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

Sara Monzón

BU-ISCIII

Unidades Científico Técnicas – SGAFI-ISCIII

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

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

Fastq files

Preprocesamiento

Alineamiento

Postprocesamiento: Análisis de variantes

Alineamiento

Definición:

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

```
AAB24882        TYHMCQFHCRYVNNHSGEKLYECNERSKAFSCPSHLQCHKRRQIGEKTHEHNQCGKAFPT
AAB24881        --------------------YECNQCGKAFAQHSSLKCHYRTHIGEKPYECNQCGKAFSK

AAB24882        PSHLQYHERTHTGEKPYECHQCGQAFKKCSLLQRHKRTHTGEKPYE-CNQCGKAFAQ-
AAB24881        HSHLQCHKRTHTGEKPYECNQCGKAFSQHGLLQRHKRTHTGEKPYMNVINMVKPLHNS
```

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)

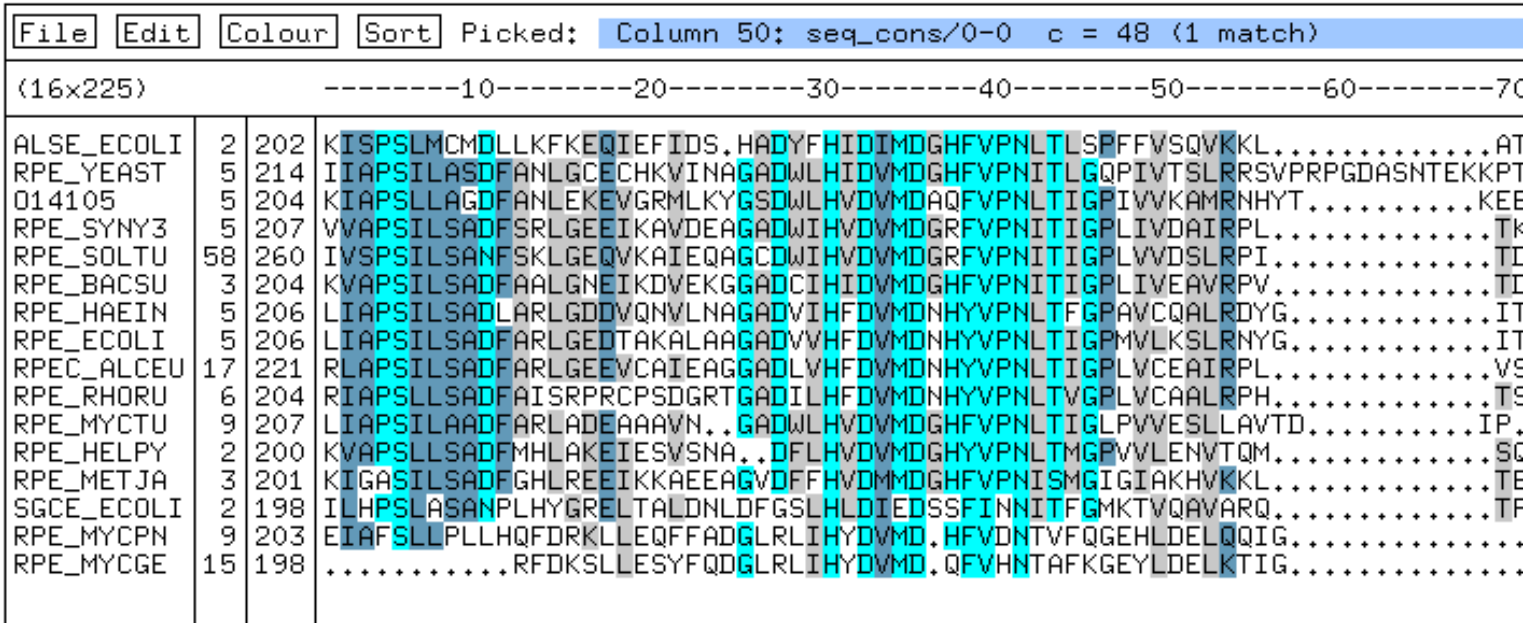
Encuentra regiones de altamente similares entre dos secuencias.

```
                    tccCAGTTATGTCAGgggacacgagcatgcagagac
                       ||||||||||||
aattgccgccgtcgttttcagCAGTTATGTCAGatc
```

