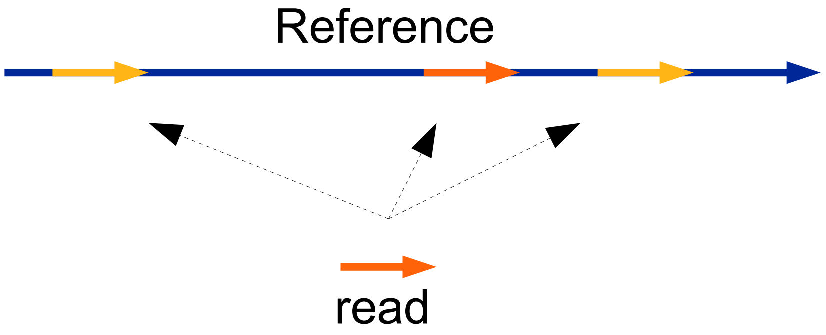
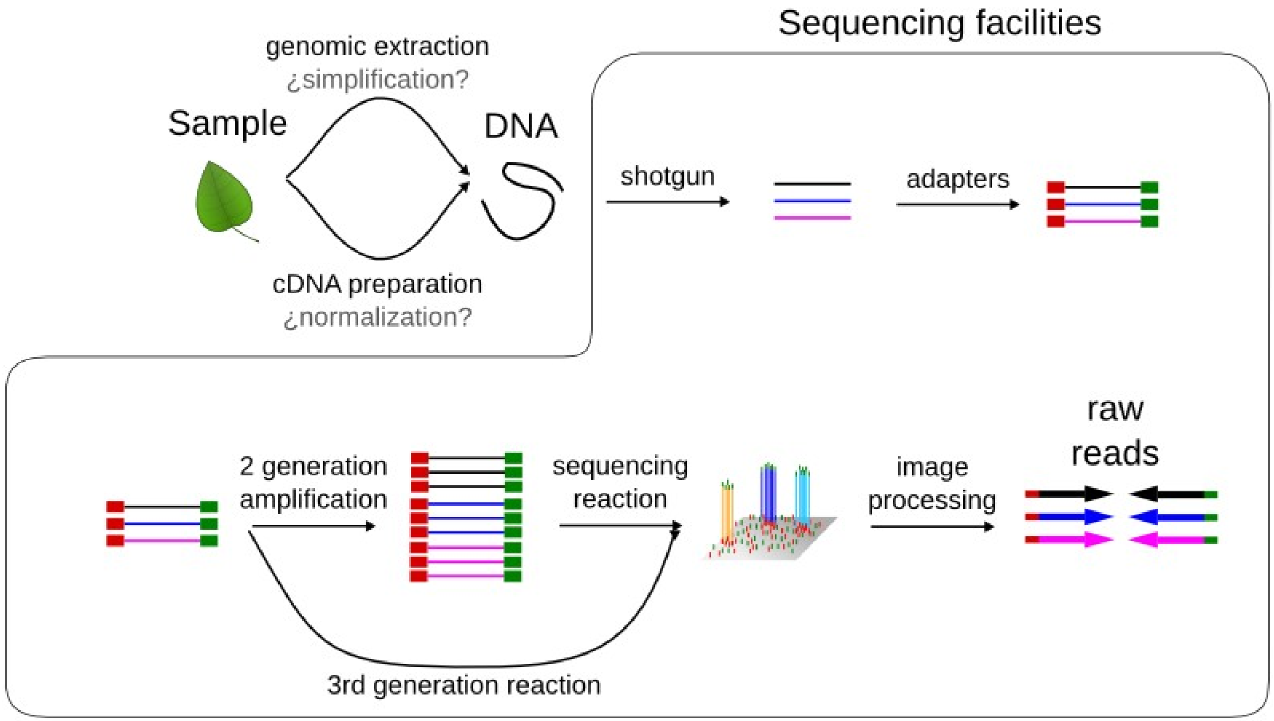
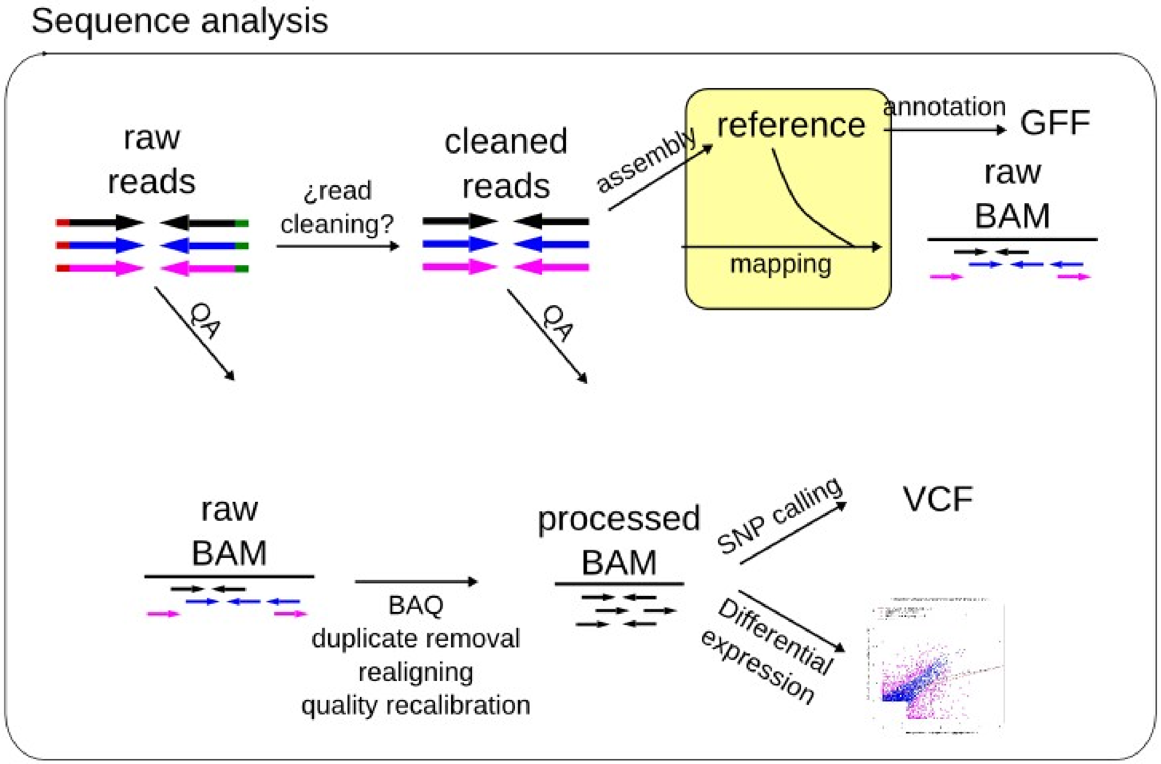
Mapping

