

II. Obtención de un dataset de Variantes a partir de datos de secuenciación - WGS

Maria José Ruiz, Lorena Lorenzo, Enrico Bazzicalupo

Buenas Practicas para un proyecto de Bioinformatica

Crear una **carpeta** para el proyecto

Utilizar **control de cambios** en la carpeta para volver a versiones anteriores

<https://rfortherestofus.com/2021/02/how-to-use-git-github-with-r>

Conservar los **datos en multiples copias** en sitios diferentes (discos duros, nubes, servidores)

Scripts bien comentados y claros - como y porque hacemos las cosas

Tener un registro del orden y en que hacemos las cosas - **markdown como un cuaderno** / material y métodos

Registro de las versiones de los programas que vamos a utilizar (util para métodos y repetibilidad)

Con mas experiencia: crear **entornos** para análisis o proyectos (conda, docker)

LECTURAS BRUTAS

```
GCTTCTGTGG
GGCTCACGTA
ACGTAAGAGG
ATAAAAGTTAC
GGACACAGAG
GCTTCTGTGG
```

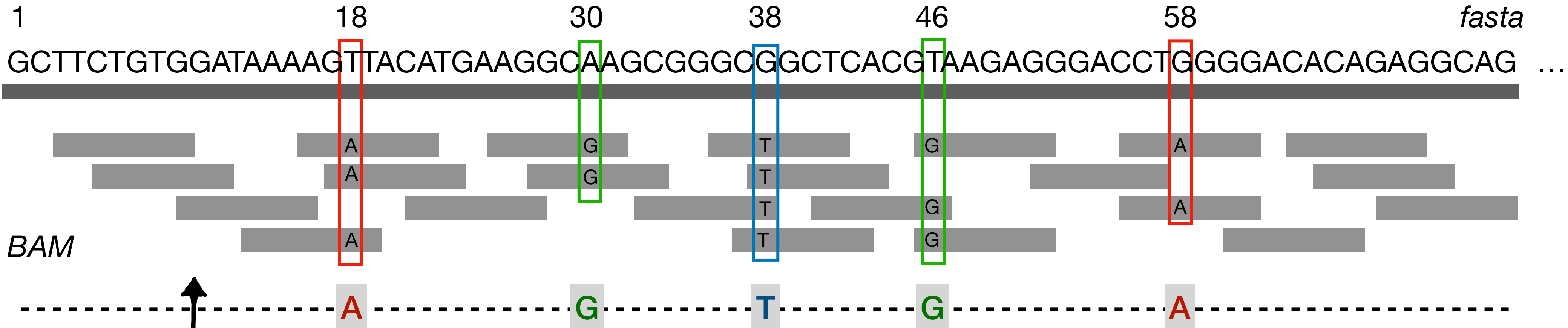
Fastq

fastp

**Filtrado y
trimming de
lecturas**

```
GCTTCTGTGG
GGCTCA
ACGTAAGAGG
AAGTTAC
GGACACAGAG
```

Fastq



**Alineamiento
de lecturas**

BWA-
MEM

**Llamada de
variantes**

GATK

**Filtrado de
variantes**

BCFtools
BEDtools
VCFtools
GATK

18	T	A
30	A	G
38	G	T
46	T	G
58	G	A

VCF

**DATASET
DE VARIANTES
FINAL**

18	T	A
38	G	T
46	T	G

VCF

LECTURAS
BRUTAS

GCTTCTGTGG

GGCTCACGTA

ACGTAAGAGG

ATAAAAGTTAC

GGACACAGAG

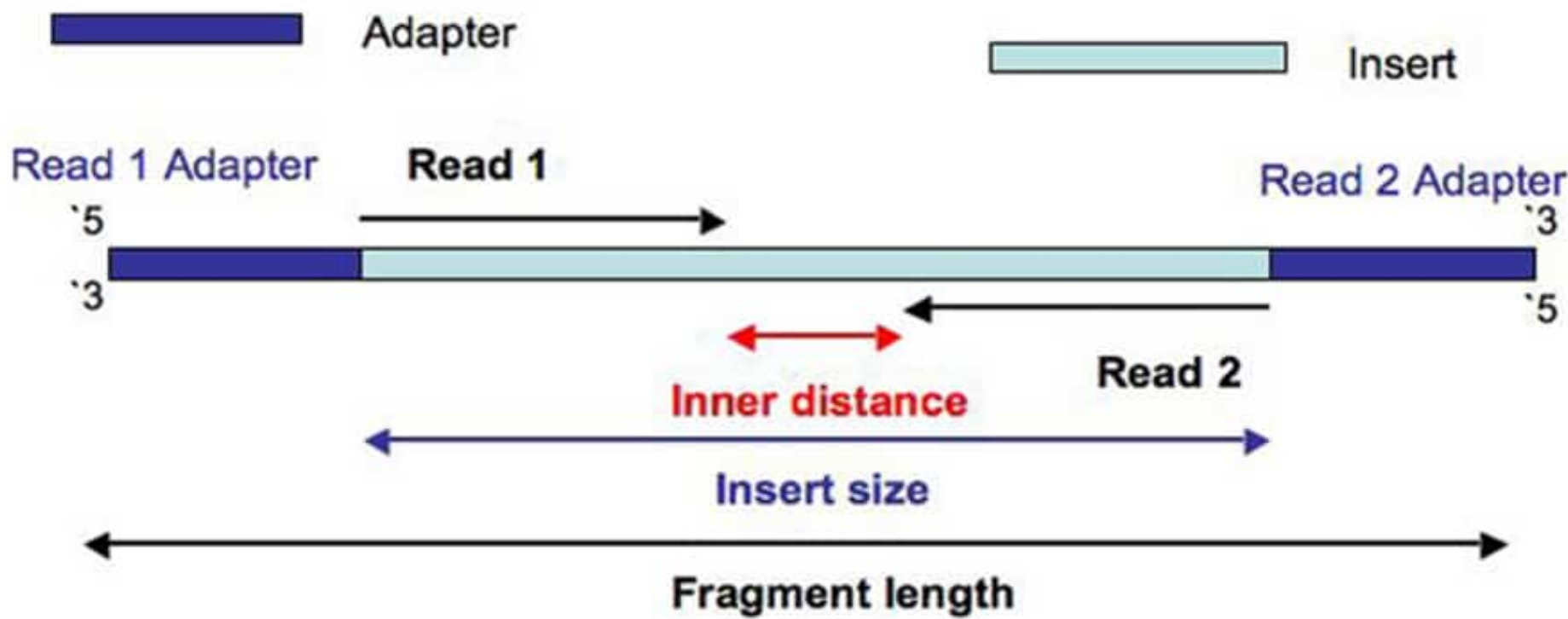
GCTTCTGTGG

Fastq

Secuenciación con Illumina

LECTURAS BRUTAS

```
GCTTCTGTGG
GGCTCACGTA
ACGTAAGAGG
ATAAAAGTTAC
GGACACAGAG
GCTTCTGTGG
```



Información sobre **Lectura 1** y **Lectura 2** obtenida del secuenciador y guardada en formato **FASTQ**

```
@ERR000589.41 EAS139_45:5:1:2:111/1
CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCCAGGGAACATCTTGTCAT
+
3IIIIIIIIIIII>1IIIF9BG08E00I%IG+&?(4)%00646.C1#&(
```

```
@ERR000589.42 EAS139_45:5:1:2:1293/1
AGTTGTTAAAATCCAAGCCAATTAAGATAGTCTTATCTTTTAAAAGAAAT
+
IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I
```

El formato FASTQ

(Fasta + Quality)

Cada secuencia está dividida en 4 filas en el FASTQ:

Primera línea:

- identificador de la secuencia
- empieza siempre con @
- contiene información sobre la secuencia

Segunda línea:

- secuencia de nucleótidos

Tercera línea:

- un + separador

Cuarta línea:

- valores de calidad de llamada de cada nucleótido

Puntuación de Calidad de los Nucleótidos

```
@ERR000589.41 EAS139_45:5:1:2:111/1
CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCAGGGAACATCTTGTCAT
+
3IIIIIIIIIIII>1IIIFF9BG08E00I%IG+&?(4)%00646.C1#&(

@ERR000589.42 EAS139_45:5:1:2:1293/1
AGTTGTTAAAATCCAAGCCAATTAAGATAGTCTTATCTTTTTTAAAAGAAAT
+
IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I
```

A cada nucleótido corresponde una puntuación de calidad asignada por la maquina de secuenciación

Codificada por uno de 41 símbolos

Q-score o ASCII una idea numerica de la calidad (Q-score mas usado ahora)

Como se traduce en Q-score en probabilidad de error?

$Q = -10\log_{10}(e)$

$Q_{10} = 1/10 = 10\%$

$Q_{20} = 1/100 = 1\%$

$Q_{30} = 1/1000 = 0.1\%$

Symbol	ASCII Code	Q-Score
!	33	0
"	34	1
#	35	2
\$	36	3
%	37	4
&	38	5
	39	6
(40	7
)	41	8
*	42	9
+	43	10
,	44	11
-	45	12
.	46	13
/	47	14
0	48	15
1	49	16
2	50	17
3	51	18
4	52	19
5	53	20

Symbol	ASCII Code	Q-Score
6	54	21
7	55	22
8	56	23
9	57	24
:	58	25
;	59	26
<	60	27
=	61	28
>	62	29
?	63	30
@	64	31
A	65	32
B	66	33
C	67	34
D	68	35
E	69	36
F	70	37
G	71	38
H	72	39
I	73	40

Controles de Calidad de un FASTQ

Qué buscamos en nuestro FASTQ?