Alineamiento múltiple (MSA)

Definición:

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



Mapeo (mapping)

Definición:

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

```
        Referencia/ genoma

...GTGGGCCGGCAATTCGATATCGCGCATATATTTCGGCGCATGCTTAGC...

Lecturas:

GCAATTCGATAT
GCGCATATATTT
TGGGCCGGCAAT
CGCATGCTTAGC
ATTCGATATCGC
GCCGGCAATTCG


        Mapeo

...GTGGGCCGGCAATTCGATATCGCGCATATATTTCGGCGCATGCTTAGC...
           GCAATTCGATAT                CGCATGCTTAGC
    TGGGCCGGCAAT        GCGCATATATTT
              ATTCGATATCGC

  GCCGGCAATTCG
```

[0]Imagen:http://personales.upv.es/jcanizar/bioinformatica/mapeo.html

Alineamiento múltiple vs Mapeo

9	t	ttt
10	a	aaaC
11	a	aaaaa
12	a	aaaaaa
13	a	aaaaaa
14	c	cccTTT
15	a	aaaaaa
16	a	aaaaaa
17	t	AAtttt
18	t	tttttt
19	a	aaaaaa
20	a	aaaaaa
21	g	Tgggg

```
coor      12345678901234     567890123456789012345678
ref       aggttttataaaac----aattaagtctacagagcaacta
sample    aggttttataaaacAAATaattaagtctacagagcaacta
read1     aggttttataaaac    aaAtaa
read2      ggttttataaaac    aaAtaaTt
read3          ttataaaacAAATaattaagtctaca
read4              CaaaT    aattaagtctacagagcaac
read5                aaT    aattaagtctacagagcaact
read6                 T     aattaagtctacagagcaacta
read1     aggttttataaaacaaataa
read2      ggttttataaaacaaataatt
read3          ttataaaacaaataattaagtctaca
read4                caaataattaagtctacagagcaac
read5                  aataattaagtctacagagcaact
read6                   taattaagtctacagagcaacta
```

[0]Heng Li Mapping, aligment 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

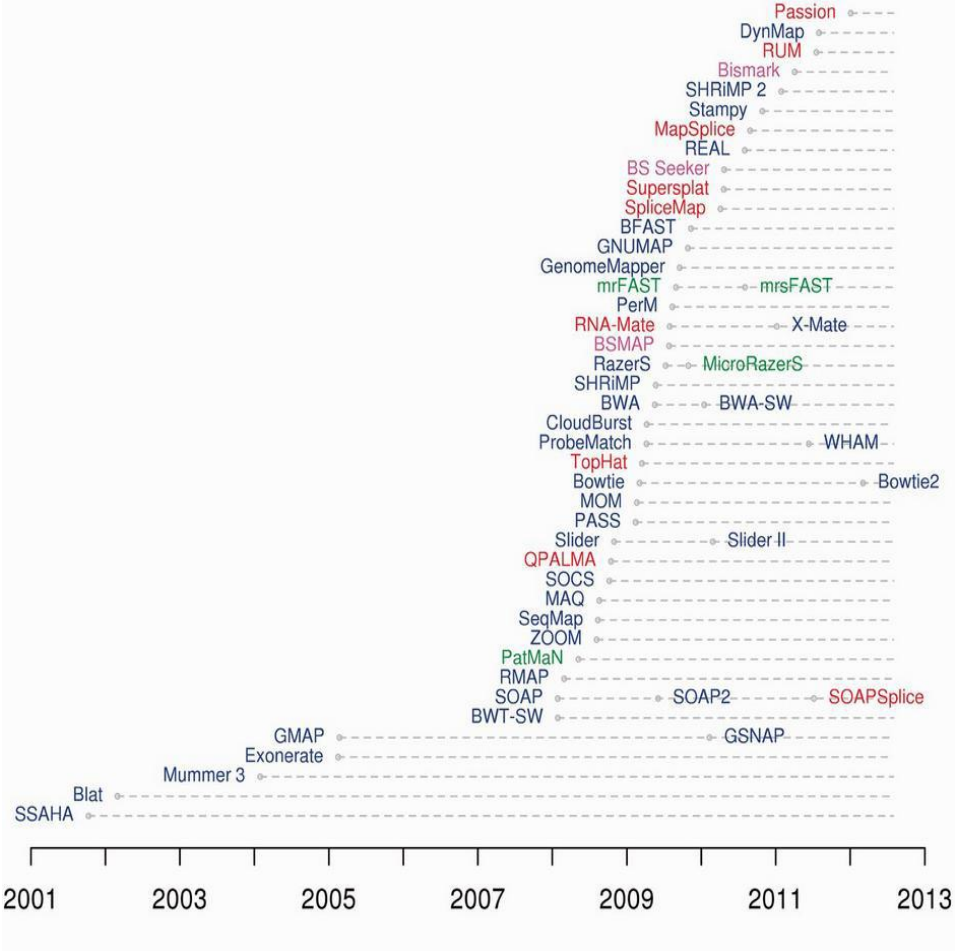
Problema:

- 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 mistmatch,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 I: Application-specific alignment features distribution among multiple aligners.

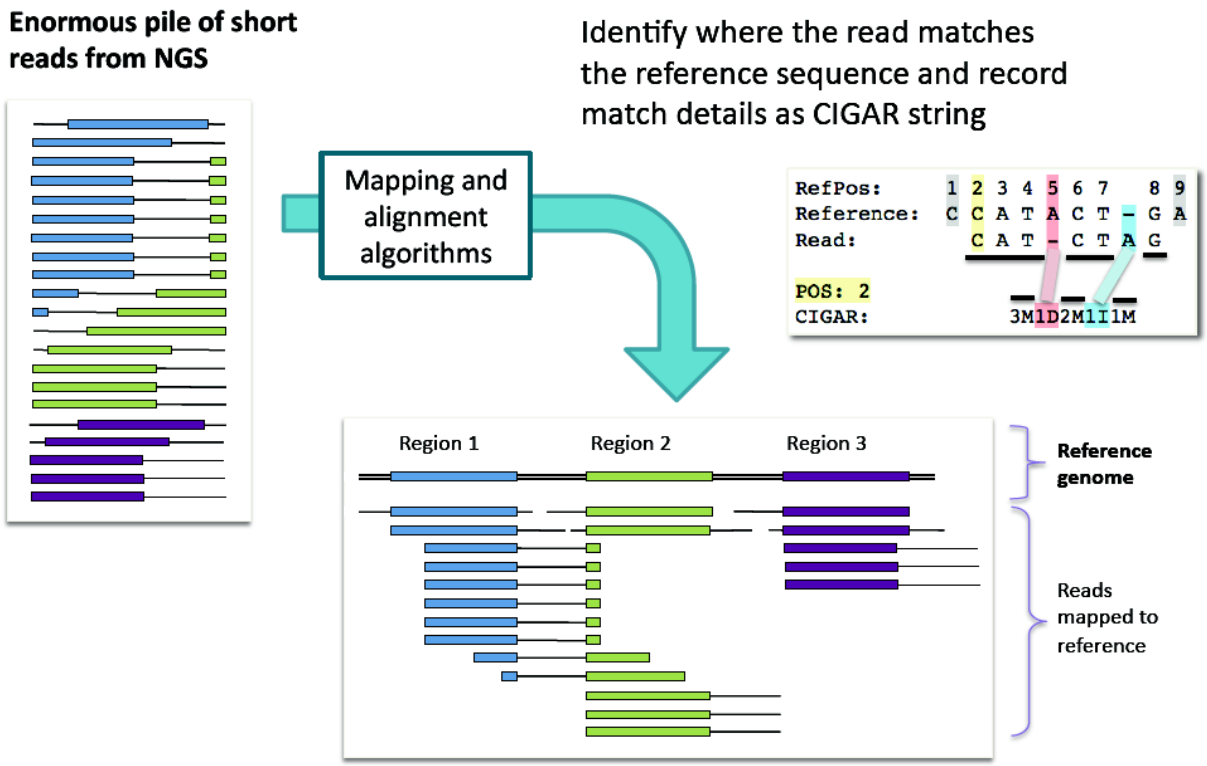
Aligners	Operate system	Programming language	Input Format[1]? (Fasta and Fastq)	Output format	Multithread?	Gapped alignment?	Paired-end alignment?	Trimming alignment?	Bisulfite alignment?	Note
Bowtie	★	C++	√	SAM	√		√	√		Maximum allowed mismatches ≤3
BWA	⊛	C++	√	SAM	√	√	√			BWA-short: 200 bp; BWA-SW: 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++	√ (sff)	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

[1] 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.

