Análisis de datos(I): Control de Calidad y Preprocesado

Miguel Juliá Molina

Unidad de Bioinformática
Unidades Comunes Científico Técnicas – SGAFI-ISCIII

17-21 Junio 2019, 7ª Edición
Programa Formación Continua, ISCIII

Dónde estamos

Fastq files

Preprocesamiento

Alineamiento

Postprocesamiento: Análisis de variantes

Ficheros de salida del secuenciador

Illumina

.fastq

454 .sff

SOLiD

.fasta

.qual

Nanopore

FAST5

PacBio RSII

Bax.h5

fasta

Formato fastq

- Fácilmente se podría decir que es un fasta con calidades.
- En NGS fasta suele contener genomas y fastq fragmentos.

```
>SEQ_ID
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAG
```

Secuencia

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

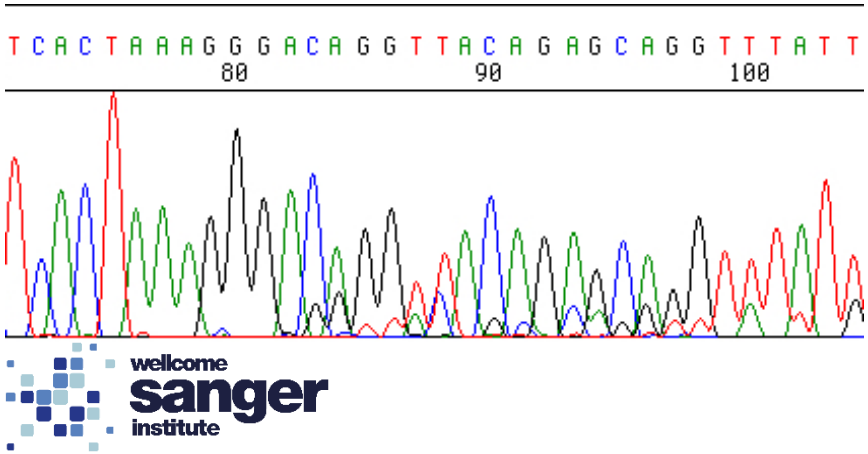
Calidades: sólo deben ocupar un bit

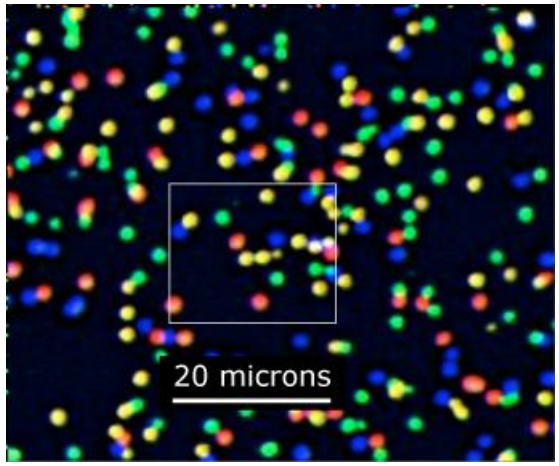
Formato fastq

- Cada base corresponde con un valor de calidad.
- ¿Cómo se codifica?

Calidad Phred y probabilidad de error

- Conversión de probabilidad de error en score de calidad Phred.
- La calidad Phred se originó como aproximación algorítmica a la calidad en secuenciación Sanger.
- La **intensidad de la señal lumínica** es utilizada para calcular la probabilidad de error.





illumina®

Calidad Phred y probabilidad de error

- Conversión de probabilidad de error en score de calidad Phred.
- La calidad phred en Sanger y en la versión 1.8+ de Illumina va de 0-40 en codificación decimal.

$$Q = -10 \log_{10} P$$

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Calidad Phred y probabilidad de error

- Conversión de score de calidad phred en código ASCIII para que ocupe un solo bit (un solo caracter)

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

- Para Sanger y versiones actuales de Illumina se usa una codificación de Phred+33 en ASCIII. Es decir una calidad de 0 es el carácter que se corresponde con el decimal 33 que se trata del símbolo !
- En versiones de Solexa y de Illumina 1.3 y 1.5 de hace unos años se utilizaba la codificación Phred+64. Lo que difería de lo acostumbrado por sanger y hacía falta una conversión para comparar las calidades.