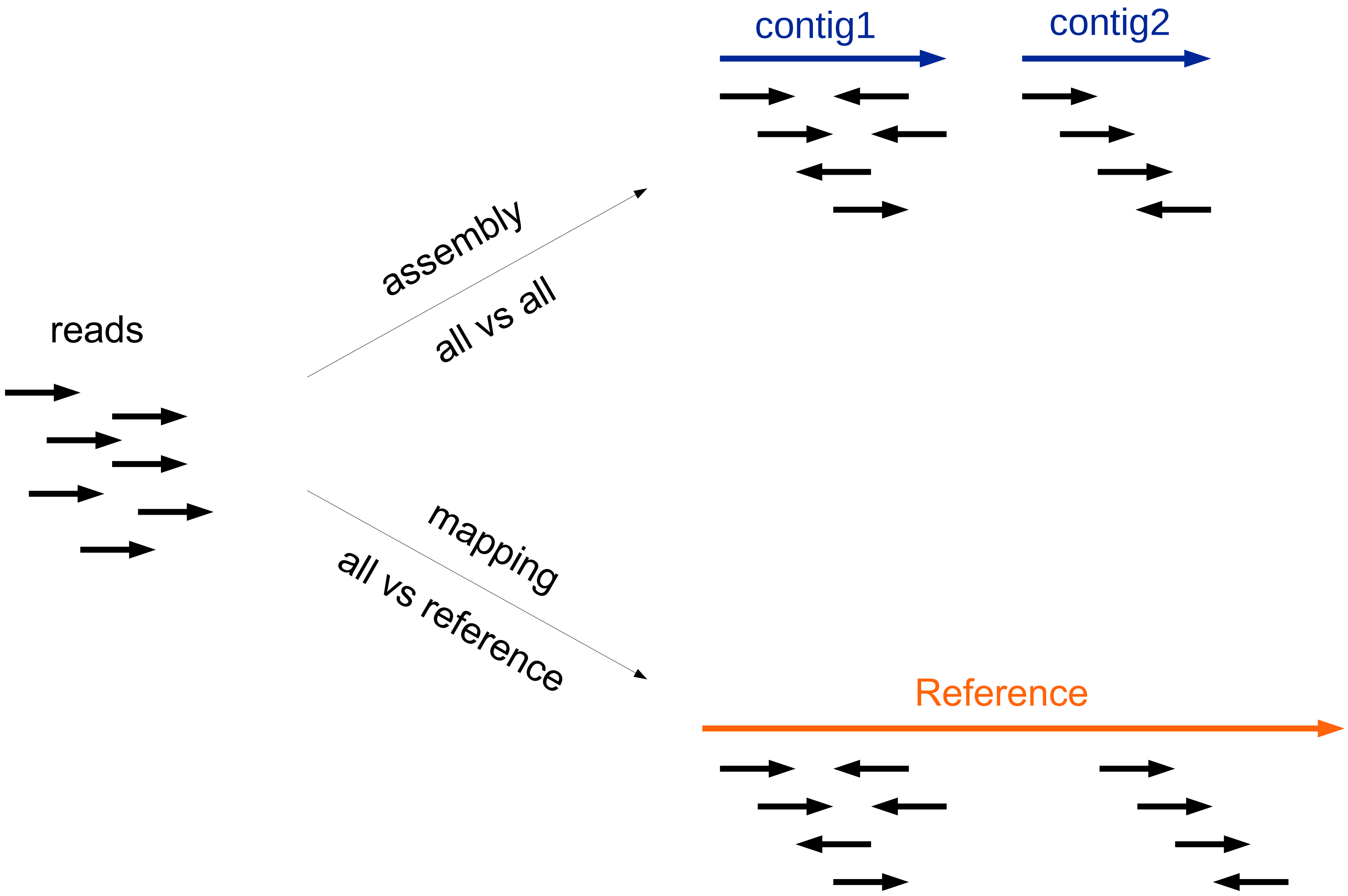


Jose Blanca
COMAV institute
bioinf.comav.upv.es


Sequencing facilities
Sample
genomic extraction
¿simplification?
cDNA preparation
¿normalization?
DNA
shotgun
adapters
2 generation
amplification
sequencing
reaction
image
processing
raw
reads
3rd generation reaction


Sequence analysis
raw
reads
¿read
cleaning?
cleaned
reads
QA
QA
assembly
reference
mapping
annotation
GFF
raw
BAM
raw
BAM
BAQ
duplicate removal
realigning
quality recalibration
processed
BAM
SNP calling
VCF
Differential
expression

Assembly vs mapping



Mapping

Reference genome / transcriptome

...GTGGGCCGGCAATTCGATATCGCGCATATATTTCGGCGCATGCTTAGC...

Reads
(unmapped)

1	GCATATATTT
2	GCATATATTT
3	TGGGCCGGCA
4	ATTCGATATC
5	ATATTTCGGC
6	CCGGCAATTC
7	TCGCGCATAT
8	CATGCTTAGC
9	GATATCGCGC

Mapping

Reference genome / transcriptome

...GTGGGCCGGCAATTCGATATCGCGCATATATTTCGGCGCATGCTTAGC...

TGGGCCGGCA

GCATATATTT

CATGCTTAGC

CCGGCAATTC

ATATTTCGGC

ATTCGATATC

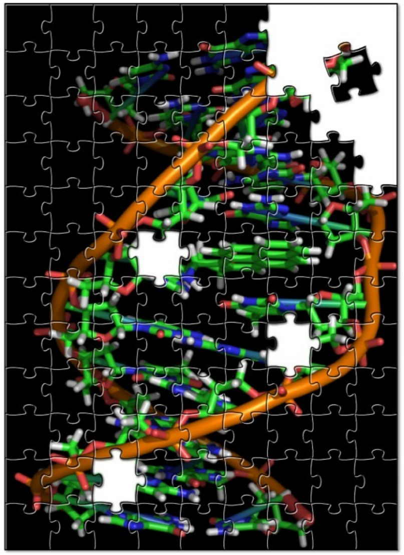
GCATATATTT

Reads
(mapped)

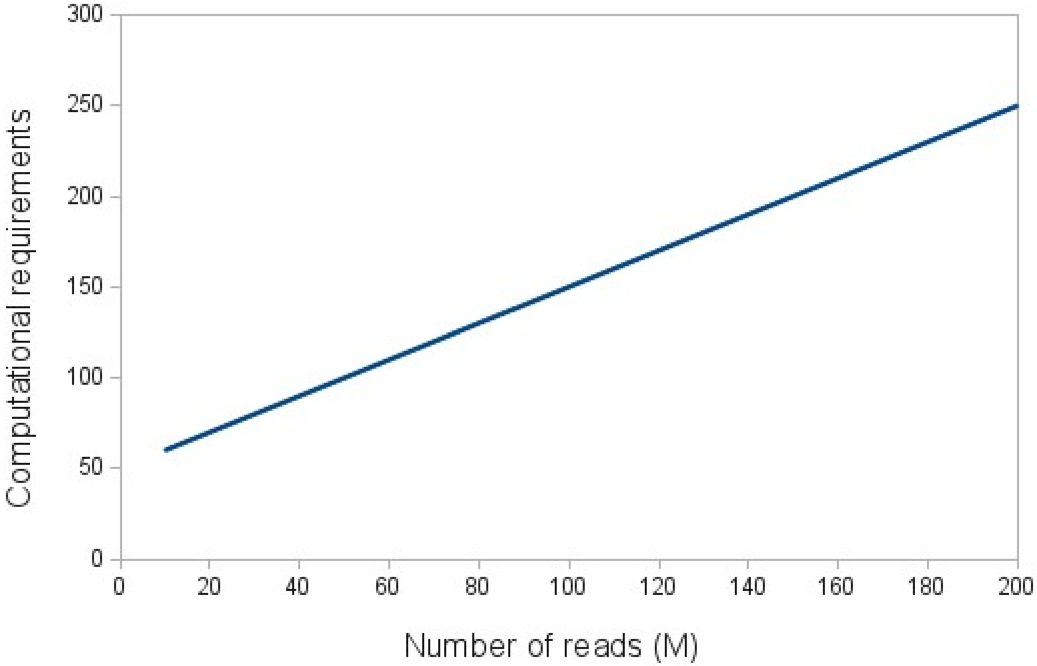
TCGCGCATAT

GATATCGCGC

Computational requirements



wikipedia

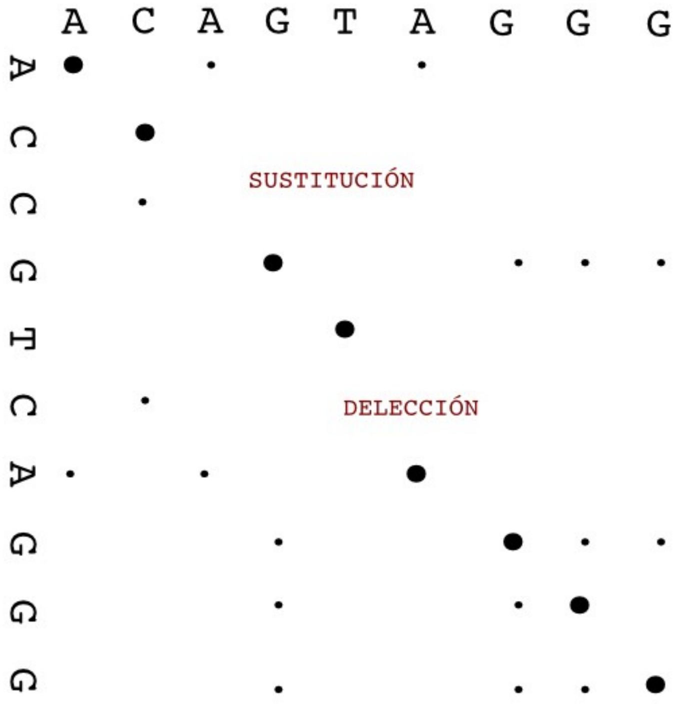


Main requirements:

- Hard drive.
- Time.