- **Calidad** de las bases
- **Longitud** de la lectura
- **Sesgos** en la secuencia
- Presencia de **adaptadores**
- Secuencias **repetidas**
- Contenido de bases **G** y **C**
- Numero de bases sin llamar (**N**)
- Lecturas **duplicadas**

```
@ERR000589.41 EAS139_45:5:1:2:111/1
CTTTCCTCCCTGCTTTCCTGGCCCCACCATTTCCAGGGAACATCTTGTCAT
+
3IIIIIIIIIIII>1IIIFF9BG08E00I%IG+&?(4)%00646.C1#&(
```

```
@ERR000589.42 EAS139_45:5:1:2:1293/1
AGTTGTTAAAATCCAAGCCAATTAAGATAGTCTTATCTTTTAAAAGAAAT
+
IIIIIGII.AIIII=?I9G-/II=+I=4?761BA2C9I+5A711+&>1$/I
```

*Siempre puede haber
excepciones en base a lo
que estamos mirando:
DNA? RAD-seq? RNA?*

Herramienta Bioinformática



FastQC

Control de calidad **estandarizado** y **sencillo** para datos de secuenciación masiva.

Proporciona una **resumen gráfico** de la información más relevante para tomar decisiones sobre cómo tratar tus lecturas.

Resultados de FASTQC



Basic Statistics

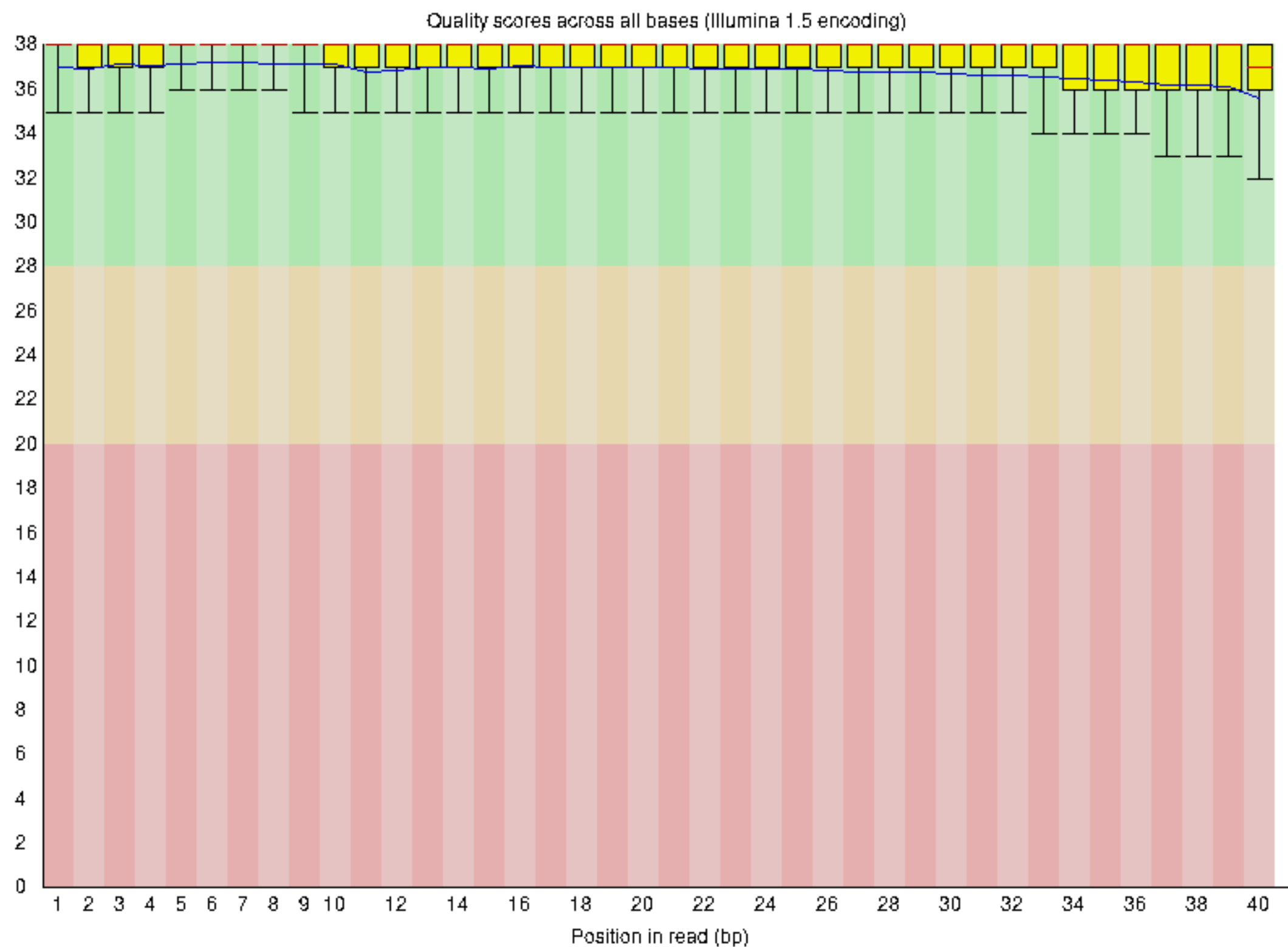
Measure	Value
Filename	RNA-Seq.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	100000
Total Bases	4 Mbp
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