Calidad Phred y probabilidad de error

- Ejemplo Phred

```
@HWI-ST731_6:1:1101:1322:1938#1@0/1
NTGACAAAGGGCTAATATCCAGAATCTACAAAGAACTTAAACAAATGTATAAGAATAAAAGTATAGTGCTAACAAT
+
#1:BDDADFDFDD@F>BGFIIIB@CFHIHICAGBC9CBCBGGIGCFF??>GGHFHIGGEGI<FECGDE=FHCHEG=
```

P=0.001 → Q=-10*log10(0.001)= 30 → ASCIII 33+30 = 63 → ?

P=0.0001 → Q=-10*log10(0.001)= 40 → ASCIII 33+40 = 73 → I

```
Quality encoding: !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                  |         |         |         |         |
   Quality score: 0........10........20........30........40
```

Formato fastq

Cabecera típica de Illumina

```
@HWUSI-EAS100R:6:73:941:1973#0/1
```

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*

```
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:2458:1027 1:N:0:ACAGTG
AGAAAAAACCTTGGANGGAAAAAAATCAGACATTTTCTAGAGGTGGAAGGCAAACTGAACAAAGAAATAATTCACA
+
DGGGEDHHHHGGGFE#CBACBCA<?HHHHBHHHHHHHHDHHHHEHEFEGGGGGG/GGDDDGHFHGFCHFHHEHEH8
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3082:1029 1:N:0:ACAGTG
GGTAATACAGACTGANATGATCAAAGGCATGCTGGAAACAAACCTATTAAAGATAAGCTTGGATCAAGCTTTCATT
+
B:B:?BB/:=55177#55877<775EDD>E=B?BBBBGGGDDAG@G>GGGGGG@)EEEEBEG>GGGGGGGAAA?<D
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3185:1033 1:N:0:ACAGTG
TCTGGGACATTGCTCNTGGCTGGGAGTCACCTGTCTGGGACATTGCTCAGGGCTGGGAGACACGTGTTGGAGGGAC
+
BC??A66;)74781<#7??;452.27'64(8,851DDG8GB?##############################
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3268:1033 1:N:0:ACAGTG
ATTCAAATTAGAAGANAGTTGATCGTTCTTCATGATGCCCAAAAATTTCACTGAGAAAACCCTTTTTTAAGCCCAC
+
IIIIIIIIIIFFFFE#ABACFEEFFIIGIIIFIHE@BIIIIIIIIIHHIIFIIF>HHIHIFGDIIIIIIGFHIEGH
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3400:1035 1:N:0:ACAGTG
TCCTGCTTTAGGAGANTCCTCATGCTCTGACAGGATGCTCTCTATGTGAGTTGAGCTGGTCTTCTCACTTTTATAG
+
IIIIIHIHIIGGEGG#AACA@?=?BHHIIIIIHHIHIIIIHIHIHHGIHIHGHGIGIHGEGGGGHG@EFGGCEFAB
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:3962:1033 1:N:0:ACAGTG
CCACCAACACAGTCTNCACCTTCTGTTGCTGGTGATAGATTTTTGCACCTTTCCATCCTCCAGGTTTCAAAATAGC
+
HHFHHDHDHH>C?CA#EEEE>?A?>HHDGHEGBGBCEEEEGHHF8HEHEEHECH,=>>==EAEE>BEBBAEAACAB
@HWUSI-EAS1752R:21:FC64JUKAAXX:3:1:4491:1028 1:N:0:ACAGTG
AGAGAGAGAGAGAGANAGAGGACTCTGGAGATGCCGAAGCACAAGCCTGCAAGAGTCCCAGCAAAGAAAATAAAAA
+
GADGGEGGEGBBB?B#@=@@72;64GGGFGB>GGGBDG<DBGB<DA??/?##########################
```

Calidad en código ASCII (0-40):

- “!"#$%” menor calidad
- “FGHI” máxima calidad

Control de Calidad

- Evaluación de la calidad de la secuenciación
  - Es el **primer paso** a realizar tras la secuenciación
  - Si realizamos una buena evaluación podremos saber cómo de fiables son nuestros resultados.
  - QC va a **determinar** el siguiente paso de **filtrado**.
  - Debemos ser consistentes con cualquier decisión de filtrado o los siguientes pasos del análisis podrían verse perjudicados.
  - Se debe realizar un control de calidad después de cada paso crítico del análisis.