jmerelo@flickr

Alignment algorithms

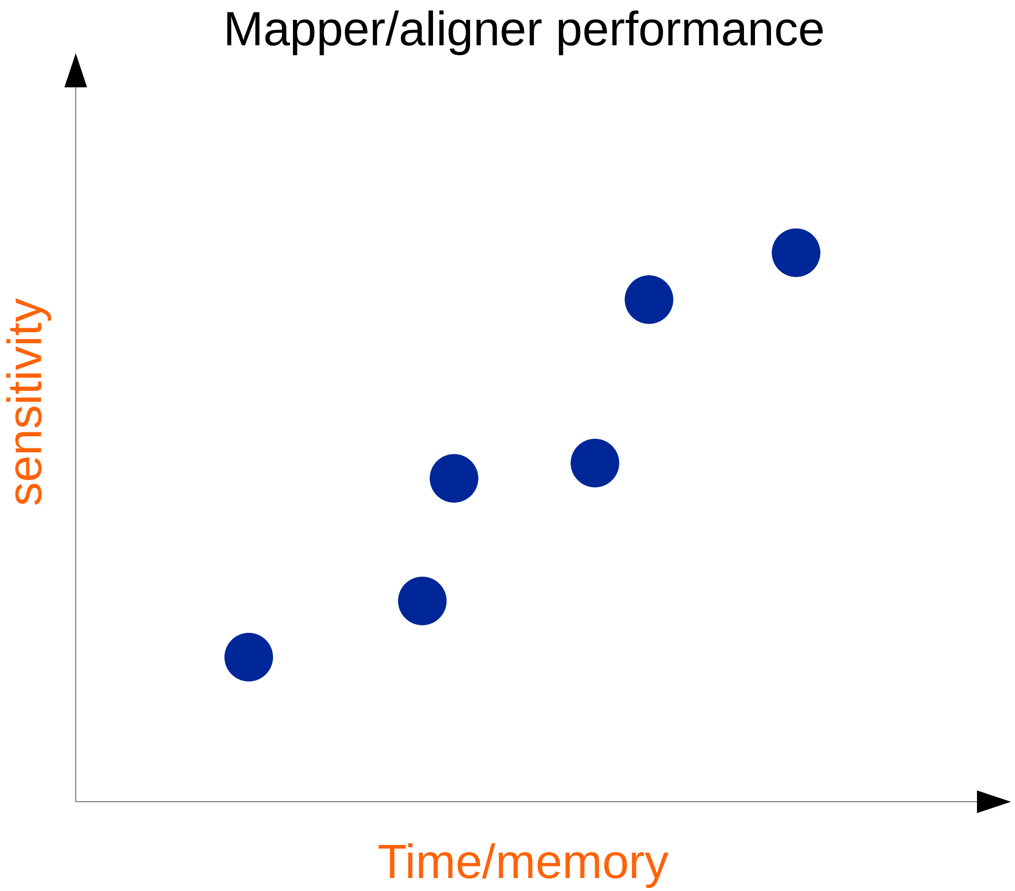


dotplot

Smith and Waterman:

- guaranteed to find the optimal local alignment (sensitive)
- Requires O(nm) time (too slow)
- Usually used to refine alignments

Mapping sensitivity



Not all reads that should be mapped (aligned) will be mapped.

Highly polymorphic regions or large insertions or deletions are difficult to detect.

NGS alignment algorithms

Seed/hash methods:

- Used by BWA mem

- Methodology:
  - find matches for short subsequences assuming that at least one seed in a read will perfectly match
  - Align with a sensitive method like SW

- Tend to be more sensitive than BWT

Burrows Wheeler Transform:

- Used by BWA and Bowtie

- Faster than hash methods at the same sensitivity level

- compact the genome into a data structure that is very efficient when searching for perfect matches

- performance decreases exponentially with number of mismatches

BWA vs Bowtie2 vs minimap2

BWA mem

- Reads from 70 bp up to few megabases
- Seeded algorithm plus Smith and Waterman
- Local alignment
- Allows gaps up to tens of bp in 100 bp reads and split alignment
- Reports chimeric alignments
- Fast, even for long reads
- Paired-end

Bowtie2

- One of the fastest alignment software for short reads (< 500pb)
- Gapped alignment, but not long gaps
- Global or local
- Paired-end

minimap2

bwa-mem replacement (same author)

- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34:3094-3100. doi:10.1093/bioinformatics/bty191

Hash seed algorithm

It does split-read alignment

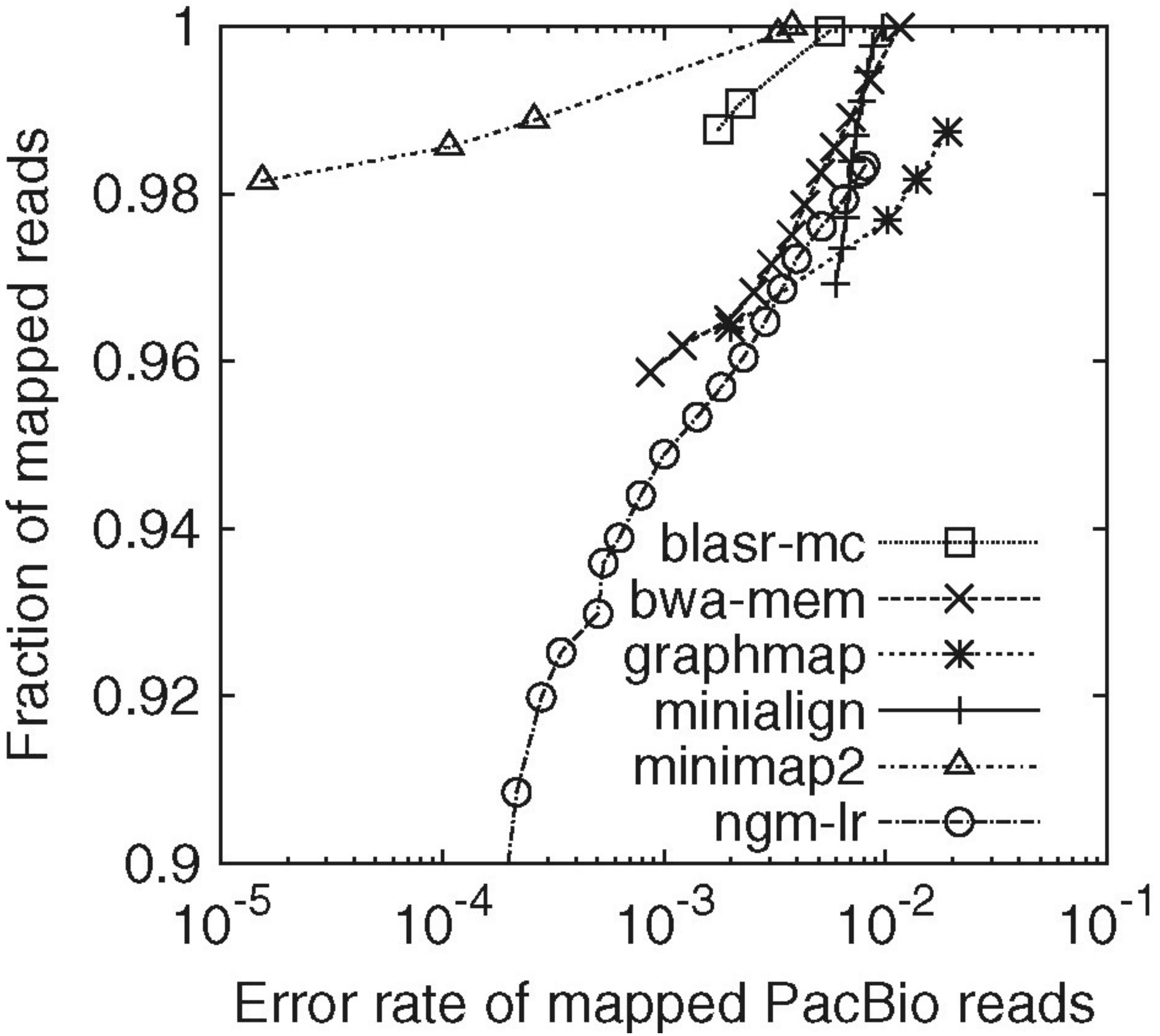
General-purpose:

- Noisy PacBio or Oxford Nanopore reads
- Illumina single- or paired-end reads
- splice-aware alignment of PacBio Iso-Seq or Nanopore cDNA or Direct RNA reads against a reference genome
- finding overlaps between long reads with error rate up to ~15%
- full-genome alignment between two closely related species with divergence below ~15%.

