

Análisis transcriptómicos de la expresión génica

Máster Universitario en Bioinformática

Sesión 6



Universidad
Internacional
de Valencia

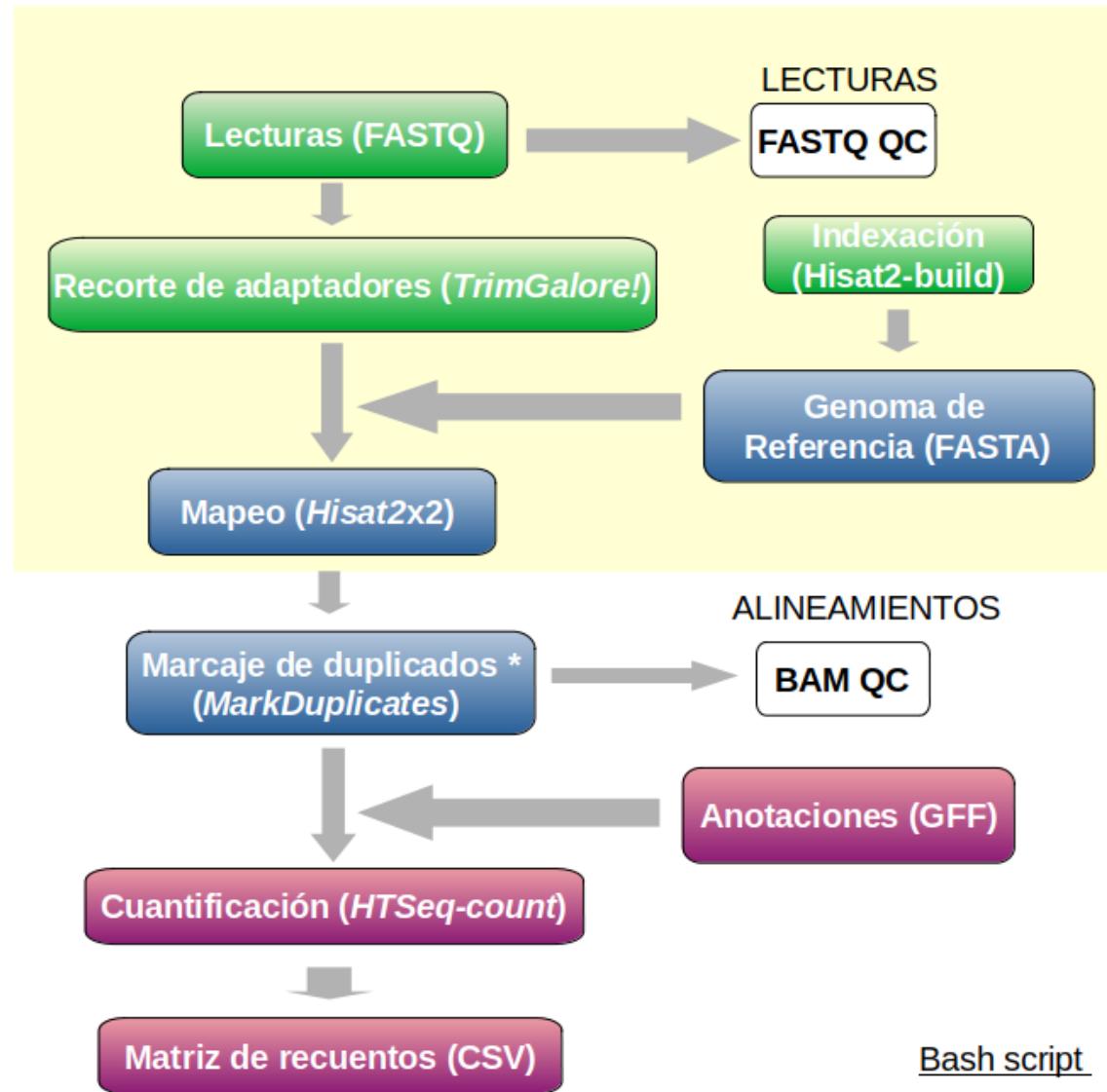
Dra. Paula Soler Vila
paula.solerv@professor.universidadviu.com

De:
 Planeta Formación y Universidades



Bloque III: Análisis de datos de NGS

Flujo de trabajo del análisis de datos de RNA-seq



Objetivos de la sesión

1

Realizar el pretratamiento de los datos crudos:

- Revisión del formato FASTQ y de los parámetros clave de calidad (*FASTQC*)
- Filtrado vs Recorte de lecturas (*Trimmomatic*)

2

Obtención del genoma de referencia (**indexación**)

3

Conocer las principales características del mapeado de lecturas.

Inspeccionando los FASTQ

- Derivado del formato FASTA
 - **Composición + Calidad**
 - Generalmente está comprimido (**.fastq.gz**)
 - Comprende **4** líneas por secuencia:
 - @ ID de la lectura + información de *run*
 - Secuencia
 - Símbolo "+"
 - Calidad (Escala **Phred** y código **ASCII**)

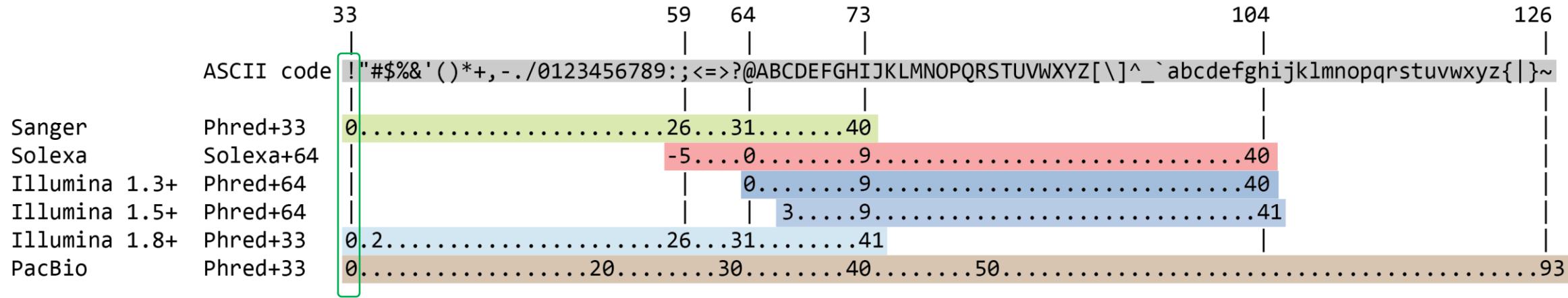
Para contar el número de secuencias: echo \$(zcat SRR1552444.fastq.gz | wc -l)/4 | bc
o
zgrep -c “^@SRR” SRR..

Inspeccionando los FASTQ: Calidad

Phred (**Q**) que se define como: $Q = -10 * \log_{10}(p)$
donde p es la probabilidad de error en la identificación de esa base.

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%

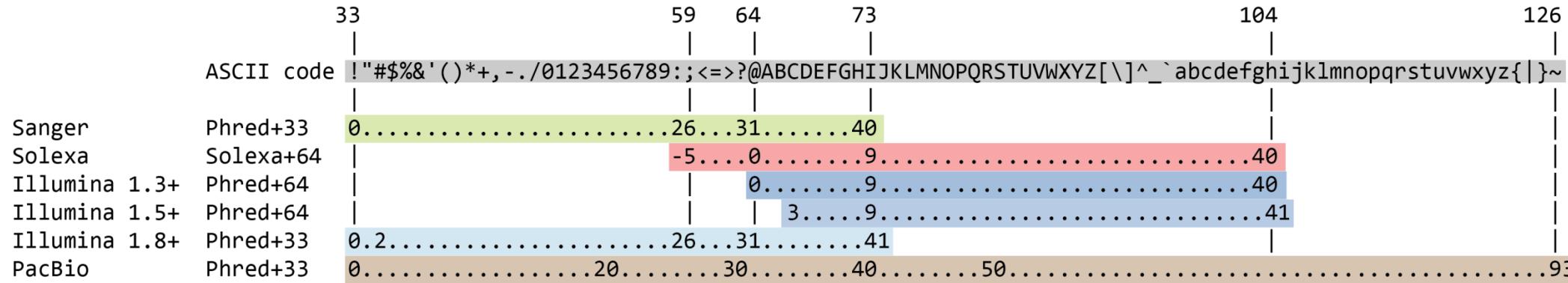
Inspeccionando los FASTQ: Calidad -> Phred



¿Qué carácter ASCII corresponde a la peor puntuación de Phred para Illumina 1.8+?

Carácter = !