Resultados de FASTQC

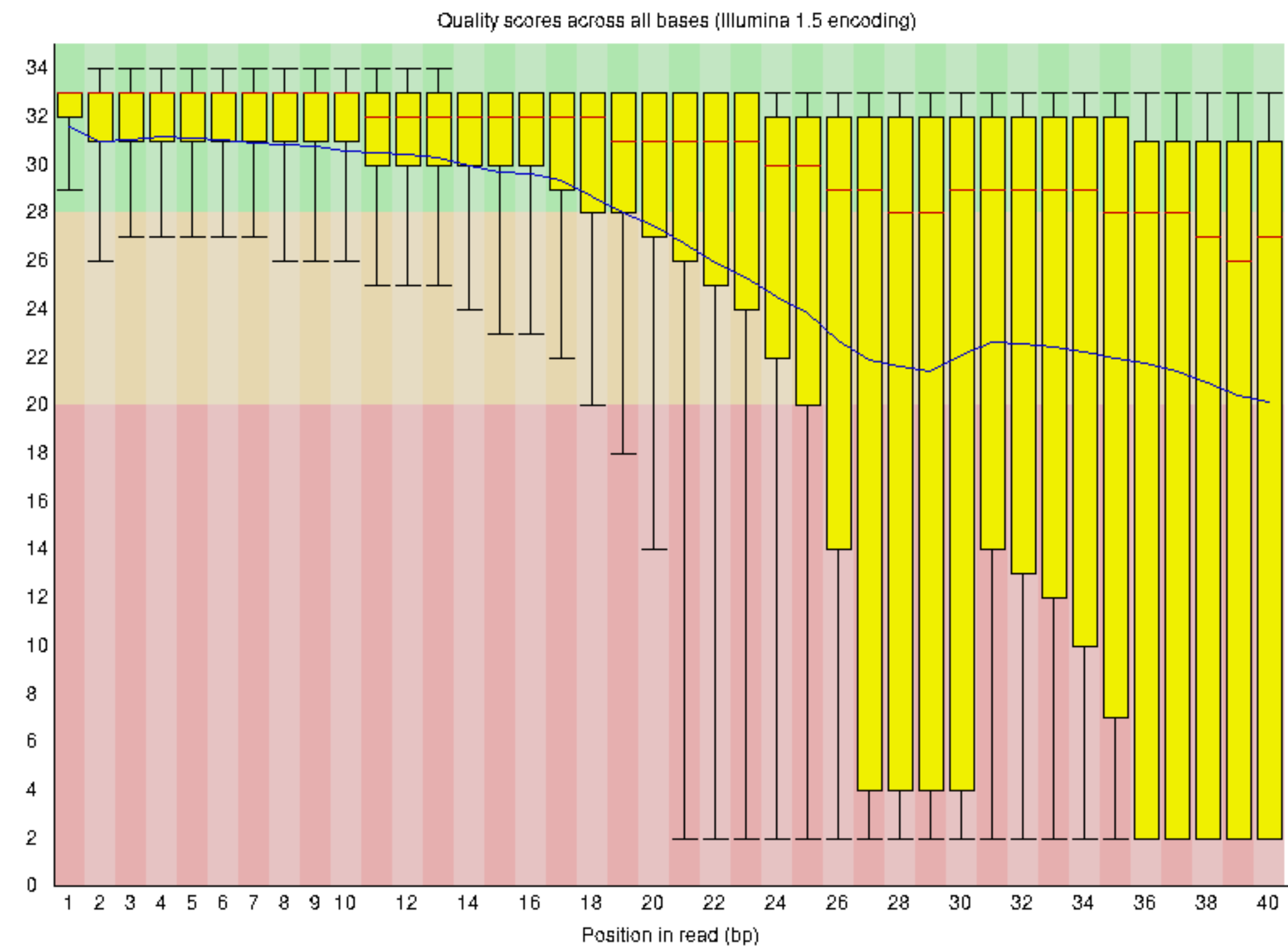
- Calidad de las bases



Per base sequence quality



Per base sequence quality

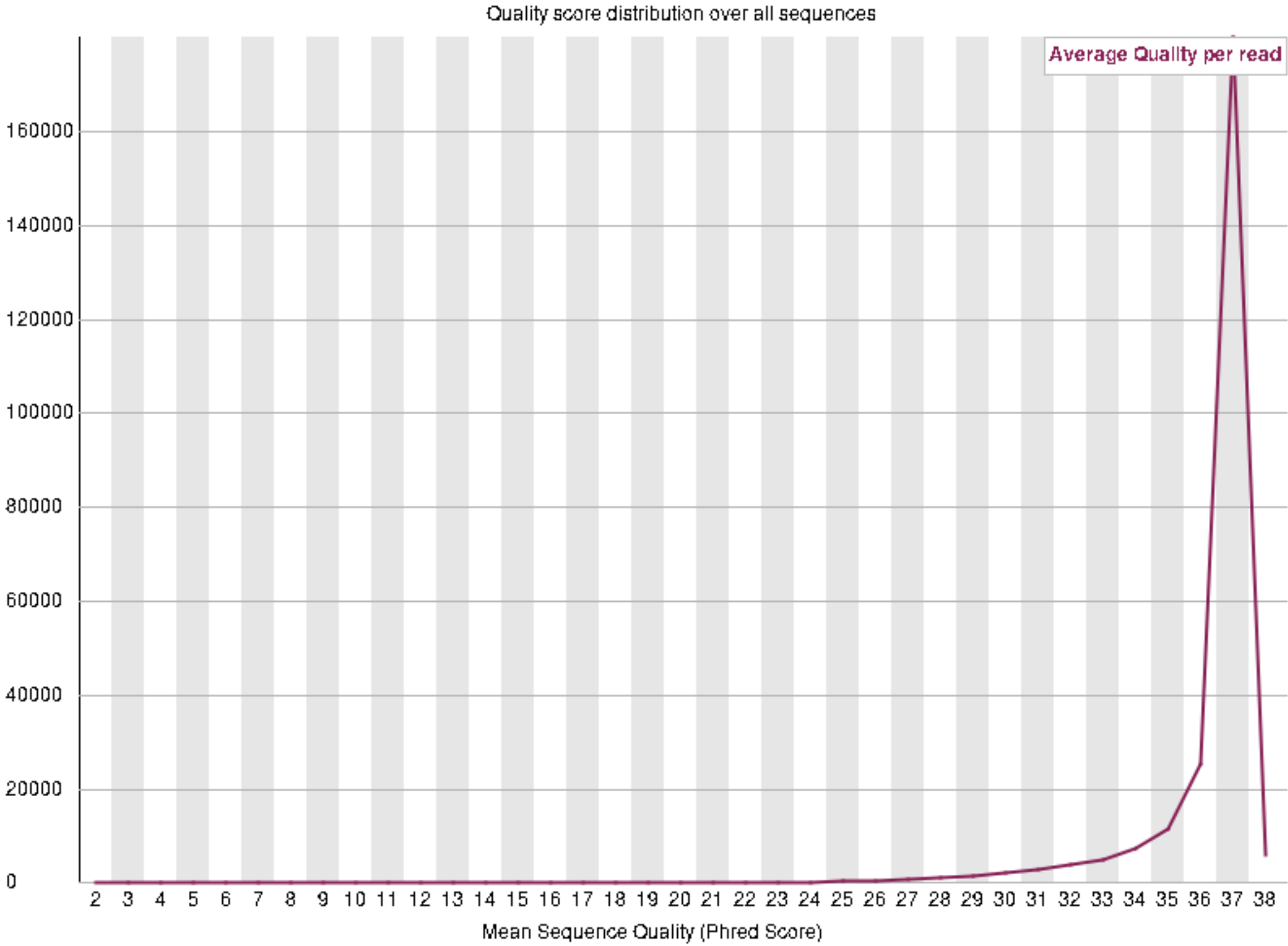