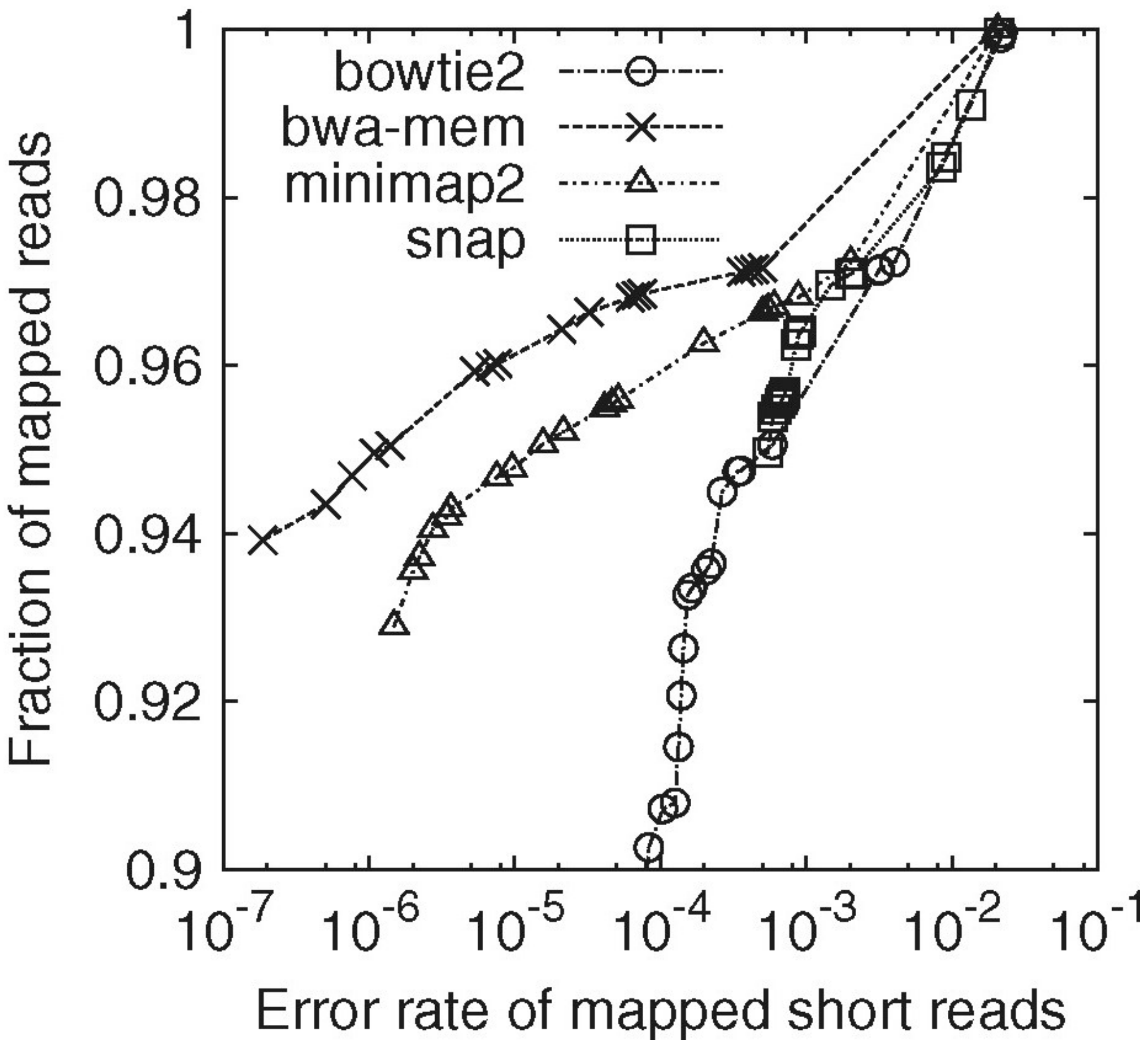
Same base algorithm for all applications, but different parameters

minimap2 long reads

tens of times faster than mainstream long-read mappers such as BLASR, BWA-MEM, NGMLR and GMAP



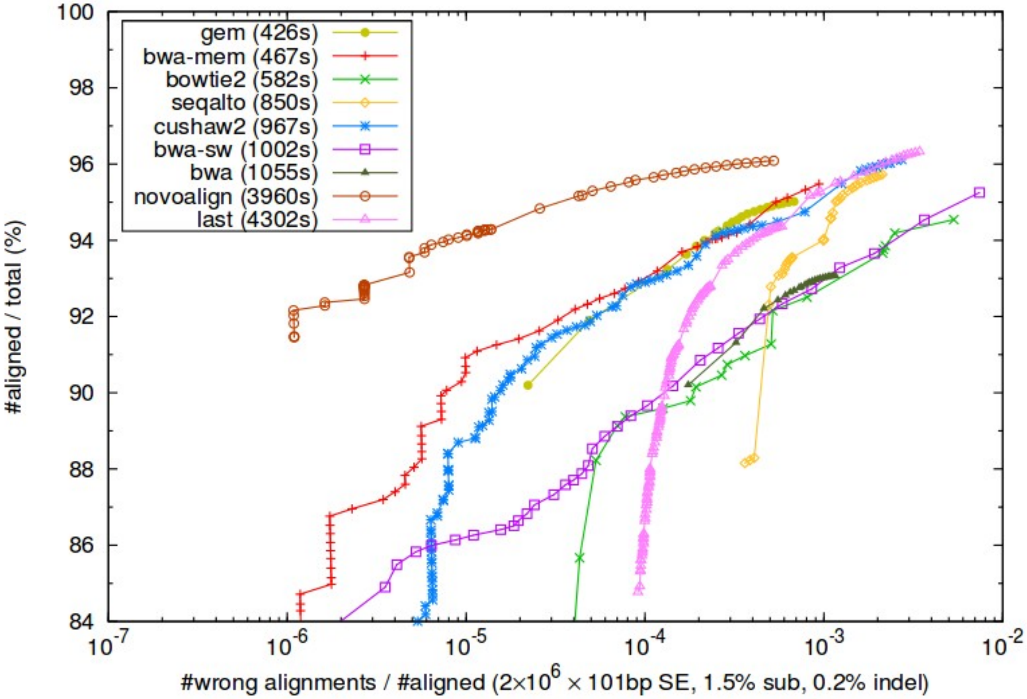
minimap2 Illumina reads



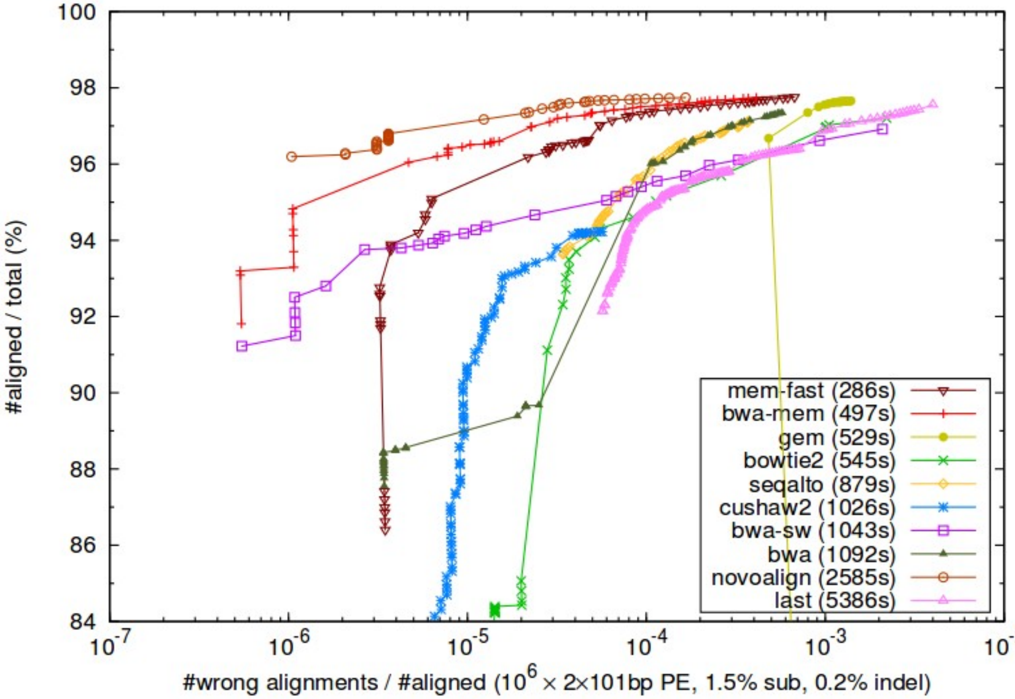
Less accurate than bwa-mem, but that might be fixed in newer versions

Mappers comparison

Single reads



Paired-ends

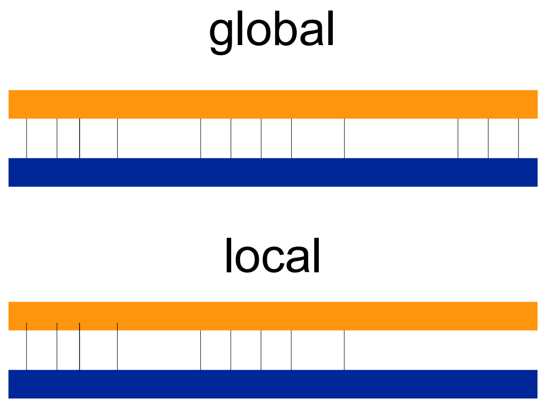


From bwa mem paper Li, H.
Simulated reads from human genome. 1.5% substitutions and 0.2% indels

Gobal vs local alignment

If the mapper does global alignments some alignments can be missed.

Features affecting global alignments:



- Bad quality tracks (common at the end of the reads).
- Non removed adapter or vector .
- EST read spanning a splice juntion.
- Read spanning a re-arrangement event.