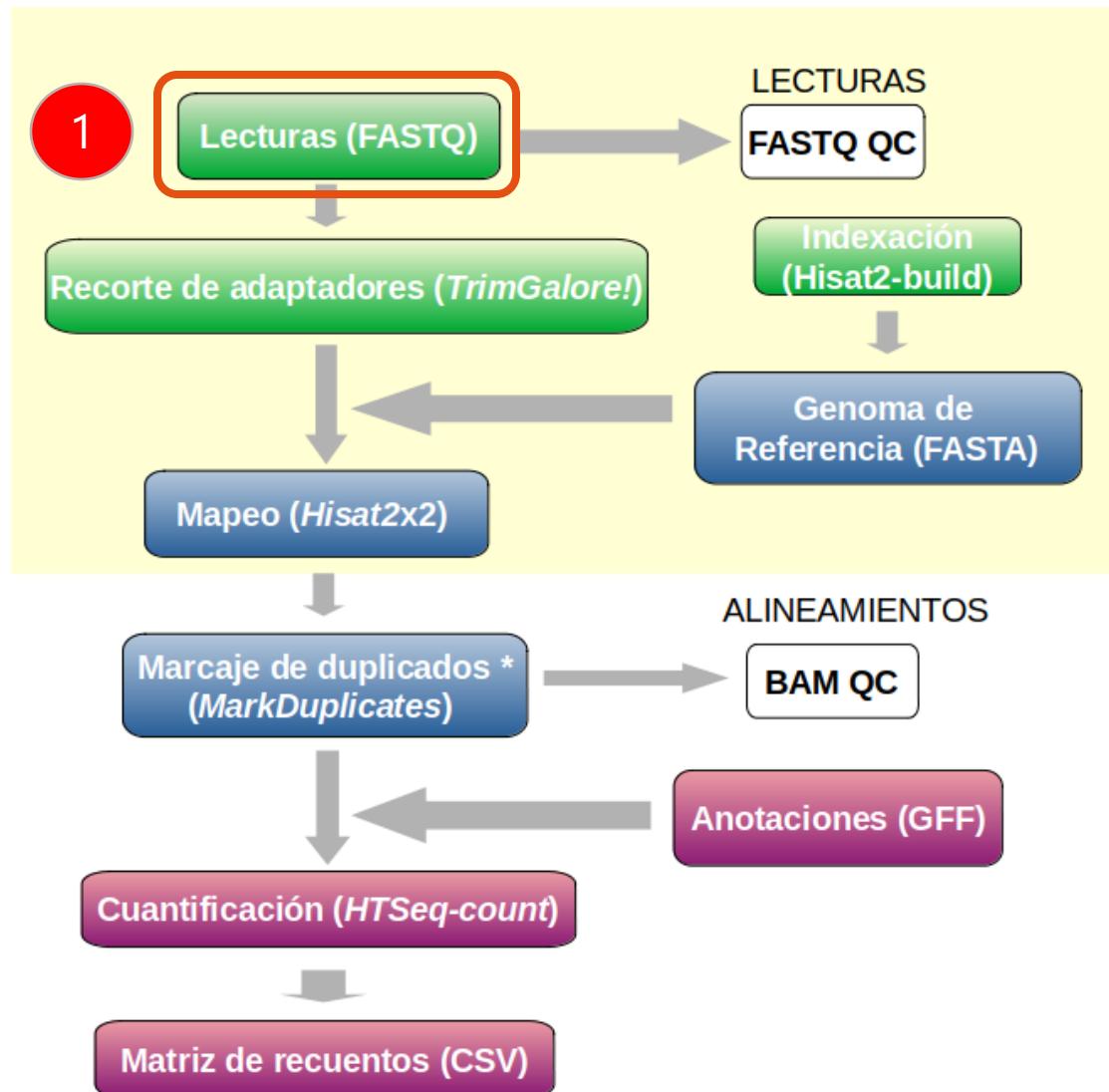
Inspeccionando los FASTQ: Calidad > Phred



¿Cuál es el puntaje de calidad de Phred+33 del tercer nucleótido de la primera secuencia (Sanger)?

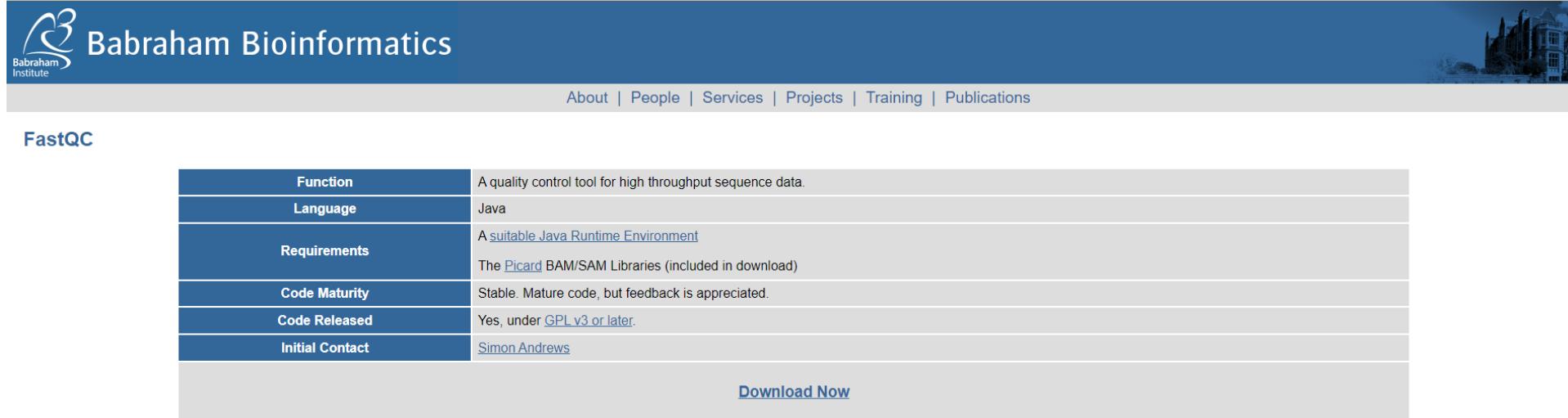
Puntaje de calidad -> 38

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



- 1 Conocer e identificar problemas potenciales en nuestros datos nos va a permitir:
1. Corregirlos antes de gastar una gran cantidad de tiempo en el análisis
 2. Tenerlos en consideración cuando interpretemos los resultados
- FASTQC, FAstX, Prinseq...**

Herramienta FASTQC



The screenshot shows the Babraham Bioinformatics website. The header includes the Babraham Institute logo and navigation links: About, People, Services, Projects, Training, Publications. Below the header, a sub-navigation bar shows the current section: FastQC. A table provides detailed information about the tool:

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment The Picard BAM/SAM Libraries (included in download)
Code Maturity	Stable. Mature code, but feedback is appreciated.
Code Released	Yes, under GPL v3 or later .
Initial Contact	Simon Andrews

[Download Now](#)

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

- Estadísticas básicas del conjunto de secuencias.
- Calidad de la secuencia por base.
- Valores de calidad por secuencia.
- Contenido de la secuencia por base.
- Contenido en GC por secuencia.
- Contenido de N por base.
- Distribución de la longitud de las secuencias.
- Secuencias duplicadas.
- Secuencias sobrerepresentadas.
- Contenido de adaptadores.

PRACTIQUEMOS



Resumen: Estadísticas básicas

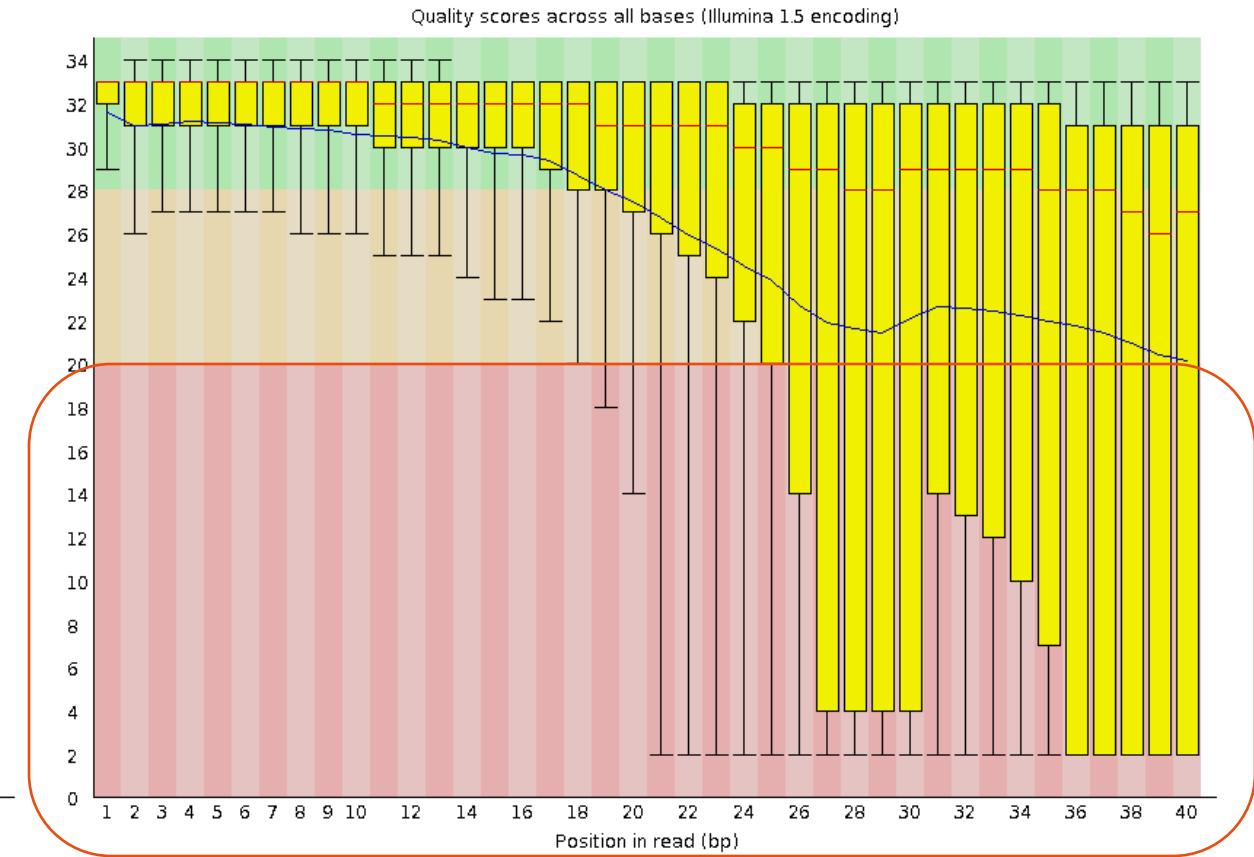
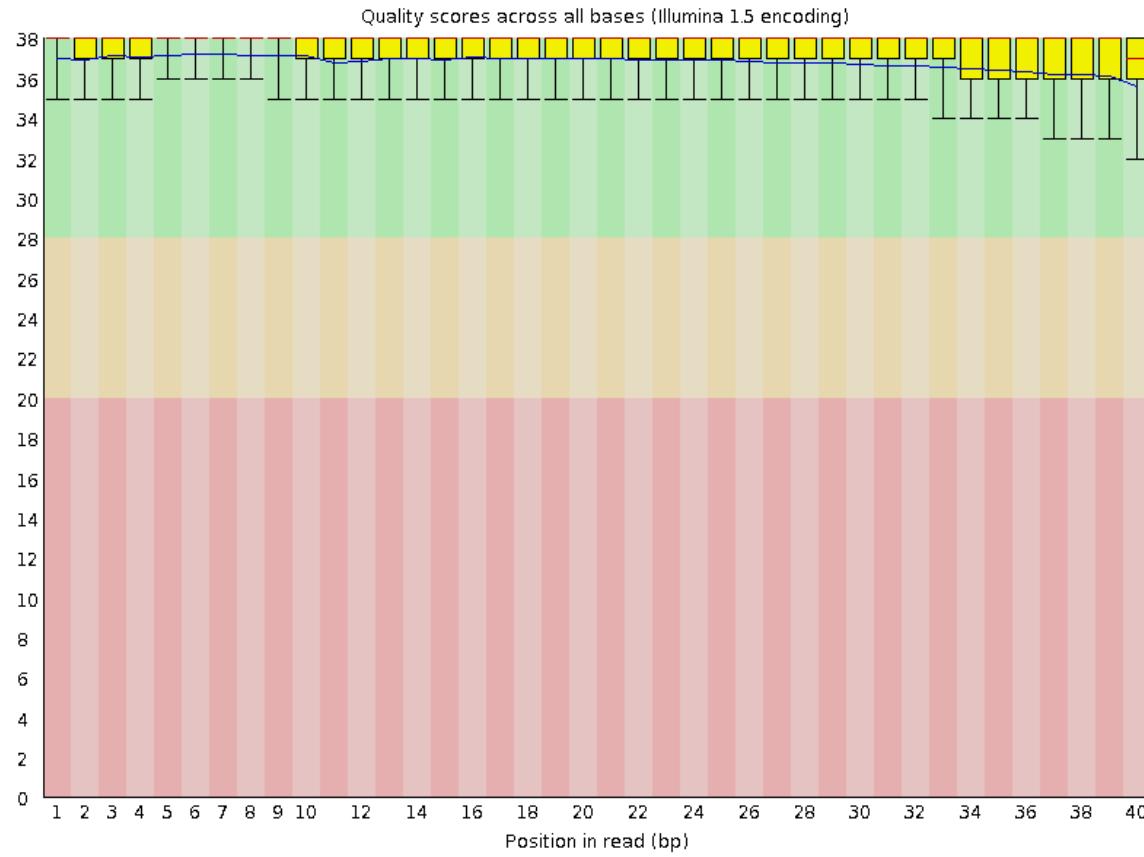
Summary