Control de Calidad

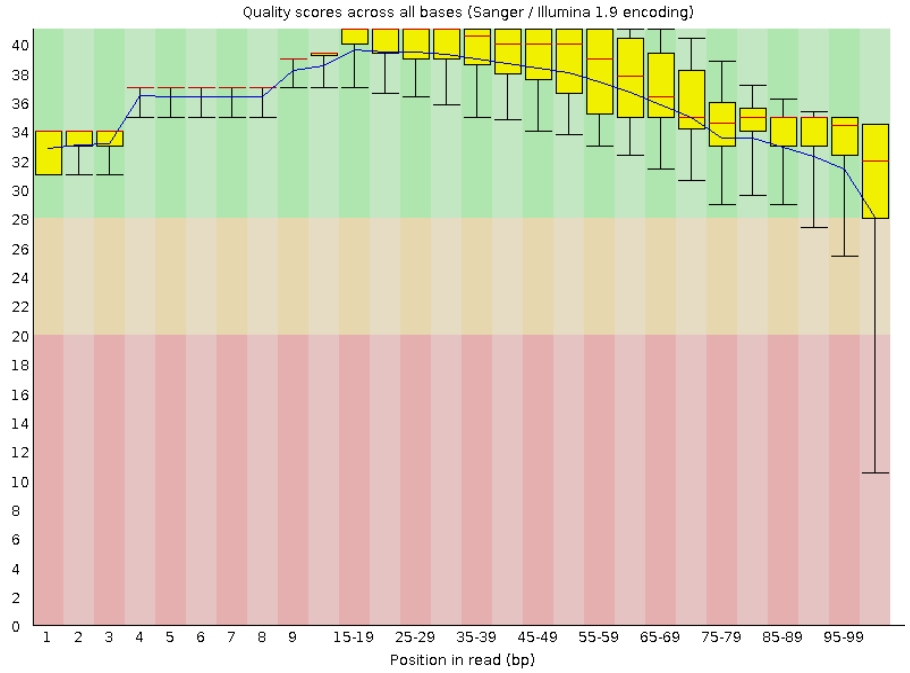
- Para realizar el control de calidad se usa la **información de calidad por base** que hemos estado viendo hasta el momento

- Otros pasos del análisis también utilizará esta calidad por base como parámetro a tener en cuenta por los distintos algoritmos que se apliquen.

