Análisis transcriptómicos de la expresión génica

Sesión 7

Máster Universitario en Bioinformática

Universidad Internacional de Valencia

Dra. Paula Soler Vila
paula.solerv@professor.universidadviu.com

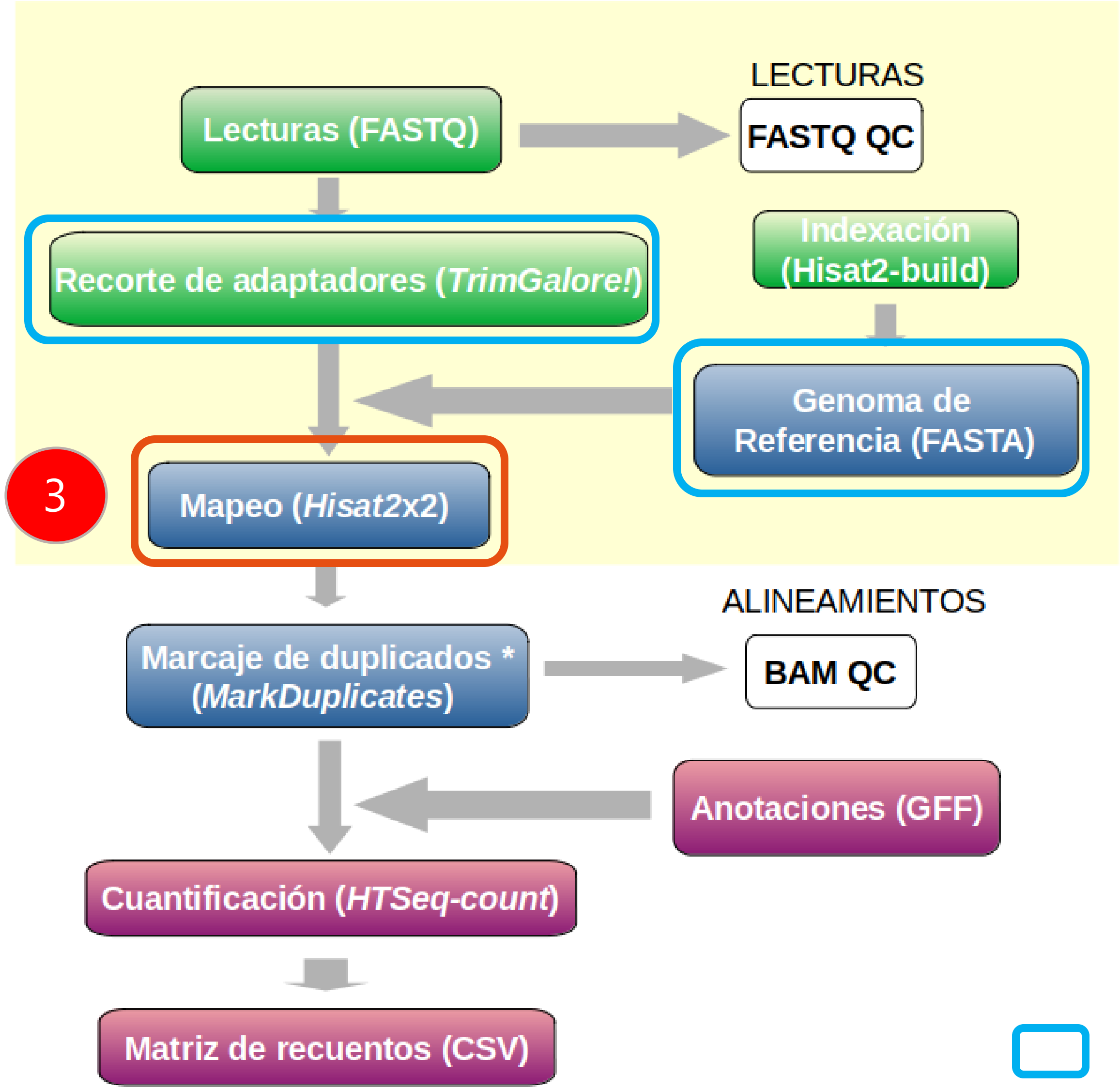
De:
Planeta Formación y Universidades

Bloque III: Análisis de datos de NGS

Objetivos de la sesión

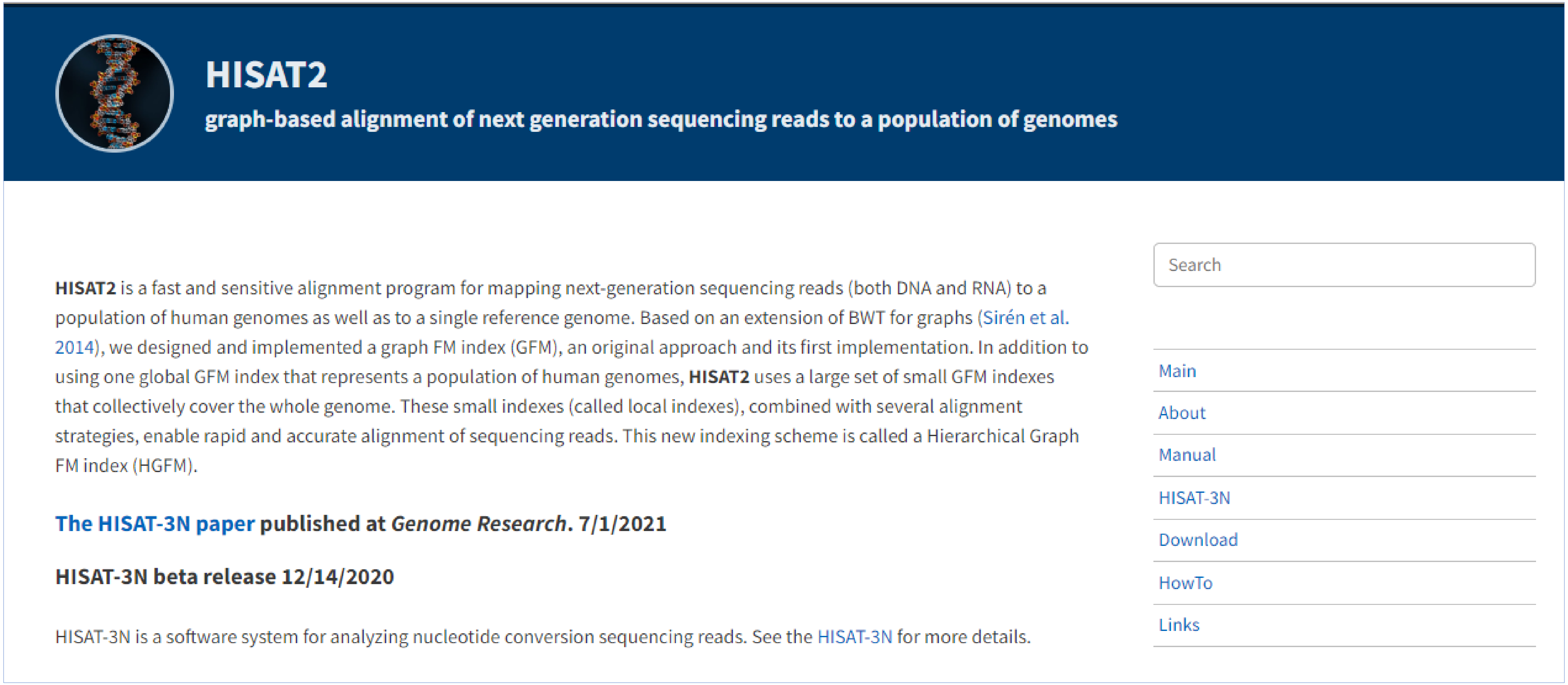
1. Comprender las principales características de los **alineadores** y ejecutar el proceso de alineamiento.
2. Analizar y comprender la información contenida en los archivos de alineamiento en formato **SAM/BAM**.
3. Analizar y comprender la información contenida en los archivos de anotaciones en formato **GFF/GTF**.
4. Realizar la **cuantificación** de los alineamientos utilizando un archivo de anotaciones como referencia.

Flujo de trabajo del análisis de datos de RNA-seq (NGS)


LECTURAS
Lecturas (FASTQ)
FASTQ QC
Indexación (Hisat2-build)
Recorte de adaptadores (*TrimGalore!*)
Genoma de Referencia (FASTA)
3
Mapeo (*Hisat2x2*)
ALINEAMIENTOS
Marcaje de duplicados * (*MarkDuplicates*)
BAM QC
Anotaciones (GFF)
Cuantificación (*HTSeq-count*)
Matriz de recuentos (CSV)

▭ Datos de entrada para el alineamiento

HISAT2 (Hierarchical Indexing for Spliced Alignments of Transcripts)



HISAT2

graph-based alignment of next generation sequencing reads to a population of genomes

HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes as well as to a single reference genome. Based on an extension of BWT for graphs (Sirén et al. 2014), we designed and implemented a graph FM index (GFM), an original approach and its first implementation. In addition to using one global GFM index that represents a population of human genomes, **HISAT2** uses a large set of small GFM indexes that collectively cover the whole genome. These small indexes (called local indexes), combined with several alignment strategies, enable rapid and accurate alignment of sequencing reads. This new indexing scheme is called a Hierarchical Graph FM index (HGFM).

The HISAT-3N paper published at *Genome Research*. 7/1/2021

HISAT-3N beta release 12/14/2020

HISAT-3N is a software system for analyzing nucleotide conversion sequencing reads. See the HISAT-3N for more details.

http://daehwankimlab.github.io/hisat2/

HISAT2 -> Opciones mínimas a considerar

HISAT2

graph-based alignment of next generation sequencing reads to a population of genomes

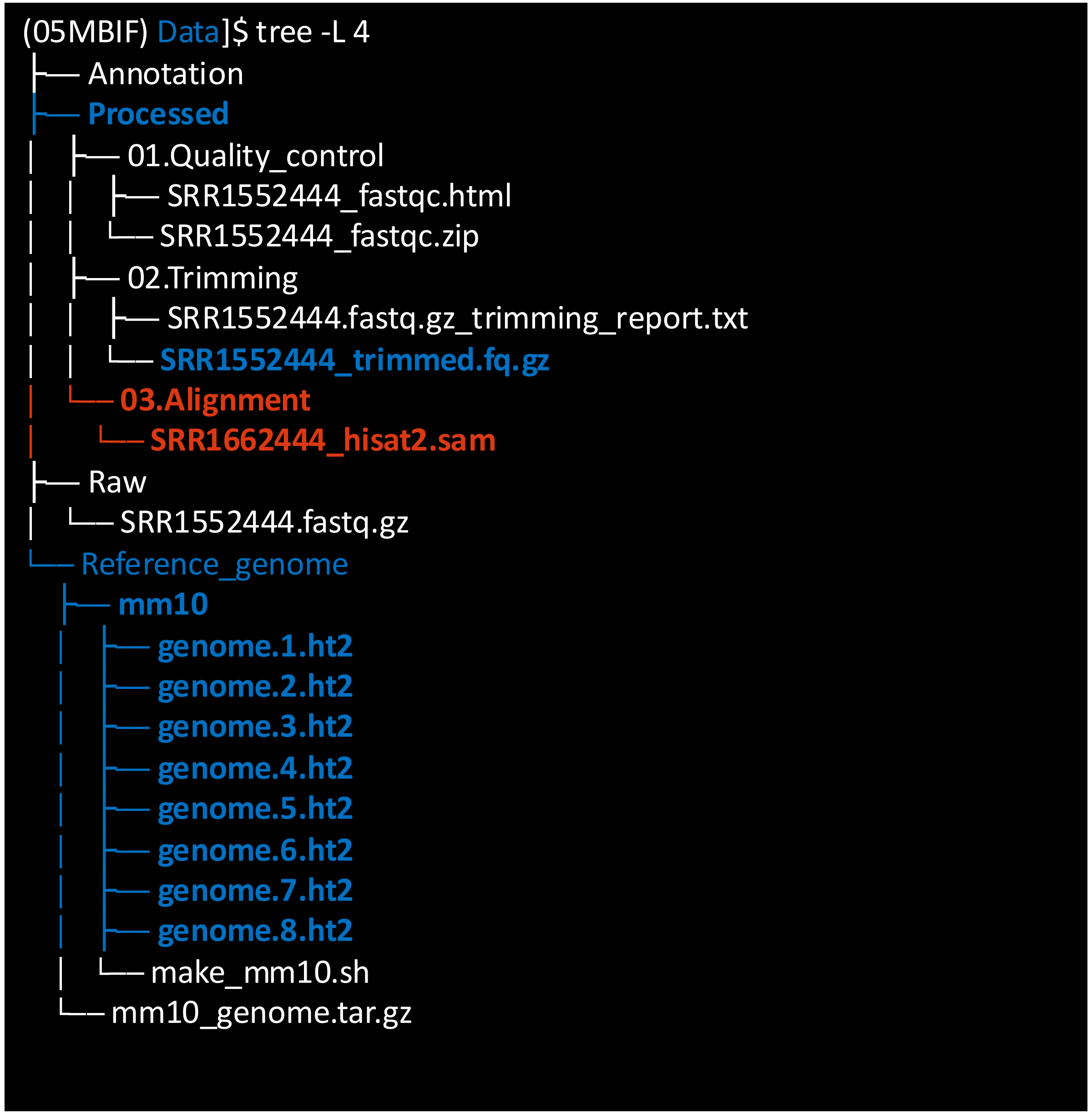
- **-x :** prefijo del índice del genoma de referencia [***genome***]

- **-U :** lista de lecturas para ser alineadas [***trimmed***]

- **-S :** archivo de salida en formato SAM

- **-k** : define el número máximo de alineamientos por lectura.

PRACTIQUEMOS



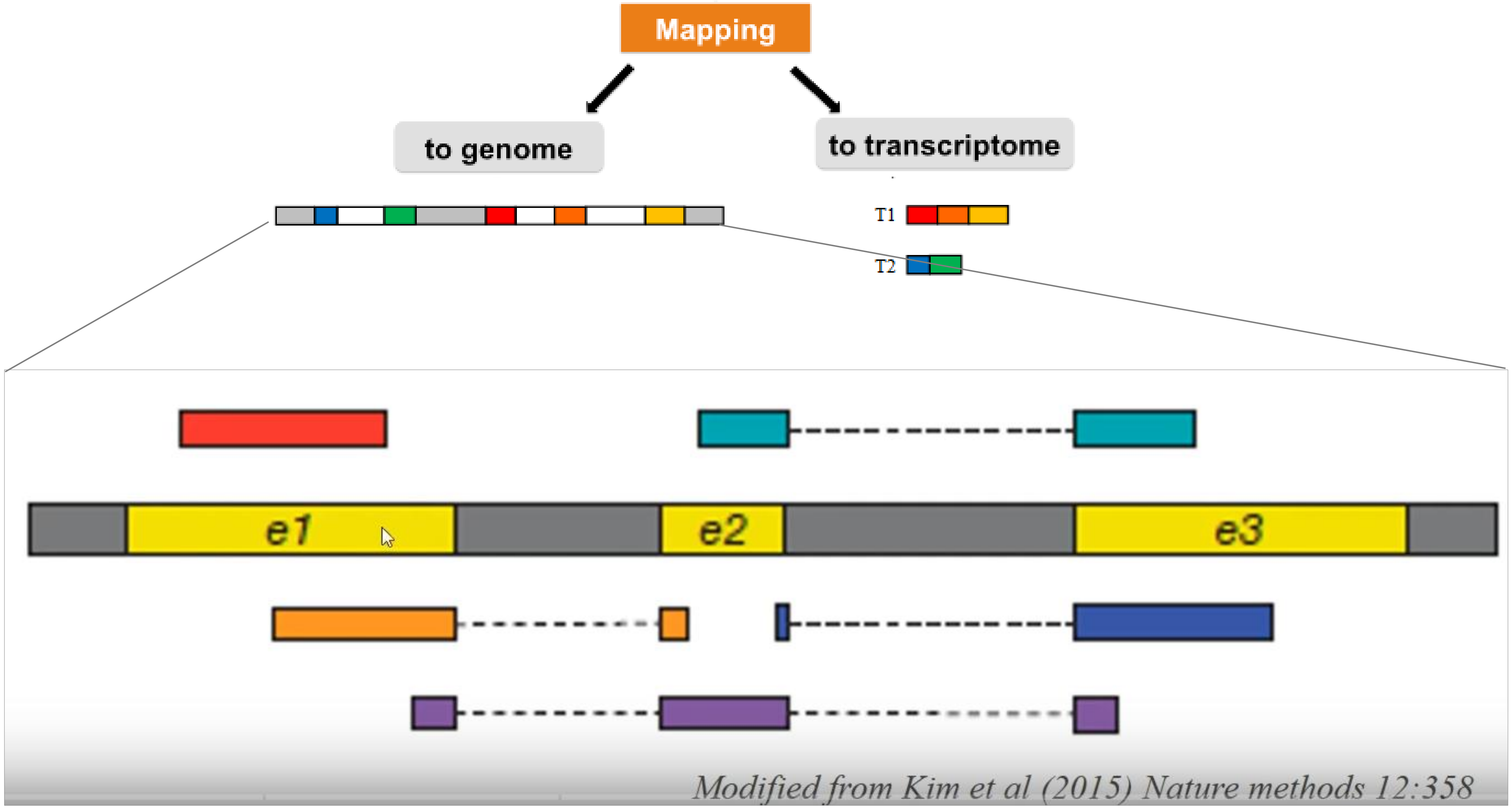
```
(05MBIF) Data]$ tree -L 4
├── Annotation
├── Processed
│   ├── 01.Quality_control
│   │   ├── SRR1552444_fastqc.html
│   │   └── SRR1552444_fastqc.zip
│   ├── 02.Trimming
│   │   ├── SRR1552444.fastq.gz_trimming_report.txt
│   │   └── SRR1552444_trimmed.fq.gz
│   └── 03.Alignment
│       └── SRR1662444_hisat2.sam
├── Raw
│   └── SRR1552444.fastq.gz
└── Reference_genome
    ├── mm10
    │   ├── genome.1.ht2
    │   ├── genome.2.ht2
    │   ├── genome.3.ht2
    │   ├── genome.4.ht2
    │   ├── genome.5.ht2
    │   ├── genome.6.ht2
    │   ├── genome.7.ht2
    │   ├── genome.8.ht2
    │   └── make_mm10.sh
    └── mm10_genome.tar.gz
```

HISAT2 -> Realizando el alineamiento

```
(05MBIF) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr 03.Alignment]$

hisat2  -k1  -U  ../02.Trimming/SRR1552444_trimmed.fq.gz  -x  ../../Reference_genome/mm10/
genome  -S SRR1552444_hisat2.sam
```

- **-x :** prefijo del índice del genoma de referencia [***genome***]
- **-U :** lista de lecturas para ser alineadas [***trimmed***]
- **-S :** archivo de salida en formato SAM
- **-k** : define el número máximo de alineamientos por lectura.

