only print records with third bit enabled (unmapped)

```
$ samtools view -f 4 aln.bam | less
```

Only print records with third bit disabled (mapped)

```
$ samtools view -F 4 aln.bam | less
```

samtools view - filter records

Only print records mapped to chromosome 3 (indexed files only)

```
$ samtools view aln.bam chr03 | less
```

Only print records mapped to chromosome 3 positions 1000-2000 (indexed files only)

```
$ samtools view aln.bam chr03:1000-2000 | less
```

Random access requires sorted and indexed bam

Samtools view - converting formats

Convert sam to bam

```
$ samtools view -bh aln.sam > aln.bam
```

Convert bam to sam

```
$ samtools view -h aln.bam > aln.sam
```

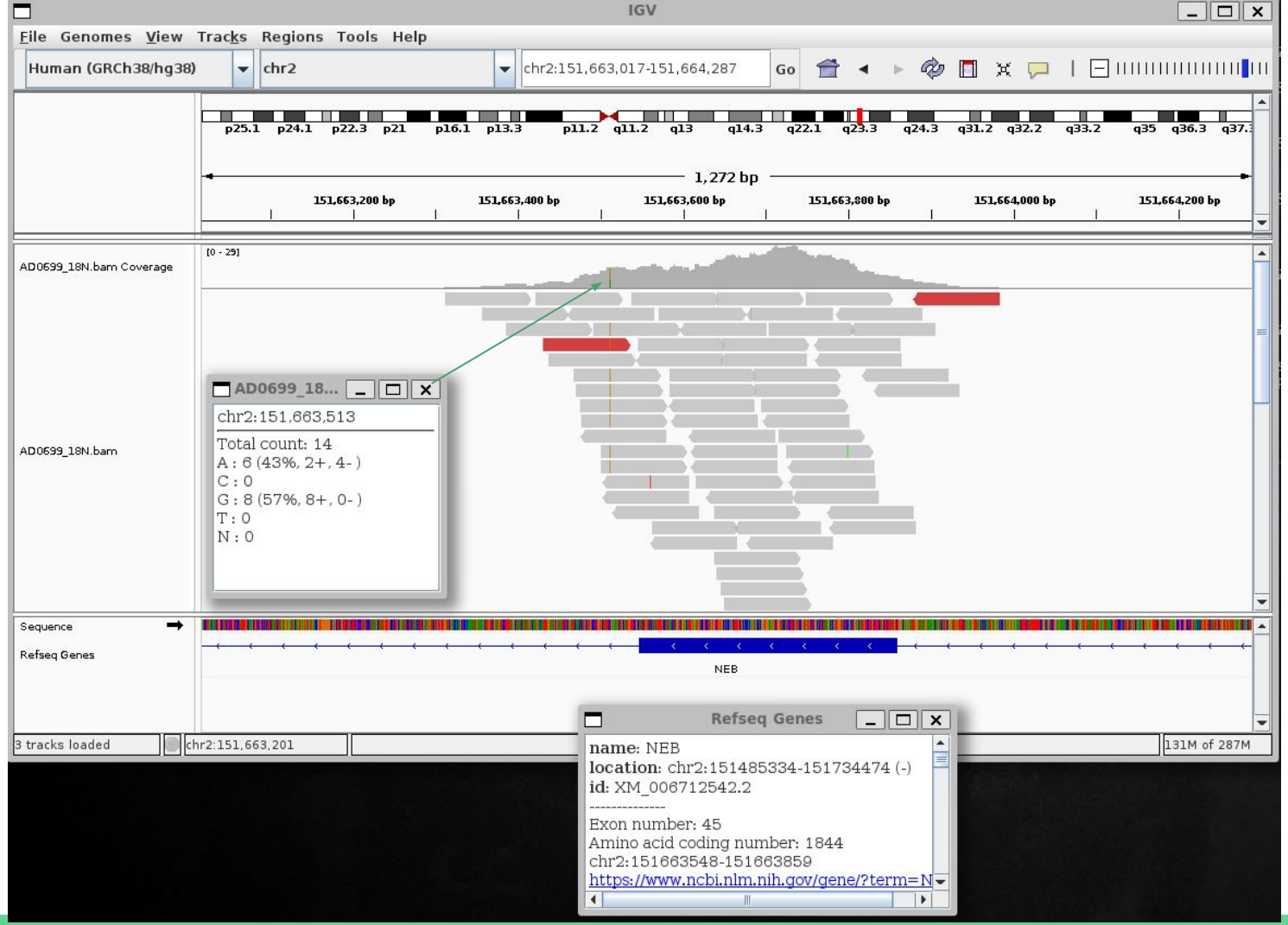
Combine with filtration

```
$ samtools view -bh -q 30 -f 2 aln.sam > aln.HQ.mapped.bam
```

Integrative Genomics Viewer (IGV)

Visualization tool for exploring and analyzing genomic data

https://software.broadinstitute.org/software/igv/


IGV
File Genomes View Tracks Regions Tools Help
Human (GRCh38/hg38)
chr2
chr2:151,663,017-151,664,287
Go
p25.1 p24.1 p22.3 p21 p16.1 p13.3 p11.2 q11.2 q13 q14.3 q22.1 q23.3 q24.3 q31.2 q32.2 q33.2 q35 q36.3
1,272 bp
151,663,200 bp
151,663,400 bp
151,663,600 bp
151,663,800 bp
151,664,000 bp
151,664,200 bp
AD0699_18N.bam Coverage
[0 - 29]
AD0699_18N.bam
AD0699_18...
chr2:151,663,513
Total count: 14
A : 6 (43%, 2+, 4-)
C : 0
G : 8 (57%, 8+, 0-)
T : 0
N : 0
Sequence
Refseq Genes
NEB
Refseq Genes
name: NEB
location: chr2:151485334-151734474 (-)
id: XM_006712542.2
Exon number: 45
Amino acid coding number: 1844
chr2:151663548-151663859
https://www.ncbi.nlm.nih.gov/gene/?term=N
3 tracks loaded
chr2:151,663,201
131M of 287M