[0]Shang et al 2014

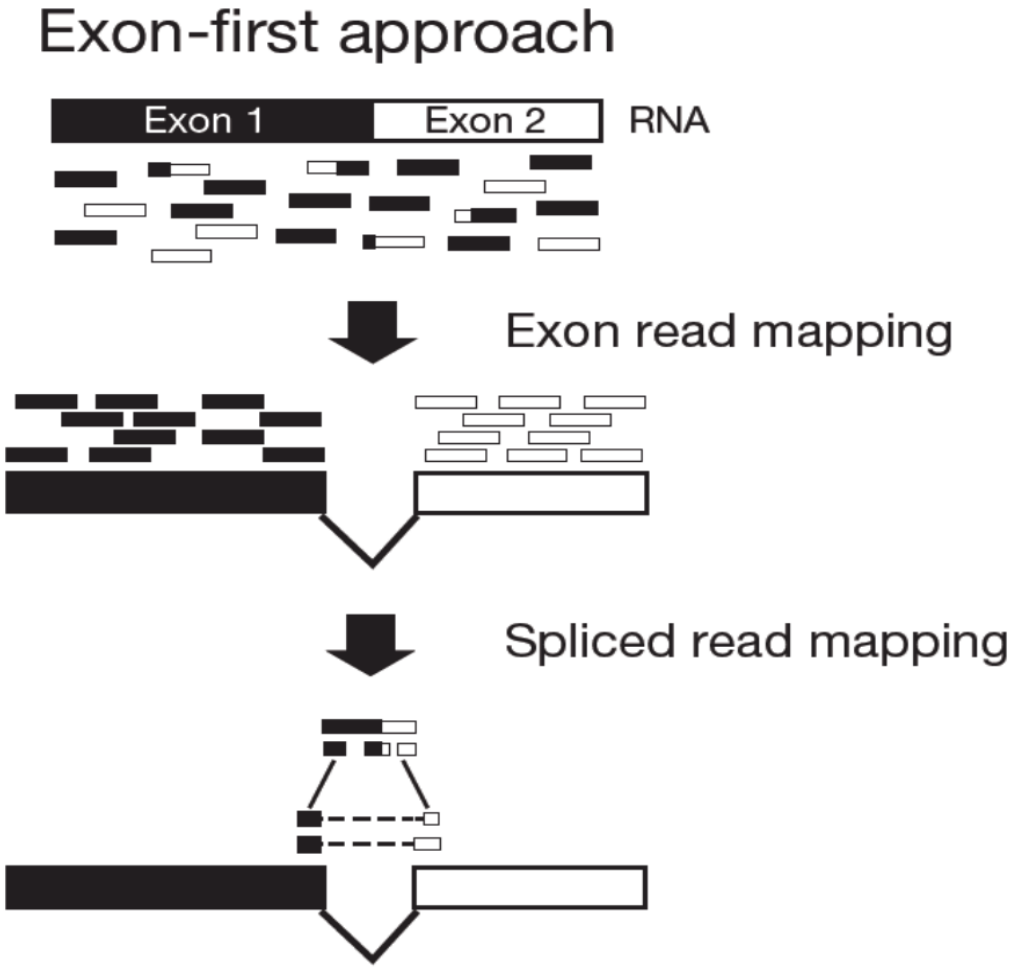
Qué alineador usar

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



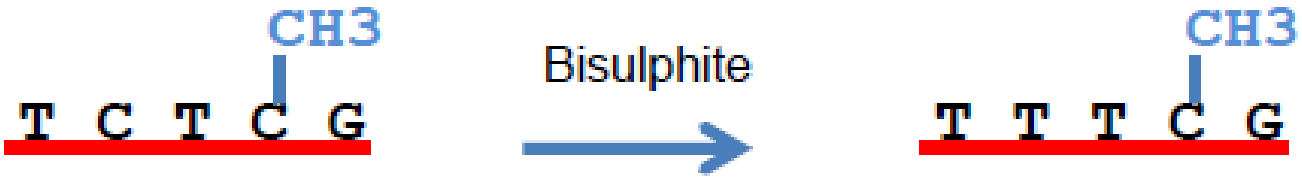
Qué alineador usar

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



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 SO: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