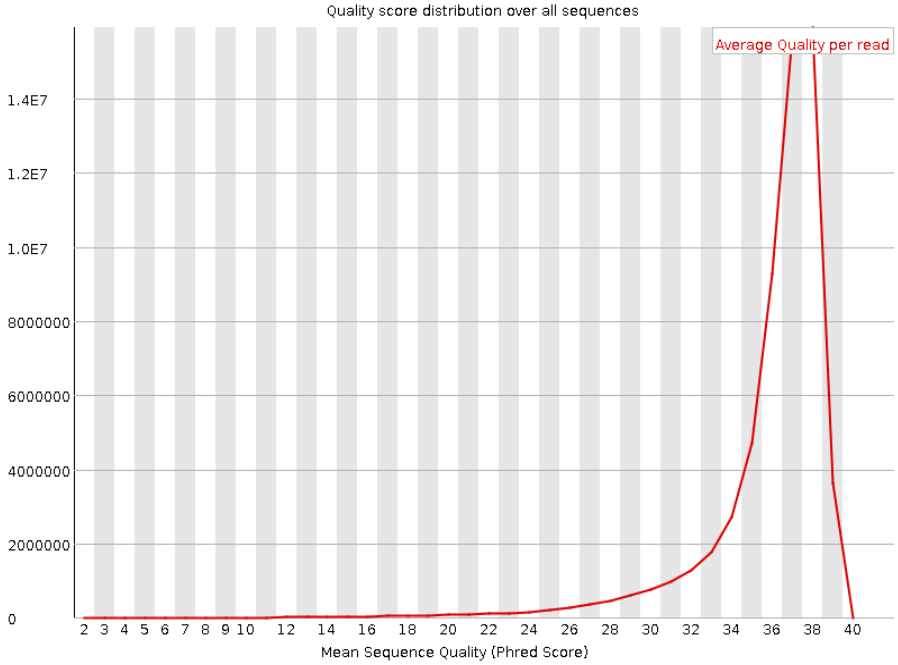
Control de Calidad

- Programas de Control de Calidad
- FastQC

Per base sequence quality

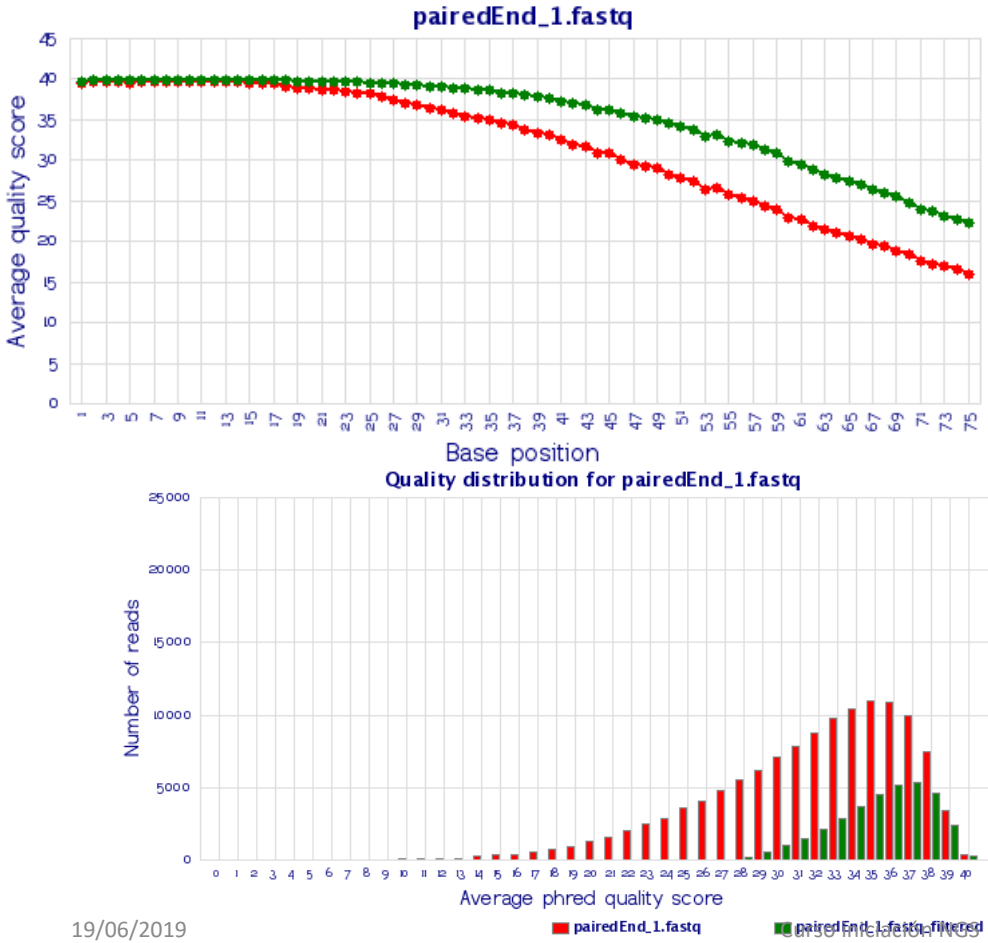


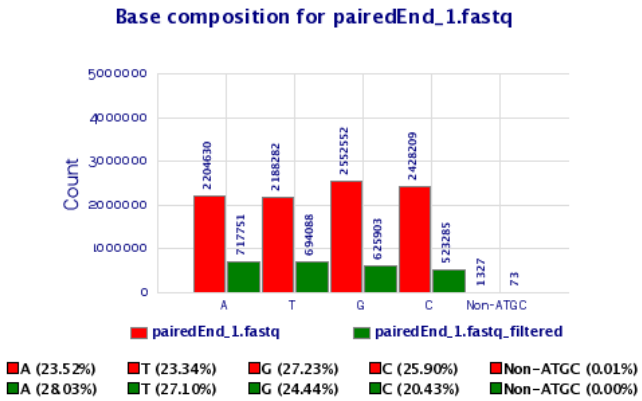
Per sequence quality scores