Resultados de FASTQC

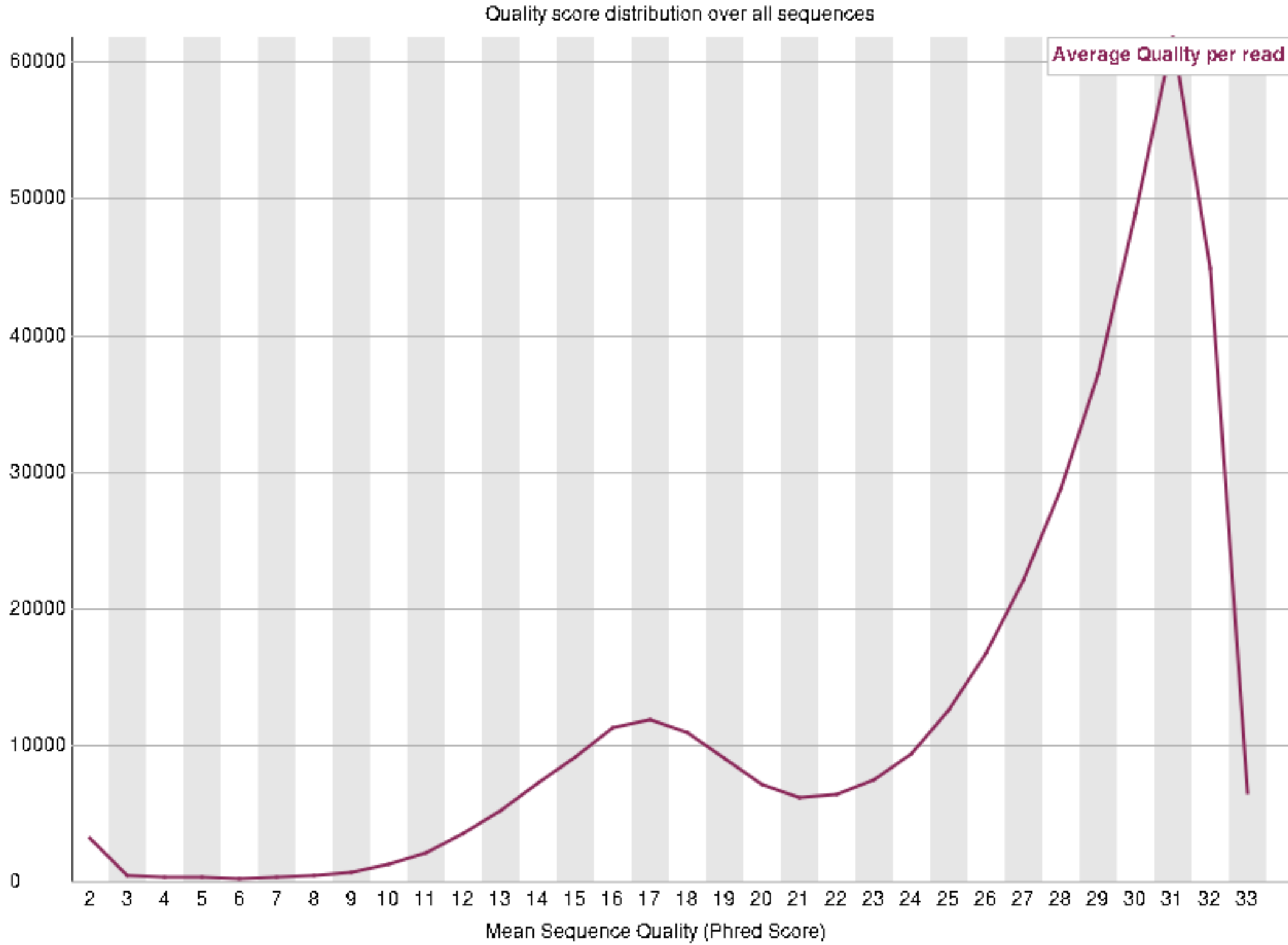
- Calidad de las bases



Per sequence quality scores

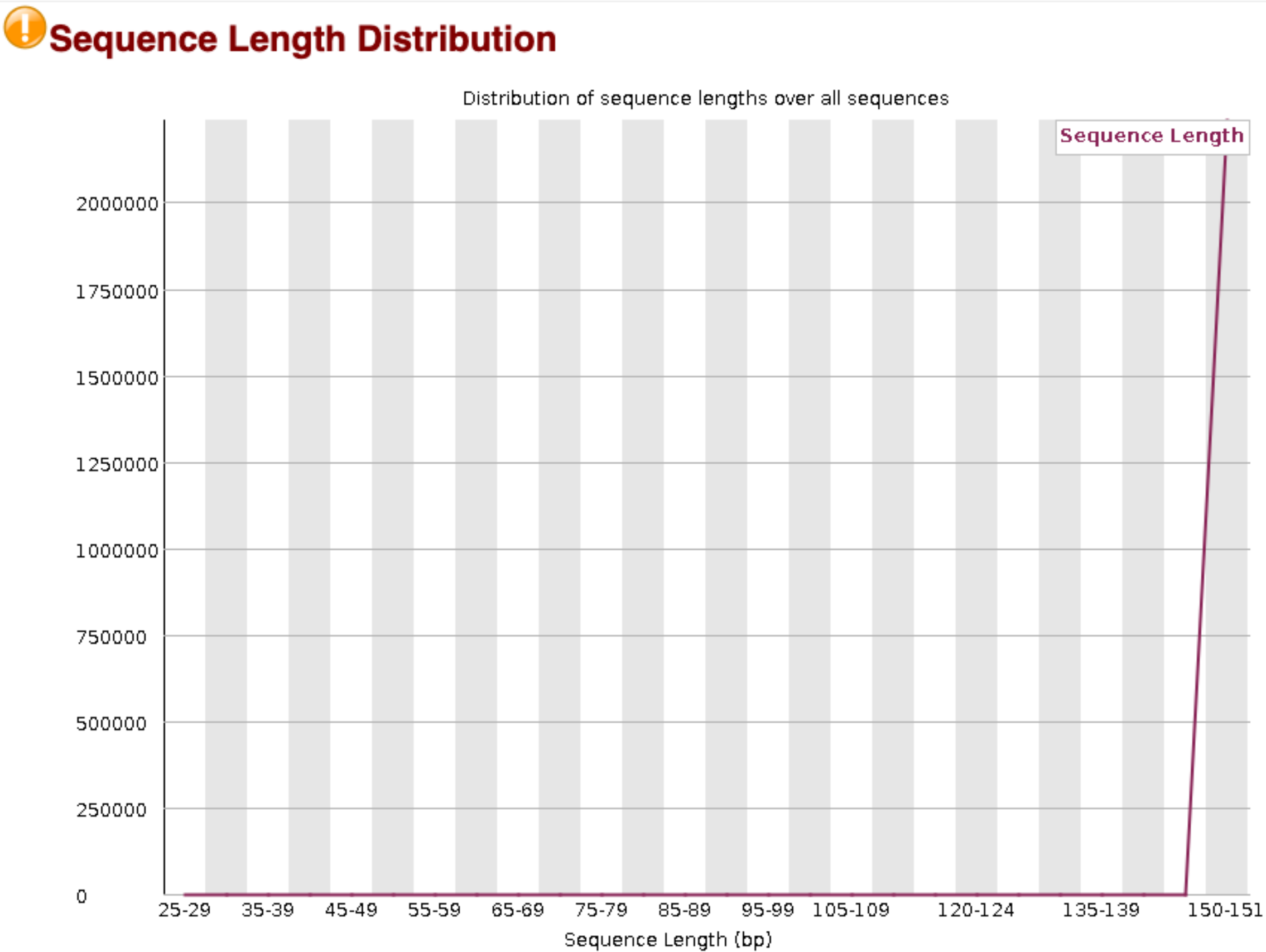
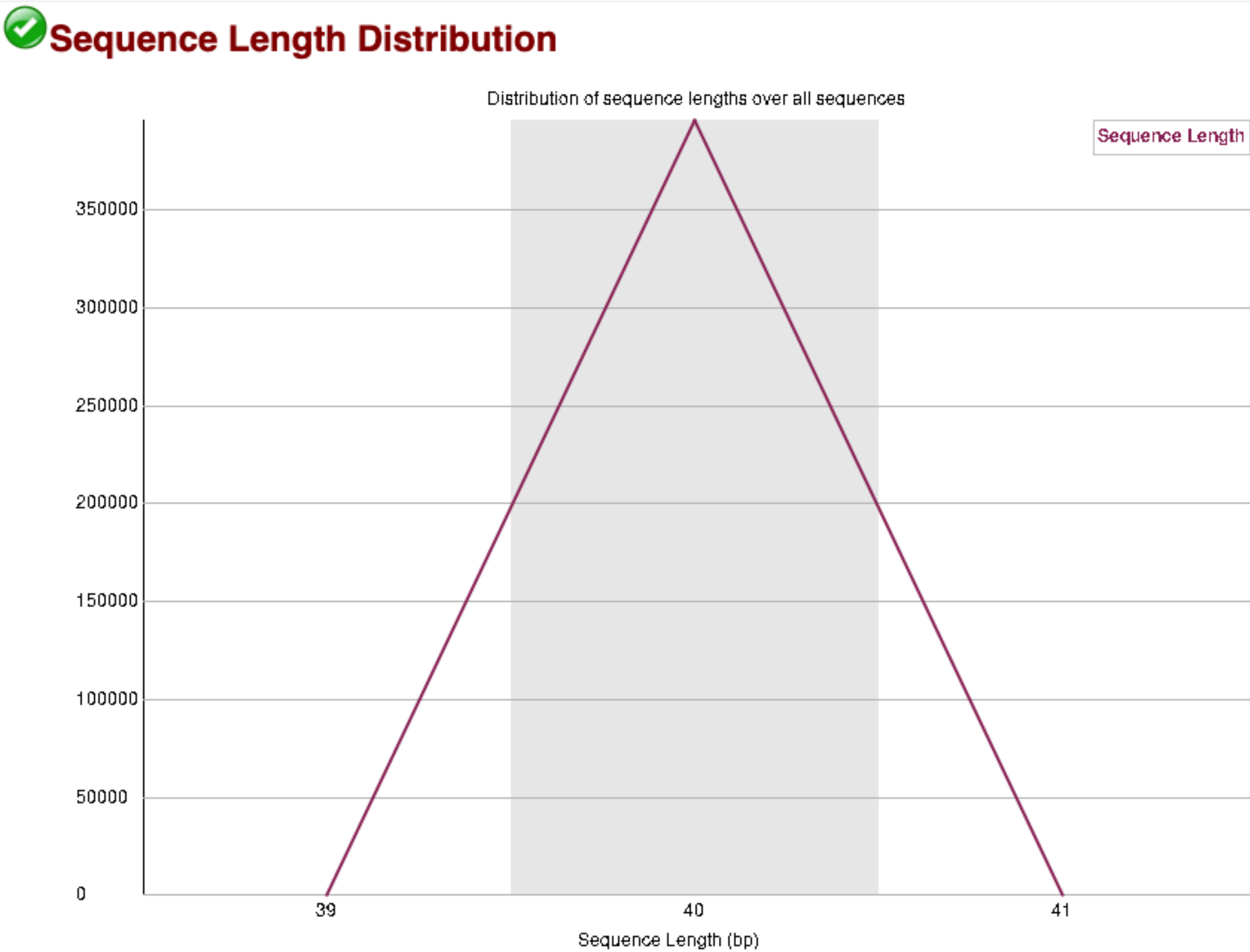