Flujo de trabajo del análisis de datos de RNA-seq (NGS)



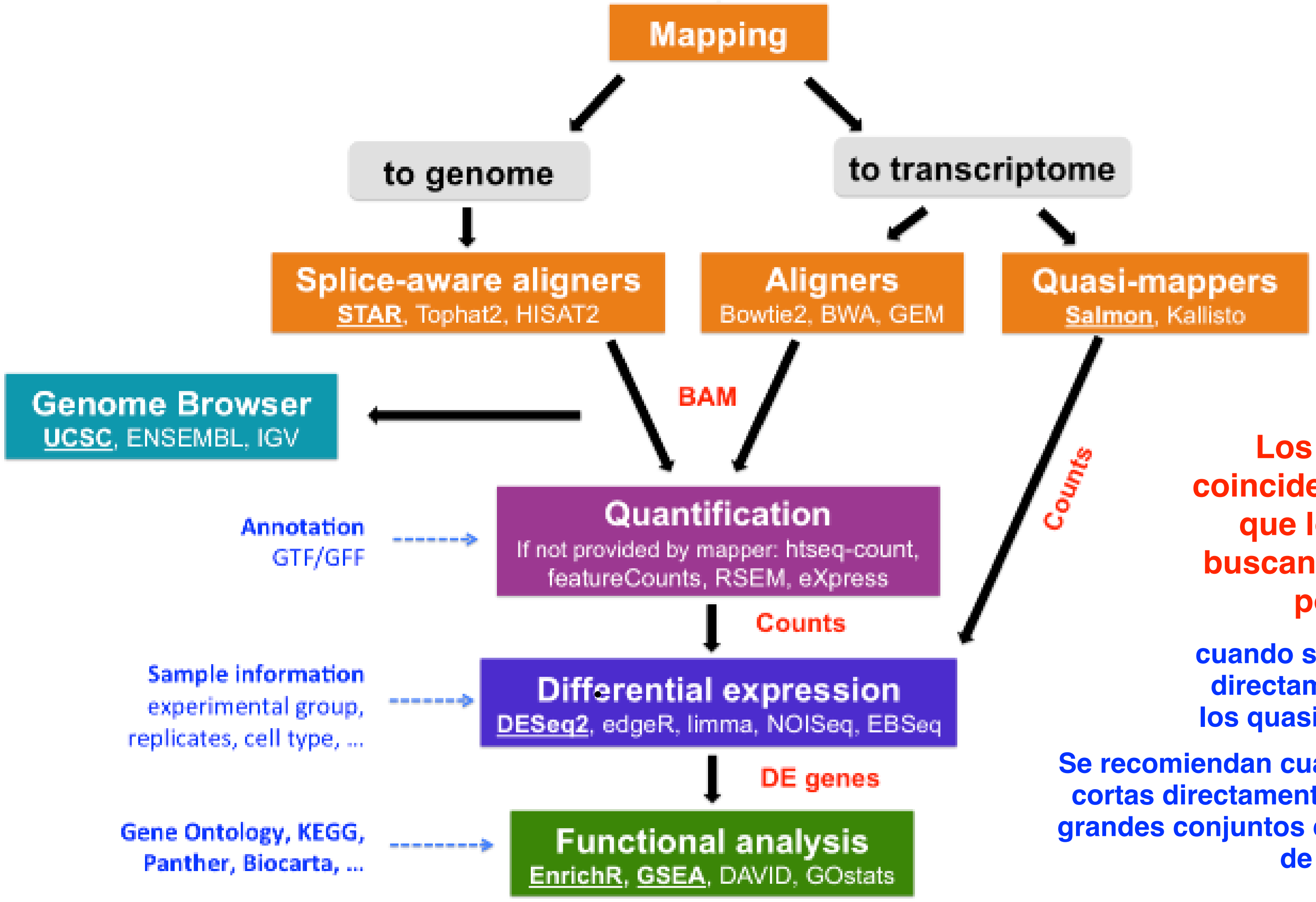
Mapping

to genome

to transcriptome

Modified from Kim et al (2015) Nature methods 12:358

Flujo de trabajo del análisis de datos de RNA-seq (NGS)



Los alineadores buscan coincidencias exactas, mientras que los quasi-mappers no buscan cosas tan exactas, van por probabilidades

cuando se quiere hacer una alineación directamente sobre el transcriptoma los quasi-mappers son muy utilizados

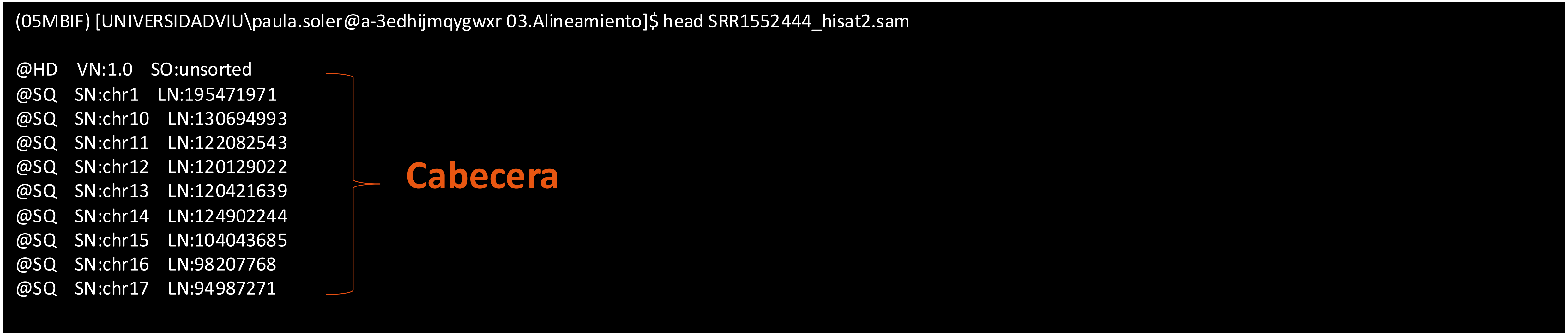
Se recomiendan cuando se trabajan con secuencias cortas directamente sobre un transcriptoma o con grandes conjuntos de RNA si se tiene una limitación de procesamiento

HISAT2 -> Realizando el alineamiento -> Resultado

```
27906762 reads; of these:

  27906762 (100.00%) were unpaired; of these:

    817651 (2.93%) aligned 0 times

    27089111 (97.07%) aligned exactly 1 time

    0 (0.00%) aligned >1 times

97.07% overall alignment rate

-rw-r--r-- 1 UNIVERSIDADVIU\paula.soler  users 8.3G Jul 27 11:02 SRR1552444_hisat2.sam
```

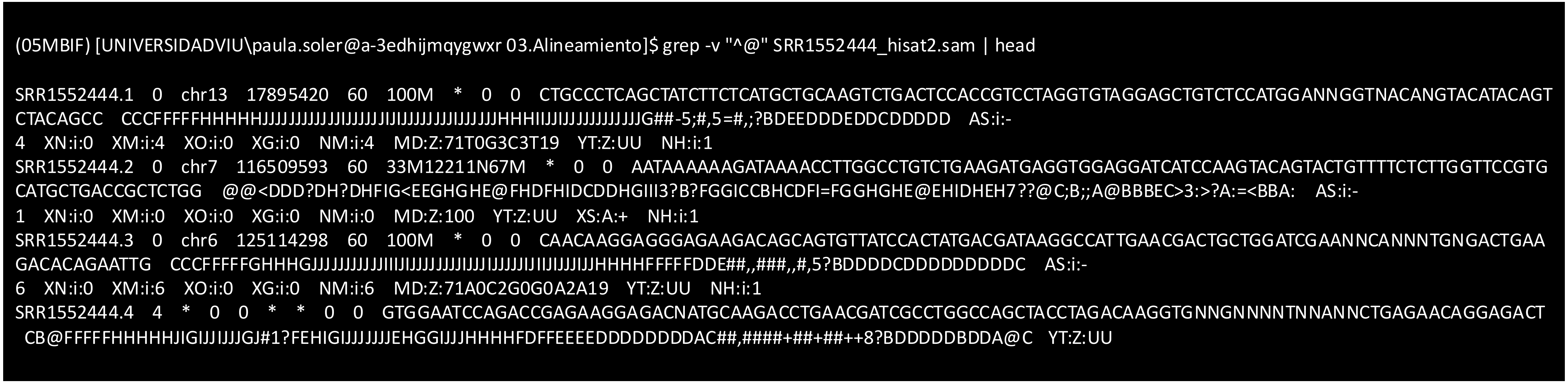
El porcentaje debería ser mayor del 80%

Archivo de alineamientos SAM



```
(05MBIF) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr 03.Alineamiento]$ head SRR1552444_hisat2.sam

@HD   VN:1.0   SO:unsorted
@SQ   SN:chr1   LN:195471971
@SQ   SN:chr10   LN:130694993
@SQ   SN:chr11   LN:122082543
@SQ   SN:chr12   LN:120129022
@SQ   SN:chr13   LN:120421639
@SQ   SN:chr14   LN:124902244
@SQ   SN:chr15   LN:104043685
@SQ   SN:chr16   LN:98207768
@SQ   SN:chr17   LN:94987271
```

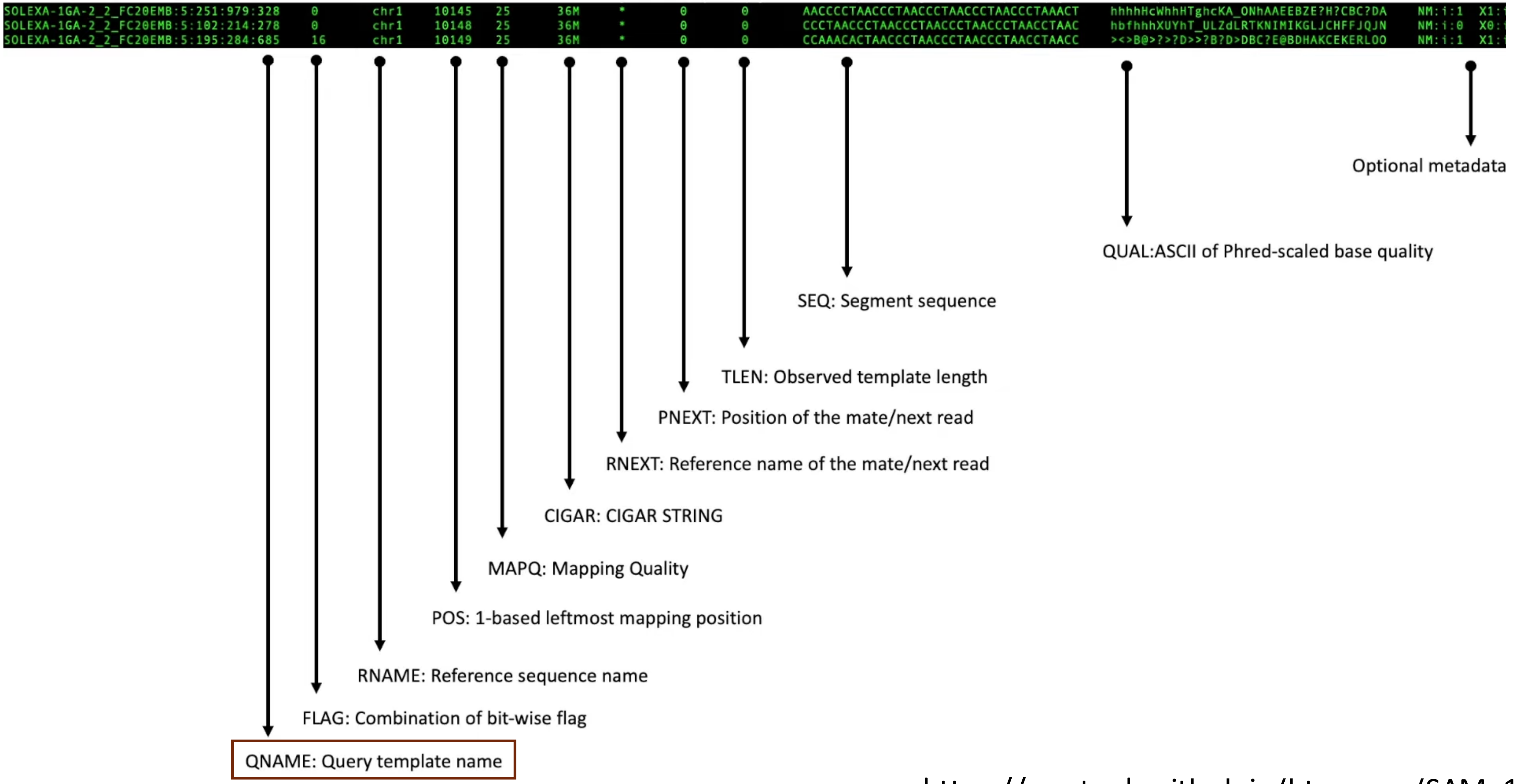
Cabecera



```
(05MBIF) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr 03.Alineamiento]$ grep -v "^@" SRR1552444_hisat2.sam | head

SRR1552444.1   0   chr13   17895420   60   100M   *   0   0   CTGCCCTCAGCTATCTTCTCATGCTGCAAGTCTGACTCCACCGTCCTAGGTGTAGGAGCTGTCTCCATGGANNGGTNACANGTACATACAGT
CTACAGCC   CCCFFFFFHHHHHJJJJJJJJJJJJJIJJJJJJIJIJJJJJJJJJIJJJJJJJHHHIIJJIJJJJJJJJJJJJG##-5;#,5=#,;?BDEEDDDEDDCDDDDD   AS:i:-
4   XN:i:0   XM:i:4   XO:i:0   XG:i:0   NM:i:4   MD:Z:71T0G3C3T19   YT:Z:UU   NH:i:1
SRR1552444.2   0   chr7   116509593   60   33M12211N67M   *   0   0   AATAAAAAAGATAAAACCTTGGCCTGTCTGAAGATGAGGTGGAGGATCATCCAAGTACAGTACTGTTTTCTCTTGGTTCCGTG
CATGCTGACCGCTCTGG   @@<DDD?DH?DHFIG<EEGHGHE@FHDFHIDCDDHGIII3?B?FGGICCBHCDFI=FGGHGHE@EHIDHEH7??@C;B;;A@BBBEC>3:>?A:=<BBA:   AS:i:-
1   XN:i:0   XM:i:0   XO:i:0   XG:i:0   NM:i:0   MD:Z:100   YT:Z:UU   XS:A:+   NH:i:1
SRR1552444.3   0   chr6   125114298   60   100M   *   0   0   CAACAAGGAGGGAGAAGACAGCAGTGTTATCCACTATGACGATAAGGCCATTGAACGACTGCTGGATCGAANNCANNNTGNGACTGAA
GACACAGAATTG   CCCFFFFFGHHHGJJJJJJJJJJJIIIJIJJJJJJJJJIJJJIJJJJJJIJIIJIJJJJIJJHHHHFFFFFDDE##,,###,,#,5?BDDDDCDDDDDDDDDC   AS:i:-
6   XN:i:0   XM:i:6   XO:i:0   XG:i:0   NM:i:6   MD:Z:71A0C2G0G0A2A19   YT:Z:UU   NH:i:1
SRR1552444.4   4   *   0   0   *   *   0   0   GTGGAATCCAGACCGAGAAGGAGACNATGCAAGACCTGAACGATCGCCTGGCCAGCTACCTAGACAAGGTGNNGNNNNTNNANNCTGAGAACAGGAGACT
 CB@FFFFFHHHHHJIGIJJIJJJGJ#1?FEHIGIJJJJJJJEHGGIJJJHHHHHFDFFEEEEDDDDDDDDAC##,####+##+##++8?BDDDDDBDDA@C   YT:Z:UU
```

Alineamiento

Archivo de alineamientos SAM (11 + 1)



```
SOLEXA-1GA-2_2_FC20EMB:5:251:979:328   0    chr1   10145   25   36M   *   0   0   AACCCCTAACCCTAACCCTAACCCTAACCCTAAACT   hhhhHcWhhHTghcKA_ONhAAEEBZE?H?CBC?DA   NM:i:1   X1:
SOLEXA-1GA-2_2_FC20EMB:5:102:214:278   0    chr1   10148   25   36M   *   0   0   CCCTAACCCTAACCCTAACCCTAACCCTAACCTAAC   hbfhhhXUYhT_ULZdLRTKNIMIKGLJCHFFJQJN   NM:i:0   X0:
SOLEXA-1GA-2_2_FC20EMB:5:195:284:685   16   chr1   10149   25   36M   *   0   0   CCAAACACTAACCCTAACCCTAACCCTAACCTAACC   ><>B@>?>?D>>?B?D>DBC?E@BDHAKCEKERLOO   NM:i:1   X1:
```