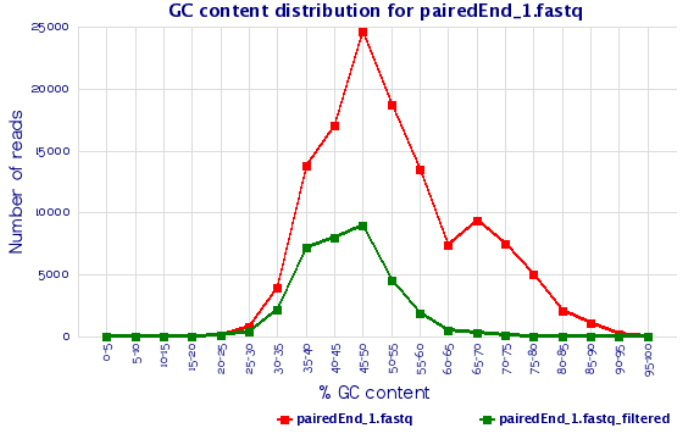


Control de Calidad

- NGSQCToolkit

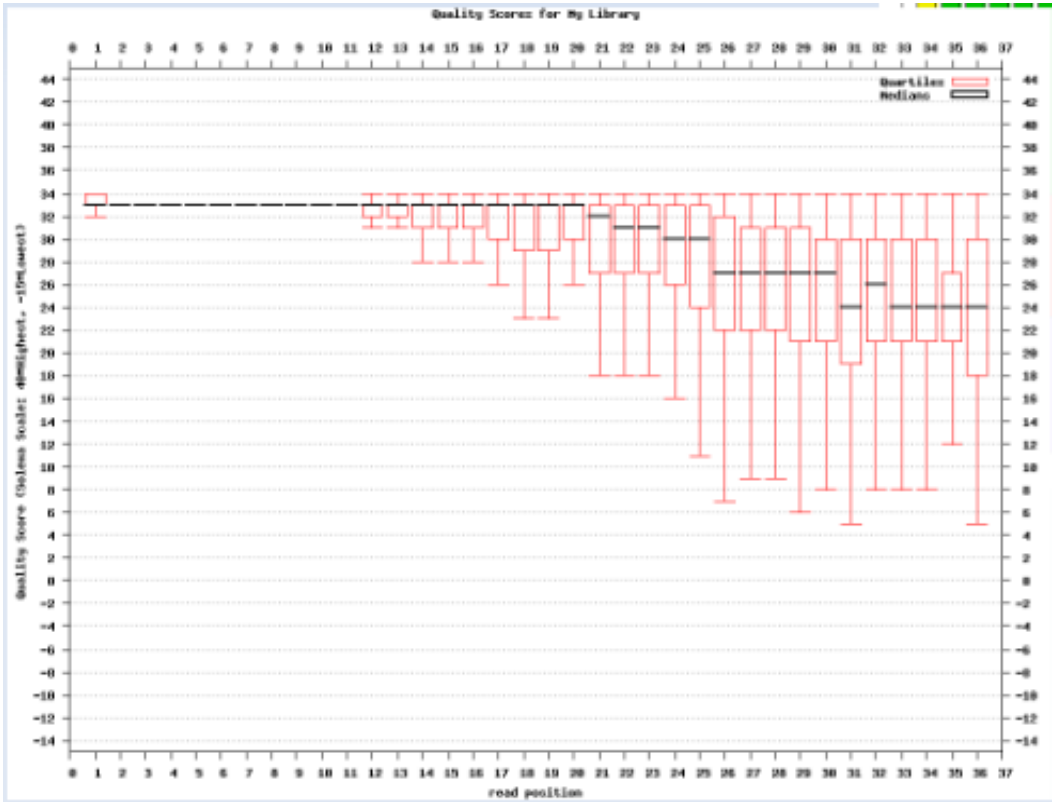






Control de Calidad

- Otros: fastx-toolkit, sfftools, etc…



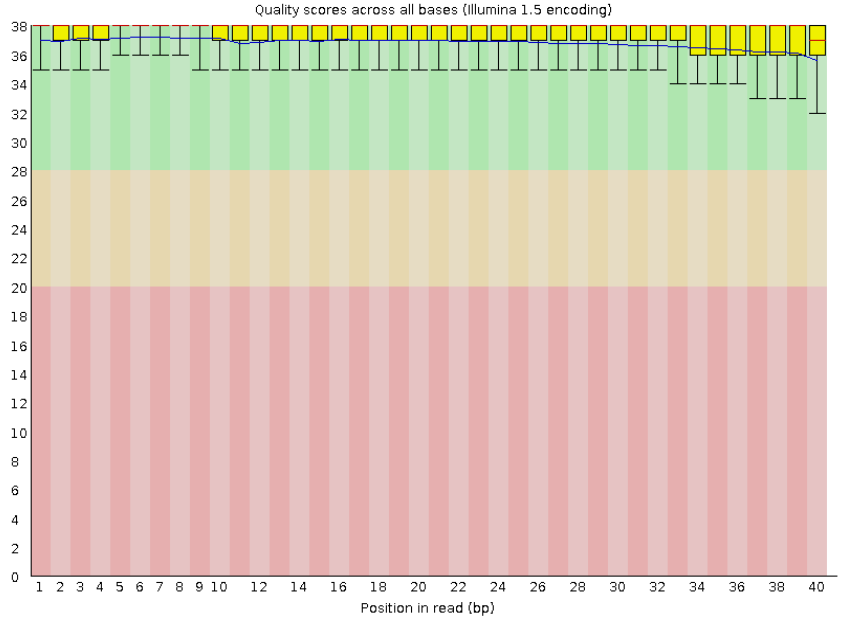
Control de Calidad

- Artefactos en preparación de librería
  - Restos de adaptadores.
  - Alto porcentaje de duplicados.
  - Sesgo en zonas GC.
- Artefactos en secuenciación
  - Baja calidad en extremos (Phasing).
  - Dificultad en determinadas zonas:
    - Repeticiones
    - Homopolímeros
    - Alto contenido GC