Per sequence quality scores



Resultados de FASTQC

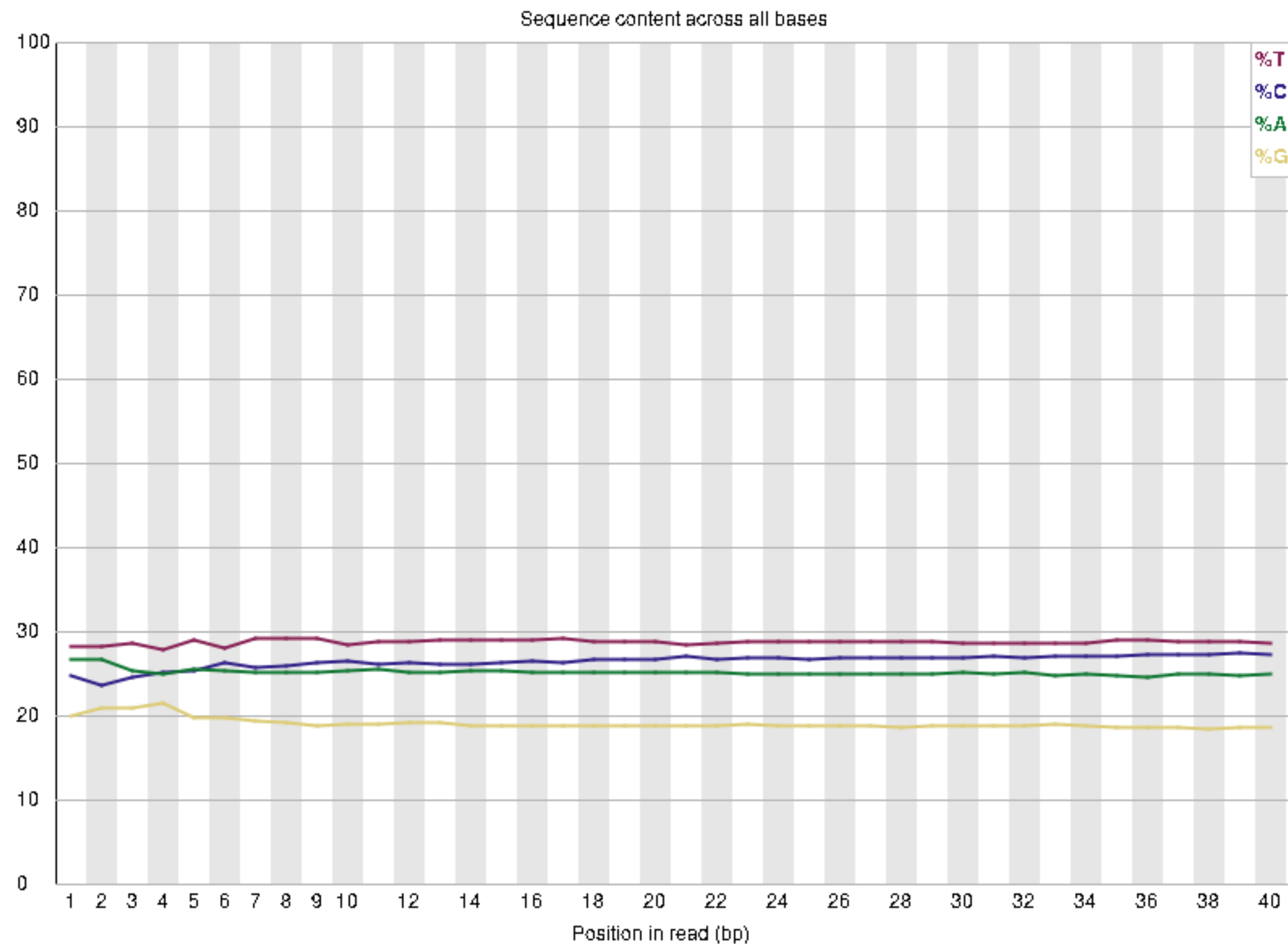
- Longitud de la lectura



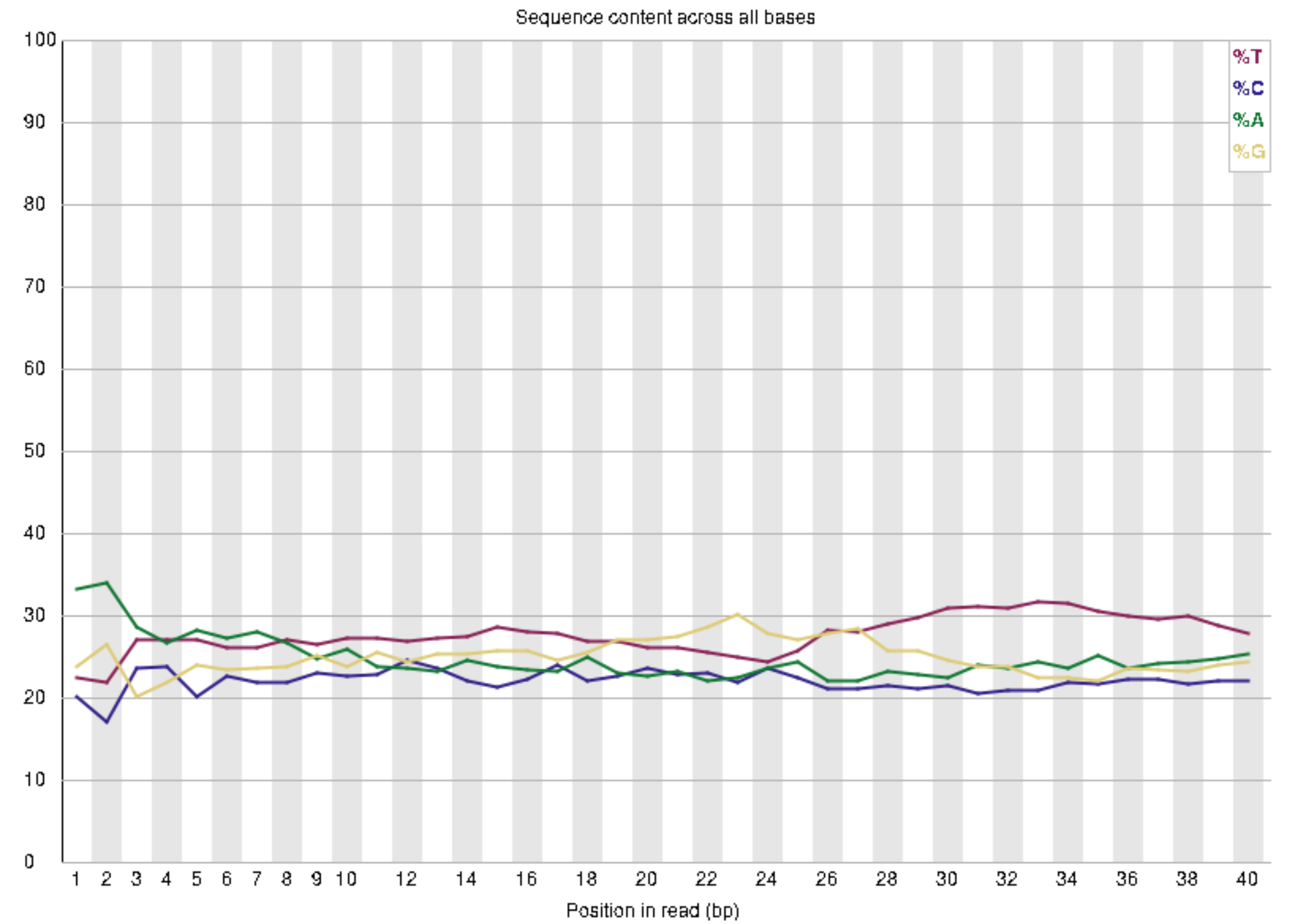
Resultados de FASTQC

- **Sesgos** en la secuencia
- Contenido de bases **G** y **C**

✅ Per base sequence content



⚠ Per base sequence content

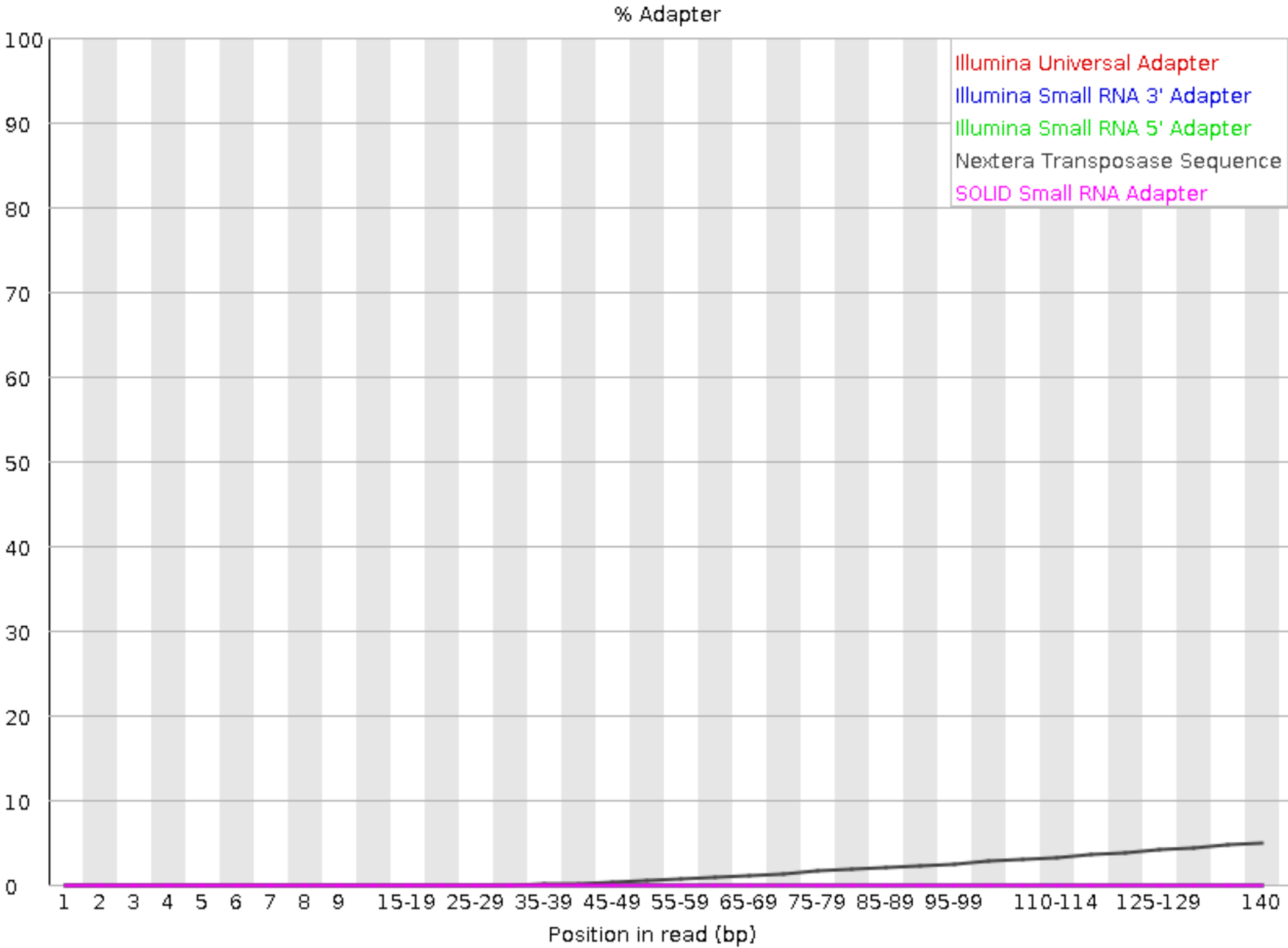


mas sesgada aun

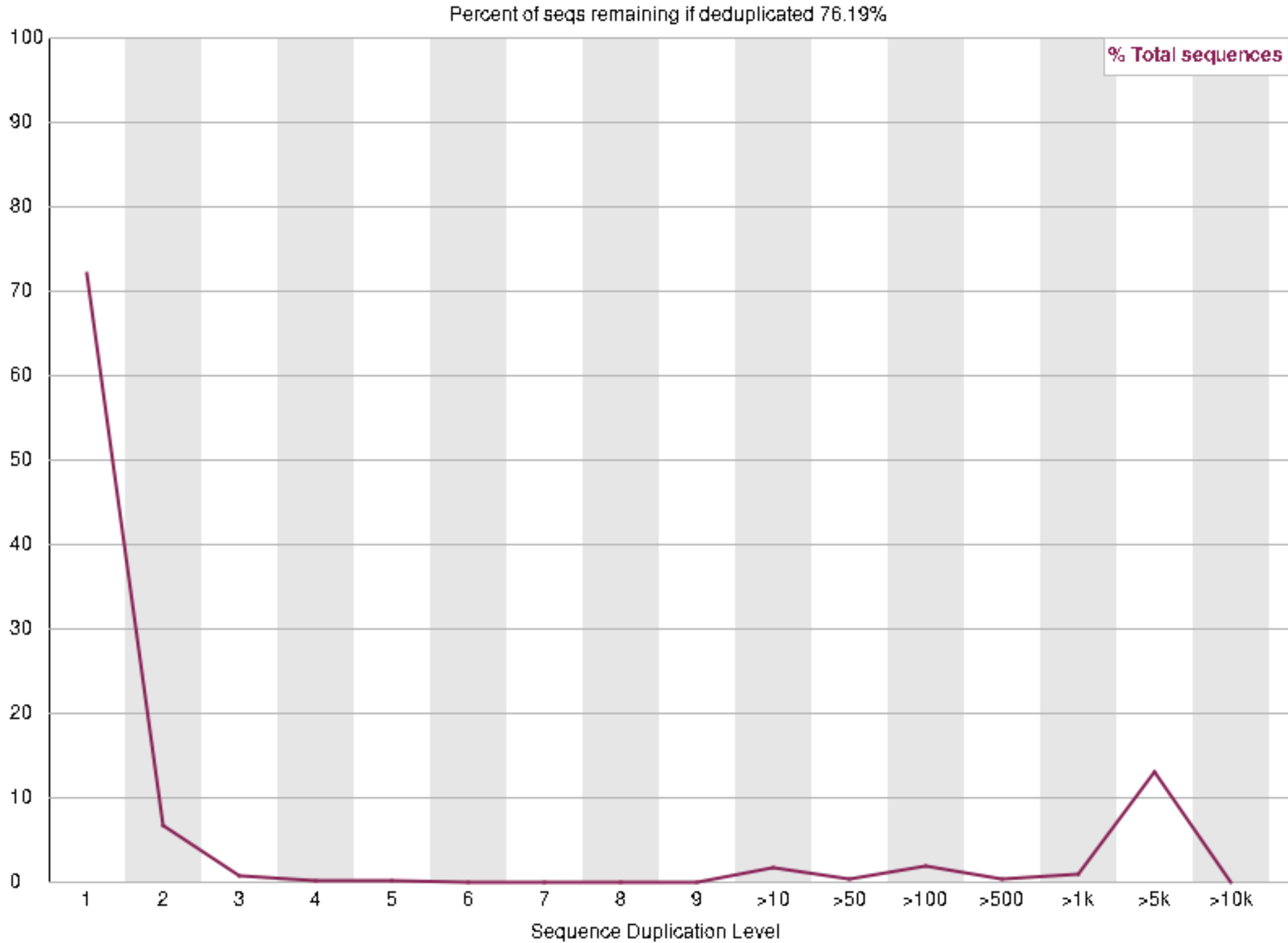
Resultados de FASTQC

- Presencia de adaptadores
- Secuencias repetidas

Adapter Content



Sequence Duplication Levels



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	5086	5.086	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	1085	1.085	Illumina Single End PCR Primer 1 (100% over 40bp)

LECTURAS
BRUTAS

GCTTCTGTGG

GGCTCACGTA

ACGTAAGAGG

ATAAAAGTTAC

AGCGGGCGTA

GCTTCTGTGG

Fastq

Hemos analizado la calidad de nuestras lecturas.
Qué hacemos ahora?

LECTURAS BRUTAS

```
GCTTCTGTGG
GGCTCACGTA
ACGTAAGAGG
ATAAAAGTTAC
GGACACAGAG
GCTTCTGTGG
```

Fastq

fastp

**Filtrado y
trimming de
lecturas**

```
GCTTCTGTGG
GGCTCA
ACGTAAGAGG
AAGTTAC
GGACACAGAG

```

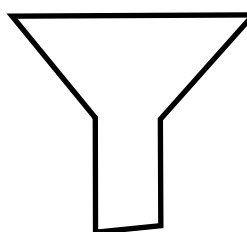
Fastq

Hemos analizado la calidad de nuestras lecturas.
Qué hacemos ahora?



Trimming de lecturas:

- adaptadores
- cadenas de polyX
- corte a los lados



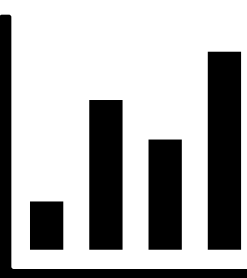
Filtrado de lecturas:

- calidad promedia
- longitud
- duplicados
- sobrerrepresentadas



Corrección de errores:

- usando solape con la pareja



Resultados resumidos en un **informe**