-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per tile sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)
-  [Kmer Content](#)

Basic Statistics

Measure	Value
Filename	Mov10_oe_1.subset.fq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	305900
Sequences flagged as poor quality	0
Sequence length	100
%GC	47

Calidad de las secuencias por base



La línea azul representa el puntuaje de calidad promedio para el nucleótido en todas las lecturas.

Pérdida de calidad de las bases -> Illumina

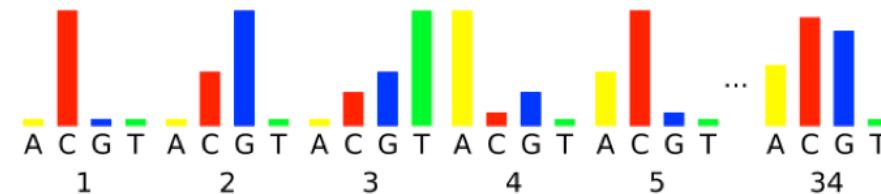
- INTENSIDAD DE LA SEÑAL FLUORESCENTE

- PUREZA DE LA SEÑAL FLUORESCENTE

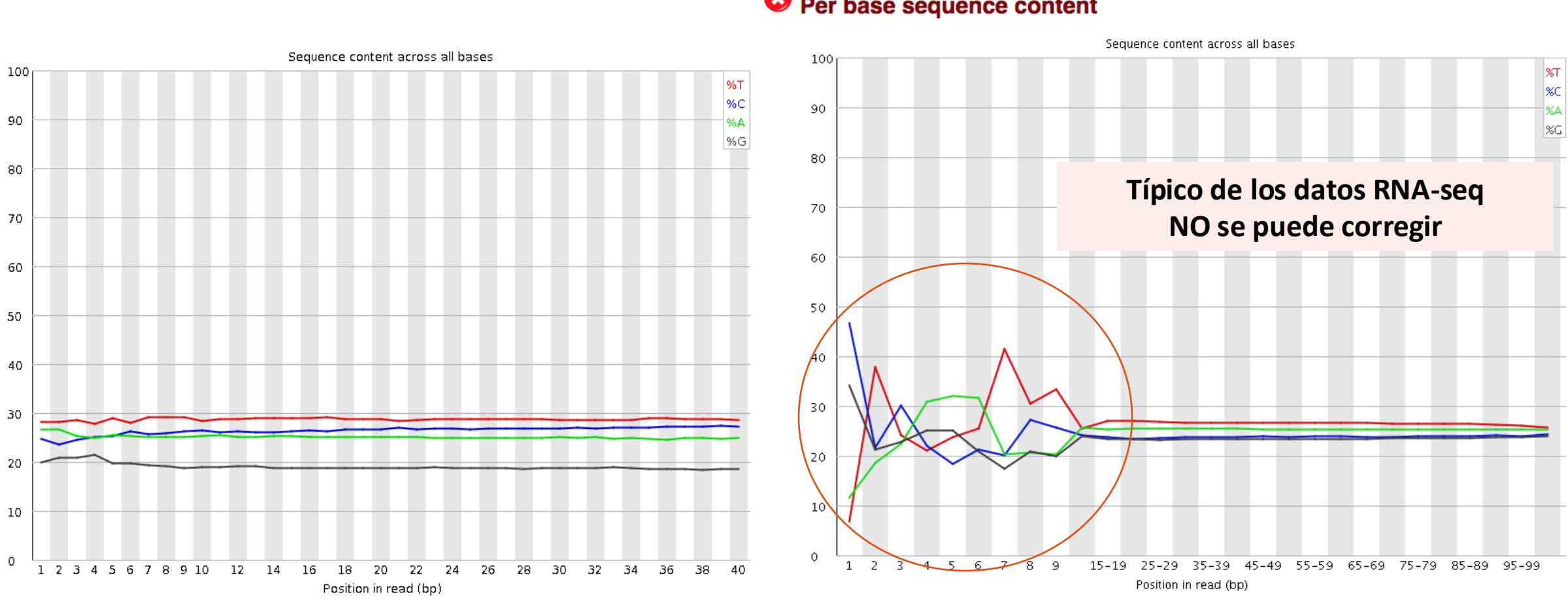
A medida que avanza la secuenciación desde el primer ciclo hasta el último, a menudo observamos una **caída en la calidad de las llamadas bases**.

Esto a menudo se debe a:

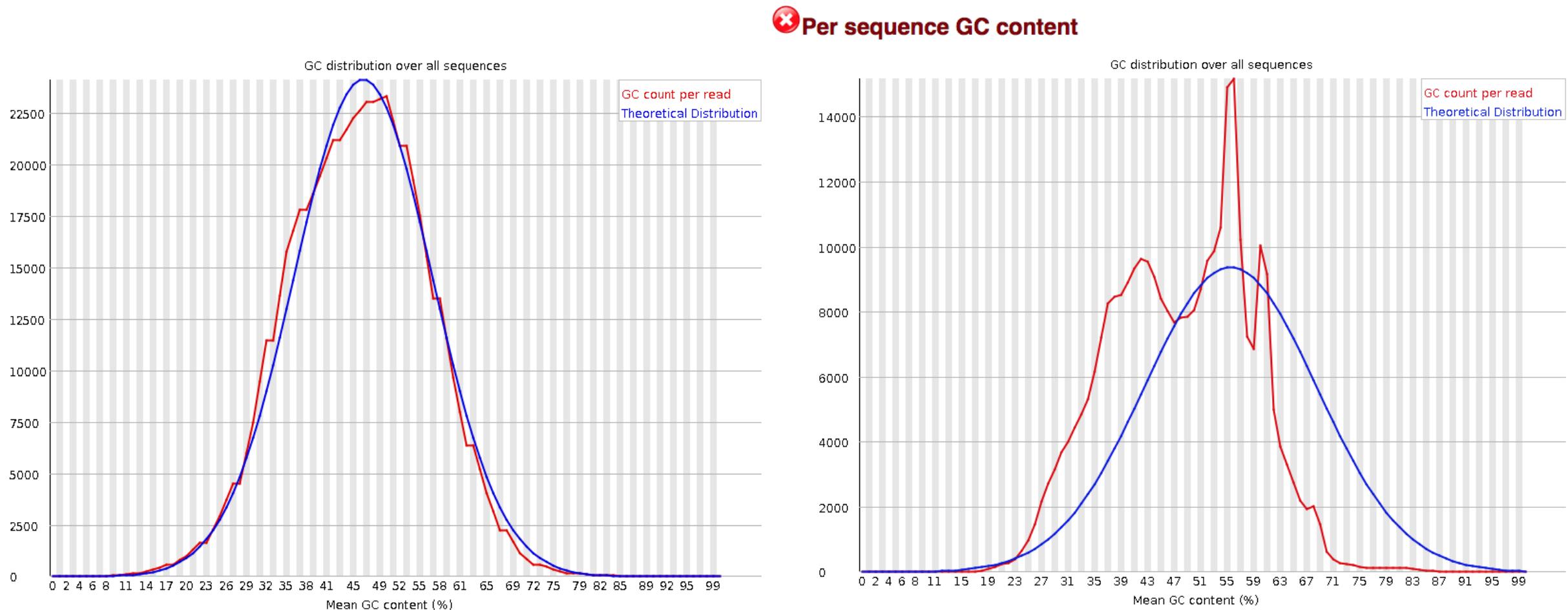
- **Decaimiento de la señal**
- **Desfase (*Phasing*)**: desfase acumulado entre el bloqueo del nucleótido que emite la señal y el siguiente que lo reemplaza



Contenido de bases (A, C, T, G)