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

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	[0,2^{16}-1]	bitwise FLAG
3	RNAME	String	*|[!-()+-<>-~][!-~]*	Reference sequence NAME
4	POS	Int	[0,2^{31}-1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2^{8}-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^{31}-1]	Position of the mate/next read
9	TLEN	Int	[-2^{31}+1,2^{31}-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 SO: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

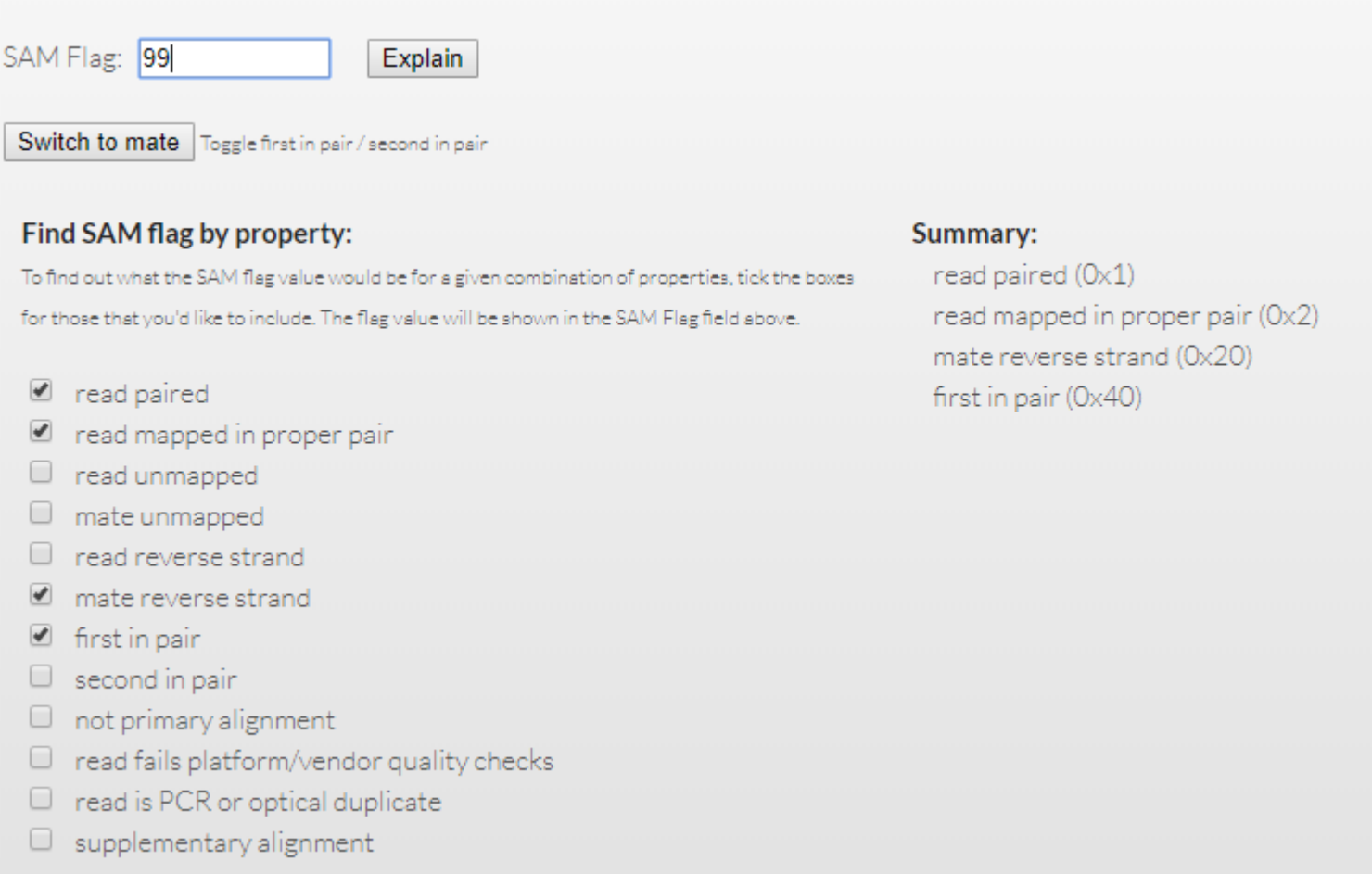
Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent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