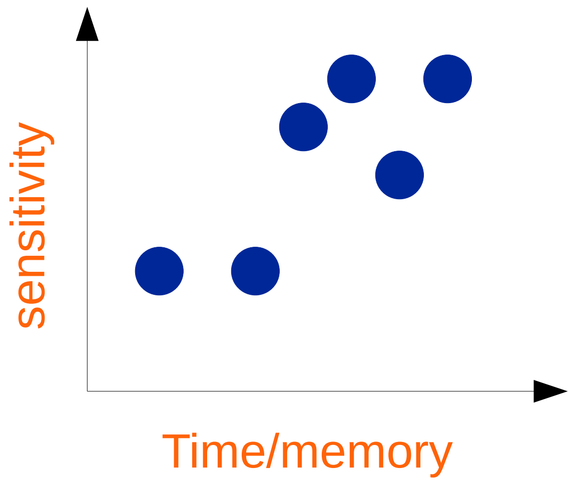
```
Reference ...ATCGACTGCGTCTAGTTACGATACGTTCATCGTATCGAT...
Read              tcaACGACTAGTTACGATACGTTacg
```

Mapping sensitivity

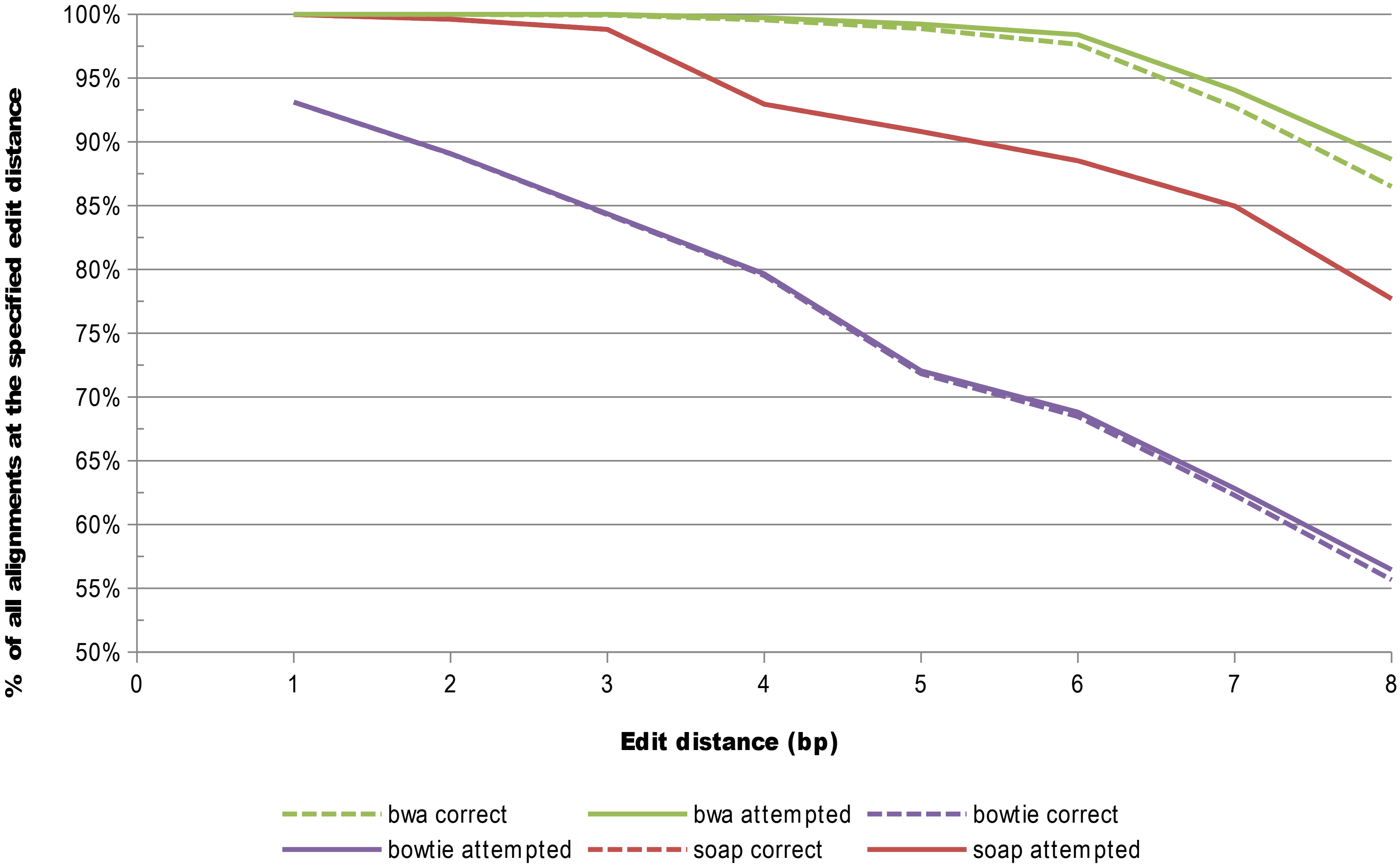
Sensitivity related mapper characteristics:

- Algorithm
- maximum edit distance (num. Mismatches)
  - Highly polymorphic regions are difficult to align
  - Interspecific mappings could be problematic
- allow large gaps
  - Introns
  - Structural variants

Mapper performance



Sensitivity vs edit distance



Michael Stromberg@bioinformatcis.ca

Sensitivity

Mapping against *A. thaliana* col. as reference

Species	Accession	SRA	% Mapped Reads
A. thaliana	Col	SRR513732	75%
	Ler	SRR392121	71%
	C24	SRR392124	72%
A. lyrata		SRR072809	69%
Brassica rapa		ERR037339	20%

Reads were preprocessed with Q20 L30. Mapping tool: Bowtie2

Taken from Aureliano Bombarely

One alignment?

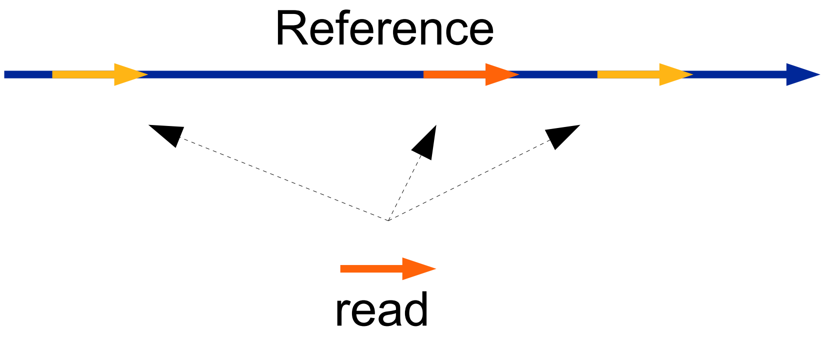
A read might be aligned to 0, 1 or more regions in the genome.

When several alignment are found we could classify them in two groups.

- Best alignments: alignments with best score
- Other alignments.

We can choose to report:

- All alignments.
- All best alignments.
- One of the best alignments at random.
- All alignments above a score threshold



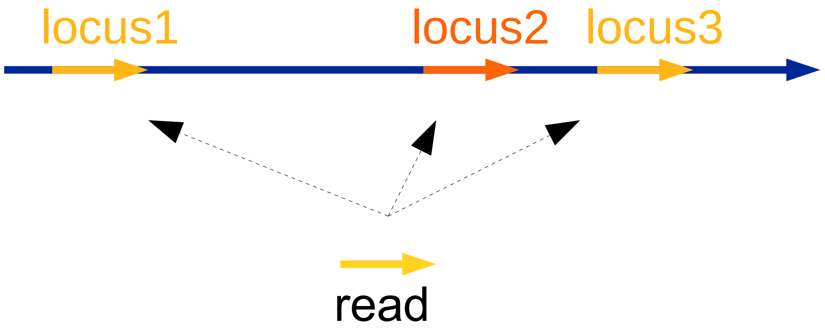
Mapping score (MAPQ)

MAPQ reflects the probability that the read originated from the region of the genome where it maps.

The mapping score of one alignment depends on:

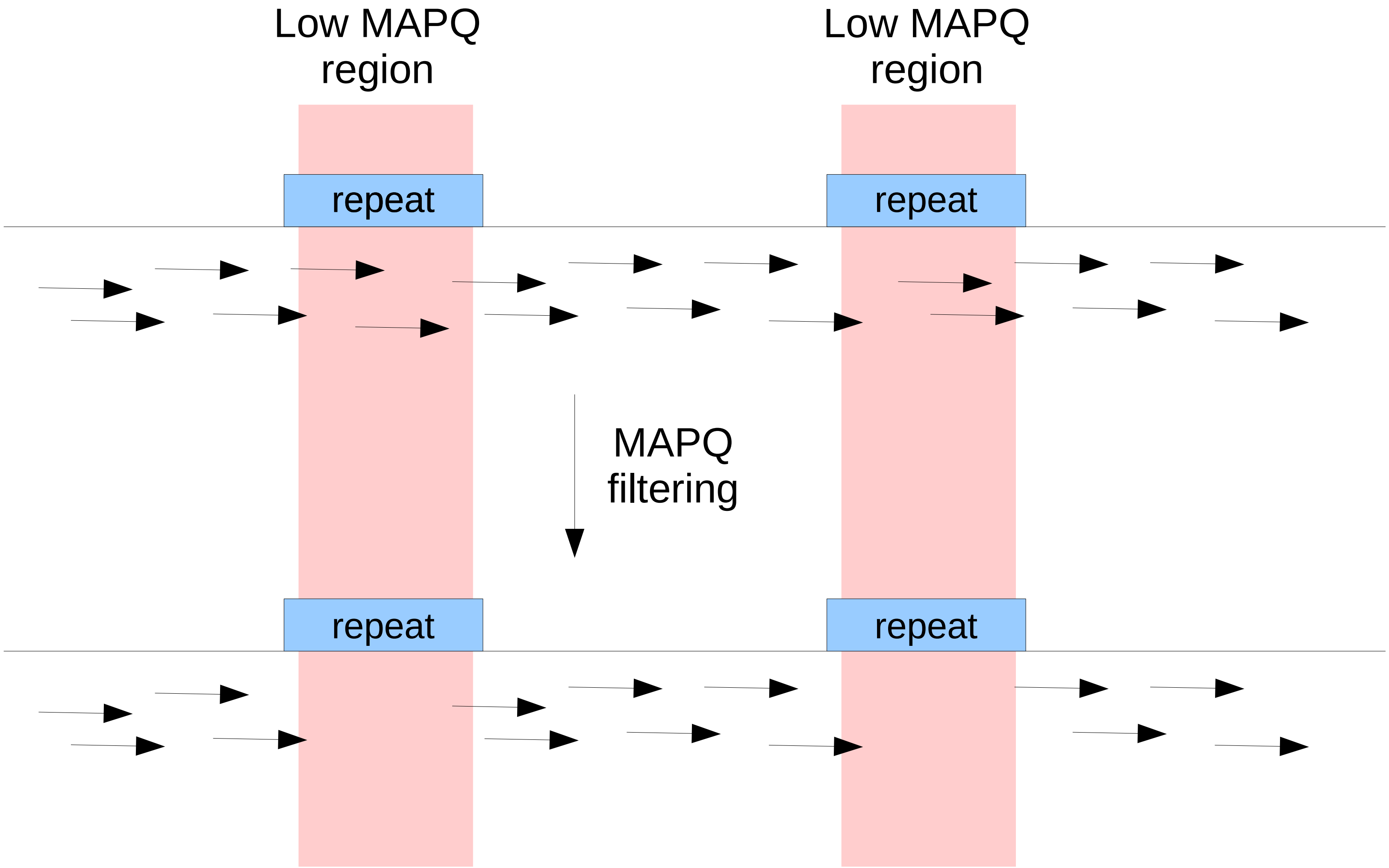
- how similar the read is to the reference and,
- how many alignments have been found.

The mapping score is usually given as a phred score.



```
Read     ACGTCTAGTTACGATACGTT
Locus1   ACGACTAGTTACGATACGTT  → score1
Locus2   ACGTCTAGCTACGCTAGGTT  → score2
Locus3   ACGACTAGTTACGATACGTT  → score1
```

MAPQ aims at removing duplicated regions



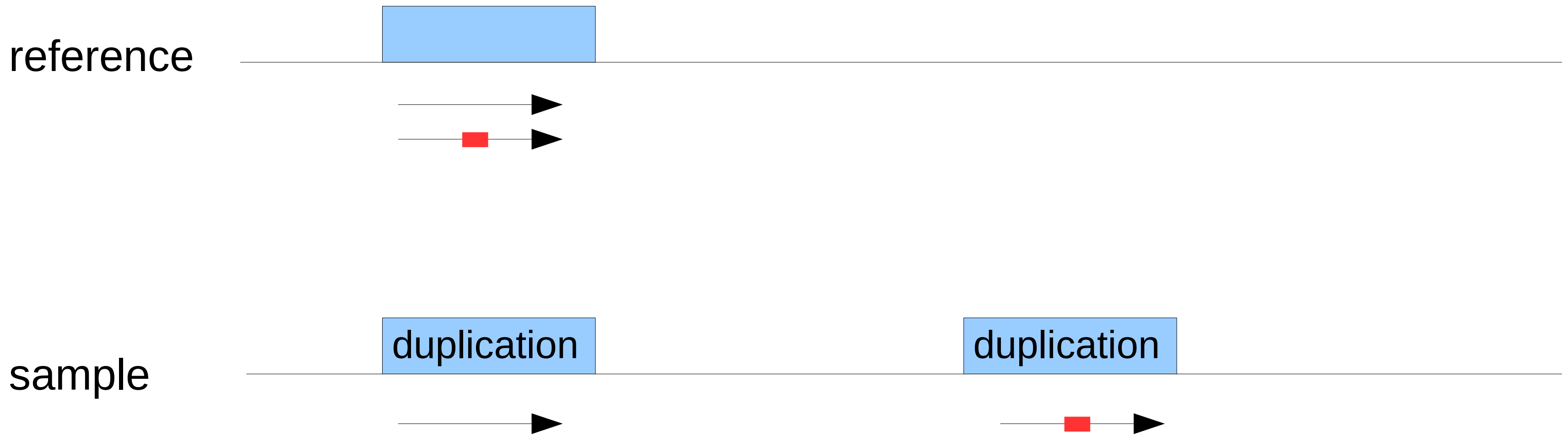
Repeat blindness

Duplicated regions are usually not analyzed:

- Repetitive elements (transposons, retrotransposons)
- Gene families
- Genes with pseudo-genes

This problem can be alleviated using pair-ends

MAPQ depends on the reference



Structrual Variants

Do not map against a region of the genome

Regions not found in the reference

Reads that correspond to regions not found in the reference won’t be mapped

- Incomplete genome reference
  - Specially relevant in transcriptomes
- Insertions in the sample relative to the reference
- Pathogens infecting the sample
- Chloroplast and mitochondrion
- Contamination

Many alignments vs multiple alignment

Mappers do many alignments, but they do not do multiple alignments.

Doing many pairwise alignments is computationally more feasible.

There's one drawback.

```
Ref      ...aggttttataaaacaattaagtctacagagcaacta...
Read1    ...aggttttataaaacaaAtaa
```

Many alignments vs multiple alignment

Mappers do many alignments, but they do not do multiple alignments.

Doing many pairwise alignments is computationally more feasible.

There's one drawback.

```
Ref       ...aggttttataaaacaattaagtctacagagcaacta...
Read1     ...aggttttataaaacaaAtaa


Ref       ...aggttttataaaac----aattaagtctacagagcaacta...
Sample    ...aggttttataaaacAAATaattaagtctacagagcaacta...
Read1     ...aggttttataaaac****aaAtaa
```

Many alignments vs multiple alignment

Mappers do many alignments, but they do not do multiple alignments.

Doing many pairwise alignments is computationally more feasible.

There's one drawback.

```
Ref      ...aggttttataaaacaattaagtctacagagcaacta...
Read1    ...aggttttataaaacaaAtaa


Ref      ...aggttttataaaac----aattaagtctacagagcaacta...
Sample   ...aggttttataaaacAAATaattaagtctacagagcaacta...
Read1    ...aggttttataaaac****aaAtaa

Ref      ...aggttttataaaac----aattaagtctacagagcaacta...
Sample   ...aggttttataaaacAAATaattaagtctacagagcaacta...
Read1    ...aggttttataaaac****aaAtaa
Read2    ....ggttttataaaac****aaAtaaTt
Read3    ........ttataaaacAAATaattaagtctaca.............
read4                CaaaT****aattaagtctacagagcaac......
read5                  aaT****aattaagtctacagagcaact.....
read6                    T****aattaagtctacagagcaacta....
```

many alignments

Many alignments vs multiple alignment

many alignments

```
Ref      ...aggttttataaaac----aattaagtctacagagcaacta...
Sample   ...aggttttataaaacAAATaattaagtctacagagcaacta...
Read1    ...aggttttataaaac****aaAtaa
Read2    ....ggttttataaaac****aaAtaaTt
Read3    ........ttataaaacAAATaattaagtctaca.............
read4                CaaaT****aattaagtctacagagcaac......
read5                  aaT****aattaagtctacagagcaact.....
read6                    T****aattaagtctacagagcaacta....
```

Multiple alignment

```
Ref      ...aggttttataaaac----aattaagtctacagagcaacta...
Sample   ...aggttttataaaacAAATaattaagtctacagagcaacta...
Read1    ...aggttttataaaacAAATaa
Read2    ....ggttttataaaacAAATaatt
Read3    ........ttataaaacAAATaattaagtctaca............
Read4                    cAAATaattaagtctacagagcaac.....
read5                      AATaattaagtctacagagcaact....
read6                        Taattaagtctacagagcaacta...
```

Many alignments vs multiple alignment

Strategies to mitigate this problem:

- Fix the problem.
  - GATK, GLIA realignment.
  - It realigns the problematic regions (lots of SNPs or some indels).
  - Computationally slow.
  - It does not fix all problems.

- Avoid using the misaligned positions.
  - Samtools BAQ (calmd).
  - For each position It calculates the probability of being misaligned.

