

# Analisis de datos (II): Mapping y Filtrado de duplicados.

Sara Monzón

BU-ISCI III

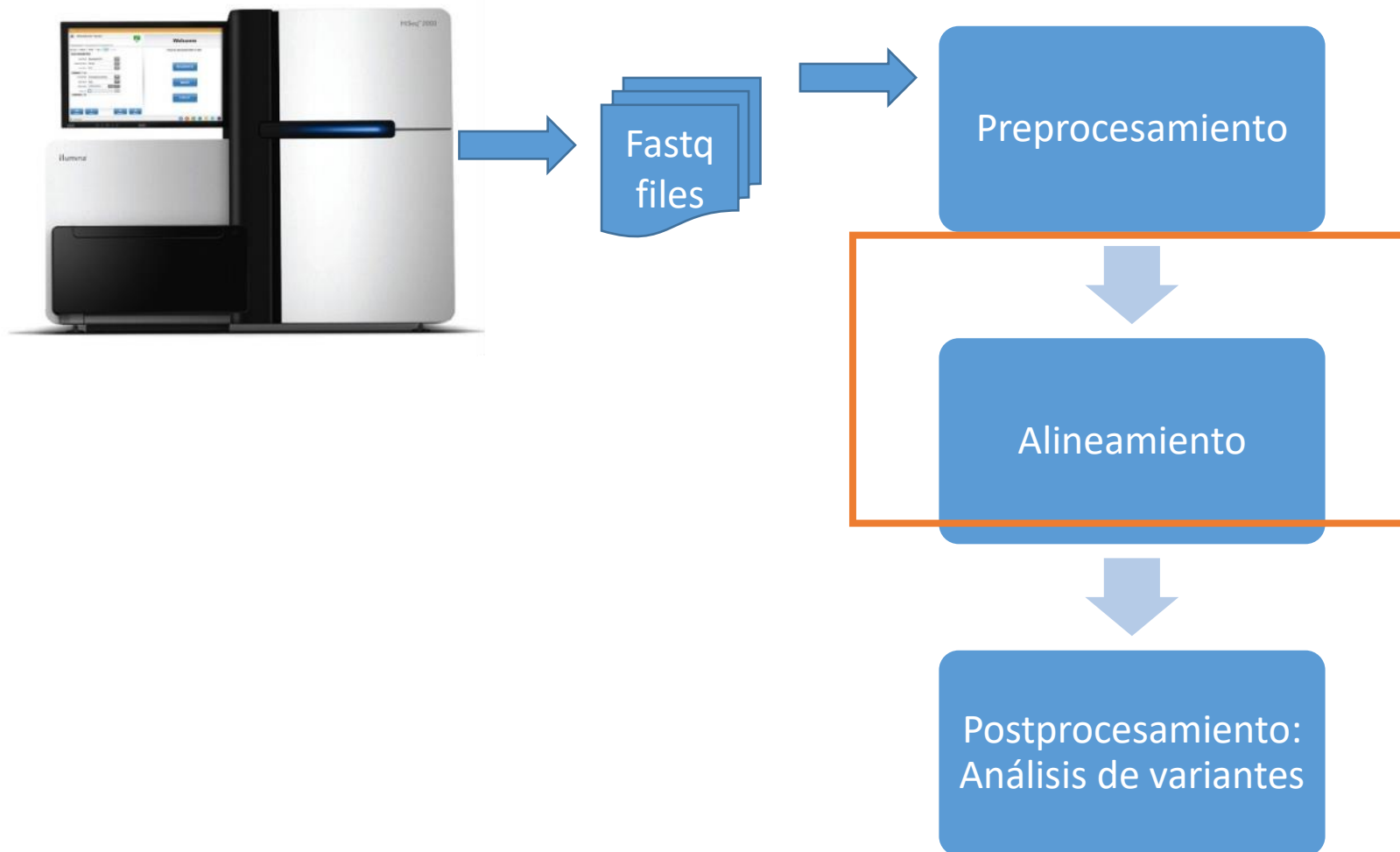
Unidades Científico Técnicas – SGAFI-ISCI III

22-26 Mayo 2023, 10ª Edición  
Programa Formación Continua, ISCI III

# Índice

- Dónde estamos
- Mapping vs Alineamiento
- Qué es el mapping
- Elección de alineador para NGS
- Formato SAM/BAM
- Filtrado de duplicados
- Objetivo de la práctica