Contenido de GC por secuencia



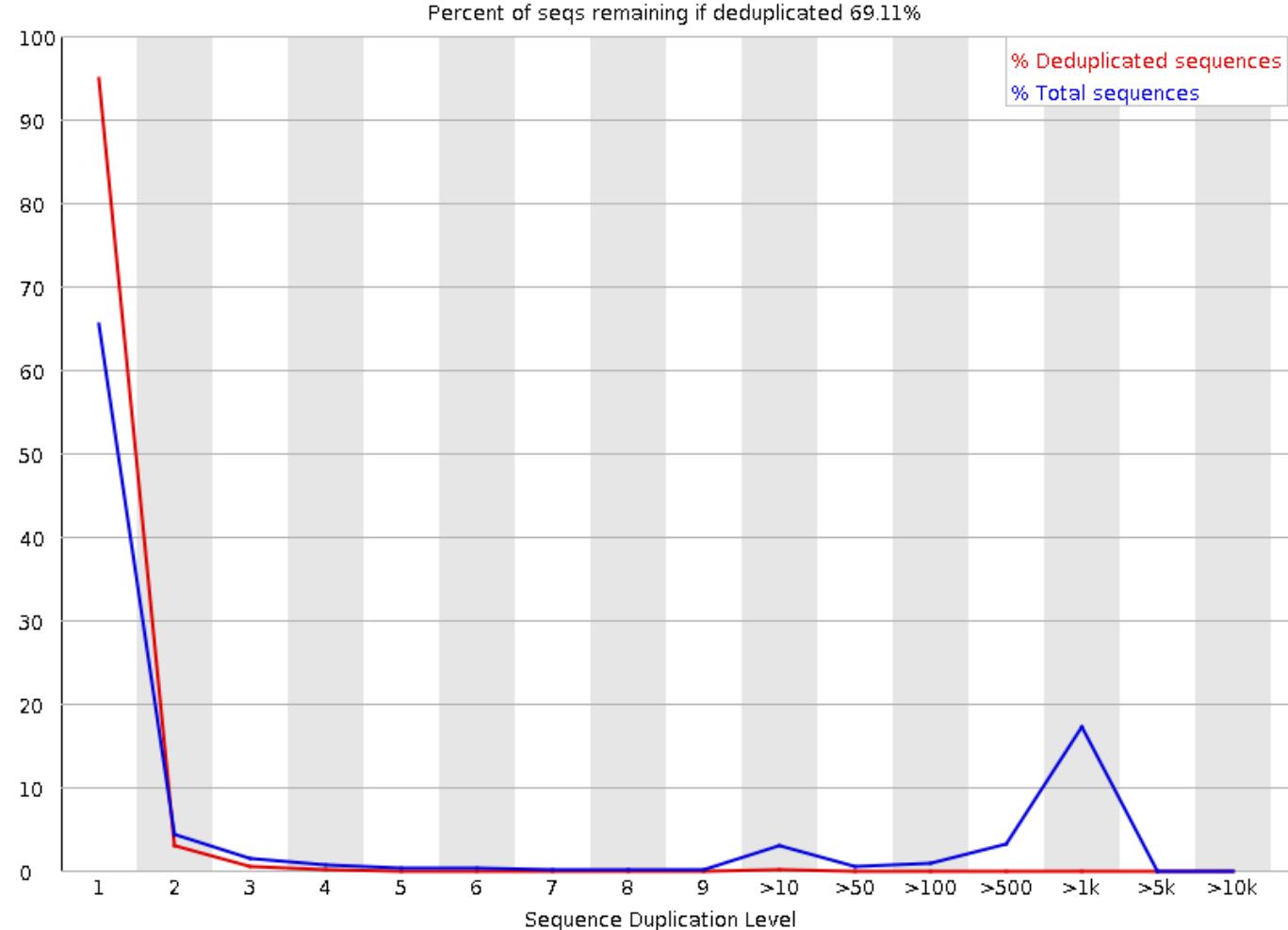
Picos altos y largos, es que hay secuencias
sobrerepresentadas, repetitivas o
contaminación

Número de secuencias duplicadas

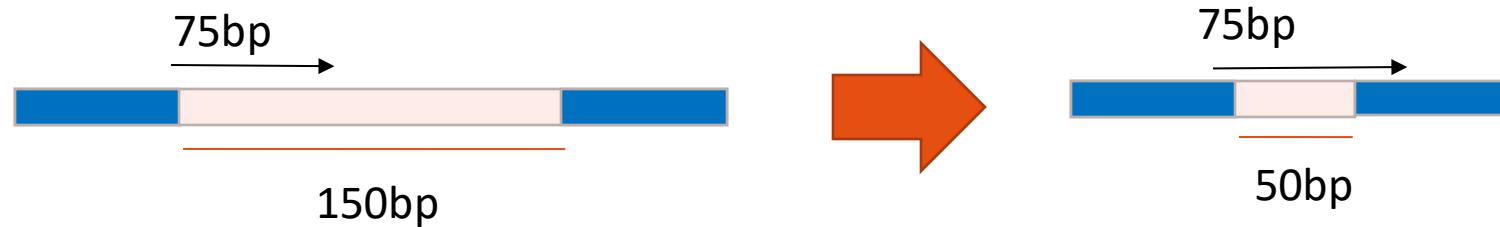
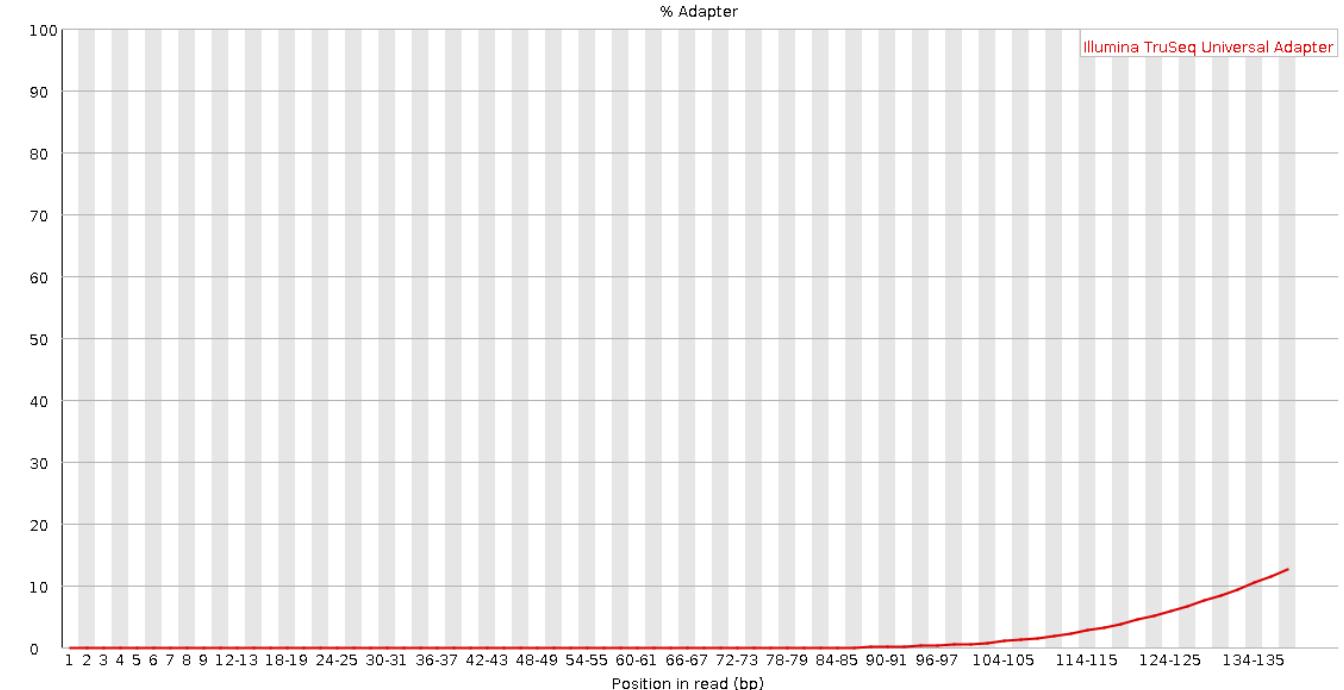
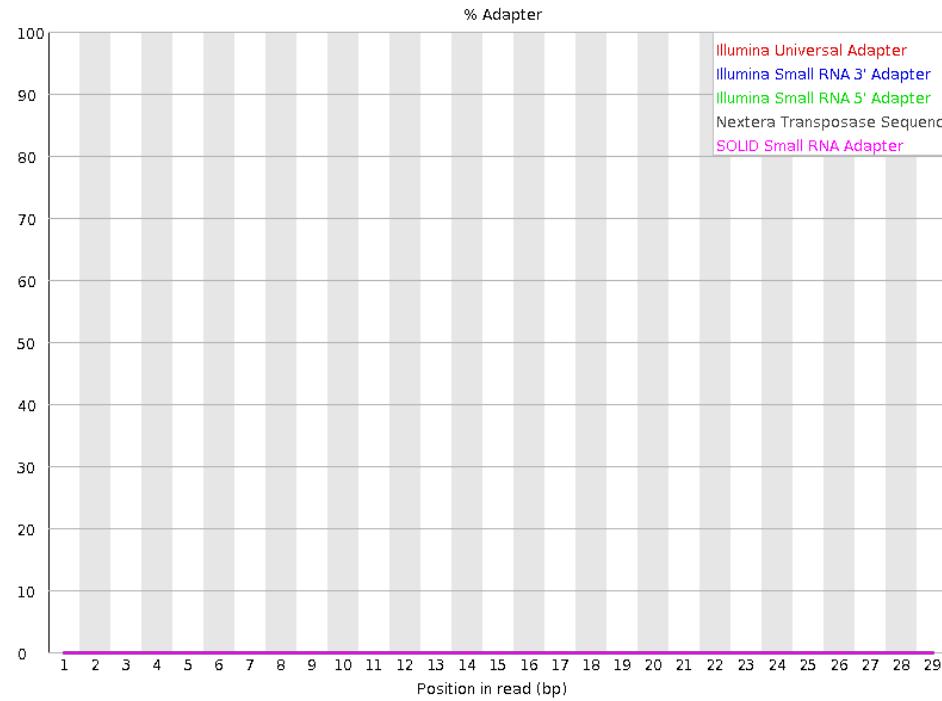
Al eliminar duplicados se eliminaría un gran porcentaje de datos, en este caso un 30% aprox ya que se quedaría con el 69,11%

- Origen de los duplicados
 - Técnico
 - Biológico
- Evaluar la complejidad de la librería
- Comparación muestral
- Valores sobreestimados
 - 100.000 secuencias
 - 50 bp de longitud