Herramienta Bioinformática



fastp: an ultra-fast all-in-one FASTQ preprocessor

Shifu Chen^{1,2,*}, Yanqing Zhou¹, Yaru Chen¹ and Jia Gu²

¹Department of Bioinformatics, HaploX Biotechnology, Shenzhen 518057, China and ²Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

*To whom correspondence should be addressed.

LECTURAS
BRUTAS

GCTTCTGTGG

GGCTCACGTA

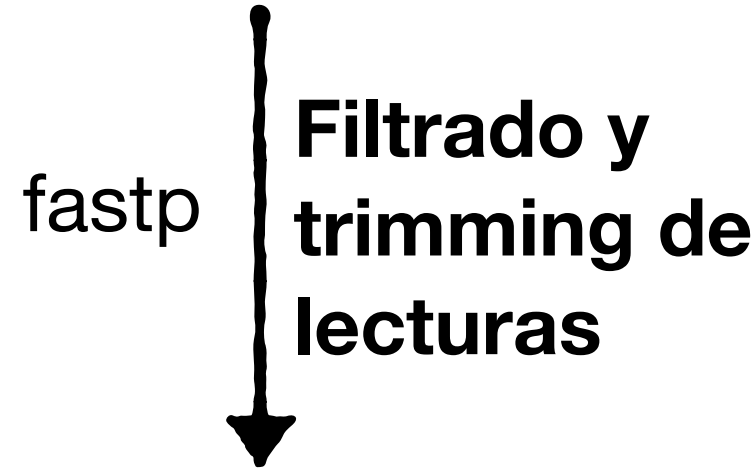
ACGTAAGAGG

ATAAAAGTTAC

GGACACAGAG

GCTTCTGTGG

Fastq



GCTTCTGTGG

GGCTCA

ACGTAAGAGG

AAGTTAC

GGACACAGAG

Fastq

Con nuestras lecturas limpias podemos proceder a su análisis!

1 18 30 38 46 58 *fasta*
GCTTCTGTGGATAAAAGTTACATGAAGGCAAGCGGGCGGCTCACGTAAGAGGGACCTGGGGACACAGAGGCAG ...

Genoma de Referencia

Secuencia de nucleótidos de un individuo
representativo de una especie

“Mosaico haploide”

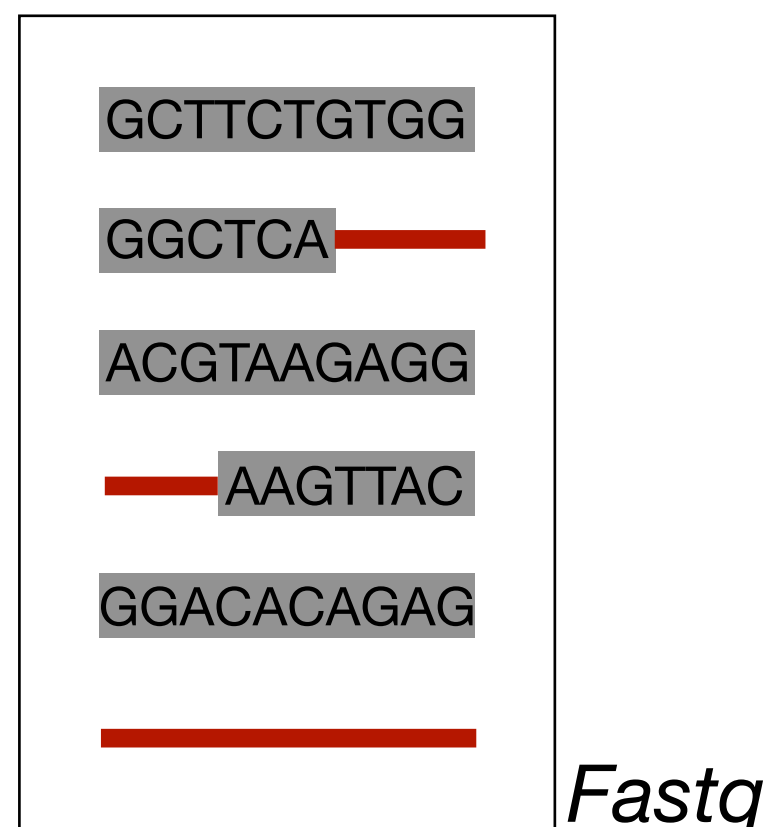
En formato **FASTA**

Dividido por contig / scaffold / cromosoma

dibujo contig scaffold cromosoma

Permite ordenar y comparar secuencias de
nuestros individuos secuenciados

se dividen en draft y high quality



11830384658

fasta

GCTTCTGTGGATAAAAGTTACATGAAGGCAAGCGGGCGGCTCACGTAAGAGGGACCTGGGGACACAGAGGCAG ...

Alineamiento

Herramienta Bioinformática



Burrows-Wheeler Aligner

Home

lh3/minimap2

A versatile pairwise aligner for genomic and spliced nucleotide sequences



GCTTCTGTGG

ACGTAAGAGG

AAGTTAC

GGACACAGAG

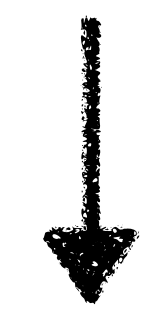
Fastq

1 18 30 38 46 58 *fasta*
GCTTCTGTGGATAAAAGTTACATGAAGGCAAGCGGGCGGCTCACGTAAGAGGGACCTGGGGGACACAGAGGCAG ...