Optional metadata

QUAL:ASCII of Phred-scaled base quality

SEQ: Segment sequence

TLEN: Observed template length

PNEXT: Position of the mate/next read

RNEXT: Reference name of the mate/next read

CIGAR: CIGAR STRING

MAPQ: Mapping Quality

POS: 1-based leftmost mapping position

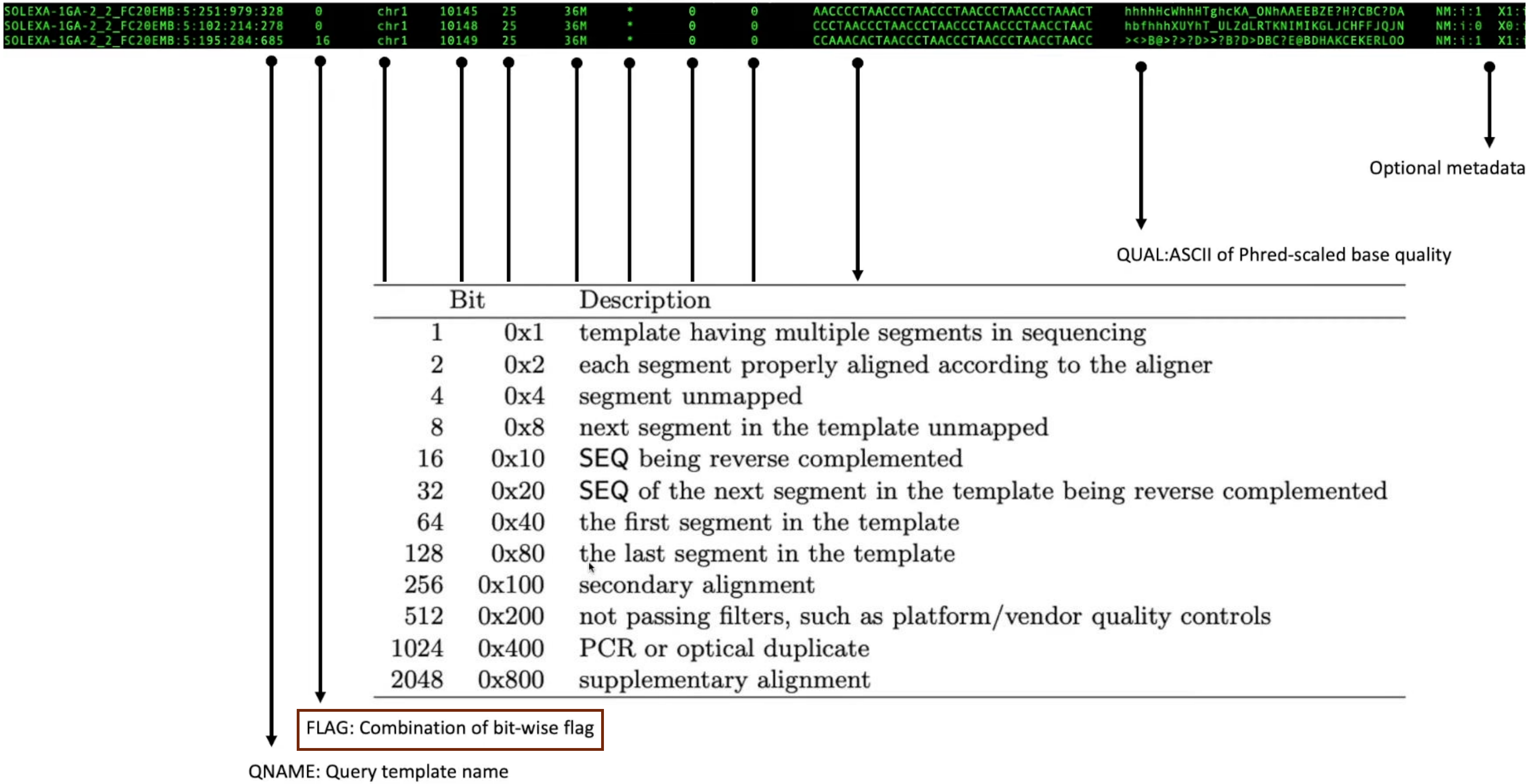
RNAME: Reference sequence name

FLAG: Combination of bit-wise flag

QNAME: Query template name

https://samtools.github.io/hts-specs/SAMv1.pdf

Archivo de alineamientos SAM



```
SOLEXA-1GA-2_2_FC20EMB:5:251:979:328   0    chr1   10145   25   36M   *   0   0   AACCCCTAACCCTAACCCTAACCCTAACCCTAAACT   hhhhHcWhhHTghcKA_ONhAAEEBZE?H?CBC?DA   NM:i:1  X1:
SOLEXA-1GA-2_2_FC20EMB:5:102:214:278   0    chr1   10148   25   36M   *   0   0   CCCTAACCCTAACCCTAACCCTAACCCTAACCTAAC   hbfhhhXUYhT_ULZdLRTKNIMIKGLJCHFFJQJN   NM:i:0  X0:
SOLEXA-1GA-2_2_FC20EMB:5:195:284:685   16   chr1   10149   25   36M   *   0   0   CCAAACACTAACCCTAACCCTAACCCTAACCTAACC   ><>B@>?>?D>>?B?D>DBC?E@BDHAKCEKERLOO   NM:i:1  X1:
```

Optional metadata

QUAL:ASCII of Phred-scaled base quality

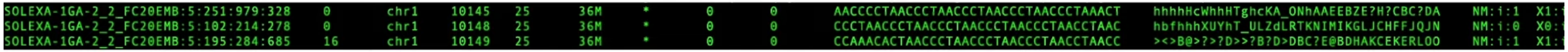
Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

FLAG: Combination of bit-wise flag

QNAME: Query template name

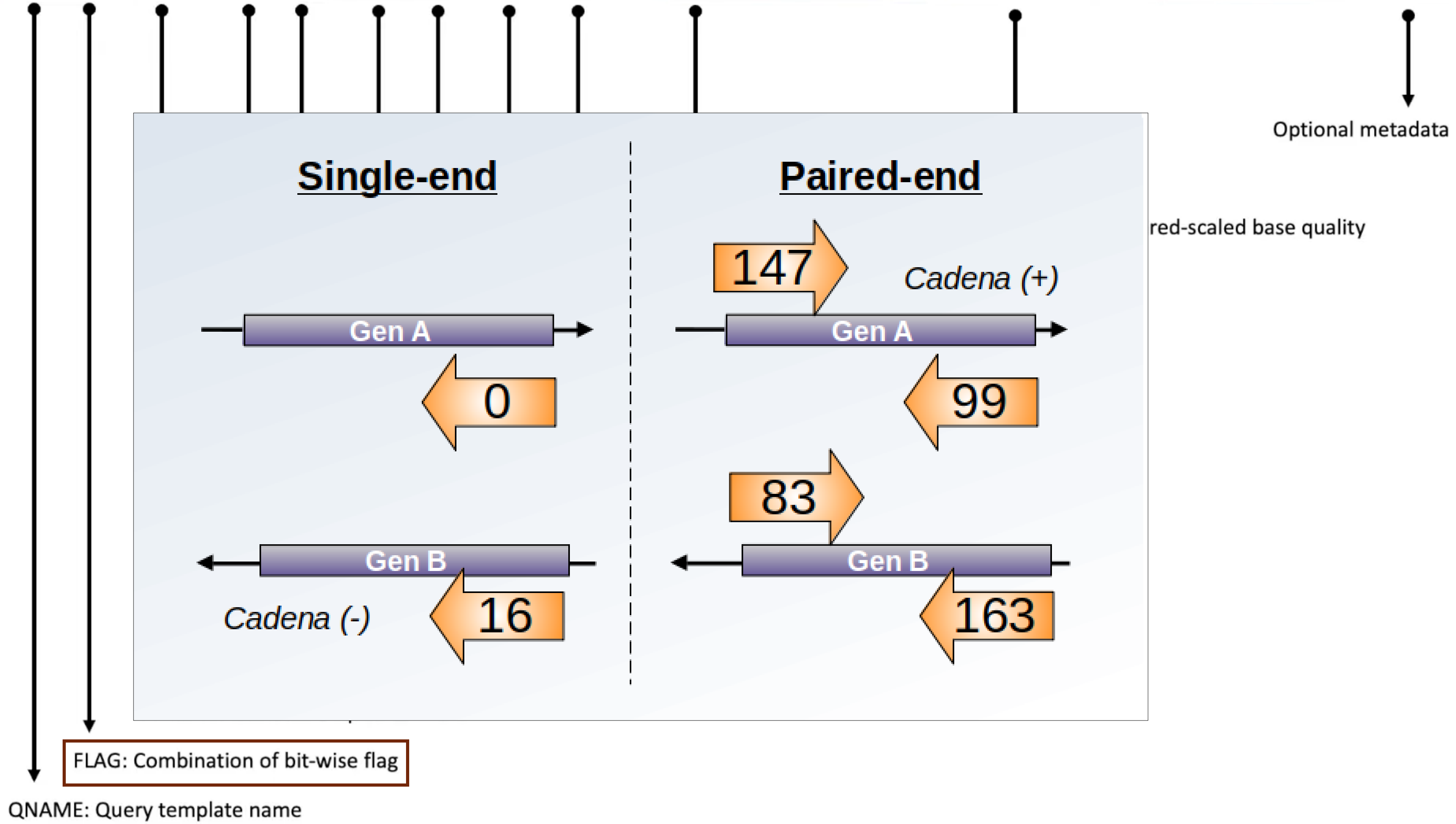
https://samtools.github.io/hts-specs/SAMv1.pdf

Archivo de alineamientos SAM



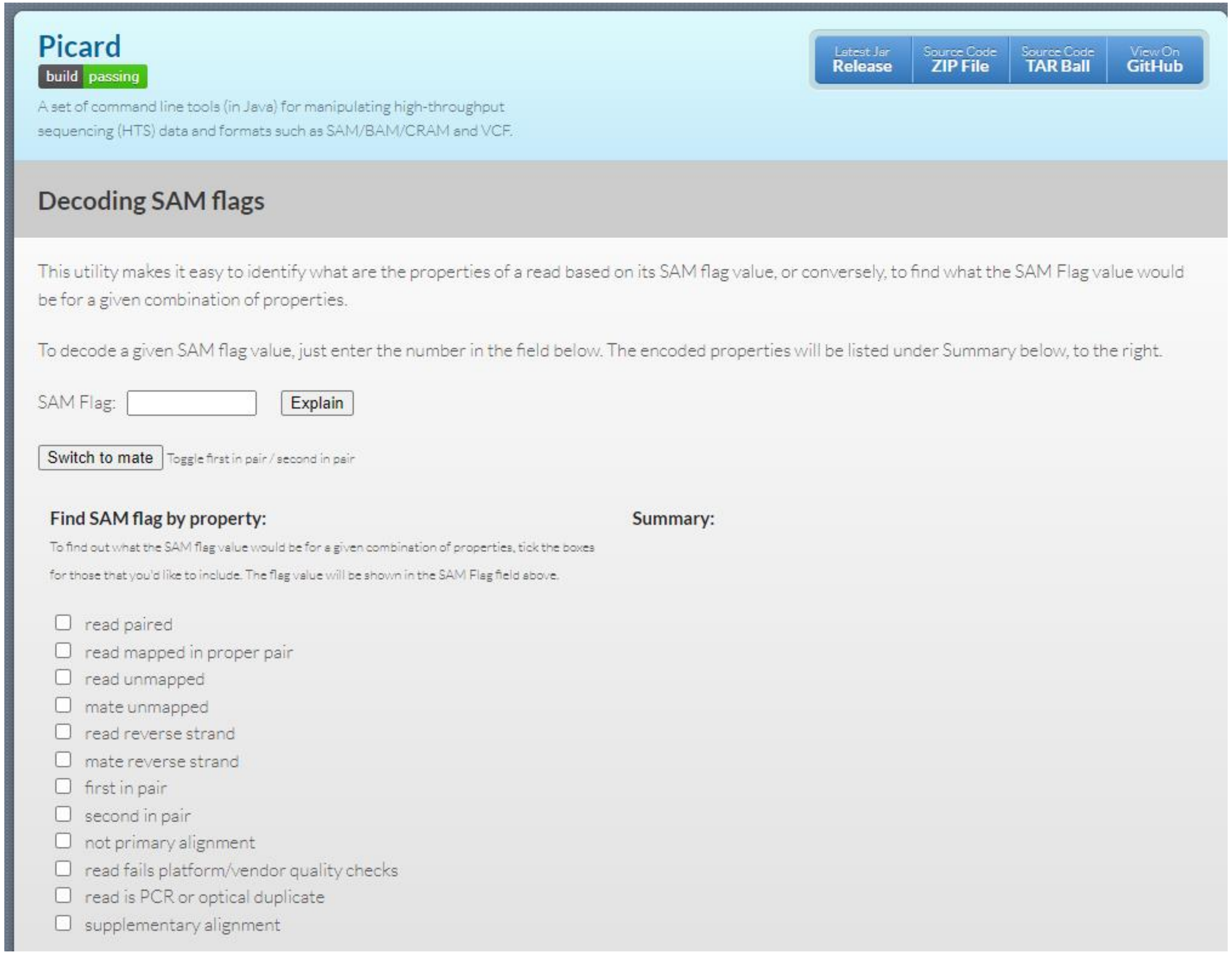
```
SOLEXA-1GA-2_2_FC20EMB:5:251:979:328   0    chr1   10145   25    36M    *    0    0    AACCCCTAACCCTAACCCTAACCCTAACCCTAAACT    hhhhHcWhhHTghcKA_ONhAAEEBZE?H?CBC?DA    NM:i:1   X1:
SOLEXA-1GA-2_2_FC20EMB:5:102:214:278   0    chr1   10148   25    36M    *    0    0    CCCTAACCCTAACCCTAACCCTAACCCTAACCTAAC    hbfhhhXUYhT_ULZdLRTKNIMIKGLJCHFFJQJN    NM:i:0   X0:
SOLEXA-1GA-2_2_FC20EMB:5:195:284:685   16   chr1   10149   25    36M    *    0    0    CCAAACACTAACCCTAACCCTAACCCTAACCTAACC    ><>B@>?>?D>>?B?D>DBC?E@BDHAKCEKERLOO    NM:i:1   X1:
```

Una flag de 0 estaría relacionada con que la lectura se mapea en la cadena +, mientras que la flag de 16 indica que se está mapeando en al reverse



https://samtools.github.io/hts-specs/SAMv1.pdf

Archivo de alineamientos SAM



Picard

A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.

Decoding SAM flags

This utility makes it easy to identify what are the properties of a read based on its SAM flag value, or conversely, to find what the SAM Flag value would be for a given combination of properties.

To decode a given SAM flag value, just enter the number in the field below. The encoded properties will be listed under Summary below, to the right.

SAM Flag: Explain

Switch to mate Toggle first in pair / second in pair

Find SAM flag by property:

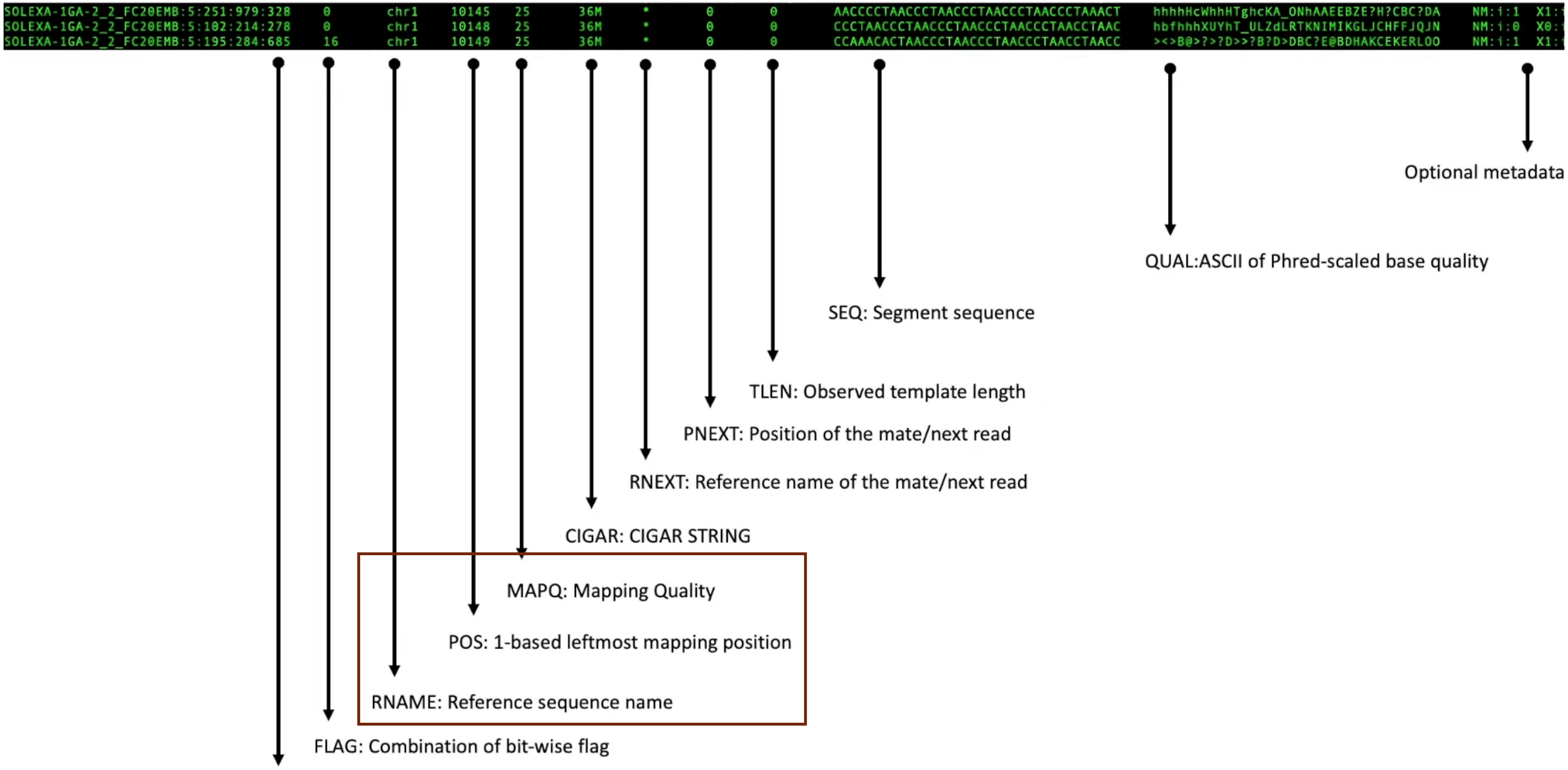
To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- □ 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:

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

Archivo de alineamientos SAM



```
SOLEXA-1GA-2_2_FC20EMB:5:251:979:328   0    chr1   10145  25   36M   *   0   0   AACCCCTAACCCTAACCCTAACCCTAACCCTAAACT   hhhhHcWhhhHTghcKA_ONhAAEEBZE?H?CBC?DA   NM:i:1  X1:
SOLEXA-1GA-2_2_FC20EMB:5:102:214:278   0    chr1   10148  25   36M   *   0   0   CCCTAACCCTAACCCTAACCCTAACCCTAACCTAAC   hbfhhhXUYhT_ULZdLRTKNIMIKGLJCHFFJQJN   NM:i:0  X0:
SOLEXA-1GA-2_2_FC20EMB:5:195:284:685   16   chr1   10149  25   36M   *   0   0   CCAAACACTAACCCTAACCCTAACCCTAACCTAACC   ><>B@>?>?D>>?B?D>DBC?E@BDHAKCEKERLOO   NM:i:1  X1:
```

Optional metadata

QUAL:ASCII of Phred-scaled base quality

SEQ: Segment sequence

TLEN: Observed template length

PNEXT: Position of the mate/next read

RNEXT: Reference name of the mate/next read

CIGAR: CIGAR STRING

MAPQ: Mapping Quality

POS: 1-based leftmost mapping position

RNAME: Reference sequence name

FLAG: Combination of bit-wise flag

QNAME: Query template name

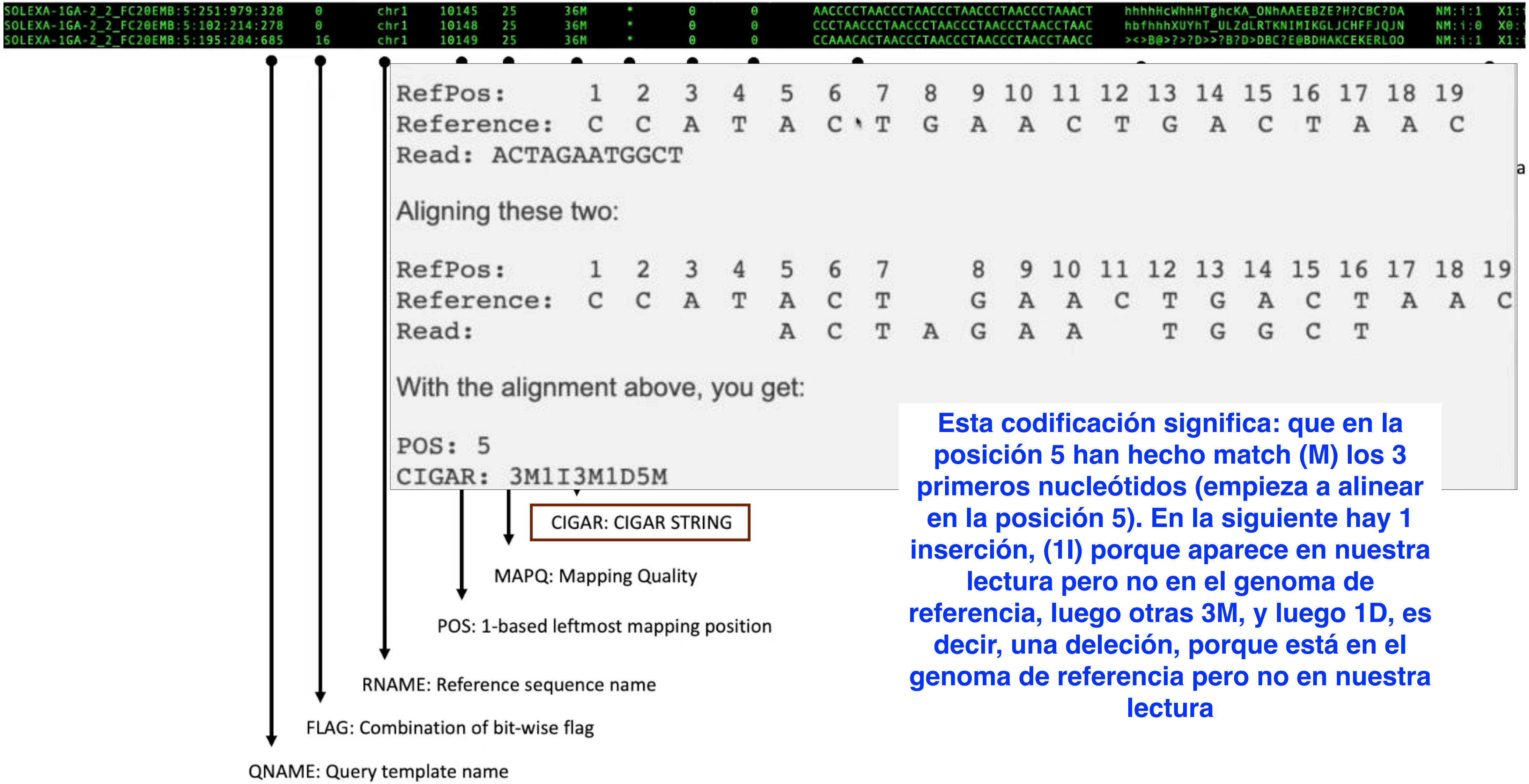
https://samtools.github.io/hts-specs/SAMv1.pdf

Calidad de mapeo (MAPQ)

¿Qué valores podemos diferenciar de MAPQ en nuestro archivo de SRR1552444_hisat2.sam ?

```
grep -v "^@" SRR1552444_hisat2.sam | cut -f5 | sort -u


0 -> No se ha podido realizar una alineación única para esa lectura.


1 -> Alineamiento único de baja calidad.


60 -> Alineamiento que mapea de forma única.
```

Archivo de alineamientos SAM



```
SOLEXA-1GA-2_2_FC20EMB:5:251:979:328   0    chr1   10145   25   36M   *   0   0   AACCCCTAACCCTAACCCTAACCCTAACCCTAAACT   hhhhHcWhhHTghcKA_ONhAAEEBZE?H?CBC?DA   NM:i:1   X1:
SOLEXA-1GA-2_2_FC20EMB:5:102:214:278   0    chr1   10148   25   36M   *   0   0   CCCTAACCCTAACCCTAACCCTAACCCTAACCTAAC   hbfhhhXUYhT_ULZdLRTKNIMIKGLJCHFFJQJN   NM:i:0   X0:
SOLEXA-1GA-2_2_FC20EMB:5:195:284:685   16   chr1   10149   25   36M   *   0   0   CCAAACACTAACCCTAACCCTAACCCTAACCTAACC   ><>B@>?>?D>>?B?D>DBC?E@BDHAKCEKERLOO   NM:i:1   X1:
```

```
RefPos:     1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T  G  A  A  C  T  G  A  C  T  A  A  C
Read: ACTAGAATGGCT
```

Aligning these two:

```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                   A  C  T  A  G  A  A     T  G  G  C  T
```

With the alignment above, you get:

```
POS: 5
CIGAR: 3M1I3M1D5M
```

CIGAR: CIGAR STRING

MAPQ: Mapping Quality

POS: 1-based leftmost mapping position

RNAME: Reference sequence name

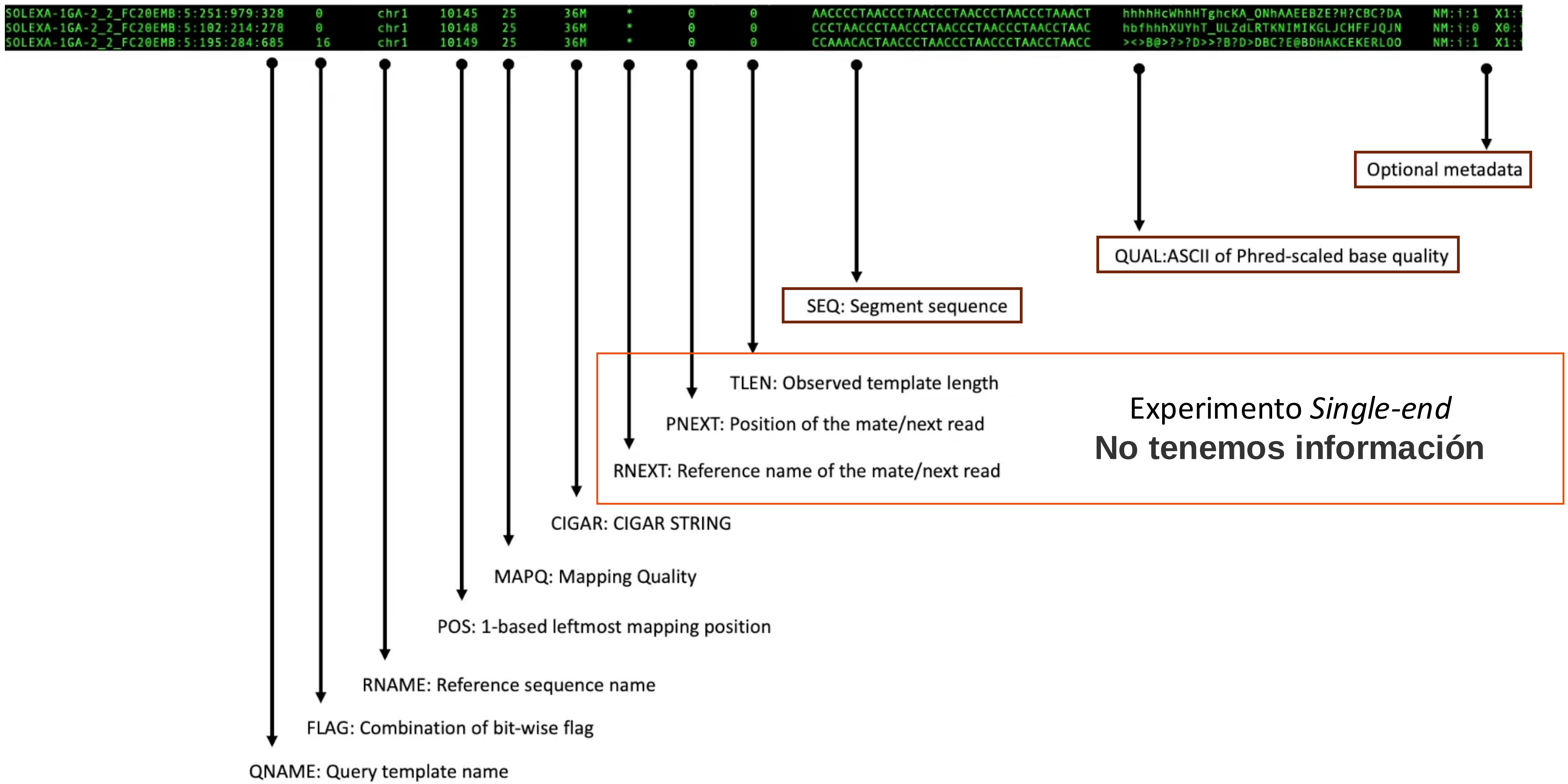
FLAG: Combination of bit-wise flag

QNAME: Query template name

Esta codificación significa: que en la posición 5 han hecho match (M) los 3 primeros nucleótidos (empieza a alinear en la posición 5). En la siguiente hay 1 inserción, (1I) porque aparece en nuestra lectura pero no en el genoma de referencia, luego otras 3M, y luego 1D, es decir, una deleción, porque está en el genoma de referencia pero no en nuestra lectura

https://samtools.github.io/hts-specs/SAMv1.pdf

Archivo de alineamientos SAM



```
SOLEXA-1GA-2_2_FC20EMB:5:251:979:328    0     chr1    10145   25     36M      *      0      0      AACCCCTAACCCTAACCCTAACCCTAACCCTAAACT    hhhhHcWhhhHTghcKA_ONhAAEEBZE?H?CBC?DA    NM:i:1   X1:
SOLEXA-1GA-2_2_FC20EMB:5:102:214:278    0     chr1    10148   25     36M      *      0      0      CCCTAACCCTAACCCTAACCCTAACCCTAACCTAAC    hbfhhhXUYhT_ULZdLRTKNIMIKGLJCHFFJQJN    NM:i:0   X0:
SOLEXA-1GA-2_2_FC20EMB:5:195:284:685    16    chr1    10149   25     36M      *      0      0      CCAAACACTAACCCTAACCCTAACCCTAACCTAACC    ><>B@>?>?D>>?B?D>DBC?E@BDHAKCEKERLOO    NM:i:1   X1:
```

Optional metadata

QUAL:ASCII of Phred-scaled base quality

SEQ: Segment sequence

TLEN: Observed template length

PNEXT: Position of the mate/next read

RNEXT: Reference name of the mate/next read

Experimento *Single-end*

No tenemos información

CIGAR: CIGAR STRING

MAPQ: Mapping Quality

POS: 1-based leftmost mapping position

RNAME: Reference sequence name

FLAG: Combination of bit-wise flag

QNAME: Query template name

https://samtools.github.io/hts-specs/SAMv1.pdf

Compresión del archivo SAM a BAM

```
conda install bioconda::samtools
conda install -c bioconda samtools
```

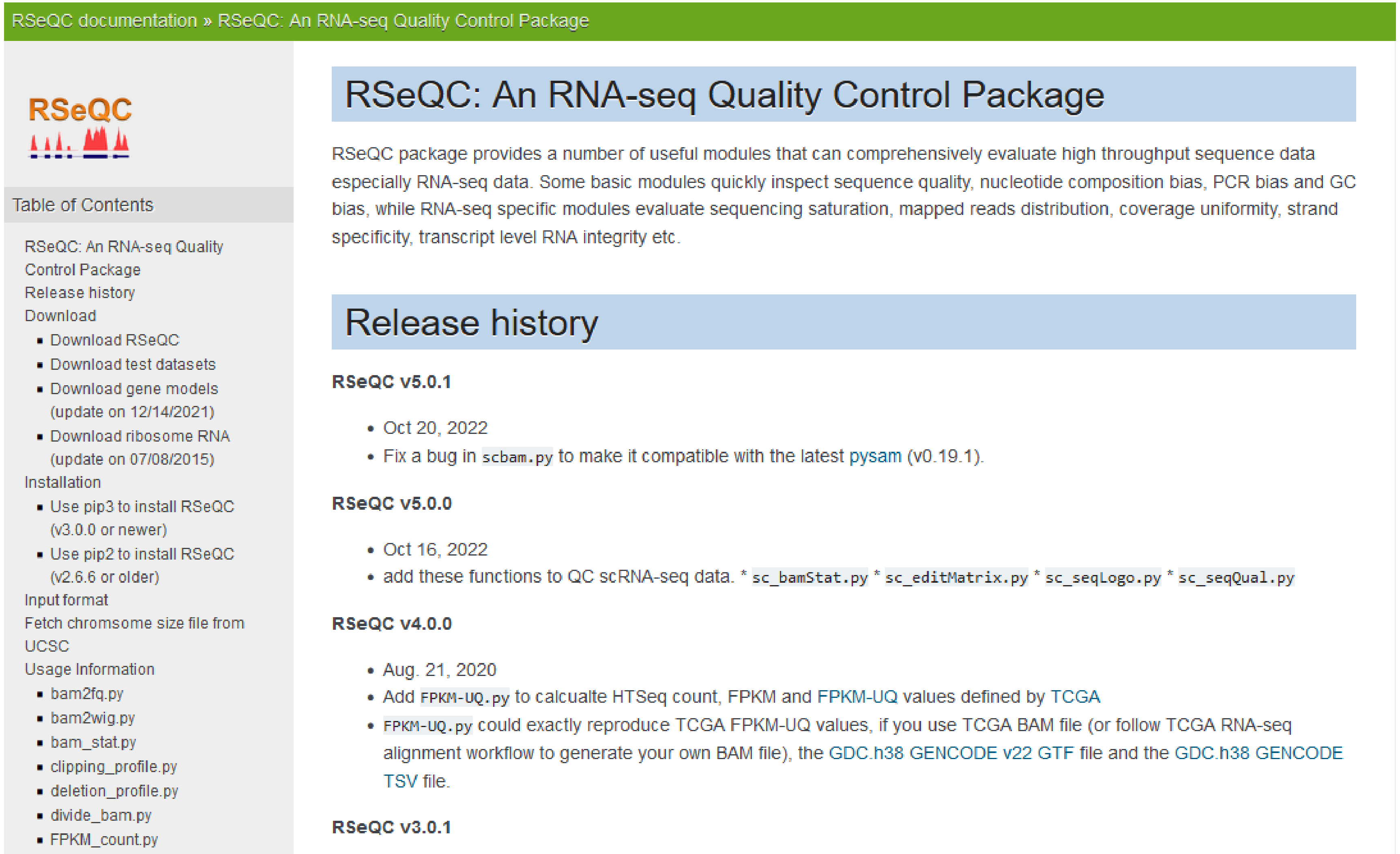
```
samtools view -Sbh SRR1552444_hisat2.sam > SRR1552444_hisat2.bam

samtools sort SRR1552444_hisat2.bam -o SRR1552444_hisat2.sorted.bam

samtools index SRR1552444_hisat2.sorted.bam


-rw-r--r-- 1 UNIVERSIDADVIU\paula.soler  users 2.7G Feb  6 11:14 SRR1552444_hisat2.bam
-rw-r--r-- 1 UNIVERSIDADVIU\paula.soler  users 8.3G Feb  5 20:31 SRR155244_hisat2.sam
-rw-r--r-- 1 UNIVERSIDADVIU\paula.soler  users 1.7G Feb  6 11:50 SRR155244_hisat2.sorted.bam
-rw-r--r-- 1 UNIVERSIDADVIU\paula.soler  users 2.3M Feb  6 13:59 SRR155244_hisat2.sorted.bam.bai
```

Samtools stats

```
$ samtools stats SRR1552444_hisat2.sorted.bam > SRR155244_hisat2.sorted.bam.stats
$ cat SRR155244.sorted.bam.stats
# This file was produced by samtools stats (1.3.1+htslib-1.3.1) and can be plotted using plot-bamstats
# This file contains statistics for all reads.
# The command line was:  stats SRR155244.sorted.bam
# CHK, Checksum    [2]Read Names    [3]Sequences    [4]Qualities
# CHK, CRC32 of reads which passed filtering followed by addition (32bit overflow)
CHK    229871e0    5d4b8724    e32ebaf0
# Summary Numbers. Use `grep ^SN | cut -f 2-` to extract this part.
SN    raw total sequences:    27906762
SN    filtered sequences:    0
SN    sequences:    27906762
SN    is sorted:    1
SN    1st fragments:    27906762
SN    last fragments:    0
SN    reads mapped:    27089111
SN    reads mapped and paired:    0    # paired-end technology bit set + both mates mapped
SN    reads unmapped:    817651
SN    reads properly paired:    0    # proper-pair bit set
SN    reads paired:    0    # paired-end technology bit set
SN    reads duplicated:    0    # PCR or optical duplicate bit set
SN    reads MQ0:    89883    # mapped and MQ=0
SN    reads QC failed:    0
SN    non-primary alignments:    0
```

RSeQC

conda install bioconda::rseqc



RSeQC documentation » RSeQC: An RNA-seq Quality Control Package

Table of Contents

RSeQC: An RNA-seq Quality Control Package

RSeQC package provides a number of useful modules that can comprehensively evaluate high throughput sequence data especially RNA-seq data. Some basic modules quickly inspect sequence quality, nucleotide composition bias, PCR bias and GC bias, while RNA-seq specific modules evaluate sequencing saturation, mapped reads distribution, coverage uniformity, strand specificity, transcript level RNA integrity etc.