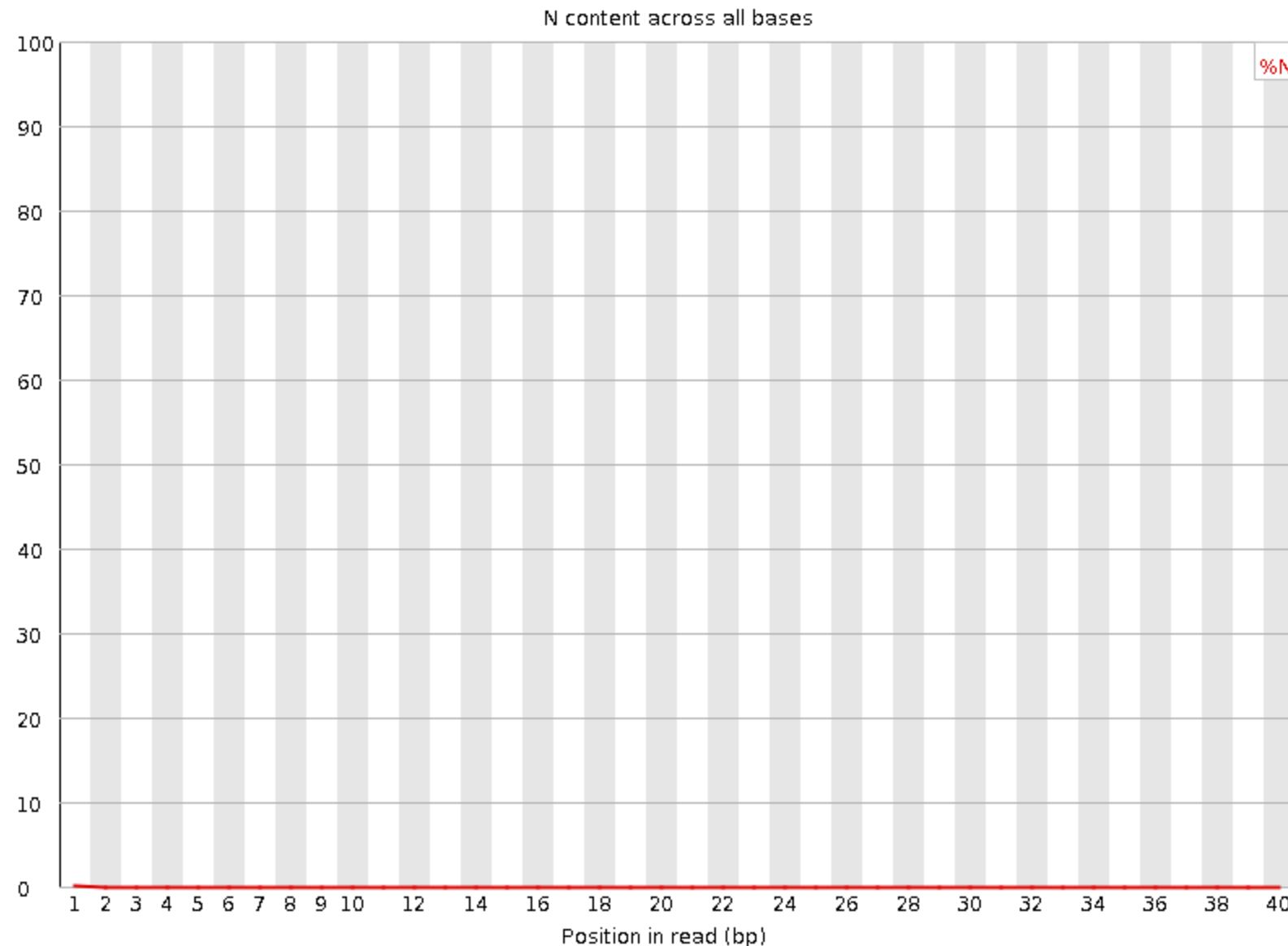
porque no mira todas las lecturas



Porcentaje de adaptadores



Porcentaje de N



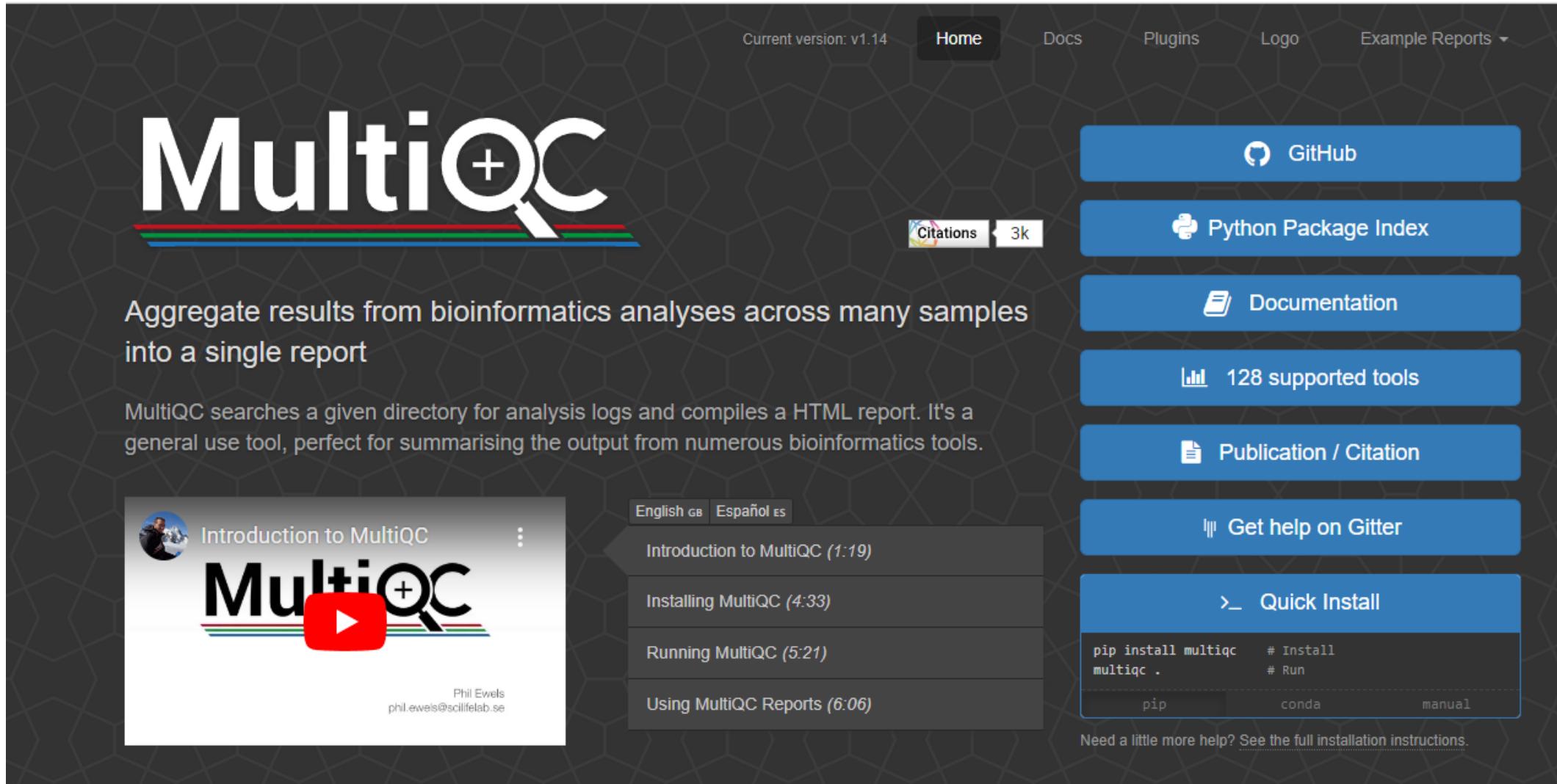
PRACTIQUEMOS



Running FASTQC

```
$ echo $(zcat SRR1552444.fastq.gz | wc -l)/4 | bc
27919481
$ fastqc SRR1552444.fastq.gz --o ..../Processed/01.Quality_Control/
Started analysis of SRR1552444.fastq.gz
Approx 5% complete for SRR1552444.fastq.gz
Approx 10% complete for SRR1552444.fastq.gz
Approx 15% complete for SRR1552444.fastq.gz
Approx 20% complete for SRR1552444.fastq.gz
Approx 25% complete for SRR1552444.fastq.gz
Approx 30% complete for SRR1552444.fastq.gz
Approx 35% complete for SRR1552444.fastq.gz
Approx 40% complete for SRR1552444.fastq.gz
Approx 45% complete for SRR1552444.fastq.gz
Approx 50% complete for SRR1552444.fastq.gz
Approx 55% complete for SRR1552444.fastq.gz
Approx 60% complete for SRR1552444.fastq.gz
Approx 65% complete for SRR1552444.fastq.gz
Approx 70% complete for SRR1552444.fastq.gz
Approx 75% complete for SRR1552444.fastq.gz
Approx 80% complete for SRR1552444.fastq.gz
Approx 85% complete for SRR1552444.fastq.gz
Approx 90% complete for SRR1552444.fastq.gz
Approx 95% complete for SRR1552444.fastq.gz
Analysis complete for SRR1552444.fastq.gz
$ firefox ..../Processed/01.Quality_Control/SRR1552444_fastqc.html
```

Inspeccionando los FASTQ: MultiQC



The screenshot shows the official MultiQC website at <https://multiqc.info/>. The page features a large "MultiQC" logo with a magnifying glass icon. Below it, a subtitle reads: "Aggregate results from bioinformatics analyses across many samples into a single report". A brief description follows: "MultiQC searches a given directory for analysis logs and compiles a HTML report. It's a general use tool, perfect for summarising the output from numerous bioinformatics tools." To the right, there's a navigation bar with links to "Home", "Docs", "Plugins", "Logo", and "Example Reports". Below the navigation are several blue call-to-action buttons: "GitHub", "Python Package Index", "Documentation", "128 supported tools", "Publication / Citation", "Get help on Gitter", and "Quick Install". The "Quick Install" section contains the command: `pip install multiqc # Install
multiqc . # Run`. At the bottom, there are links for "pip", "conda", and "manual". On the left, there's a video player showing a YouTube video titled "Introduction to MultiQC" by Phil Ewels, with a duration of 6:06.

Current version: v1.14

Home Docs Plugins Logo Example Reports

 GitHub

 Python Package Index

 Documentation

 128 supported tools

 Publication / Citation

 Get help on Gitter

 Quick Install

```
pip install multiqc # Install  
multiqc . # Run
```

Need a little more help? See the full installation instructions.