## Dónde estamos



# Alineamiento

## Definición:

Colocar dos o mas secuencias de nucleótidos o de aminoácidos para identificar las regiones de similitud.

```
AAB24882      TYHMCQFHCERYVNNHSGEKLIECNERSKAFSCPSHLQCHKRRQIGKTHEHNQCGKAFPT
AAB24881      -----YECNQCGKAFAQHSSLKCHYRTHIGKPYECNQCGKAFSK

AAB24882      PSHLQYHERTHTGKPYECHQCGQAFKKCSLLQRHKRTHTGKPYE-CNQCGKAFAQ-
AAB24881      HSHLQCHKRTHTGKPYECNQCGKAFSQHGLLQRHKRTHTGKPYMNVINMVKPLHNS
```

## Alineamiento

### Alineamiento global: Needleman-Wunch (1970)

Encuentra el mejor posible alineamiento de dos secuencias a lo largo de toda su longitud .

```
--T--CC-C-AGT--TATGT-CAGGGGACACG-A-GCATGCAGA-GAC
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG-T-CAGAT--C
```

### Alineamiento local: Smith-Waterman (1981)

Encuentra regiones de altamente similares entre dos secuencias.

```
          tccCAGTTATGTCAGgggacacgagcatgcagagac
          | | | | | | | | | |
aattgccgccgtcggttttcagCAGTTATGTCAGatc
```

## Alineamiento múltiple (MSA)

### Definición:

Un alineamiento múltiple es una colección de tres o mas secuencias de aminoácidos o nucleótidos parcial o completamente alineados.

File		Edit	Colour	Sort	Picked:	Column 50: seq_cons/0-0	c = 48 (1 match)
(16x225)		-----10-----20-----30-----40-----50-----60-----70					
ALSE_ECOLI	2	202	KISPSLMCDLLKFKEQIEFIDS.HADYFHIDIMDGHFVFNLTLSFFVVSQVKKL.....AT				
RPE_YEAST	5	214	IIAPSIASDFANLGCECHKVINAGADWLHIDVMDGHFVFNITLGQPIVTSLRSVPRPGDASNTEKKPT				
O14105	5	204	KIAPSLLAGDFANLEKEVGRMLKYGSDWLHVDVMDAQFVFNLTIGPIVVKAMRNHYT.....KEE				
RPE_SYNY3	5	207	VVAPSILSADFSRLGEEIKAVDEAGADWIHVDVMDGRFVFNITIGPLIVDAIRPL.....Tk				
RPE_SOLTU	58	260	IVSPSILSANFSKLGEQVKAIEQAGCDWIHVDVMDGRFVFNITIGPLVVDSLRPI.....TI				
RPE_BACSU	3	204	KVAPSILSADFAALGNEIKDVEKGADCIHIDVMDGHFVFNITIGPLIVEAVRPV.....TI				
RPE_HAEIN	5	206	LIAPSILSADLARLGDDVQNVNLNAGADVIHFDVMDNHYVFNLTFGPAVCQALRDYG.....IT				
RPE_ECOLI	5	206	LIAPSILSADFARLGEDTAKALAAAGADVVHFDVMDNHYVFNLTIGPMVLKSLRNYG.....IT				
RPEC_ALCEU	17	221	RLAPSILSADFARLGEEVCAIEAGADLVHFDVMDNHYVFNLTIGPLVCEAIRPL.....VS				
RPE_RHORU	6	204	RIAPSLLSADFAISRPRCPSDGRGADILHFDVMDNHYVFNLTIGPLVCAALRPH.....TS				
RPE_MYCTU	9	207	LIAPSILAADFARLADGAAAVN..GADWLHVDVMDGHFVFNLTIGLPVVESLLAVTD.....IP.				
RPE_HELPY	2	200	KVAPSLLSADFMHLAKEIESVSN..DFLHVDVMDGHFVFNLTIGPVLLENVTQM.....SG				
RPE_METJA	3	201	KIGASILSADFGHLREEIKKAEAGVDFHVDMDGHFVFNITIGMIGIAKHVKKL.....TE				
SGCE_ECOLI	2	198	ILHPSILASANPLHYGRELTALDNLDGSLHLDIEDSSFINNITFGMKTQAVARQ.....TF				
RPE_MYCPN	9	203	ETAFSLPLLLHQFDRKLLQGFADGLRLIHYDVMD.HFVDNTVFQGEHLDELQIG.....				
RPE_MYCGE	15	198	.....RFDKSLLESYFQDGLRLIHYDVMD.QFVNHTAFKGEYLDELKTIG.....				