- Most problematic regions are:
  - Low complexity
  - At the ends of reads

SAM format

Sequence Alignment/Map (http://samtools.sourceforge.net/)

File describing reads aligned to a reference genome.

Standard file format.

Not meant for human consumption, although can be opened with a text editor:

Its binary version is more common (BAM)

Input for genome browsers (e.g., IGV) and SNP callers.

It is usually found with the reads sorted along the reference

There are some differences in the output between mappers. For instance bwa represent multiple hits with an optional tag (XA) and bowtie with multiple lines (one per hit).

SAM format

Structure:

```
HEADER
 Version
 Program parameters
 Genome:
  Chrom1 size
  Chrom2 size
  Chrom3 size
 Groups:
  Group1: sample 1, library 1, platform
  Group2: sample 2, library 2, platform
BODY
 Read 1, group1
 Read 2, group1
 Read 3, group2
```

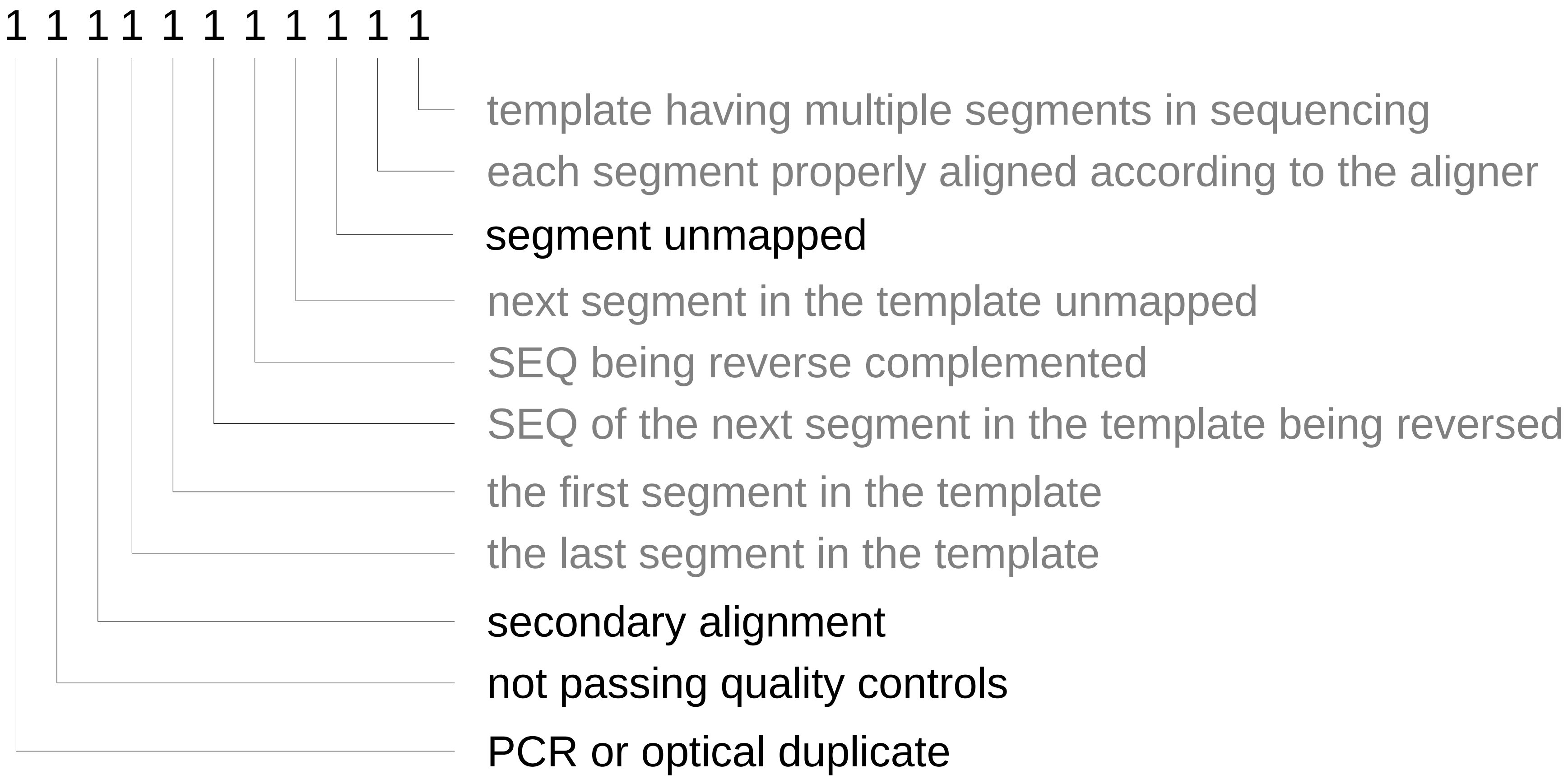
Example:

```
@HD VN:1.0
@SQ SN:chr20 LN:62435964
@RG ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99 chr20 28833 20 10M1D25M = 28993 195 AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG \
     \ <<<<<<<<<<<<<<<<<<<<<:<9/,&,22;;<<< NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M= 28701 -168 ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA \\
<<<<<;<<<<7;:<<<6;<<<<<<<<<<<<7<<<< MF:i:18 RG:Z:L2
```

Alignment section fields

Col	Field	Brief description
1	QNAME	Query template NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost mapping POSition
5	MAPQ	MAPping Quality
6	CIGAR	CIGAR string
7	RNEXT	Ref. Name of the mate/next read
8	PNEXT	Position of the mate/next read
9	TLEN	Observed Template LENgth
10	SEQ	segment SEQuence
11	QUAL	ASCII of Phred-scaled base QUALity+33

SAM flag



1 1 1 1 1 1 1 1 1 1 1

template having multiple segments in sequencing

each segment properly aligned according to the aligner

segment unmapped

next segment in the template unmapped

SEQ being reverse complemented

SEQ of the next segment in the template being reversed

the first segment in the template

the last segment in the template

secondary alignment

not passing quality controls

PCR or optical duplicate

Template: DNA/RNA which is sequenced on a sequencing machine or assembled from raw sequences.

Segment: contiguous (sub)sequence on a template which is sequenced or assembled.

Read: raw sequence that comes off a sequencing machine. A read may consist of multiple segments.

SAM QA

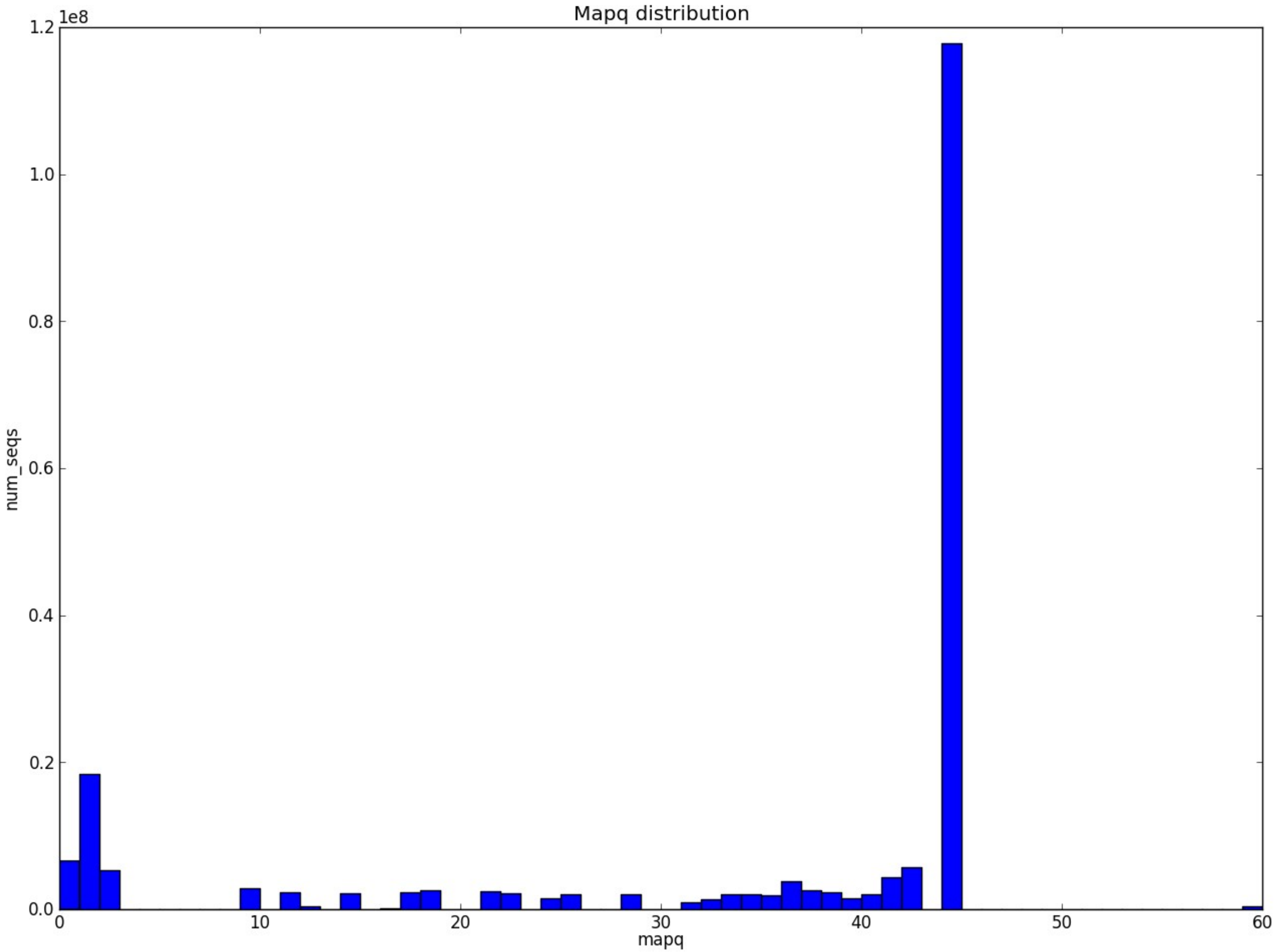
Flag statistics (samtools flagstats):