English GB Español ES

Introduction to MultiQC (1:19)

Installing MultiQC (4:33)

Running MultiQC (5:21)

Using MultiQC Reports (6:06)

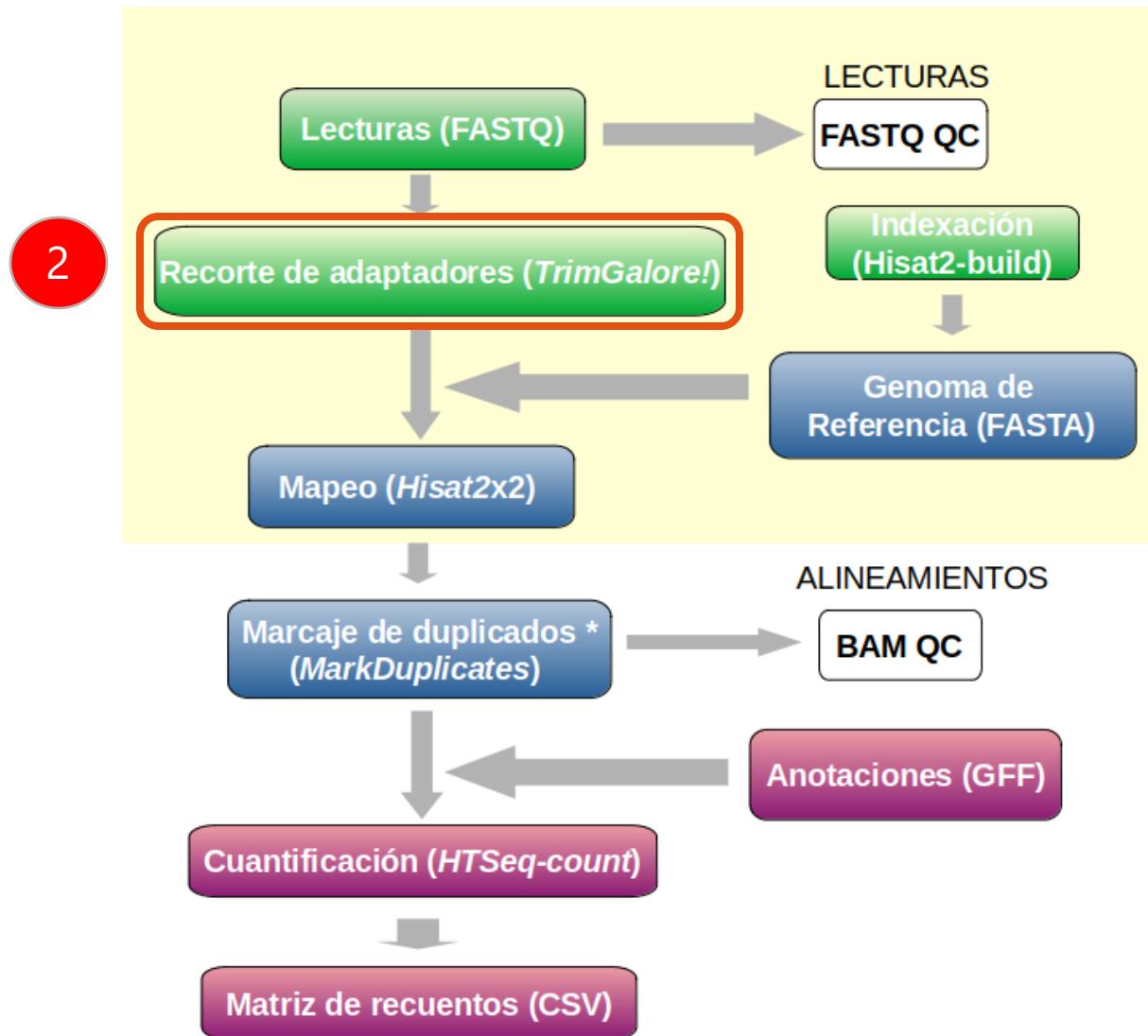
Introduction to MultiQC

MultiQC

Phil Ewels
phil.ewels@scilifelab.se

<https://multiqc.info/>

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



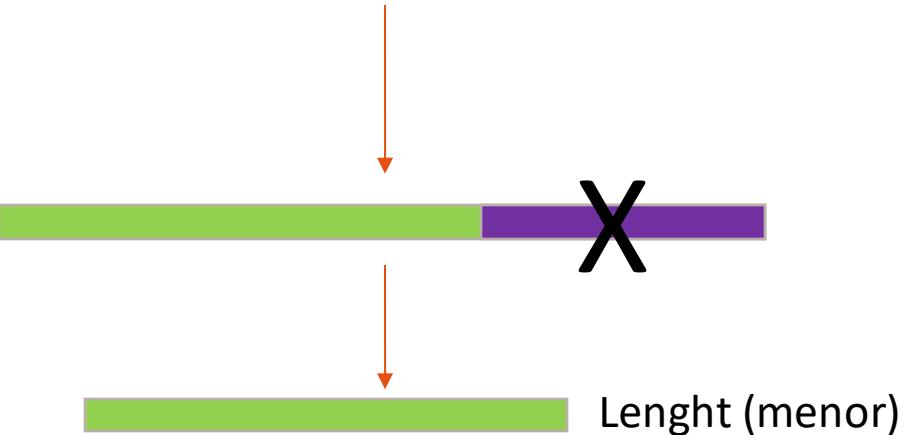
Filtering vs Trimming

Filtrar/eliminar secuencias completas

- Puntuación de calidad baja (Phred < 20)
- Demasiado cortas (Length < 20bp)
- Con demasiadas bases ambiguas (N)
- Lecturas impares en el caso de un protocolo de lecturas emparejadas

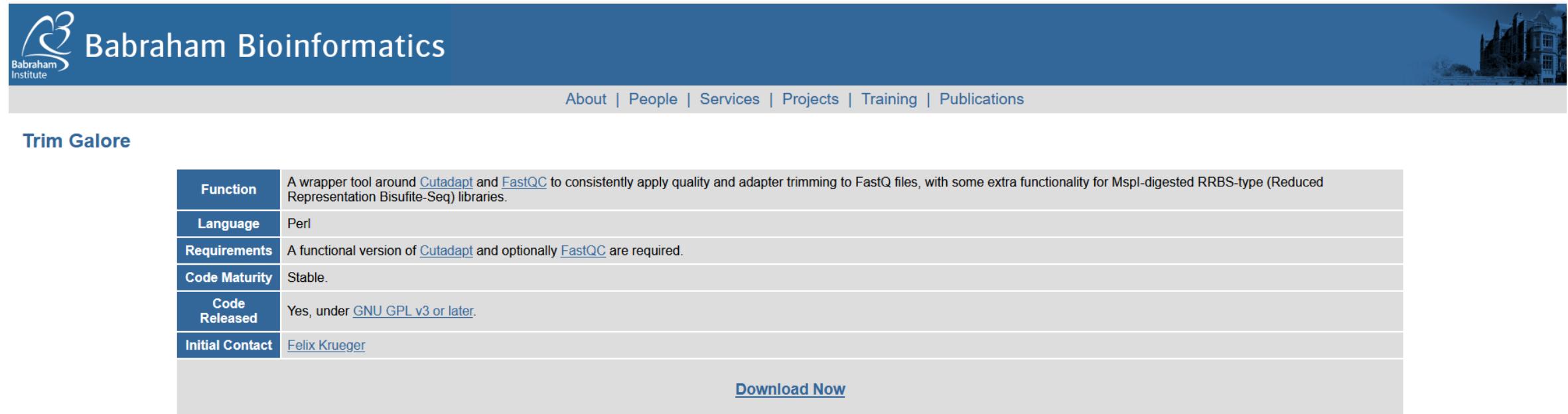
Recortar parcialmente las secuencias

- Regiones con bajo nivel de calidad
- Eliminación de adaptadores



Calidad adaptativa y recorte de adaptadores con TrimGalore!

FastX, Prinseq, TagCleaner, Trimmomatic, Cutadapt, TrimGalore!...



The image shows a screenshot of the Babraham Bioinformatics website. The header features the Babraham Institute logo (a stylized 'B' and '3') and the text "Babraham Bioinformatics". Below the header is a navigation bar with links: About | People | Services | Projects | Training | Publications. The main content area is titled "Trim Galore" and contains a table with the following data:

Function	A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for Mspl-digested RRBS-type (Reduced Representation Bisulfite-Seq) libraries.
Language	Perl
Requirements	A functional version of Cutadapt and optionally FastQC are required.
Code Maturity	Stable.
Code Released	Yes, under GNU GPL v3 or later .
Initial Contact	Felix Krueger

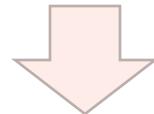
Below the table is a button labeled "Download Now".

https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

<https://github.com/FelixKrueger/TrimGalore>

Calidad adaptativa y recorte de adaptadores con TrimGalore!

Las bases que presenten una baja calidad se recortan desde el extremo 3' de las lecturas (Phred = 20)



Eliminan los adaptadores desde el extremo 3'

Cutadapt

42 40 26 27 8 7 11 4 2 3

Umbral = 10

Restar

32 30 16 17 -2 -3 1 -6 -8 -7

Sumar (3')

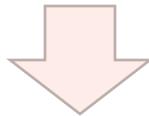
(70) (38) 8 -8 -25 -23 -20 -21 -15 -7

Valor mínimo

42 40 26 27

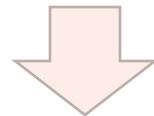
Calidad adaptativa y recorte de adaptadores con TrimGalore!

Las bases que presenten una baja calidad se recortan desde el extremo 3' de las lecturas (Phred = 20)



Eliminan los adaptadores desde el extremo 3'

Cutadapt

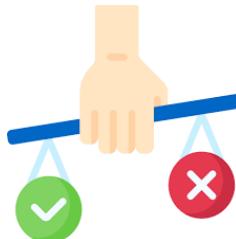


Filtrado de las lecturas recortadas en función de su secuencia (predeterminado 20bp)

-- paired/ --retain_unpaired

Validación y eliminación por pares

Calidad adaptativa y recorte de adaptadores con TrimGalore!



- Actualmente **NO** hay un consenso
- ***Trade-off*** entre tener una buena calidad de las lecturas manteniendo un número suficiente de secuencias para analizar

An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis

Cristian Del Fabbro¹*, Simone Scalabrin²*, Michele Morgante¹, Federico M. Giorgi^{1,3}*

¹ Institute of Applied Genomics, Udine, Italy, ² IGA Technology Services, Udine, Italy, ³ Center for Computational Biology and Bioinformatics, Columbia University, New York, New York, United States of America

frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE

published: 31 January 2014
doi: 10.3389/fgene.2014.00013

On the optimal trimming of high-throughput mRNA sequence data

Matthew D. MacManes^{1,2}*

¹ Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham, NH, USA

² Hubbard Center for Genome Studies, Durham, NH, USA

Calidad adaptativa y recorte de adaptadores con TrimGalore!