## Mapeo (mapping)

### Definición:

Situar una secuencia dentro de una secuencia mucho más larga. Por ejemplo, determinar la posición de una lectura dentro de un genoma.

```
Referencia/ genoma
...GTGGCCGGCAATTCGATATCGCGCATATTTTCGGCGCATGCTTAGC...

Lecturas:
GCAATTCGATAT
GCGCATATATTT
TGGCCGGCAAT
CGCATGCTTAGC
ATTCGATATCGC
GCCGGCAATTCG

Mapeo
...GTGGCCGGCAATTCGATATCGCGCATATTTTCGGCGCATGCTTAGC...
      GCAATTCGATAT          CGCATGCTTAGC
TGGCCGGCAAT    GCGCATATATTT
      ATTCGATATCGC

GCCGGCAATTCG
```

<sup>0</sup>Imagen:<http://personales.upv.es/jcanizar/bioinformatica/mapeo.html>

## Alineamiento múltiple vs Mapeo

			coord	12345678901234	5678901234567890123456
9	t	ttt	ref	aggtttttataaaaac----	aattaagtctacagagcaacta
10	a	aaaC	sample	aggtttttataaaaacAAAT	aattaagtctacagagcaacta
11	a	aaaaa	read1	aggtttttataaaaac	<u>aaA</u> taa
12	a	aaaaaa	read2	ggtttttataaaaac	<u>aaA</u> taaTt
13	a	aaaaaa	read3	ttataaaaac	<u>AAAT</u> aattaagtctaca
14	c	cccTTT	read4	<u>CaaaT</u>	aattaagtctacagagcaac
15	a	aaaaaa	read5	<u>aaT</u>	aattaagtctacagagcaact
16	a	aaaaaa	read6	<u>T</u>	aattaagtctacagagcaacta
17	t	AAtttt	read1	aggtttttataaaaac	<u>aaat</u> aa
18	t	tttttt	read2	ggtttttataaaaac	<u>aaat</u> aatt
19	a	aaaaaa	read3	ttataaaaac	<u>aaat</u> aattaagtctaca
20	a	aaaaaa	read4		<u>caaat</u> aattaagtctacagagcaac
21	g	Tgggg	read5		<u>aat</u> aattaagtctacagagcaact
			read6		<u>t</u> aattaagtctacagagcaacta

<sup>0</sup>Heng Li Mapping, alignment and SNP calling. MPG Next Gen Workshop 2011.



## Alineamiento múltiple vs Mapeo

### Mapeo:

- Esta bien si la secuencia solapa la región correcta
- Cada secuencia se mapea independientemente
- De miles a millones de secuencias

### Alineamiento múltiple

- Está bien si cada base se sitúa correctamente
- Minimiza las diferencias entre las secuencias
- De decenas a centenares de secuencias

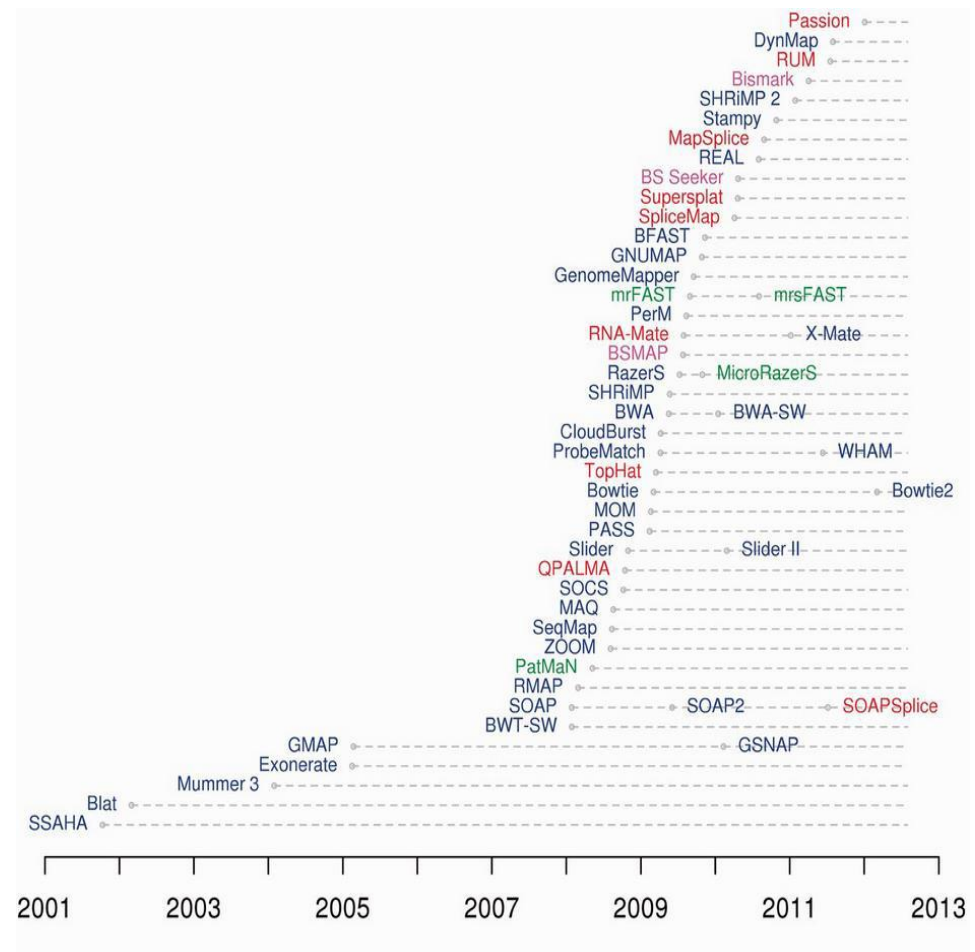
### Problema:

- Un algoritmo puede ser bueno mapando pero no necesariamente alineando
- Un buen alineamiento minimiza las diferencias entre lecturas mientras que un mapador solo ve la referencia

## Qué alineador usar

### Alineadores:

- Más de 60 alineadores disponibles.
- Muchos papers con reviews comparando características y rendimiento.



## Qué alineador usar

### Cosas a tener en cuenta:

- Recursos de computación vs sensibilidad
- Plataforma y tipo de experimento (Illumina/454/etc, paired-end, DNA/RNA/etc)
- Variación (permite indels, número de mismatch, etc.)
- Repeticiones (todas las regiones, best match, random, user defined number)

### Importante:

- Las opciones por defecto no tienen porqué ser las mejores
- “... there is no tool that outperforms all of the others in all the tests. Therefore, the end user should clearly specify his needs in order to choose the tool that provides the best results.” - Hatem et al *BMC Bioinformatics* 2013, **14**:184

# Qué alineador usar

TABLE 1: Application-specific alignment features distribution among multiple aligners.

Aligners	Operate system	Programming language	Input Format <sup>1</sup> ? (Fasta and Fastq)	Output format	Multithread?	Gapped alignment?	Paired-end alignment?	Trimming alignment?	Bisulfite alignment?	Note
Bowtie	★	C++	✓	SAM	✓		✓	✓		Maximum allowed mismatches ≤3
BWA	⊗	C++	✓	SAM	✓	✓	✓			BWA-short: 200 bp; BWA-SW: 100 kbp
BOAT	⊗	C	✓	*	✓	✓				Maximum allowed mismatches ≤3
GASSST	⊗	C++	Fasta	SAM	✓	✓				Merely Fasta format required for reads
Gnumap	⊗	C	✓ ( prb)	SAM	✓	✓		✓	✓	Maximum read length <1000 bp
GenomeMapper	⊗	C	✓	BED	✓	✓				Maximum read length < 2000 bp
mrFAST	★	C	✓	SAM		✓	✓			Maximum read length <300 bp
mrsFAST	★	C	✓	SAM			✓		✓	Maximum read length <200 bp
MAQ	⊗	C++	Fastq	map			✓			Maximum read length ≤128 bp
NovoAlign	●	C++	✓	SAM	✓	✓	✓	✓	✓	Restrictions for academic version
PASS	⊗	C++	✓ ( stf )	GFF3	✓	✓	✓			Maximum read length <1000 bp
PerM	⊗	C++	✓	SAM	✓		✓	✓		Maximum read length ≤128 bp
RazerS	★	C++	✓ ( prb)	Eland, GFF		✓	✓	✓		Arbitrary read length
RMAP	⊗	C++	✓	BED			✓		✓	Fixed-length reads required
SeqMap	★	C++	Fasta	Eland		✓				Maximum allowed mismatches ≤5
SOAPv2	⊗	C++	✓	*	✓	✓	✓			Maximum read length <1000 bp
SHRiMAP2	⊗	Python	Fasta	SAM	✓	✓	✓			Parallel computing supported
Segemehl	⊗	C	Fasta	*	✓	✓	✓	✓	✓	Large memory usage required
SSAHA2	●	NA	✓	GFF, SAM			✓			For long reads mapping

<sup>1</sup>We here only consider short-reads input format.