GCTTCTGTGG AAGTTAC GGCTCA ACGTAAGAGG GGACACAGAG

Alineamiento

Herramienta Bioinformática



Burrows-Wheeler Aligner

Home

GCTTCTGTGG

GGCTCA

ACGTAAGAGG

AAGTTAC

GGACACAGAG

Fastq

lh3/minimap2

A versatile pairwise aligner for genomic and spliced nucleotide sequences

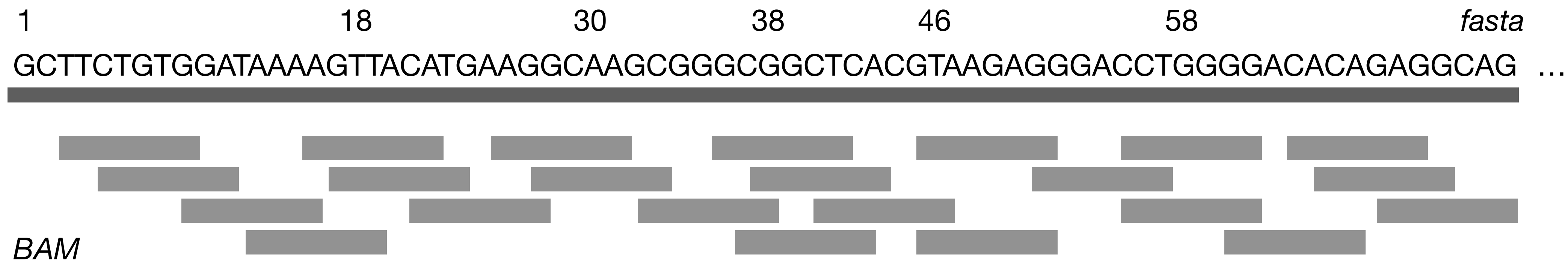


BOW

TIE

Bowtie

An ultrafast memory-efficient short read aligner



El formato BAM

Versión comprimida de un **SAM** (Sequence Alignment Map)

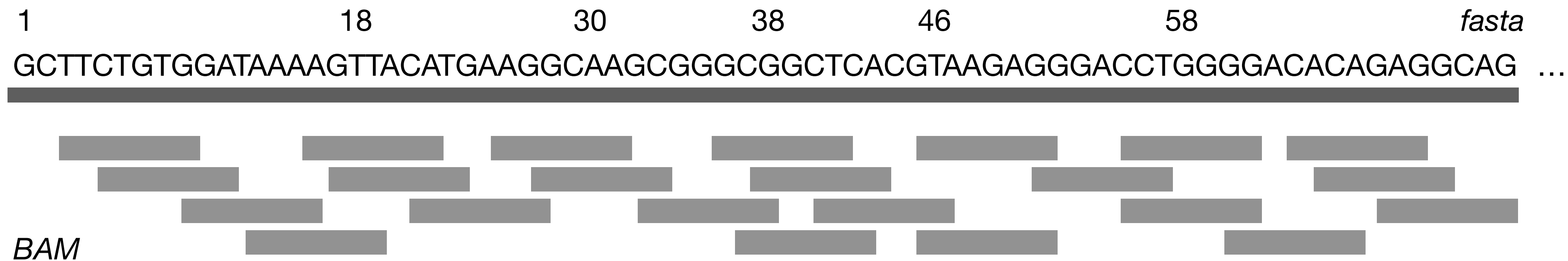
Almacena información sobre donde secuencias alinean a un genoma de referencia

Cabecero:

- filas que empiezan con @
- metadatos sobre el alineamiento

Alineamiento:

- mínimo de **11 columnas** separadas por tabulador
- columnas adicionales facultativas



El formato BAM

Versión comprimida de un **SAM** (Sequence Alignment Map)

Almacena información sobre donde secuencias alinean a un genoma de referencia

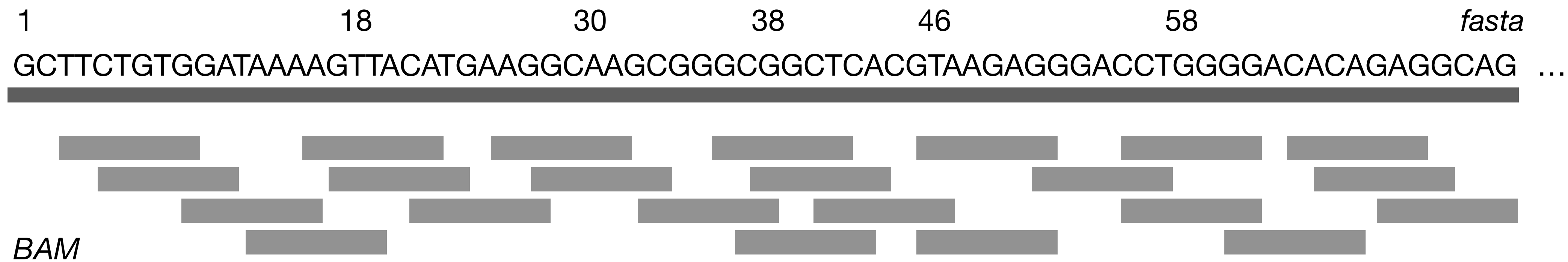
Cabecero:

- filas que empiezan con @
- metadatos sobre el alineamiento

Alineamiento:

- mínimo de **11 columnas** separadas por tabulador
- columnas adicionales facultativas

1. **QNAME**: Nombre codificado del fragmento de ADN. Las lecturas/segmentos con el mismo QNAME se consideran provenientes del mismo fragmento: r1 y r2, la misma lectura alineada en otro sitio del genoma.
2. **FLAG**: Suma de números que describen características de la lectura alineada. [https://en.wikipedia.org/wiki/SAM_\(file_format\)#Bitwise_flags](https://en.wikipedia.org/wiki/SAM_(file_format)#Bitwise_flags)
3. **RNAME**: Nombre de la secuencia de referencia del alineamiento. Un segmento no alineado (sin coordenadas) tiene un '*' en este campo.
4. **POS**: Coordenada de alineamiento de la base más a la izquierda en la secuencia de referencia. Lecturas no alineadas tienen POS=0
5. **MAPQ**: Calidad de mapeo. Es igual a $-10 \log_{10} \text{Pr} \{ \text{la posición de mapeo es incorrecta} \}$, redondeado al entero más cercano. Un valor 255 indica que la calidad de mapeo no está disponible.
6. **CIGAR**: **C**oncise **I**diosyncratic **G**apped **A**lignment **R**eport. Parejas de número-operador que describen el alineamiento <https://www.drive5.com/usearch/manual/cigar.html>
7. **RNEXT**: Nombre de la secuencia asociada a esta ('=' si es el mismo)
8. **PNEXT**: Posición de la secuencia asociada a esta.
9. **TLEN**: Longitud del fragmento de ADN, el número de bases desde la base mapeada más a la izquierda hasta la base mapeada más a la derecha. La secuencia más a la izquierda es positiva, la más a la derecha es negativa.
10. **SEQ**: Secuencia de la lectura.
11. **QUAL**: ASCII de la calidad de las bases, como en el FASTQ



El formato BAM

Versión comprimida de un **SAM** (Sequence Alignment Map)

Almacena información sobre donde secuencias alinean a un genoma de referencia

Cabecero:

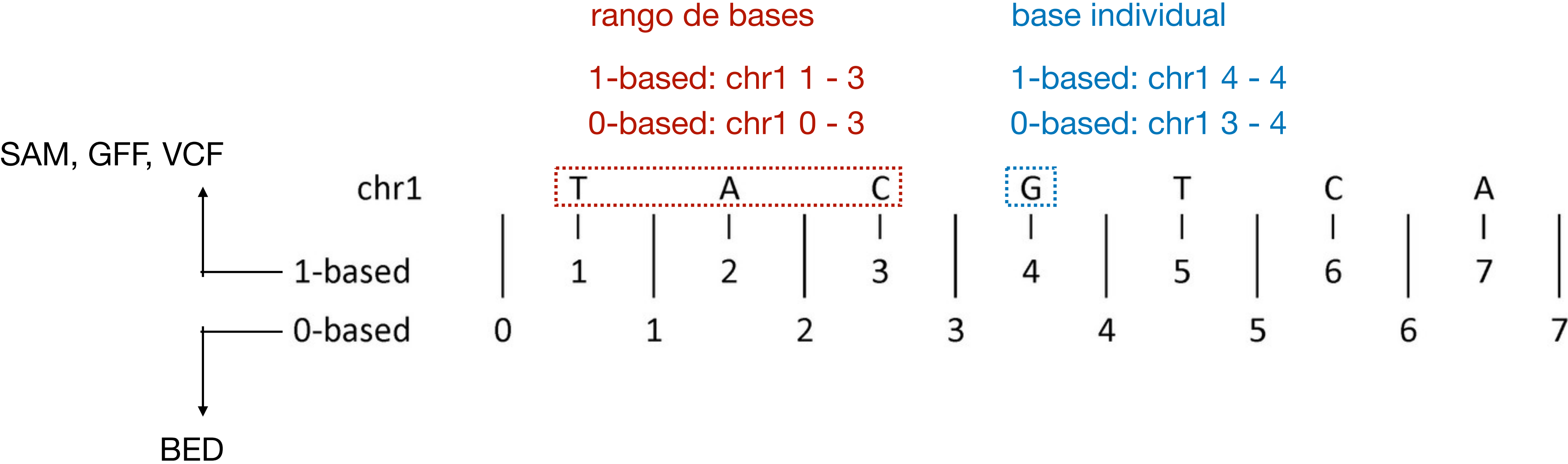
- filas que empiezan con @
- metadatos sobre el alineamiento