- Actualmente **NO** hay un consenso
- ***Trade-off*** entre tener una buena calidad de las lecturas manteniendo un número suficiente de secuencias para analizar

An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis

Cristian Del Fabbro¹*, Simone Scalabrin²*, Michele Morgante¹, Federico M. Giorgi^{1,3*}

Inicialmente siempre realizar un *trimming/filtering poco restrictivo* y evaluar su impacto en los informes FASTQC

On the optimal trimming of high-throughput mRNA sequence data

Matthew D. MacManes^{1,2*}

¹ Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham, NH, USA

² Hubbard Center for Genome Studies, Durham, NH, USA

PRACTIQUEMOS



Instalación -> TRIM_GALORE

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

```
# Install Trim Galore
curl -fsSL https://github.com/FelixKrueger/TrimGalore/archive/0.6.10.tar.gz -o trim_galore.tar.gz
tar xvzf trim_galore.tar.gz
cd TrimGalore-0.6.10/
```

```
(05MBIF) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr TrimGalore-0.6.10]$ ls -lh
total 212K
-rw-r--r-- 1 UNIVERSIDADVIU\paula.soler UNIVERSIDADVIU\domain users 22K Feb  2 12:25 CHANGELOG.md
drwxr-xr-x 3 UNIVERSIDADVIU\paula.soler UNIVERSIDADVIU\domain users 4.0K Feb  2 12:25 Docs
-rw-r--r-- 1 UNIVERSIDADVIU\paula.soler UNIVERSIDADVIU\domain users 35K Feb  2 12:25 LICENSE
-rw-r--r-- 1 UNIVERSIDADVIU\paula.soler UNIVERSIDADVIU\domain users 1.8K Feb  2 12:25 README.md
drwxr-xr-x 2 UNIVERSIDADVIU\paula.soler UNIVERSIDADVIU\domain users 4.0K Feb  2 12:25 test_files
-rwxr-xr-x 1 UNIVERSIDADVIU\paula.soler UNIVERSIDADVIU\domain users 139K Feb  2 12:25 trim_galore
```

```
# Run Trim Galore
./trim_galore
```

```
(05MBIF) [UNIVERSIDADVIU\paula.soler@a-3edhijmqygwxr TrimGalore-0.6.10]$ export PATH=$PATH:$PWD
```

Running TRIM_GALORE

```
(05MBIF)[Raw]$ trim_galore SRR1552444.fastq.gz -o ../Processed/02.Trimming/
```

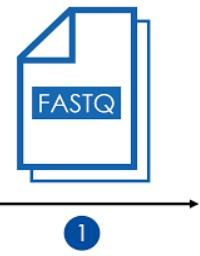
```
(05MBIF) [02.Trimming]$ ls -lh
```

```
-rw-r--r-- 1 UNIVERSIDADADVIU\paula.soler UNIVERSIDADADVIU\domain users 3.8K Jul 25 16:12 SRR1552444.fastq.gz_trimming_report.txt
```

```
-rw-r--r-- 1 UNIVERSIDADADVIU\paula.soler UNIVERSIDADADVIU\domain users 660K Jul 23 19:34 SRR1552444_trimmed_fastqc.html
```

```
-rw-r--r-- 1 UNIVERSIDADADVIU\paula.soler UNIVERSIDADADVIU\domain users 445K Jul 23 19:34 SRR1552444_trimmed_fastqc.zip
```

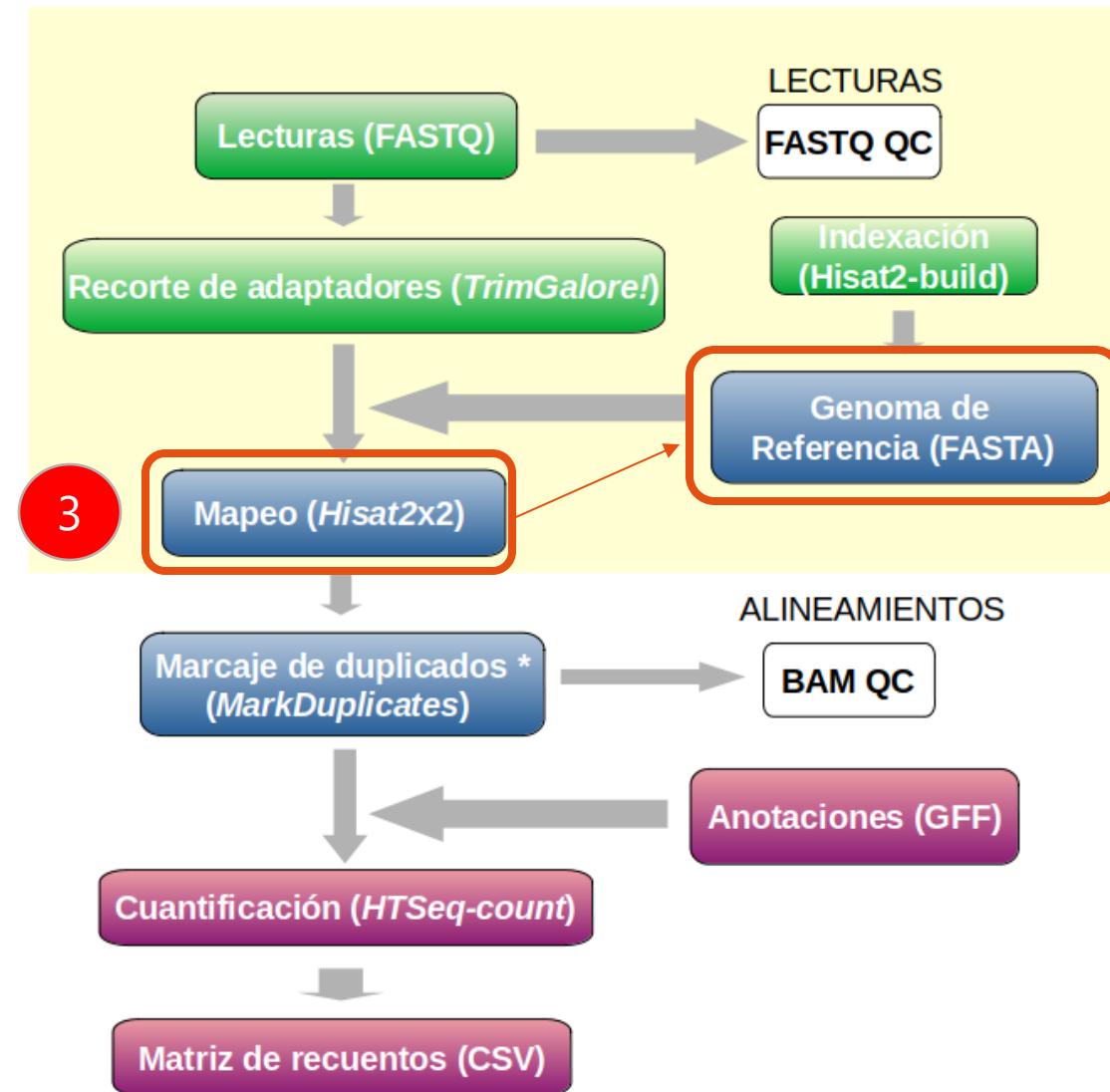
```
-rw-r--r-- 1 UNIVERSIDADADVIU\paula.soler UNIVERSIDADADVIU\domain users 2.2G Jul 25 16:12 SRR1552444_trimmed.fq.gz
```



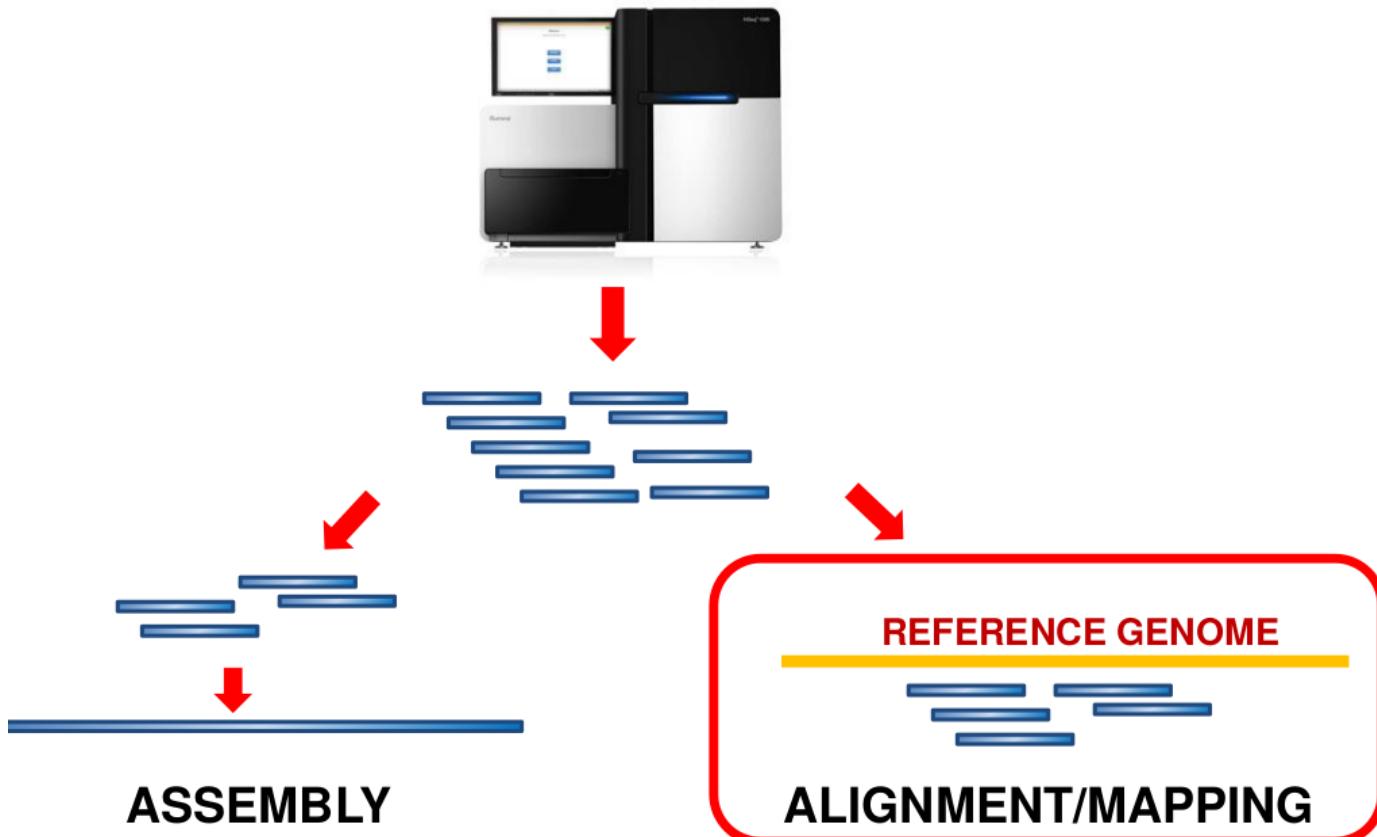
Revisar de nuevo los informes FASTQC tras el *trimming*



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



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



Alineamiento de Lecturas

- **Objetivo:** Determinar la correspondencia exacta base a base en el genoma.
 - La lectura X probablemente se originó en el cromosoma 1, posiciones 123 a 140.