Release history

RSeQC v5.0.1

- Oct 20, 2022
- Fix a bug in `scbam.py` to make it compatible with the latest pysam (v0.19.1).

RSeQC v5.0.0

- Oct 16, 2022
- add these functions to QC scRNA-seq data. * `sc_bamStat.py` * `sc_editMatrix.py` * `sc_seqLogo.py` * `sc_seqQual.py`

RSeQC v4.0.0

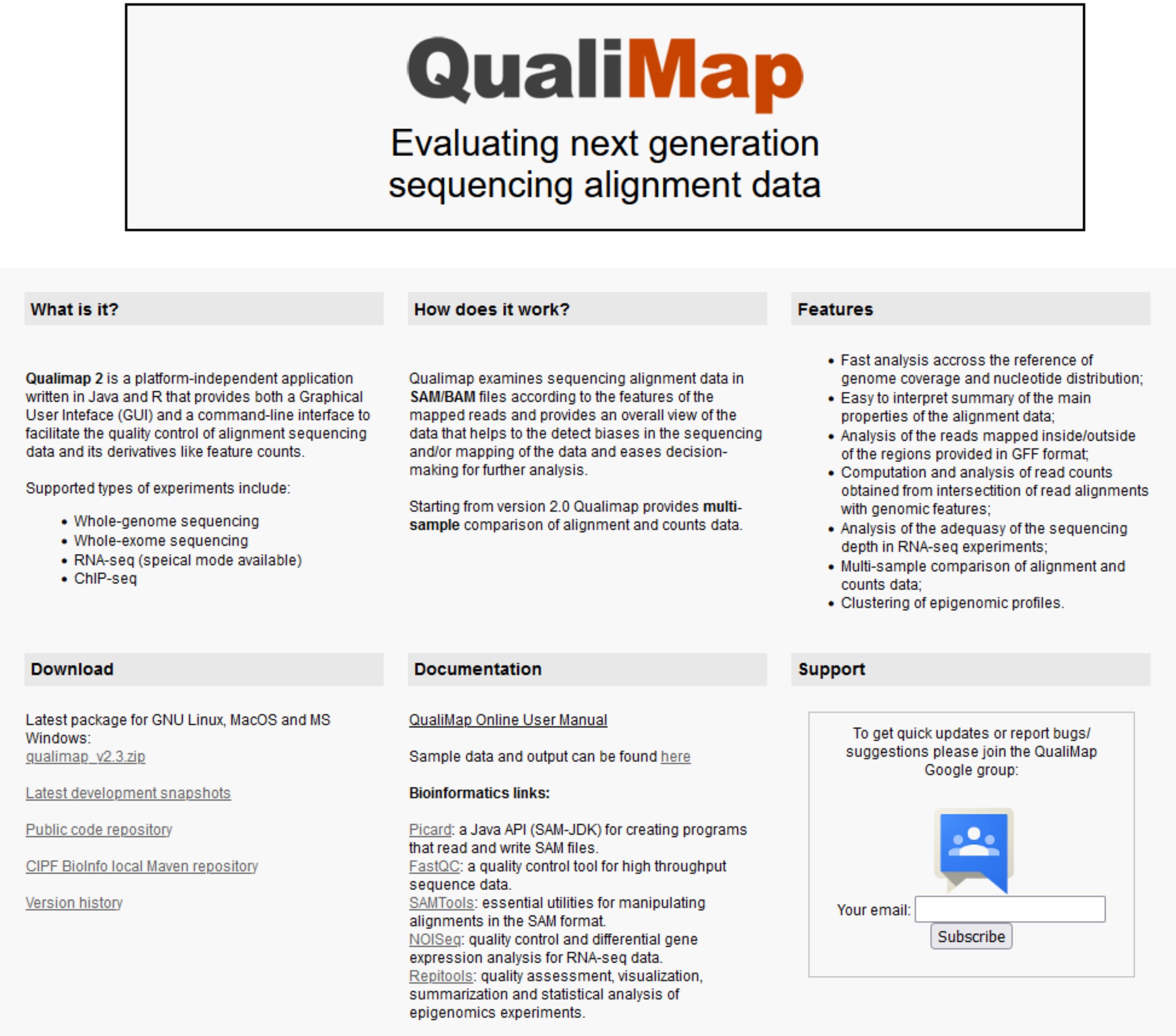
- Aug. 21, 2020
- Add `FPKM-UQ.py` to calcualte HTSeq count, FPKM and FPKM-UQ values defined by TCGA
- `FPKM-UQ.py` could exactly reproduce TCGA FPKM-UQ values, if you use TCGA BAM file (or follow TCGA RNA-seq alignment workflow to generate your own BAM file), the GDC.h38 GENCODE v22 GTF file and the GDC.h38 GENCODE TSV file.

RSeQC v3.0.1

https://rseqc.sourceforge.net/

QualiMap

conda install bioconda::qualimap



QualiMap

Evaluating next generation sequencing alignment data

What is it?

Qualimap 2 is a platform-independent application written in Java and R that provides both a Graphical User Inteface (GUI) and a command-line interface to facilitate the quality control of alignment sequencing data and its derivatives like feature counts.

Supported types of experiments include:

- Whole-genome sequencing
- Whole-exome sequencing
- RNA-seq (speical mode available)
- ChIP-seq

How does it work?

Qualimap examines sequencing alignment data in **SAM/BAM** files according to the features of the mapped reads and provides an overall view of the data that helps to the detect biases in the sequencing and/or mapping of the data and eases decision-making for further analysis.

Starting from version 2.0 Qualimap provides **multi-sample** comparison of alignment and counts data.

Features

- Fast analysis accross the reference of genome coverage and nucleotide distribution;
- Easy to interpret summary of the main properties of the alignment data;
- Analysis of the reads mapped inside/outside of the regions provided in GFF format;
- Computation and analysis of read counts obtained from intersectition of read alignments with genomic features;
- Analysis of the adequasy of the sequencing depth in RNA-seq experiments;
- Multi-sample comparison of alignment and counts data;
- Clustering of epigenomic profiles.

Download

Latest package for GNU Linux, MacOS and MS Windows:
qualimap_v2.3.zip

Latest development snapshots

Public code repository

CIPF BioInfo local Maven repository

Version history

Documentation

QualiMap Online User Manual

Sample data and output can be found here

Bioinformatics links:

Picard: a Java API (SAM-JDK) for creating programs that read and write SAM files.
FastQC: a quality control tool for high throughput sequence data.
SAMTools: essential utilities for manipulating alignments in the SAM format.
NOISeq: quality control and differential gene expression analysis for RNA-seq data.
Repitools: quality assessment, visualization, summarization and statistical analysis of epigenomics experiments.

Support

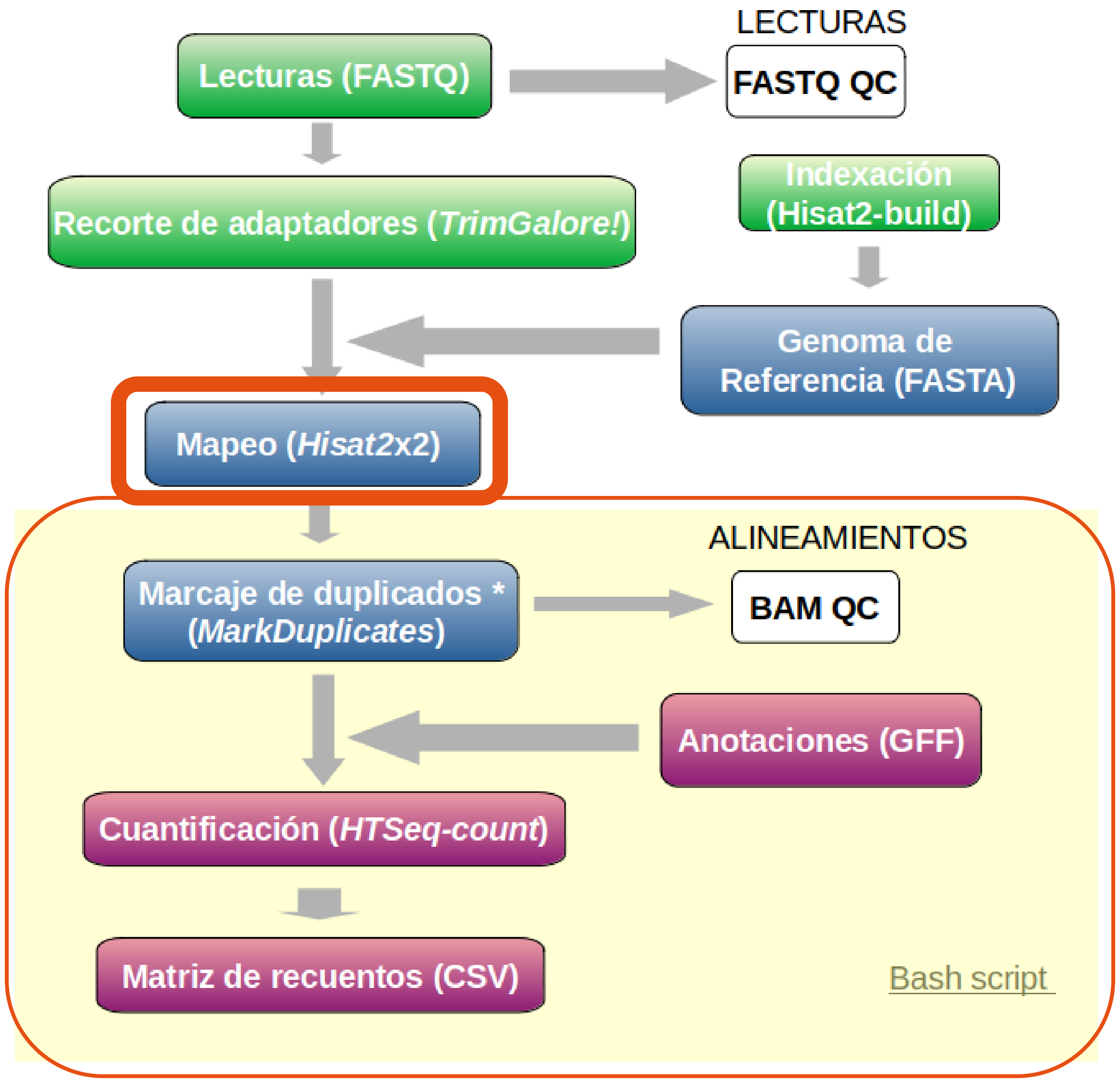
To get quick updates or report bugs/ suggestions please join the QualiMap Google group:

Your email:

Subscribe

http://qualimap.conesalab.org/

Flujo de trabajo del análisis de datos de RNA-seq (NGS)



LECTURAS

Lecturas (FASTQ)

FASTQ QC

Recorte de adaptadores (*TrimGalore!*)

Indexación (Hisat2-build)

Genoma de Referencia (FASTA)

Mapeo (*Hisat*2x2)

ALINEAMIENTOS

Marcaje de duplicados * (*MarkDuplicates*)

BAM QC

Anotaciones (GFF)

Cuantificación (*HTSeq-count*)

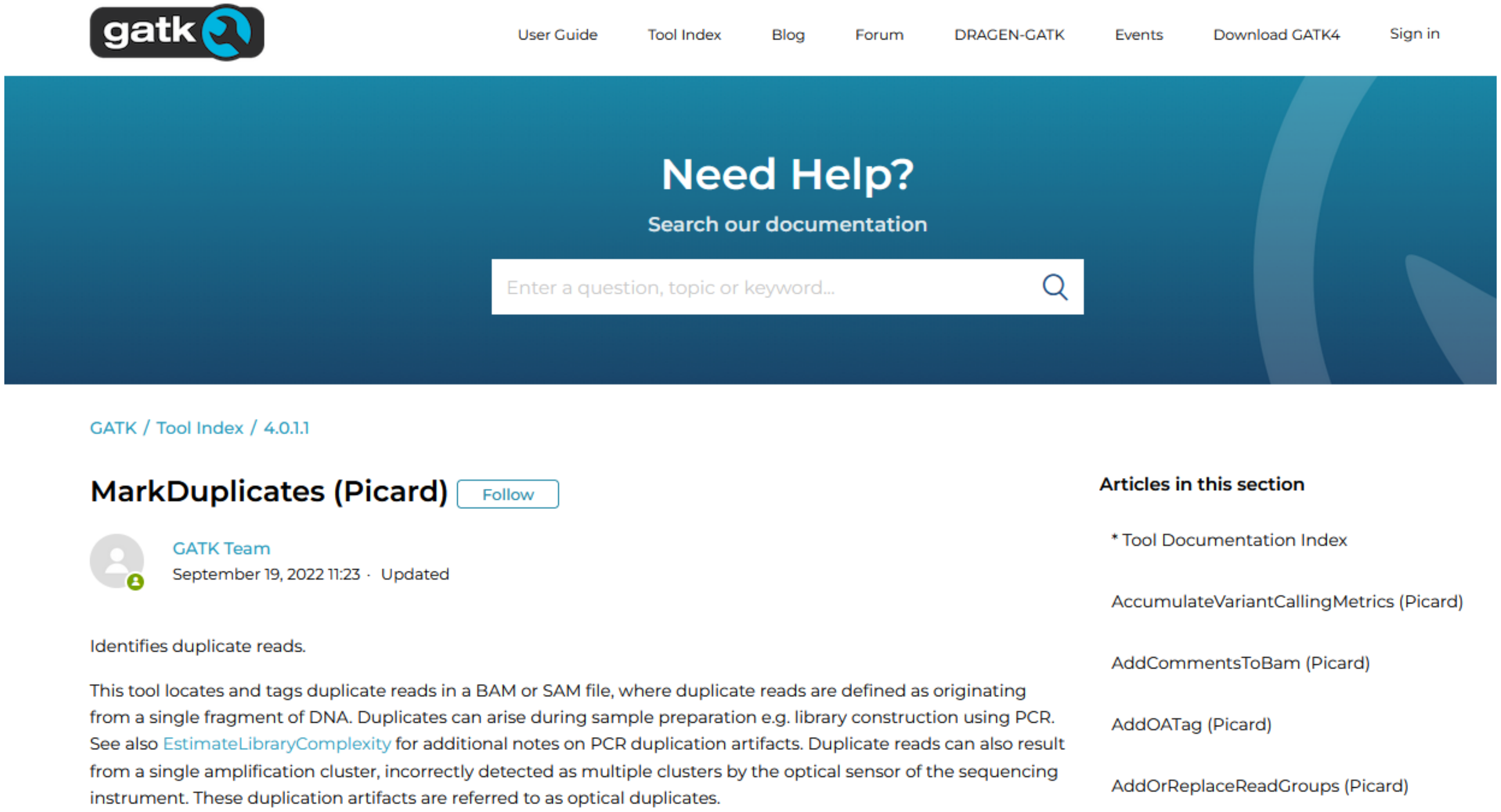
Matriz de recuentos (CSV)

Bash script

GATK: Marcaje de duplicados

conda install bioconda::picard

Permite marcar y eliminar duplicados, aunque normalmente se van a mantener esos duplicados.



gatk
User Guide
Tool Index
Blog
Forum
DRAGEN-GATK
Events
Download GATK4
Sign in
Need Help?
Search our documentation
Enter a question, topic or keyword...
GATK / Tool Index / 4.0.1.1
Follow
Articles in this section
* Tool Documentation Index
AccumulateVariantCallingMetrics (Picard)
AddCommentsToBam (Picard)
AddOATag (Picard)
AddOrReplaceReadGroups (Picard)