★Windows, Linux, or Unix operating system.

⊗Windows, Linux, Unix, or Mac X operating system.

●Linux, Unix, or Mac X operating system.

⊗Linux or Unix operating system.

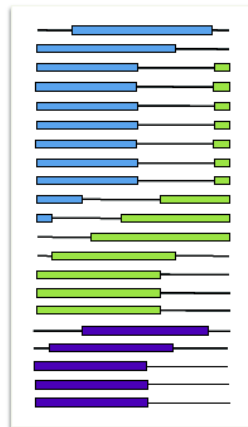
\* The short-read aligning algorithms' own output format.

<sup>0</sup>Shang et al 2014

## Qué alineador usar

- DNA
  - Whole Genome
  - Whole Exome
  - Amplicon
- Alineador: bowtie, bwa, bfast...

Enormous pile of short reads from NGS

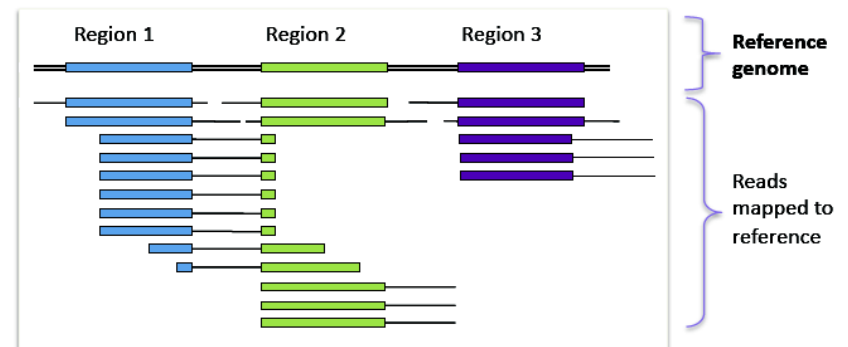


Mapping and alignment algorithms

Identify where the read matches the reference sequence and record match details as CIGAR string

RefPos:	1	2	3	4	5	6	7	8	9
Reference:	C	C	A	T	A	C	T	-	G
Read:		C	A	T	-	C	T	A	G
POS: 2									
CIGAR:									