Formato SAM

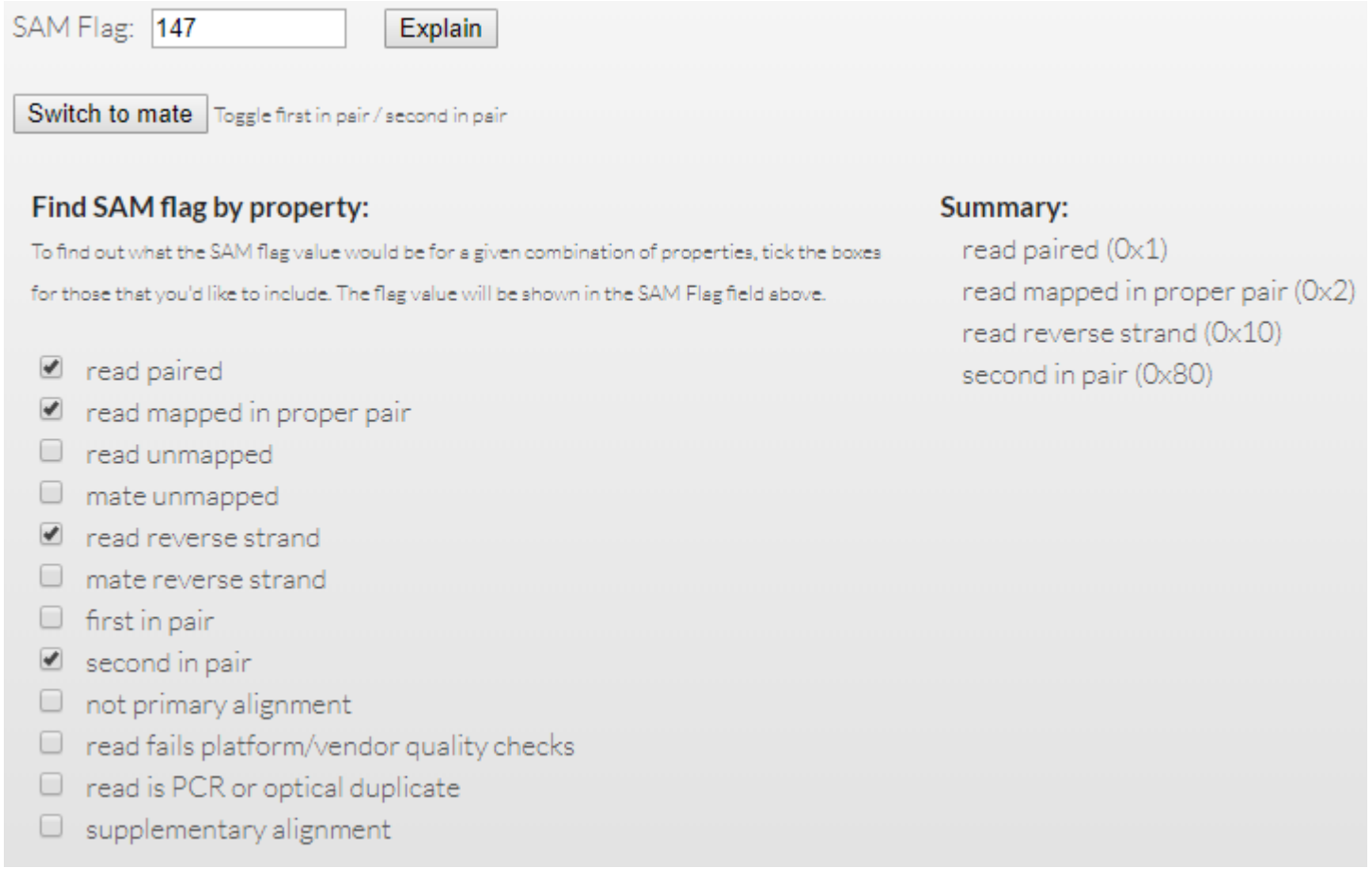
Bit	Description
0x1	template having multiple segments in sequencing
0x2	each segment properly aligned according to the aligner
0x4	segment unmapped
0x8	next segment in the template unmapped
0x10	SEQ being reverse complemented
0x20	SEQ of the next segment in the template being reversed
0x40	the first 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