- Number of mapped and unmapped reads per read group

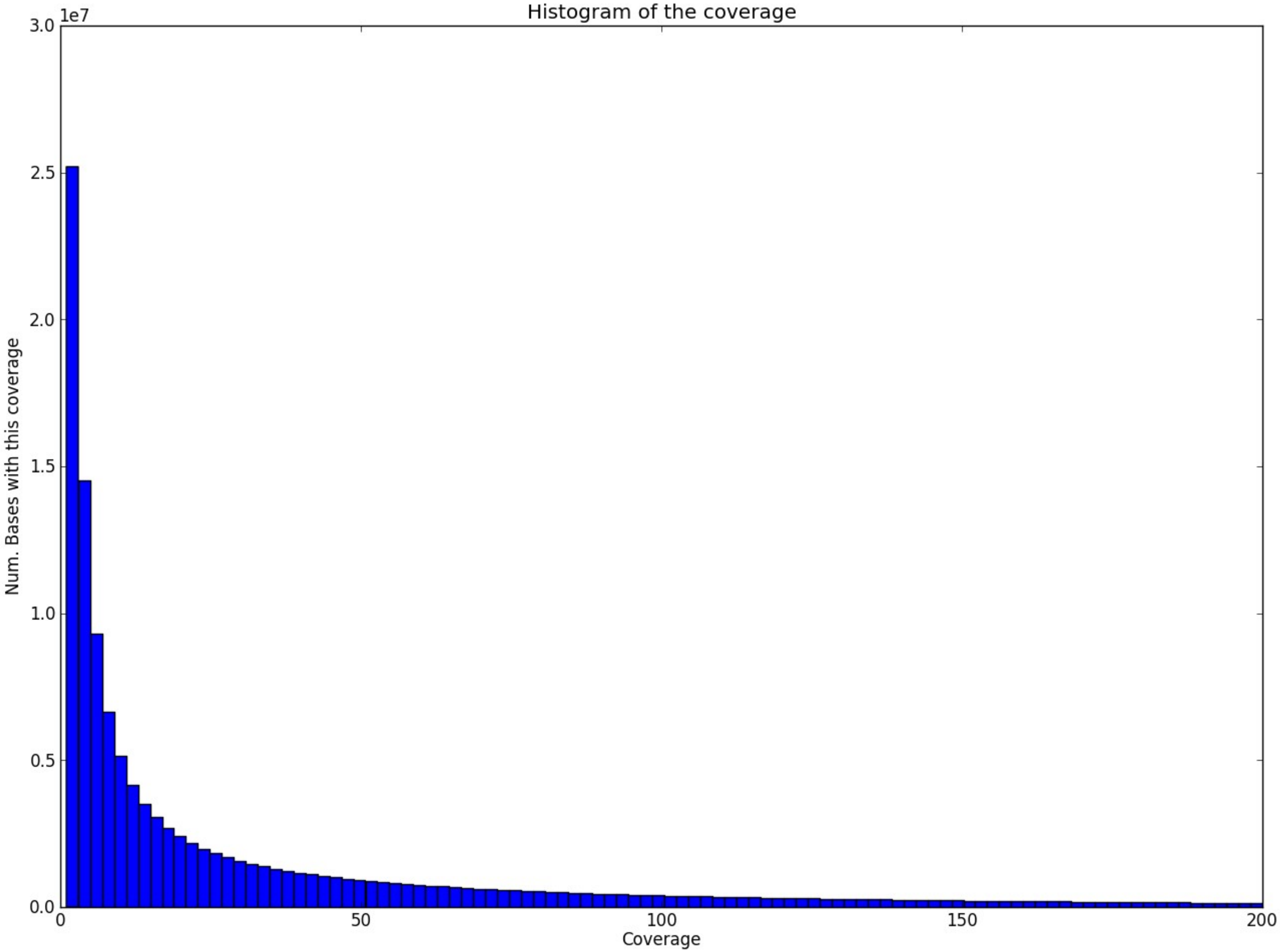
MAPQ distribution

Coverage distribution

MAPQ distribution



Coverage


Histogram of the coverage
1e7
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Num. Bases with this coverage
0
50
100
150
200
Coverage

SAM processing

Algorithms:

- Sorting
- Indexing
- Filtering
- Merging
- Read group modifications
- Duplicate location and removal
- Realignment
- BAQ

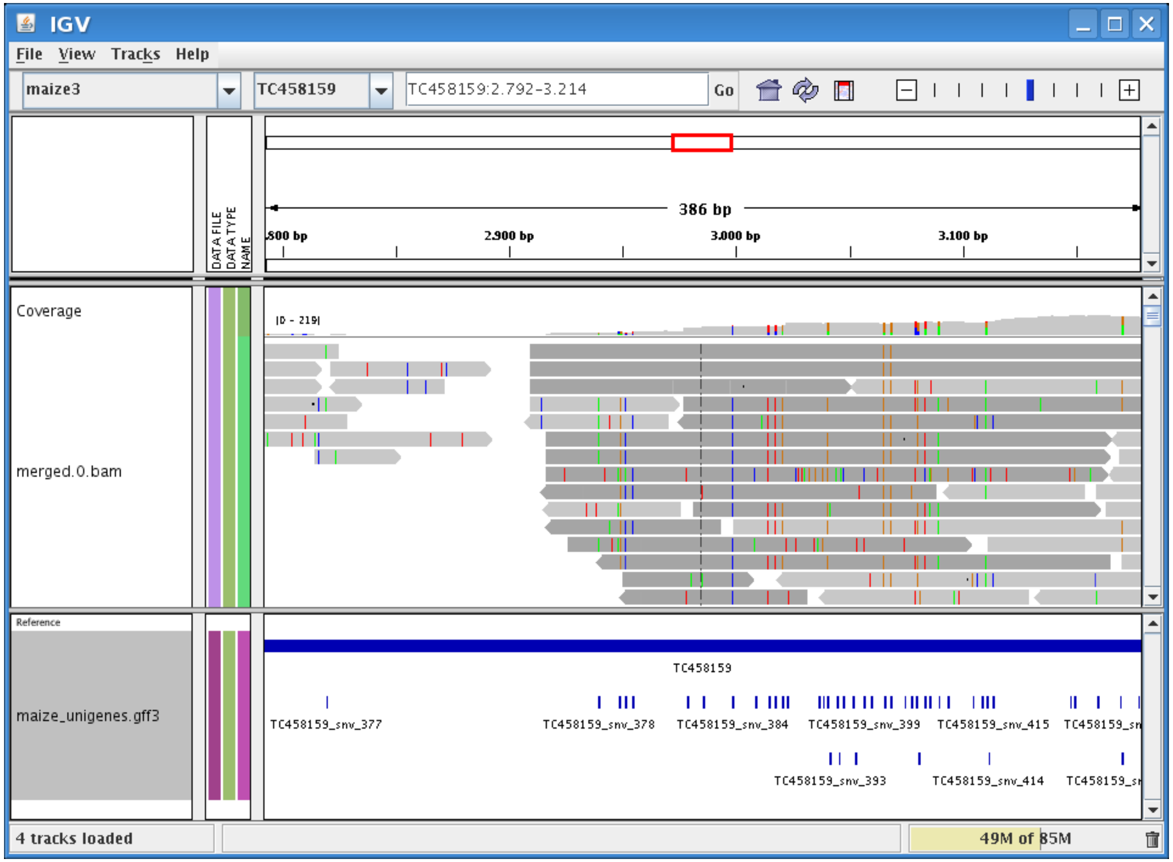
Software:

- Samtools
- Picard
- GATK

IGV viewer

Visualization tool for interactive exploration of large, integrated datasets.

Supports a wide variety of data types including: alignments, microarrays, and genomic annotations.



maize3

TC458159

TC458159:2.792-3.214

Jose Blanca
COMAV institute
bioinf.comav.upv.es