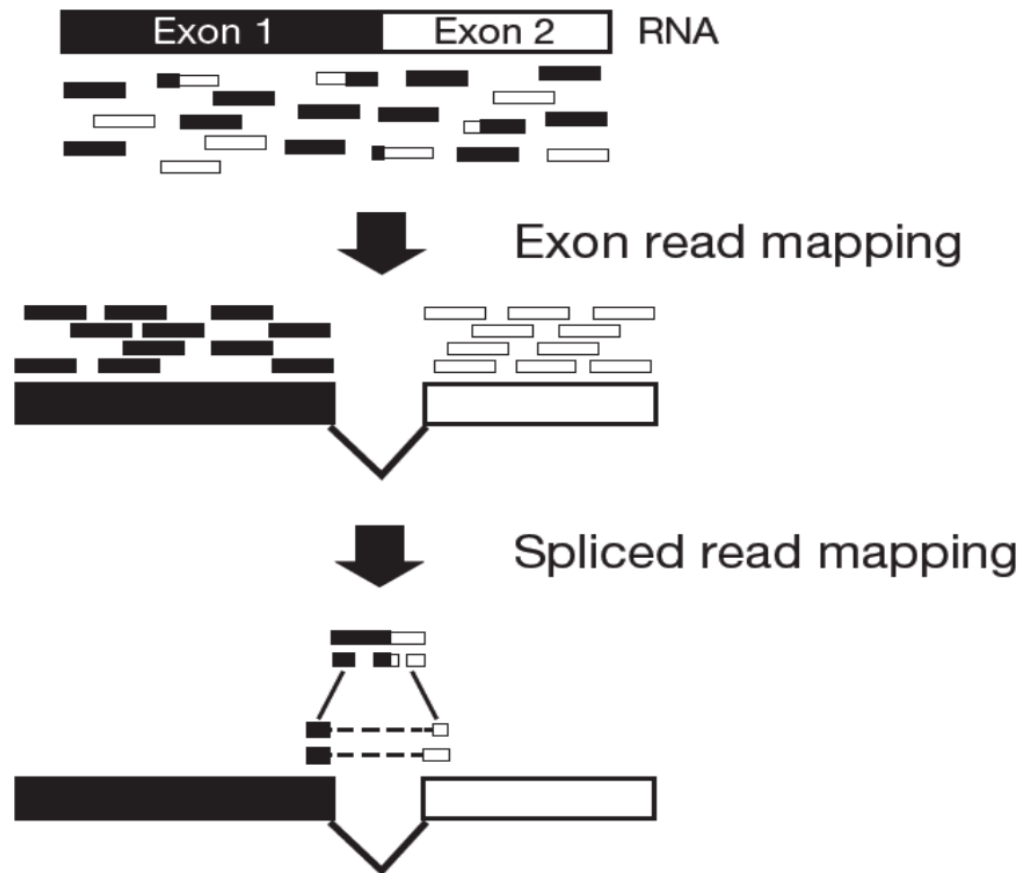
3M1D2M1I1M



## Qué alineador usar

- RNA
  - RNA-Seq
- Alineador: tophat, start...

### Exon-first approach



## Qué alineador usar

- Bisulphite sequencing
- Alineador: Bismark, BSMAP, BSeeker2



## Formato SAM

### Definición:

Es una especificación que define un formato genérico para representar alineamiento de nucleótidos. Describe el alineamiento de una secuencia query a una secuencia de referencia o ensamblaje.

```
@HD VN:1.5 S0:coordinate
@SQ SN:ref LN:45
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```



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

## Formato SAM

Col	Field	Type	Regex/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	[0,2 <sup>16</sup> -1]	bitwise FLAG
3	RNAME	String	\*  [!-( )+-<>-~] [!-~]*	Reference sequence NAME
4	POS	Int	[0,2 <sup>31</sup> -1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2 <sup>8</sup> -1]	MAPping Quality
6	CIGAR	String	\*  ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	\* =  [!-( )+-<>-~] [!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2 <sup>31</sup> -1]	Position of the mate/next read
9	TLEN	Int	[-2 <sup>31</sup> +1,2 <sup>31</sup> -1]	observed Template LENgth
10	SEQ	String	\*  [A-Za-z=.]+	segment SEQUENCE
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33

@HD VN:1.5 S0:coordinate

@SQ SN:ref LN:45

```
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

## Formato SAM

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

## Formato SAM

Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first segment in the template
0x80	the last segment in the template
0x100	secondary alignment
0x200	not passing quality controls
0x400	PCR or optical duplicate
0x800	supplementary alignment

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

## Flag explanation example 1

SAM Flag:

Toggle first in pair / second in pair

**Find SAM flag by property:**

To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- ☒ read paired
- ☒ read mapped in proper pair
- ☐ read unmapped
- ☐ mate unmapped
- ☐ read reverse strand
- ☒ mate reverse strand
- ☒ first in pair
- ☐ second in pair
- ☐ not primary alignment
- ☐ read fails platform/vendor quality checks
- ☐ read is PCR or optical duplicate
- ☐ supplementary alignment

**Summary:**

- read paired (0x1)
- read mapped in proper pair (0x2)
- mate reverse strand (0x20)
- first in pair (0x40)

## Flag explanation example 2

SAM Flag:

Toggle first in pair / second in pair

**Find SAM flag by property:**  
To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- ☒ read paired
- ☒ read mapped in proper pair
- ☐ read unmapped
- ☐ mate unmapped
- ☒ read reverse strand
- ☐ mate reverse strand
- ☐ first in pair
- ☒ second in pair
- ☐ not primary alignment
- ☐ read fails platform/vendor quality checks
- ☐ read is PCR or optical duplicate
- ☐ supplementary alignment