Flag explanation example 2



Formato SAM

Formato texto - SAM

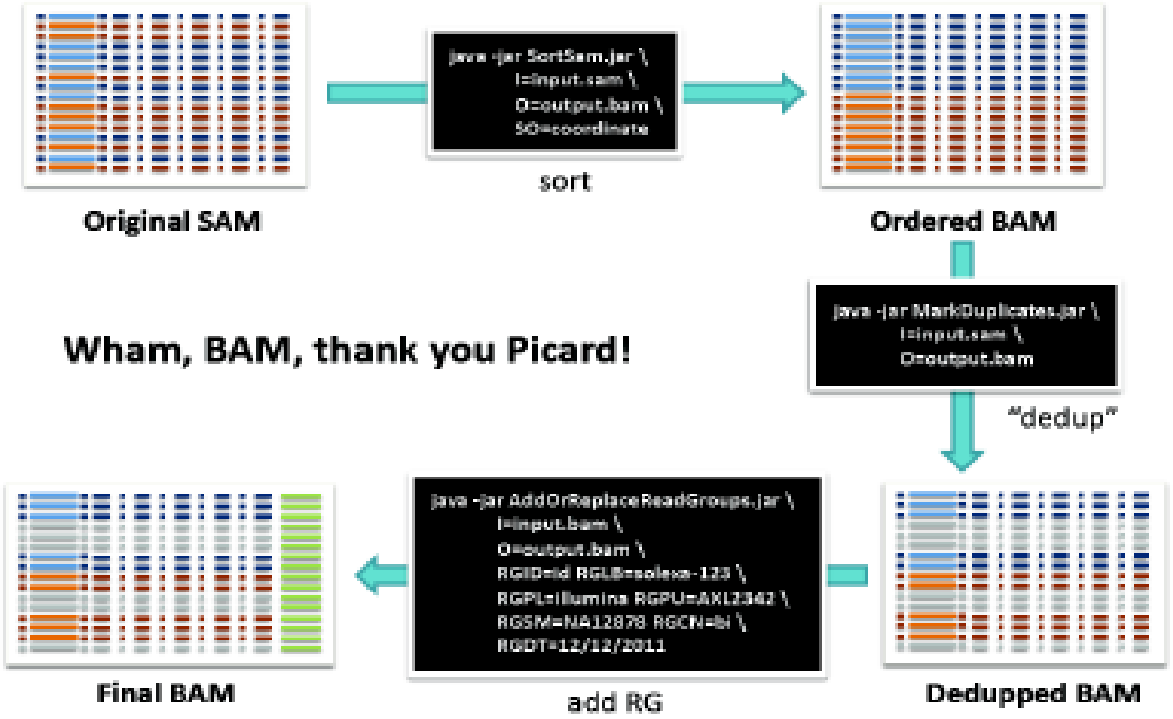
- Delimitado por tabuladores
- Es lo suficientemente sencillo para ser generado por los programas de alineamiento o ser convertido desde otros formatos existentes
- Es simple de *parsear* y puede ser producido al vuelo (*streaming*) desde un BAM
- Es adecuado para un análisis exploratorio o para conectar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