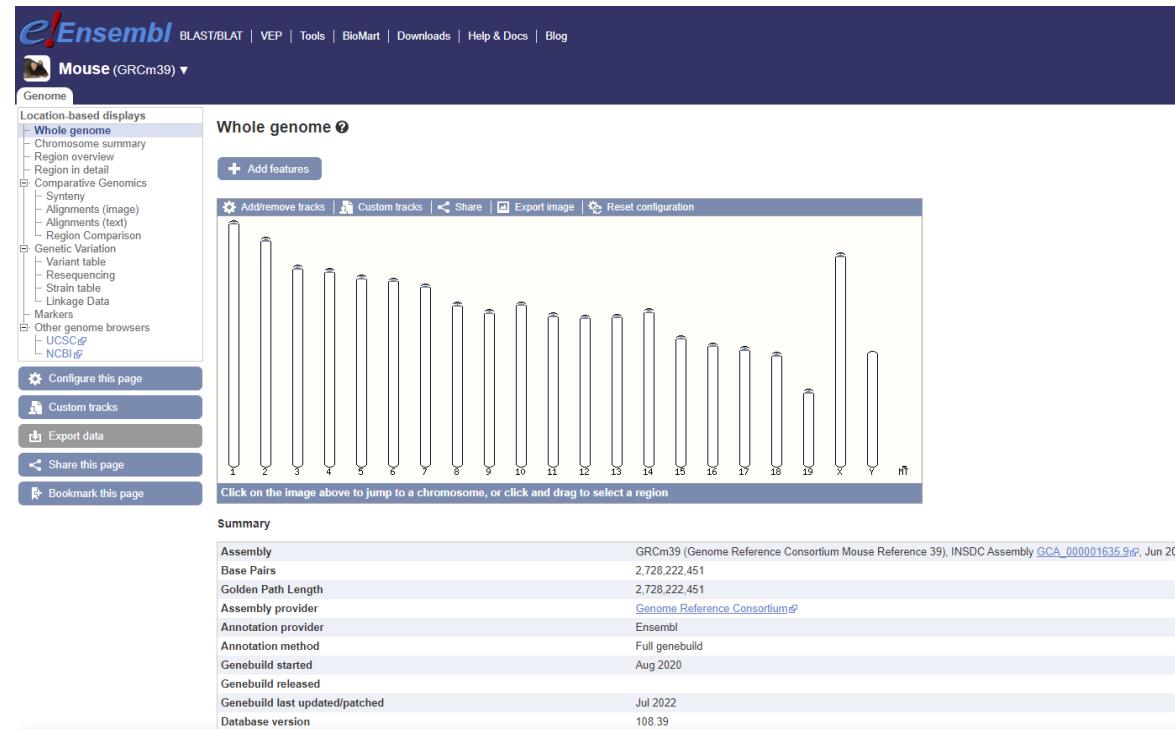
Mapeo de Lecturas

- **Objetivo:** Determinar de dónde proviene la lectura.
- **Importancia de la Correspondencia Exacta:** No es necesario conocer la alineación exacta entre la lectura y su origen.

Genoma de referencia de *Mus Musculus*



- Un conjunto de secuencias representativas de un genoma
- Útil como estándar para comparar y analizar información genética
- Existen múltiples versiones



http://www.ensembl.org/Mus_musculus/Location/Genome

Genoma de referencia indexado

- Paso previo al alineamiento
- Mejora la eficiencia computacional
- Son específicos según el alineador empleado



Download

Please cite:

Kim, D., Paggi, J.M., Park, C. et al. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* 37, 907–915 (2019). <https://doi.org/10.1038/s41587-019-0201-4>

- Index
 - H. sapiens
 - M. musculus
 - R. norvegicus
 - D. melanogaster
 - C. elegans
 - S. cerevisiae
- Binaries

Search

Main

About

Manual

HISAT-3N

Download

HowTo

Links

Funding

This work was supported in part by the National Human Genome Research Institute under grants R01-HG006102 and

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

Genoma de referencia indexado

M. musculus

- GRCm38

genome <https://cloud.biohpc.swmed.edu/index.php/s/grcm38/download>

genome_snp <https://cloud.biohpc.swmed.edu/index.php/s/grcm38.snp/download>

genome_tran https://cloud.biohpc.swmed.edu/index.php/s/grcm38_tran/download

genome_snp_tran https://cloud.biohpc.swmed.edu/index.php/s/grcm38.snp_tran/download

- UCSC mm10

genome https://genome-idx.s3.amazonaws.com/hisat/mm10_genome.tar.gz

Genoma de referencia previamente indexado de *Mus Musculus*

```
(05MBIF) [UNIVERSIDADADVIU\paula.soler@a-3edhijmqygwxr Reference_genome]$ ls  
mm10_genome.tar.gz
```

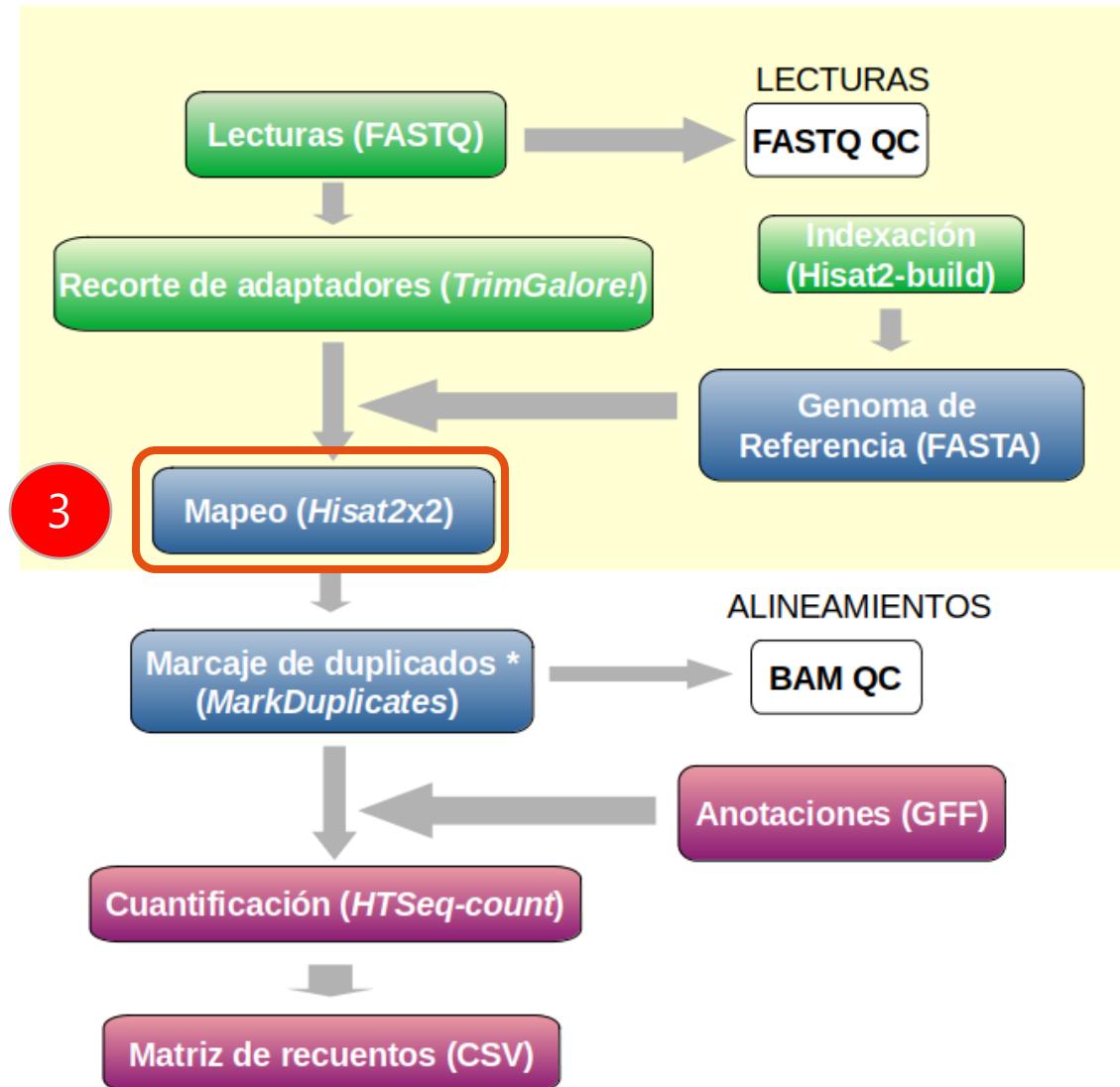
```
(05MBIF) [UNIVERSIDADADVIU\paula.soler@a-3edhijmqygwxr Reference_genome]$ tar -xvf mm10_genome.tar.gz  
mm10/  
mm10/genome.8.ht2  
mm10/genome.5.ht2  
mm10/make_mm10.sh  
mm10/genome.7.ht2  
mm10/genome.6.ht2  
mm10/genome.4.ht2  
mm10/genome.3.ht2  
mm10/genome.1.ht2  
mm10/genome.2.ht2
```

```
(05MBIF) [UNIVERSIDADADVIU\paula.soler@a-3edhijmqygwxr Reference_genome]$ ls  
mm10 mm10_genome.tar.gz
```

```
(05MBIF) [UNIVERSIDADADVIU\paula.soler@a-3edhijmqygwxr Reference_genome]$ cd mm10
```

```
(05MBIF) [UNIVERSIDADADVIU\paula.soler@a-3edhijmqygwxr mm10]$ ls  
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
```

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



3
3
Retos

1. Las lecturas contienen variaciones genómicas y errores de secuenciación.
2. Los genomas incluyen secuencias repetitivas e intrones.
3. Coste computacional : millones de lecturas



viu

Universidad
Internacional
de Valencia

universidadviu.com

De:
 Planeta Formación y Universidades