**Summary:**  
read paired (0x1)  
read mapped in proper pair (0x2)  
read reverse strand (0x10)  
second in pair (0x80)

## Formato SAM

### Formato texto - SAM

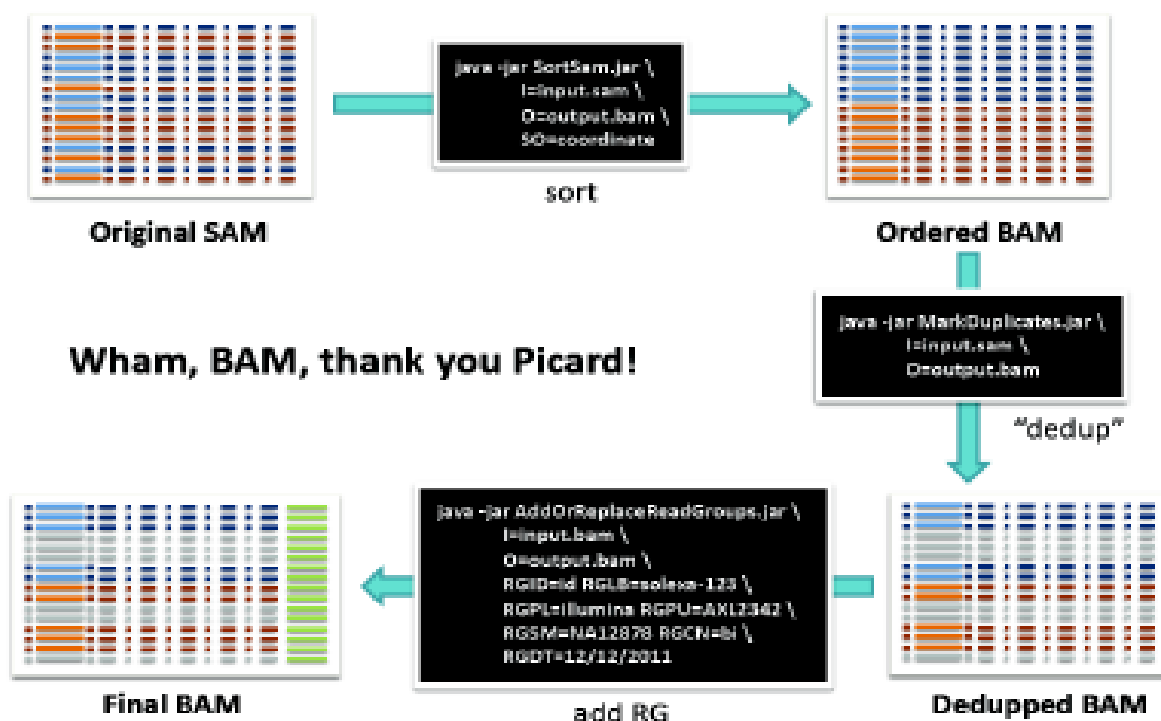
- Delimitado por tabuladores
- Es lo suficientemente sencillo para ser generado por los programas de alineamiento o ser convertido desde otros formatos existentes
- Es simple de *parsear* y puede ser producido al vuelo (*streaming*) desde un BAM
- Es adecuado para un análisis exploratorio o para conectar con otras aplicaciones

### Formato binario - BAM

- Utiliza una compresión BGZF
- Sus valores numéricos son independientes del sistema base
- Es lo suficientemente sencillo para ser generado por los programas de alineamiento o ser convertido desde otros formatos existentes
- Permite ser indexado para proporcionar un acceso rápido a las lecturas que solapan un determinado *locus*
- Debe ordenarse por coordenadas antes de indexar

## Duplicados de PCR

Utilización de *Picard* para finalizar la preparación del fichero SAM/BAM



## Mapping statistics

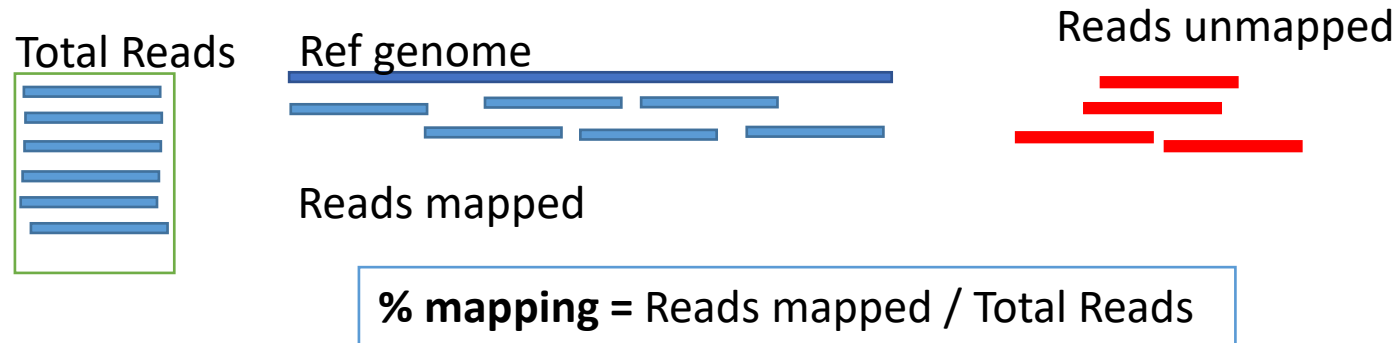
- % mapped:  $\text{reads mapped} / \text{total reads}$
- % unmapped:  $\text{reads unmapped} / \text{total reads}$
- % duplicates:  $\text{reads belonging to same template} / \text{total reads}$
- Mean depth of coverage
- Coverage: % genome with at least one read mapped.