MarkDuplicates (Picard)

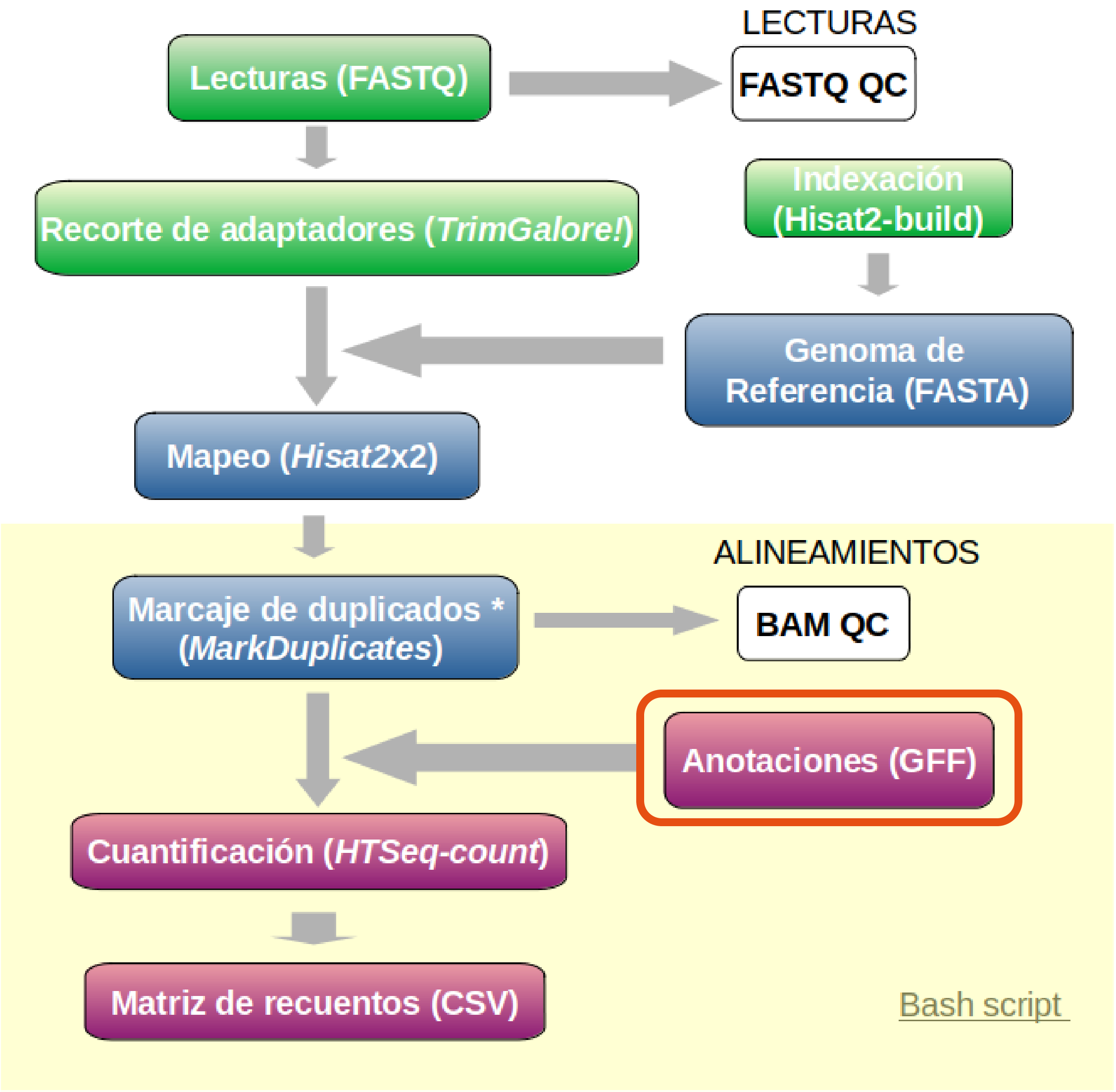
GATK Team
September 19, 2022 11:23 · Updated

Identifies duplicate reads.

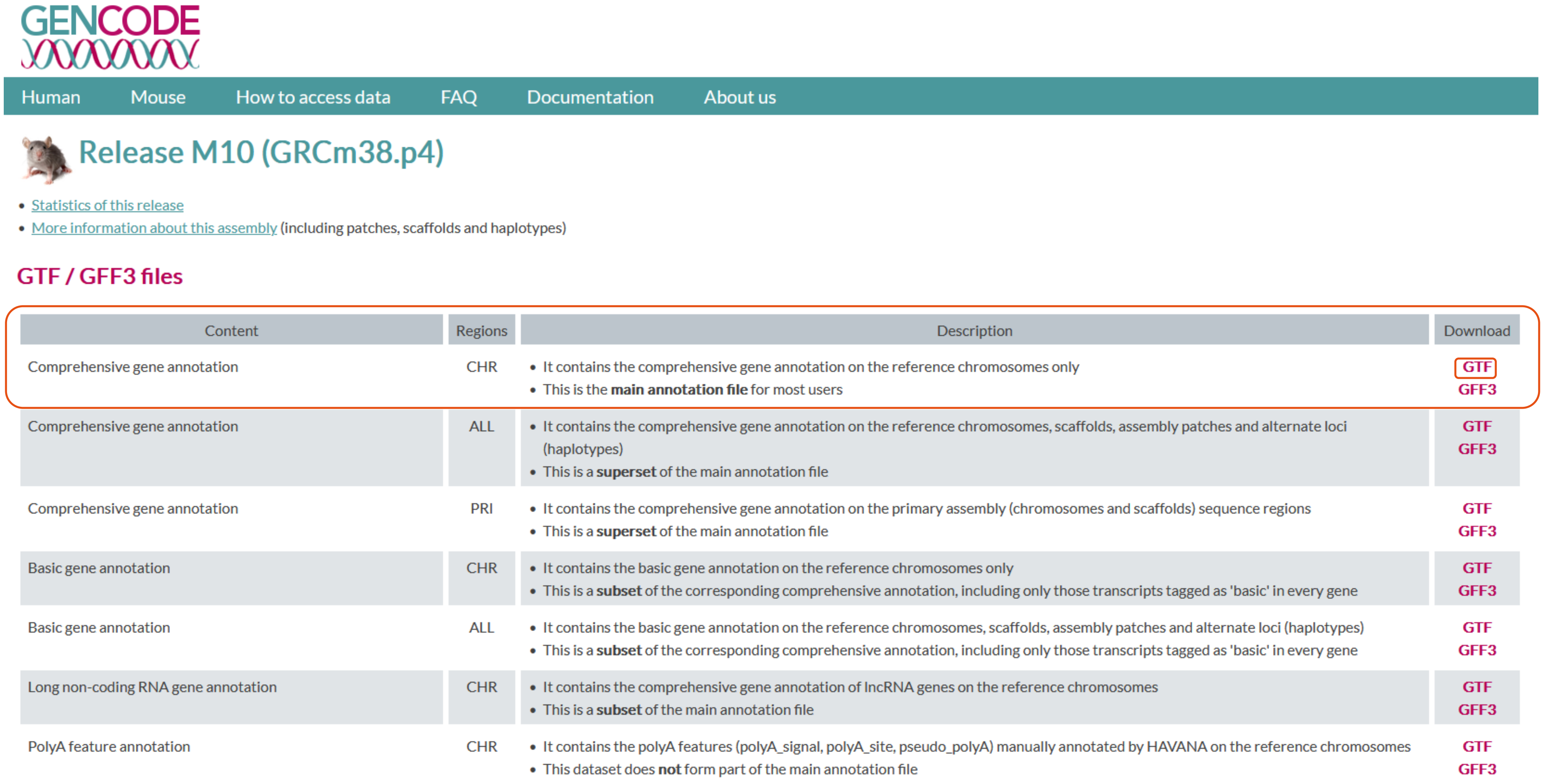
This tool locates and tags duplicate reads in a BAM or SAM file, where duplicate reads are defined as originating from a single fragment of DNA. Duplicates can arise during sample preparation e.g. library construction using PCR. See also EstimateLibraryComplexity for additional notes on PCR duplication artifacts. Duplicate reads can also result from a single amplification cluster, incorrectly detected as multiple clusters by the optical sensor of the sequencing instrument. These duplication artifacts are referred to as optical duplicates.

https://gatk.broadinstitute.org/hc/en-us/articles/360037052812-MarkDuplicates-Picard-

Flujo de trabajo del análisis de datos de RNA-seq (NGS)


LECTURAS
Lecturas (FASTQ)
FASTQ QC
Recorte de adaptadores (*TrimGalore!*)
Indexación (Hisat2-build)
Genoma de Referencia (FASTA)
Mapeo (*Hisat2*x2)
ALINEAMIENTOS
Marcaje de duplicados * (*MarkDuplicates*)
BAM QC
Anotaciones (GFF)
Cuantificación (*HTSeq-count*)
Matriz de recuentos (CSV)
Bash script

Archivo de anotación de *Mus musculus (formato GTF)*



Release M10 (GRCm38.p4)

- Statistics of this release
- More information about this assembly (including patches, scaffolds and haplotypes)

GTF / GFF3 files

Content	Regions	Description	Download
Comprehensive gene annotation	CHR	• It contains the comprehensive gene annotation on the reference chromosomes only • This is the **main annotation file** for most users	GTF GFF3
Comprehensive gene annotation	ALL	• It contains the comprehensive gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes) • This is a **superset** of the main annotation file	GTF GFF3
Comprehensive gene annotation	PRI	• It contains the comprehensive gene annotation on the primary assembly (chromosomes and scaffolds) sequence regions • This is a **superset** of the main annotation file	GTF GFF3
Basic gene annotation	CHR	• It contains the basic gene annotation on the reference chromosomes only • This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene	GTF GFF3
Basic gene annotation	ALL	• It contains the basic gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes) • This is a **subset** of the corresponding comprehensive annotation, including only those transcripts tagged as 'basic' in every gene	GTF GFF3
Long non-coding RNA gene annotation	CHR	• It contains the comprehensive gene annotation of lncRNA genes on the reference chromosomes • This is a **subset** of the main annotation file	GTF GFF3
PolyA feature annotation	CHR	• It contains the polyA features (polyA_signal, polyA_site, pseudo_polyA) manually annotated by HAVANA on the reference chromosomes • This dataset does **not** form part of the main annotation file	GTF GFF3

https://www.gencodegenes.org/mouse/release_M10.html

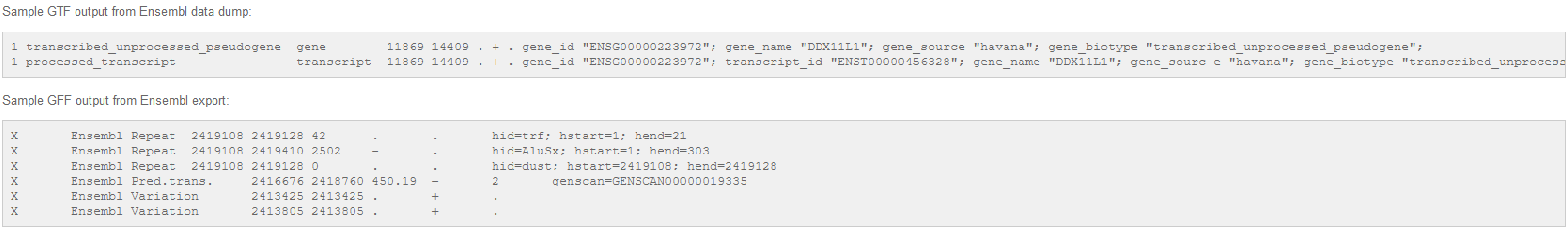
Archivo de anotación de *Mus musculus (formato GTF)*

```
(05MBIF) Data]$ tree -L 4
├── Annotation
├── Processed
│   ├── 01.Quality_control
│   │   ├── SRR1552444_fastqc.html
│   │   └── SRR1552444_fastqc.zip
│   ├── 02.Trimming
│   │   ├── SRR1552444.fastq.gz_trimming_report.txt
│   │   └── SRR1552444_trimmed.fq.gz
│   └── 03.Alignment
│       └── SRR1662444_hisat2.sam
├── Raw
│   └── SRR1552444.fastq.gz
└── Reference_genome
    ├── mm10
    │   ├── genome.1.ht2
    │   ├── genome.2.ht2
    │   ├── genome.3.ht2
    │   ├── genome.4.ht2
    │   ├── genome.5.ht2
    │   ├── genome.6.ht2
    │   ├── genome.7.ht2
    │   ├── genome.8.ht2
    │   └── make_mm10.sh
    └── mm10_genome.tar.gz
```

```
$ gunzip gencode.vM10.annotation.gtf.gz


-rw-r--r-- 1 UNIVERSIDADVIU\paula.soler  765M May  1  2015
gencode.vM10.annotation.gtf.gz
```

General Feature Format (GFF3) / Gene Transfer format (GTF)

- Ambos formatos presentan en general la misma estructura.
- Ambos presentan 9 columnas de datos.
- Los campos se separan por tabulaciones (todos los campos deben contener un valor).



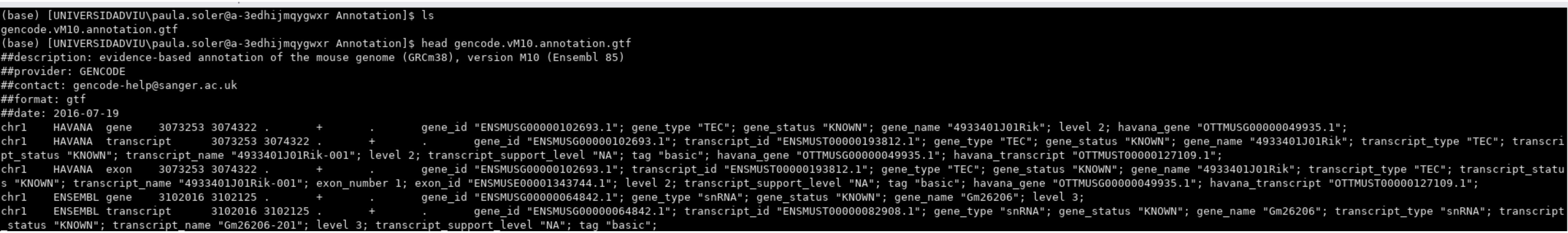
Sample GTF output from Ensembl data dump:

```
1 transcribed_unprocessed_pseudogene  gene        11869 14409 . + . gene_id "ENSG00000223972"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene";
1 processed_transcript                transcript  11869 14409 . + . gene_id "ENSG00000223972"; transcript_id "ENST00000456328"; gene_name "DDX11L1"; gene_sourc e "havana"; gene_biotype "transcribed_unprocess
```

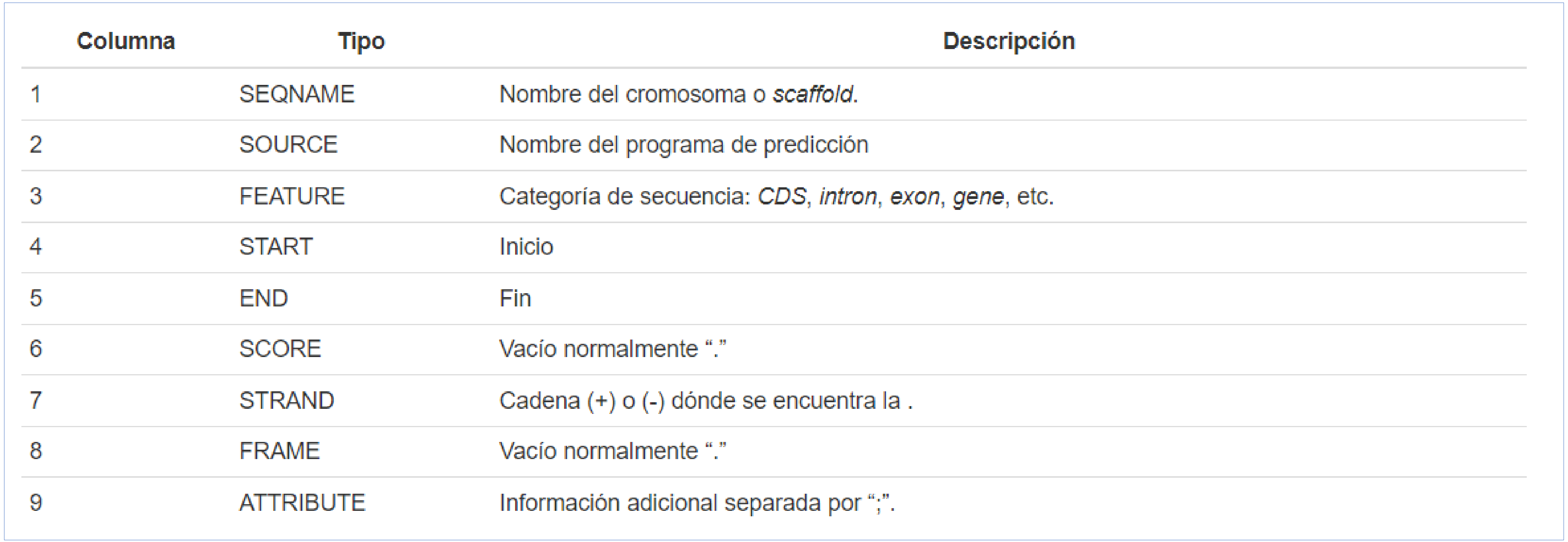
Sample GFF output from Ensembl export:

```
X       Ensembl Repeat  2419108 2419128 42      .       .       hid=trf; hstart=1; hend=21
X       Ensembl Repeat  2419108 2419410 2502    -       .       hid=AluSx; hstart=1; hend=303
X       Ensembl Repeat  2419108 2419128 0       .       .       hid=dust; hstart=2419108; hend=2419128
X       Ensembl Pred.trans.     2416676 2418760 450.19  -       2       genscan=GENSCAN00000019335
X       Ensembl Variation       2413425 2413425 .       +       .
X       Ensembl Variation       2413805 2413805 .       +       .
```