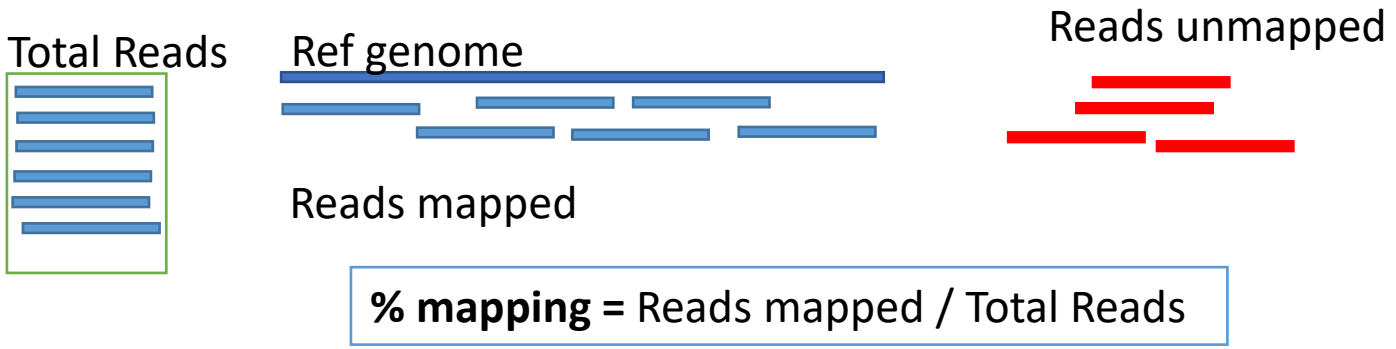
Mapping statistics

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

Mapping quality control

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

Picard
Samtools



% mapping = Reads mapped / Total Reads

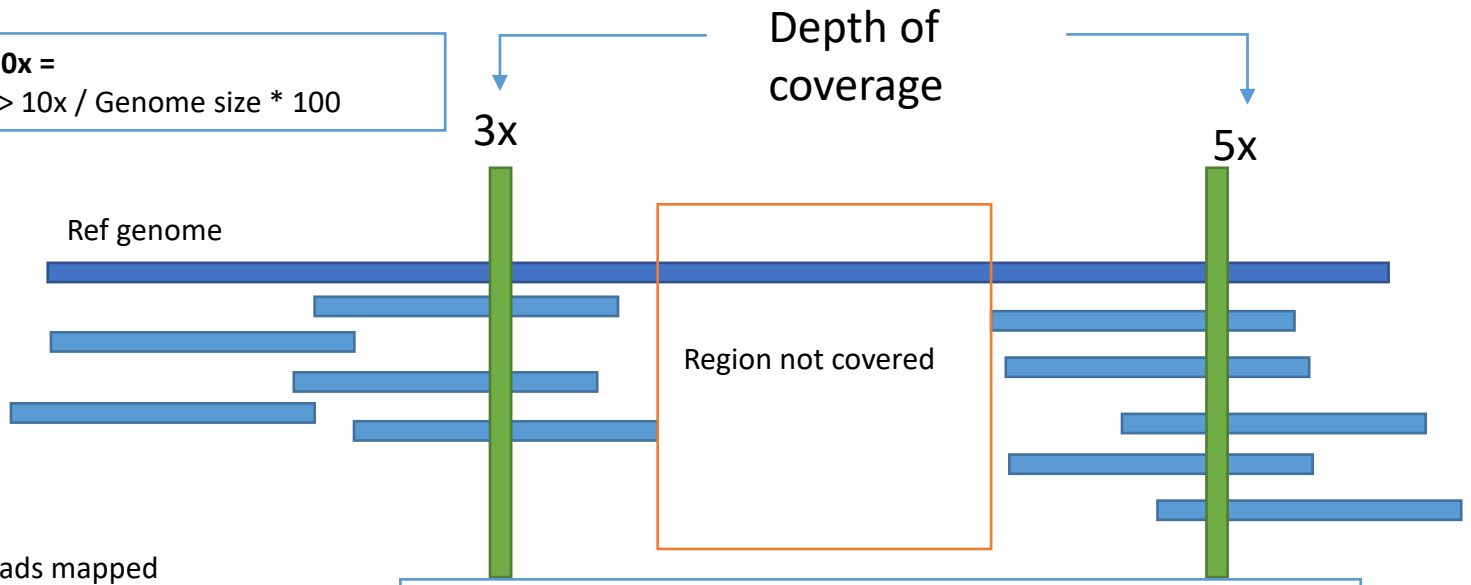
Mandatory parameter for microbial genomics!! It indicates us how many reads we have from our organism of interest. In human genomics this is almost always 99.99% unless something terrible happens. Not here!!!

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

% genome > 10x =
Genome size > 10x / Genome size * 100



Mean Depth of coverage =
Depth of coverage in position 1, 2, 3 / Genome size

¿Preguntas?