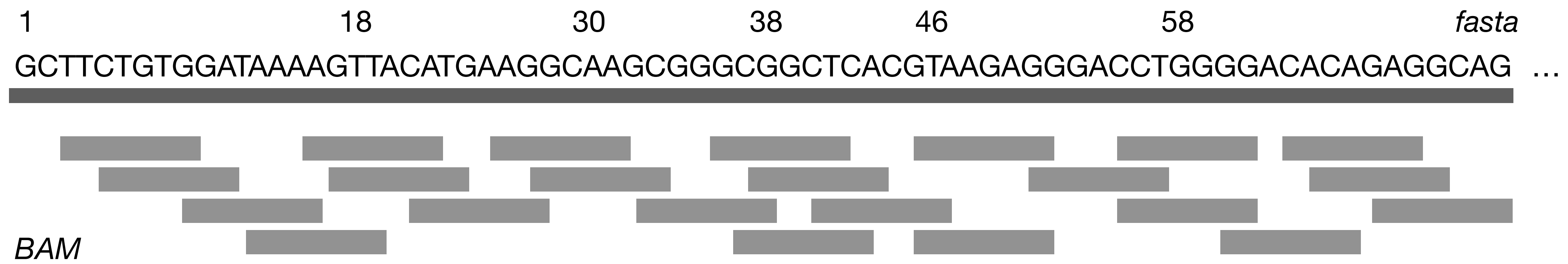
Alineamiento:

- mínimo de **11 columnas** separadas por tabulador
- columnas adicionales facultativas

1. **NM**: número de mismatches (cambios)
2. **MD**: cadena descriptiva de los cambios
3. **MC**: CIGAR de la pareja
4. **MQ**: calidad de mapeo de la pareja
5. **XA**: alineamientos alternativos (chr,pos,CIGAR,NM;)

Formatos 0-based vs Formatos 1-based

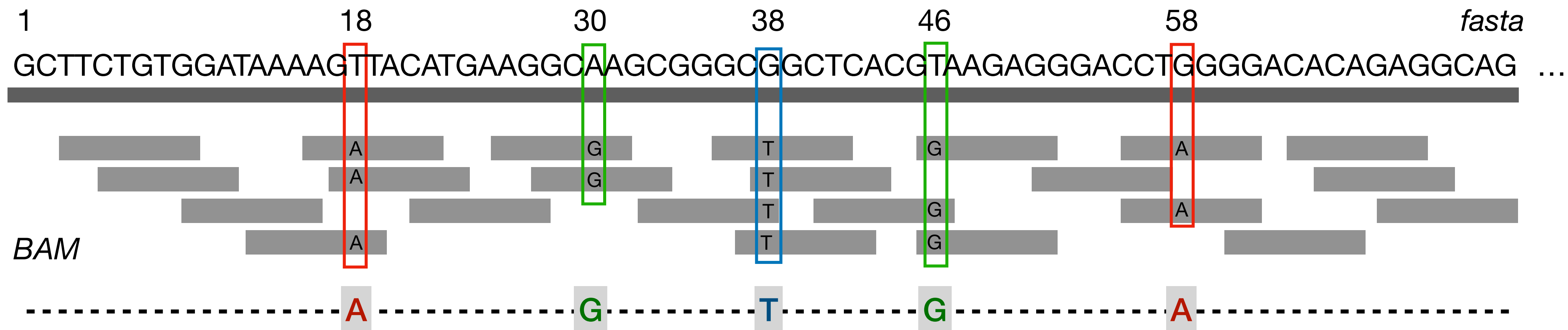




Como visualizarlo?

Herramienta Bioinformática





Llamada de Variantes

Objetivo: identificar regiones del genoma donde nuestros individuos difiere del genoma de referencia

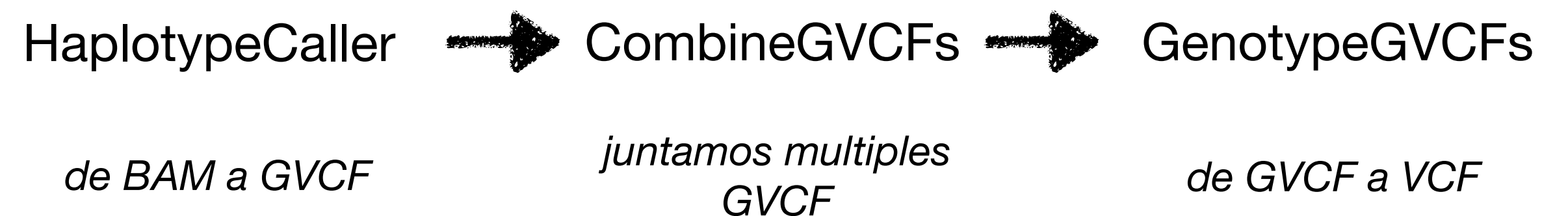
Procedemos en dos pasos:

- Identificación de **variantes** del genoma **de cada individuo**, conservadas en formato *GVCF*
- Unión de las variantes individuales en un **única base de datos**, conservadas en formato *VCF*

Herramienta Bioinformática



freebayes
deepvariant
...



Formatos GVCF y VCF

Genome Variant Calling Format:

Informa sobre **si y dónde cambian** las secuencias de uno o más individuos con respecto al genoma de referencia.

Las regiones no variantes (bloques) se forman juntando intervalos consecutivos para los cuales la calidad del genotipo está dentro de un rango específico.

Cabecero:

- filas que empiezan con **#**
- **describe las columnas** de la tabla que sigue
- información adicional sobre **cómo se ha generado** y **qué contiene** el file (comando, contigs, etc.)
- en GVCF, información sobre como se juntan **bloques** en función de su **calidad**

Tabla:

- **9** columnas descriptiva de la **región genómica** (bloque o variante)
- **1** columna **para cada individuo** con información sobre **su genotipo** en la región

Variant Calling Format

Informa sobre **dónde cambian** las secuencias de uno o más individuos con respecto al genoma de referencia.

1. **CHROM**: nombre del contig/scaffold/cromosoma
2. **POS**: posición dentro del contig/scaffold/cromosoma (1-based)
3. **ID**: identificador específico de la región
4. **REF**: secuencia de referencia en la posición
5. **ALT**: secuencias alternativas observadas en la posición
6. **QUAL**: calidad de la inferencia de la región
7. **FILTER**: informa sobre los filtros no pasados por la región o PASS si todos los filtros se han pasado
8. **INFO**: columna extensible con información adicional sobre la región genómica. Sub-columnas separadas por ‘;’, identificadas al principio (e.g. AF=0.05;AN=5) y descritas en el cabecero
9. **FORMAT**: columna extensible que informa sobre el formato y el contenido de las columnas de genotipo de los individuos (10+)
- 10.+ **SAMPLEs**: valores definidos en la columna FORMAT para cada individuo (genotipo, numero de lecturas con cada alelo etc.)

LECTURAS BRUTAS

```
GCTTCTGTGG
GGCTCACGTA
ACGTAAGAGG
ATAAAAGTTAC
GGACACAGAG
GCTTCTGTGG
```

Fastq

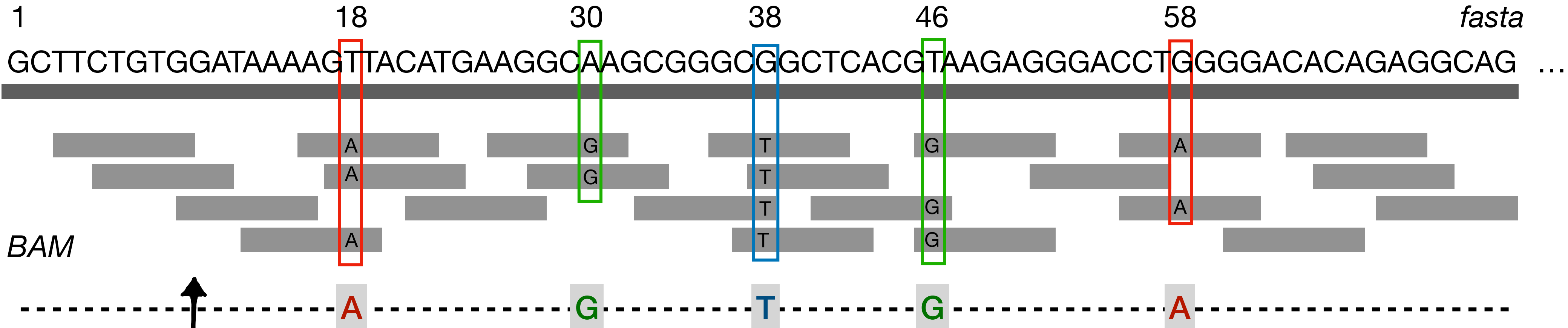
fastp

**Filtrado y
trimming de
lecturas**

```
GCTTCTGTGG
GGCTCA
ACGTAAGAGG
AAGTTAC
GGACACAGAG

```

Fastq



**Alineamiento
de lecturas**

BWA-
MEM

**Llamada de
variantes**

GATK

18	T	A
30	A	G
38	G	T
46	T	G
58	G	A

VCF

Filtrado de variantes

Ahora que tenemos el VCF con nuestras variantes, podemos proceder a varios pasos de filtrado para eliminar las que no nos interesan.

Filtrado de variantes

Regiones:

- baja complejidad / repetitivas
- paralogos
- genes

Tipos de variante:

- INDELs
- no-bialelicas
- fijadas

Valores cuantitativos:

- calidad
- missing
- profundidad
- heterozigosidad / hardy-weinberg