General Feature Format (GFF3) / **Gene Transfer format (GTF)**

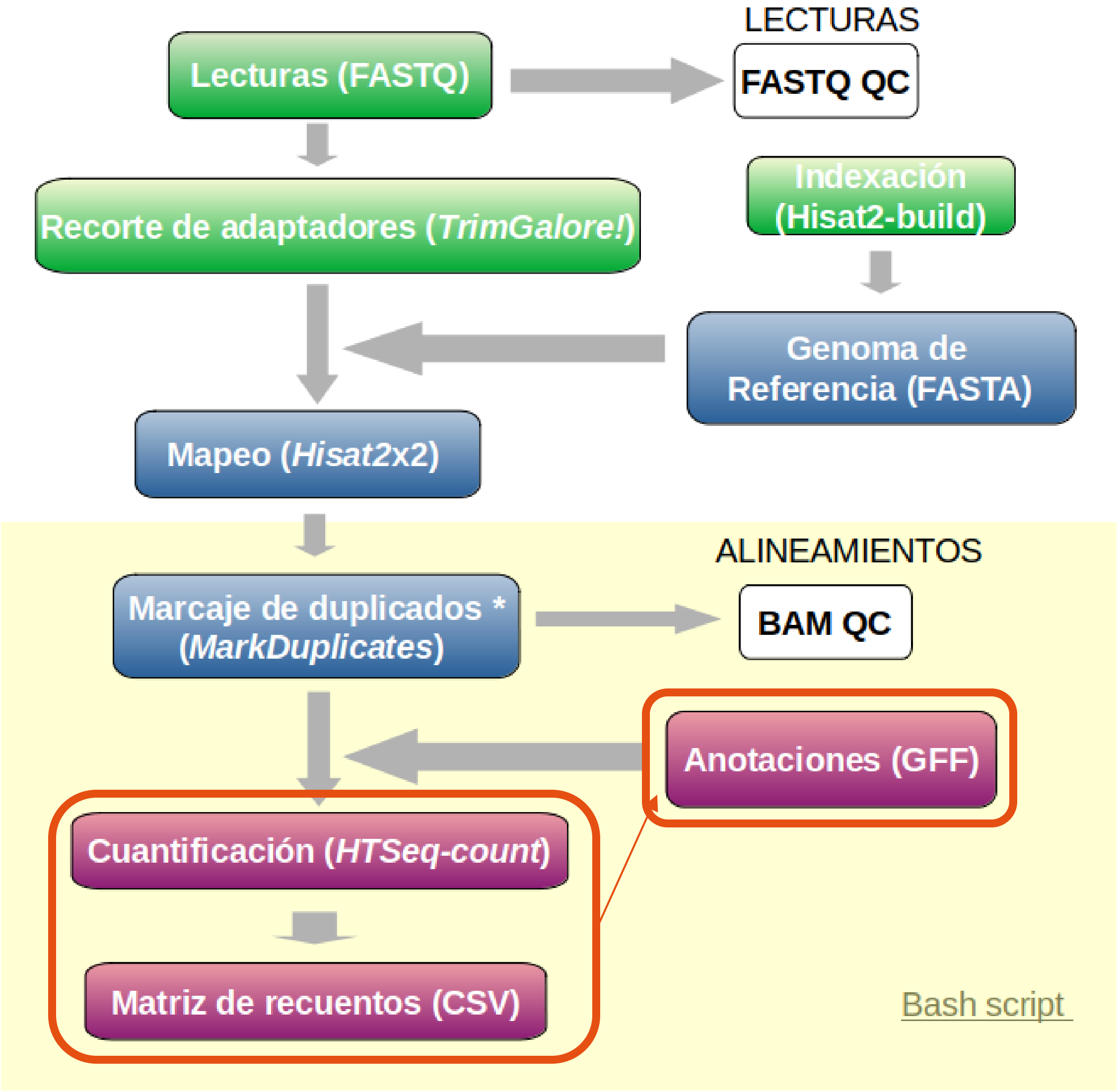


```
(base) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Annotation]$ ls
gencode.vM10.annotation.gtf
(base) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Annotation]$ head gencode.vM10.annotation.gtf
##description: evidence-based annotation of the mouse genome (GRCm38), version M10 (Ensembl 85)
##provider: GENCODE
##contact: gencode-help@sanger.ac.uk
##format: gtf
##date: 2016-07-19
chr1    HAVANA  gene    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; gene_type "TEC"; gene_status "KNOWN"; gene_name "4933401J01Rik"; level 2; havana_gene "OTTMUSG00000049935.1";
chr1    HAVANA  transcript      3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUST00000193812.1"; gene_type "TEC"; gene_status "KNOWN"; gene_name "4933401J01Rik"; transcript_type "TEC"; transcri
pt_status "KNOWN"; transcript_name "4933401J01Rik-001"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_transcript "OTTMUST00000127109.1";
chr1    HAVANA  exon    3073253 3074322 .       +       .       gene_id "ENSMUSG00000102693.1"; transcript_id "ENSMUST00000193812.1"; gene_type "TEC"; gene_status "KNOWN"; gene_name "4933401J01Rik"; transcript_type "TEC"; transcript_statu
s "KNOWN"; transcript_name "4933401J01Rik-001"; exon_number 1; exon_id "ENSMUSE00001343744.1"; level 2; transcript_support_level "NA"; tag "basic"; havana_gene "OTTMUSG00000049935.1"; havana_transcript "OTTMUST00000127109.1";
chr1    ENSEMBL gene    3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; gene_type "snRNA"; gene_status "KNOWN"; gene_name "Gm26206"; level 3;
chr1    ENSEMBL transcript      3102016 3102125 .       +       .       gene_id "ENSMUSG00000064842.1"; transcript_id "ENSMUST00000082908.1"; gene_type "snRNA"; gene_status "KNOWN"; gene_name "Gm26206"; transcript_type "snRNA"; transcript
_status "KNOWN"; transcript_name "Gm26206-201"; level 3; transcript_support_level "NA"; tag "basic";
```



Columna	Tipo	Descripción
1	SEQNAME	Nombre del cromosoma o *scaffold*.
2	SOURCE	Nombre del programa de predicción
3	FEATURE	Categoría de secuencia: *CDS*, *intron*, *exon*, *gene*, etc.
4	START	Inicio
5	END	Fin
6	SCORE	Vacío normalmente “.”
7	STRAND	Cadena (+) o (-) dónde se encuentra la .
8	FRAME	Vacío normalmente “.”
9	ATTRIBUTE	Información adicional separada por “;”.

Flujo de trabajo del análisis de datos de RNA-seq (NGS)



LECTURAS

Lecturas (FASTQ)

FASTQ QC

Recorte de adaptadores (*TrimGalore!*)

Indexación (Hisat2-build)

Genoma de Referencia (FASTA)

Mapeo (*Hisat2*x2)

ALINEAMIENTOS

Marcaje de duplicados * (*MarkDuplicates*)

BAM QC

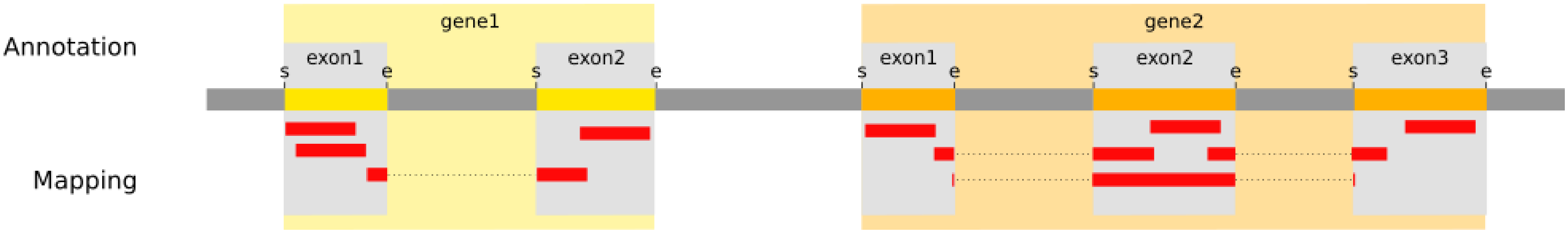
Anotaciones (GFF)

Cuantificación (*HTSeq-count*)

Matriz de recuentos (CSV)

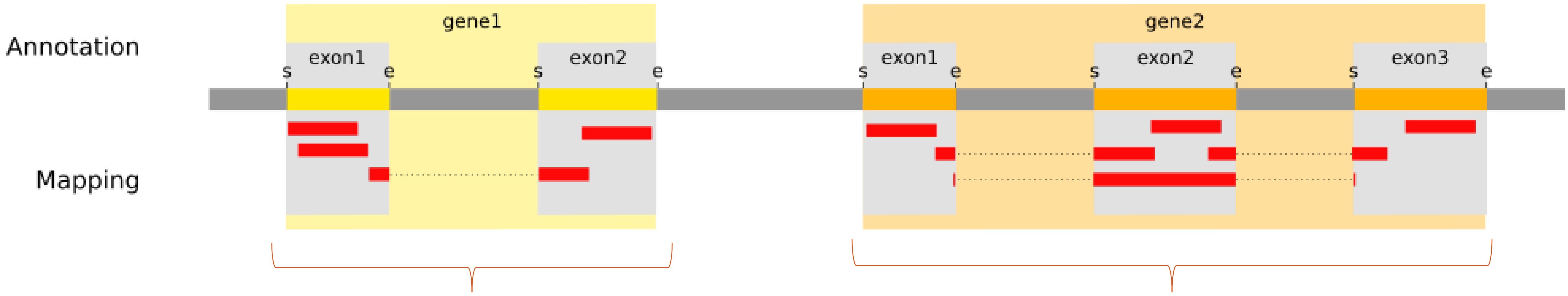
Bash script

Recuento del número de lecturas por gen anotado



Exon	Number of reads
gene1 - exon1	3
gene1 - exon2	2
gene2 - exon1	3
gene2 - exon2	4
gene2 - exon3	3

Recuento del número de lecturas por gen anotado

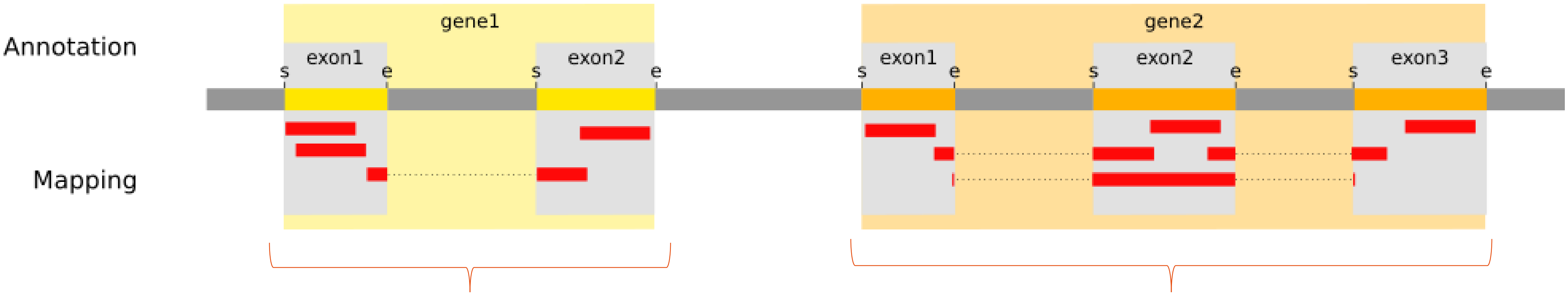


gene1 tiene **4** lecturas

gene2 tiene **6** lecturas

- Número de lecturas por exón
- Número final de lecturas por gen, entiendo que este es la unión de todos los exones.

Recuento del número de lecturas por gen anotado



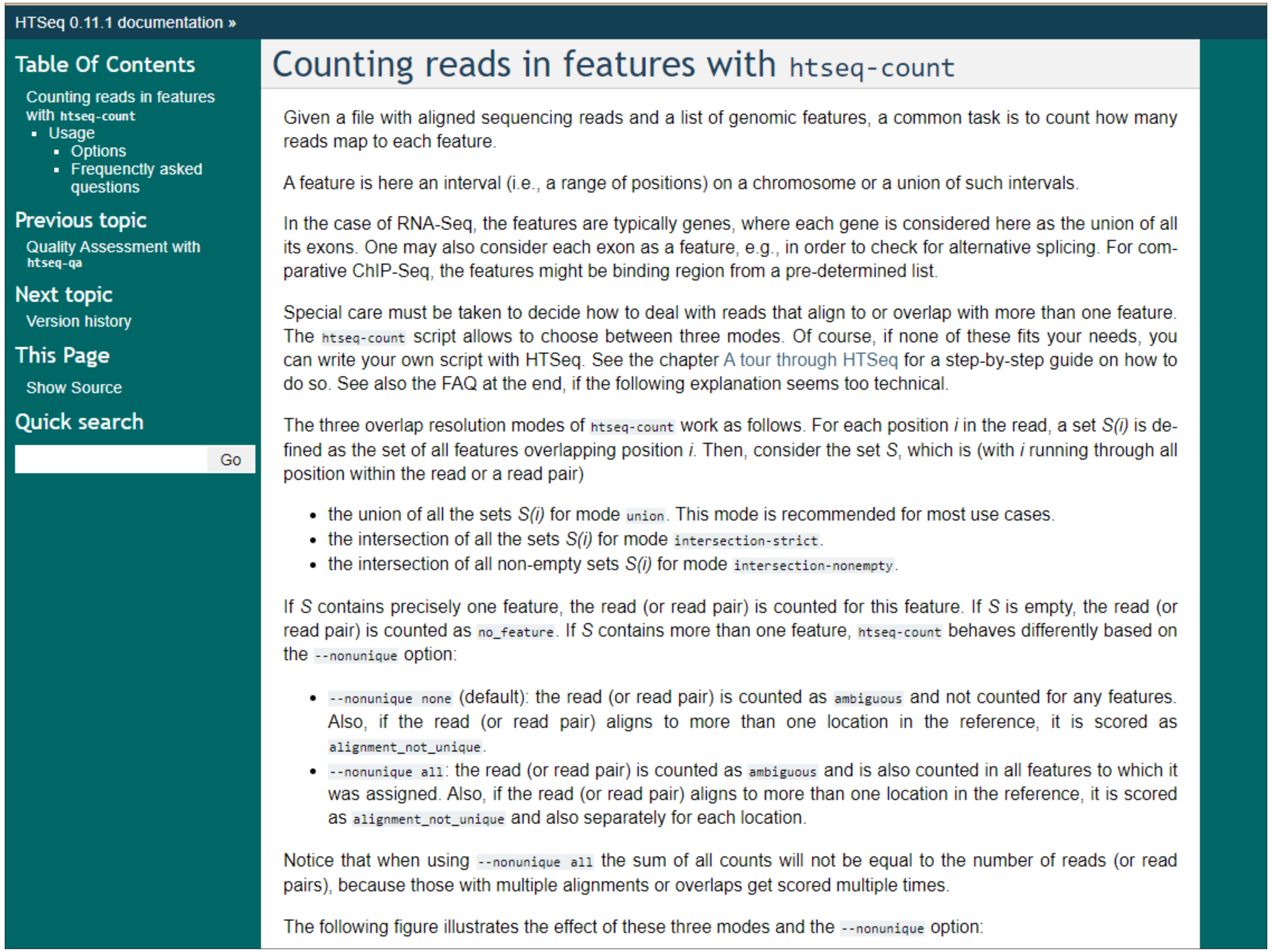
gene1 tiene **4** lecturas

gene2 tiene **6** lecturas

HTSeq-count (Anders *et al.* 2015)

featureCounts (Liao et al. 2013)

Recuento de los alineamientos (HTseq-count)



HTSeq 0.11.1 documentation »

Previous topic

Quality Assessment with htseq-qa

Next topic

Version history

This Page

Show Source

Quick search

Counting reads in features with `htseq-count`

Given a file with aligned sequencing reads and a list of genomic features, a common task is to count how many reads map to each feature.

A feature is here an interval (i.e., a range of positions) on a chromosome or a union of such intervals.

In the case of RNA-Seq, the features are typically genes, where each gene is considered here as the union of all its exons. One may also consider each exon as a feature, e.g., in order to check for alternative splicing. For comparative ChIP-Seq, the features might be binding region from a pre-determined list.

Special care must be taken to decide how to deal with reads that align to or overlap with more than one feature. The `htseq-count` script allows to choose between three modes. Of course, if none of these fits your needs, you can write your own script with HTSeq. See the chapter A tour through HTSeq for a step-by-step guide on how to do so. See also the FAQ at the end, if the following explanation seems too technical.

The three overlap resolution modes of `htseq-count` work as follows. For each position *i* in the read, a set *S(i)* is defined as the set of all features overlapping position *i*. Then, consider the set *S*, which is (with *i* running through all position within the read or a read pair)

- the union of all the sets *S(i)* for mode `union`. This mode is recommended for most use cases.
- the intersection of all the sets *S(i)* for mode `intersection-strict`.
- the intersection of all non-empty sets *S(i)* for mode `intersection-nonempty`.

If *S* contains precisely one feature, the read (or read pair) is counted for this feature. If *S* is empty, the read (or read pair) is counted as `no_feature`. If *S* contains more than one feature, `htseq-count` behaves differently based on the `--nonunique` option:

- `--nonunique none` (default): the read (or read pair) is counted as `ambiguous` and not counted for any features. Also, if the read (or read pair) aligns to more than one location in the reference, it is scored as `alignment_not_unique`.
- `--nonunique all`: the read (or read pair) is counted as `ambiguous` and is also counted in all features to which it was assigned. Also, if the read (or read pair) aligns to more than one location in the reference, it is scored as `alignment_not_unique` and also separately for each location.