# Mapping quality control

- **% mapping:** number of reads mapping againsts reference genome.

Picard  
Samtools

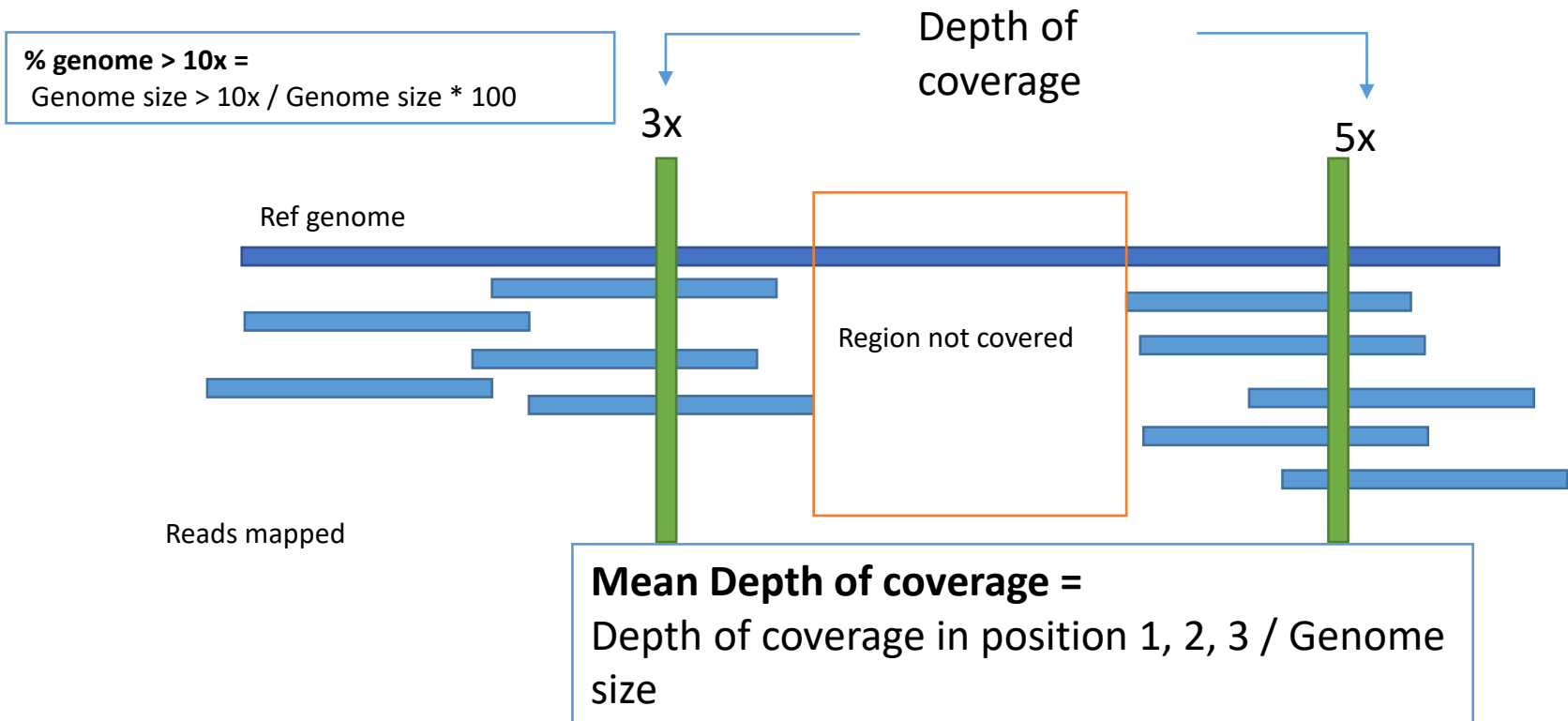


Mandatory parameter for microbial genomics!! It indicates us how many reads we have from our organism of interest. In human genomics this is almost always 99.99% unless something terrible happens. Not here!!!

# Mapping quality control

- **% genome > 10x**: percentage of genome covered with more than 10 reads.
- **Mean Depth of coverage**: mean of reads covering a genome position.

Picard  
Samtools



# ¿Preguntas?

---