Notice that when using `--nonunique all` the sum of all counts will not be equal to the number of reads (or read pairs), because those with multiple alignments or overlaps get scored multiple times.

The following figure illustrates the effect of these three modes and the `--nonunique` option:

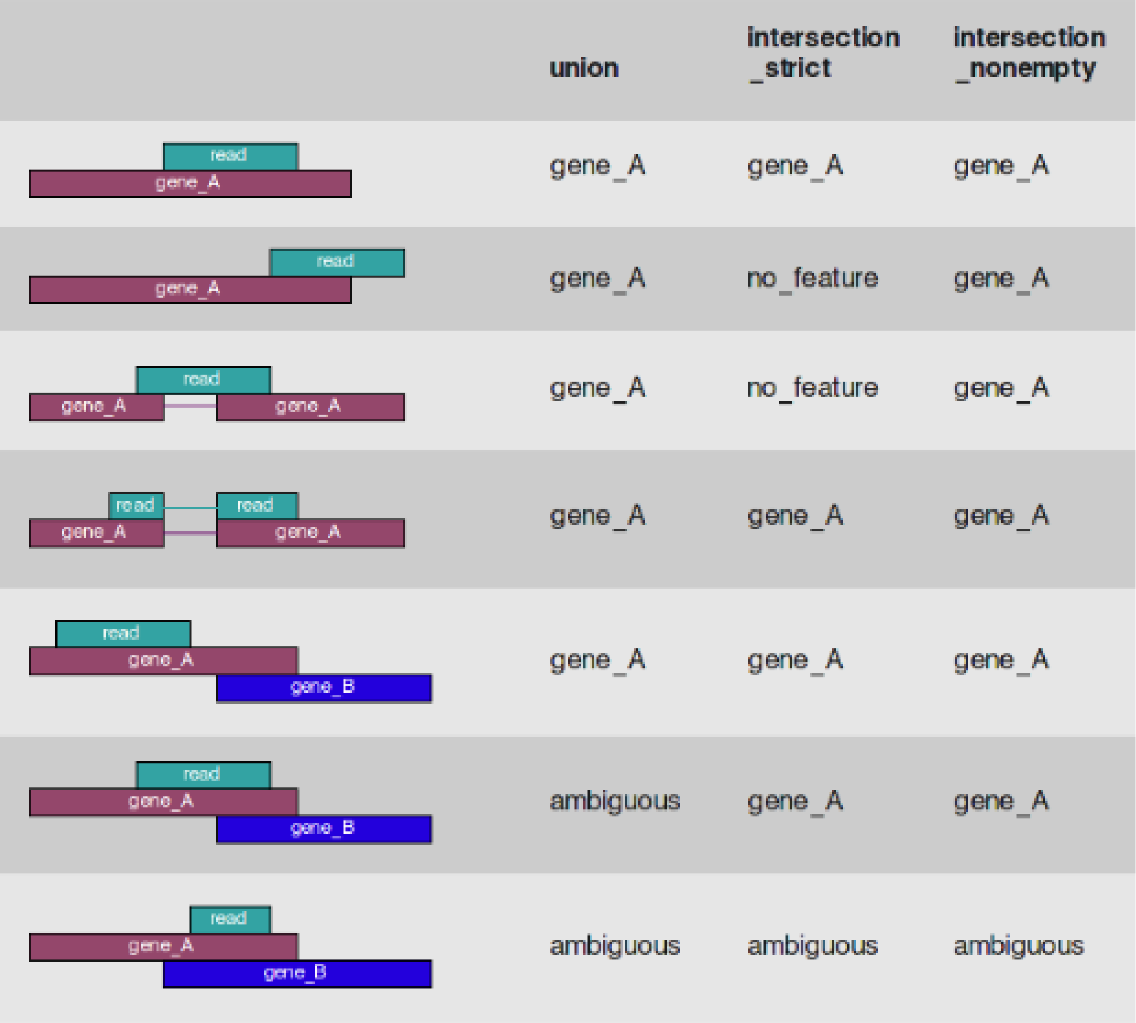
https://htseq.readthedocs.io/en/release_0.11.1/count.html

Recuento de los alineamientos (HTseq-count).

¿Qué contar?
⇒ *id_attr=**gene_id***
Para recuentos por gen

¿Dónde contar?
⇒ *tipo=**exón***
Contar alineaciones superpuestas exones

¿Como contar?
⇒ *modo=**unión***
Para la expresión diferencial en el nivel genético



	union	intersection_strict	intersection_nonempty
read gene_A	gene_A	gene_A	gene_A
read gene_A	gene_A	no_feature	gene_A
read gene_A gene_A	gene_A	no_feature	gene_A
read read gene_A gene_A	gene_A	gene_A	gene_A
read gene_A gene_B	gene_A	gene_A	gene_A
read gene_A gene_B	ambiguous	gene_A	gene_A
read gene_A gene_B	ambiguous	ambiguous	ambiguous

Recuento de los alineamientos (HTseq-count).

```
conda install bioconda::htseq
```

```
(05MBIF) $ htseq- (+TAB)
htseq-count        htseq-count-barcodes  htseq-qa
```

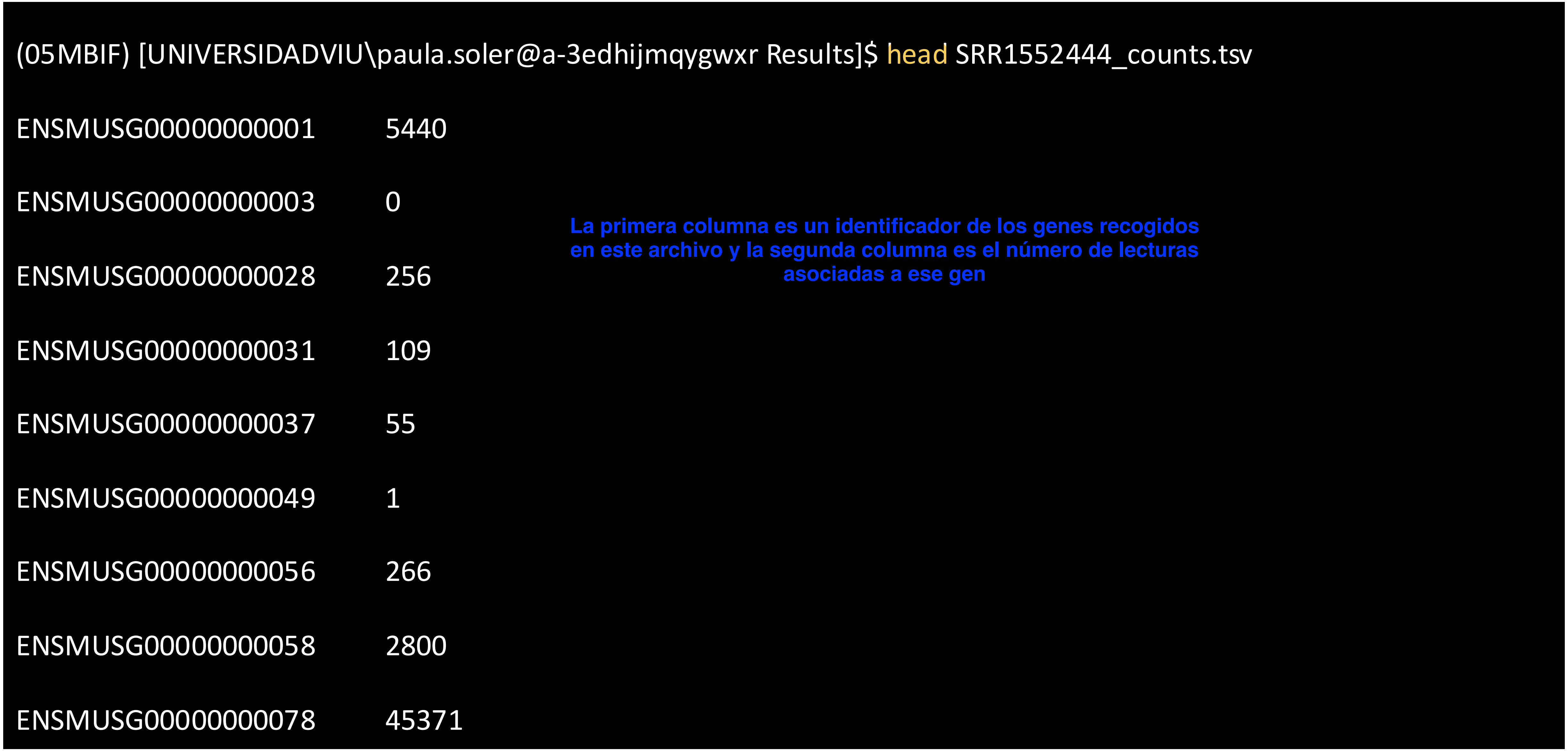
https://anaconda.org/bioconda/htseq

con -t le decimos que busque los exones, -i que agrupe las lecturas que pertenecen a un mismo gen con el id, —stranded le estamos diciendo que NO tiene información de hebra específica, -f el formato del archivo, -r que lo ordene por la posición, -s va junto con la opción stranded, que básicamente es para especificar que no tenemos información de orientación específica.

Lanzar la instrucción

```
(05MBIF) [03.Alignment]$ htseq-count -t exon -i gene_id --stranded=no -f bam -r pos -s no SRR1552444_hisat2.sorted.bam

../../Annotation/gencode.vM10.annotation.gtf > ../../../Results/SRR1552444_counts.tsv



100000 GFF lines processed.

200000 GFF lines processed.

300000 GFF lines processed.
```

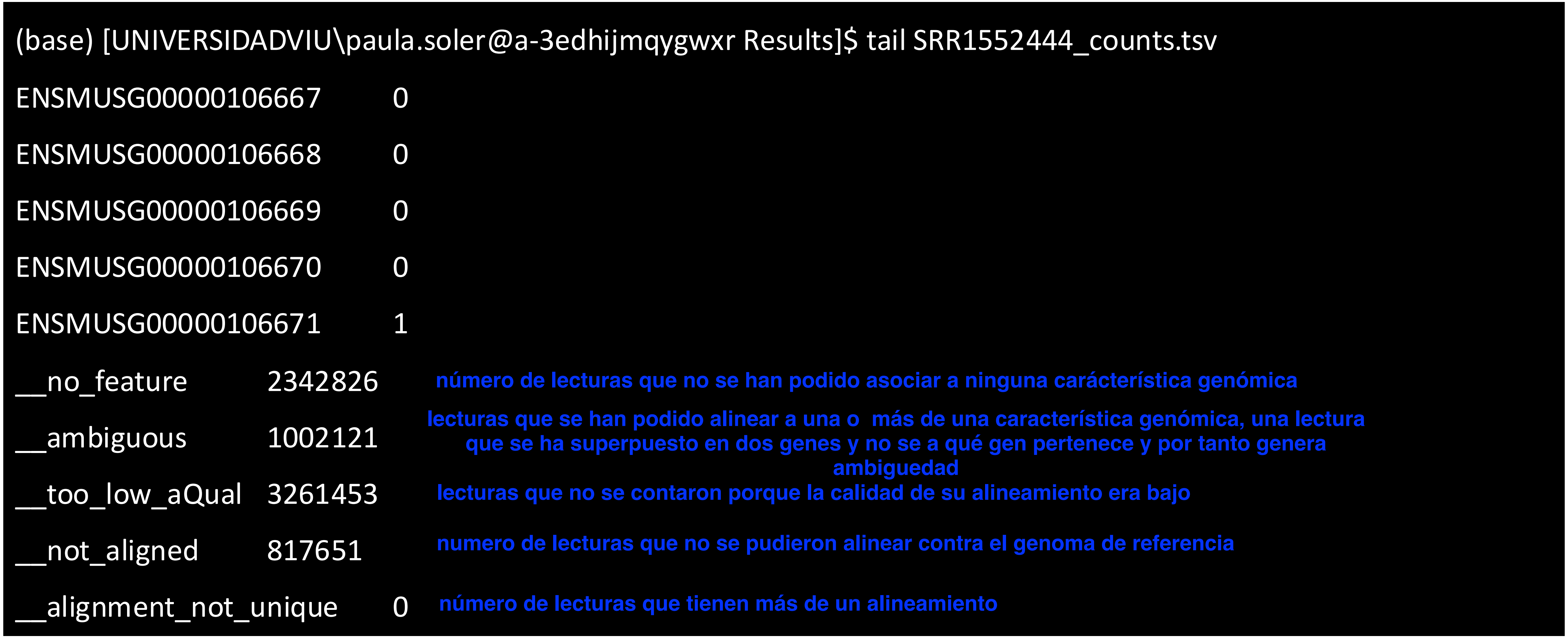
Archivo de recuentos (formato tsv)



```
(05MBIF) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Results]$ head SRR1552444_counts.tsv
ENSMUSG00000000001	5440
ENSMUSG00000000003	0
ENSMUSG00000000028	256
ENSMUSG00000000031	109
ENSMUSG00000000037	55
ENSMUSG00000000049	1
ENSMUSG00000000056	266
ENSMUSG00000000058	2800
ENSMUSG00000000078	45371
```

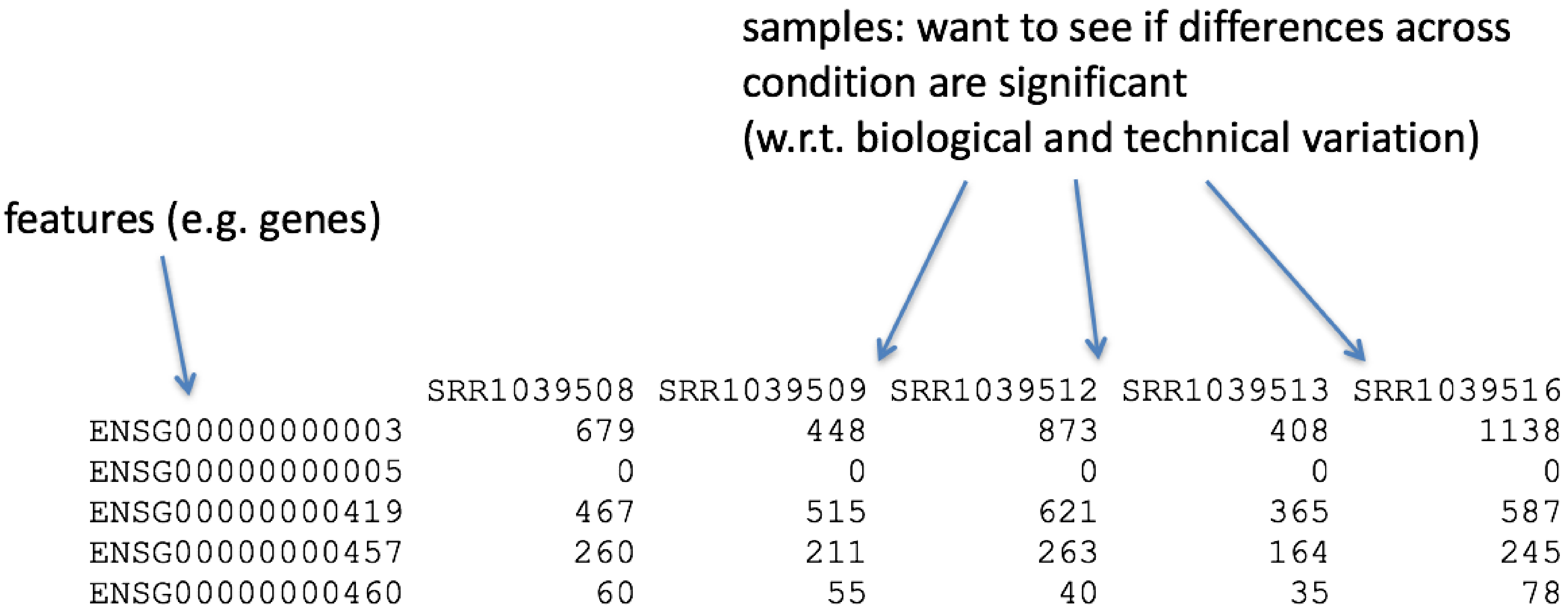
La primera columna es un identificador de los genes recogidos en este archivo y la segunda columna es el número de lecturas asociadas a ese gen

Archivo de recuentos (formato tsv)



```
(base) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr Results]$ tail SRR1552444_counts.tsv
ENSMUSG00000106667      0
ENSMUSG00000106668      0
ENSMUSG00000106669      0
ENSMUSG00000106670      0
ENSMUSG00000106671      1
__no_feature    2342826     número de lecturas que no se han podido asociar a ninguna carácterística genómica
__ambiguous     1002121    lecturas que se han podido alinear a una o  más de una característica genómica, una lectura
                              que se ha superpuesto en dos genes y no se a qué gen pertenece y por tanto genera
                                                          ambiguedad
__too_low_aQual 3261453     lecturas que no se contaron porque la calidad de su alineamiento era bajo
__not_aligned   817651      numero de lecturas que no se pudieron alinear contra el genoma de referencia
__alignment_not_unique  0   número de lecturas que tienen más de un alineamiento
```

Merging de todas las muestras obtenidas: *Count matrix*



samples: want to see if differences across condition are significant (w.r.t. biological and technical variation)

features (e.g. genes)

```
                SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516
ENSG00000000003        679        448        873        408       1138
ENSG00000000005          0          0          0          0          0
ENSG00000000419        467        515        621        365        587
ENSG00000000457        260        211        263        164        245
ENSG00000000460         60         55         40         35         78
```

Package "EdgeR"

viu

Universidad Internacional de Valencia

universidadviu.com

De:
Planeta Formación y Universidades