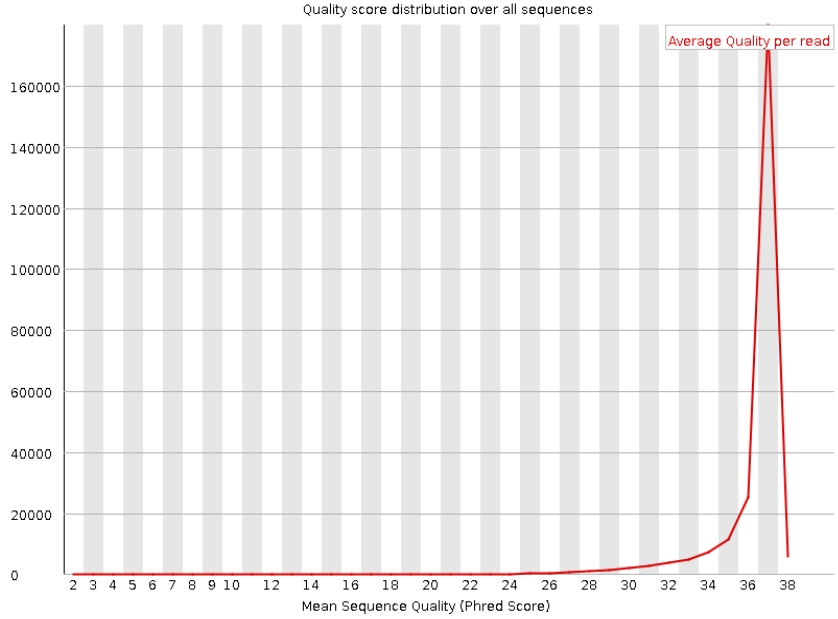
Control de Calidad

- Ejemplos
- Buena secuenciación de Illumina

Per base sequence quality

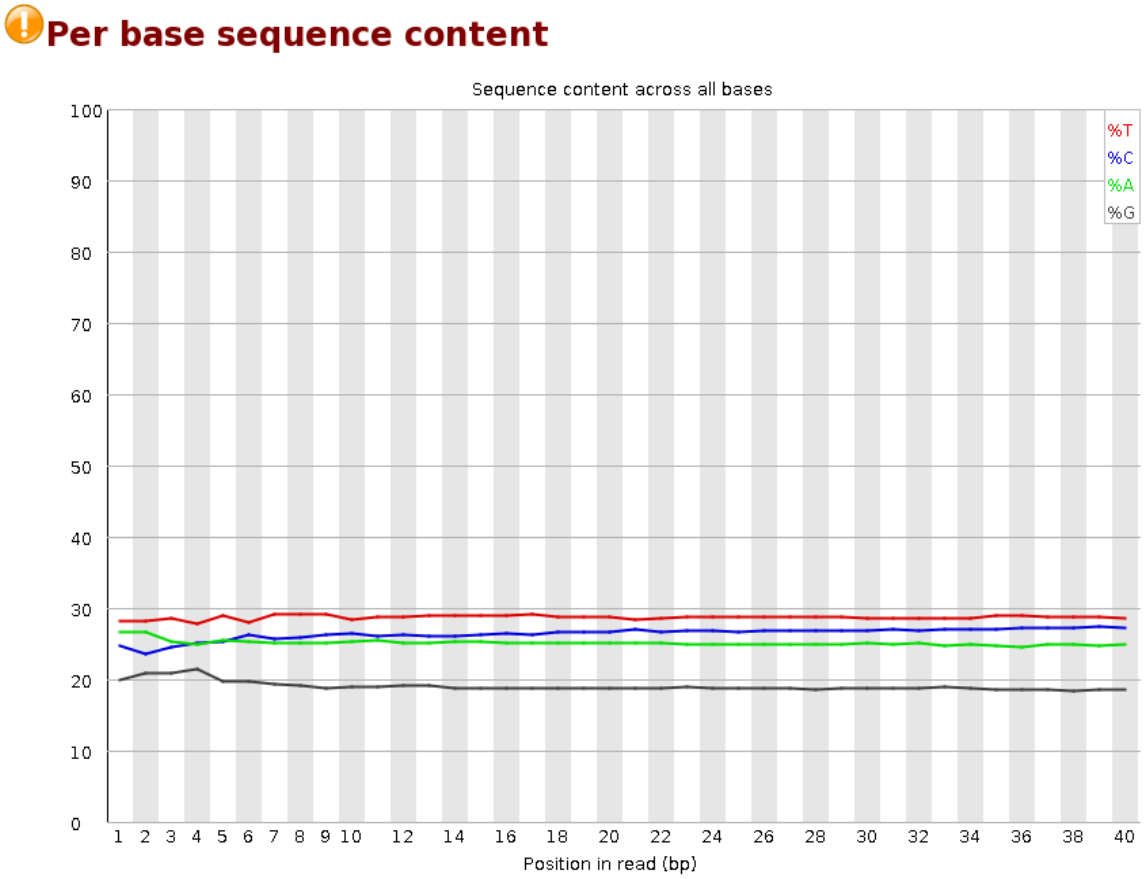


Per sequence quality scores



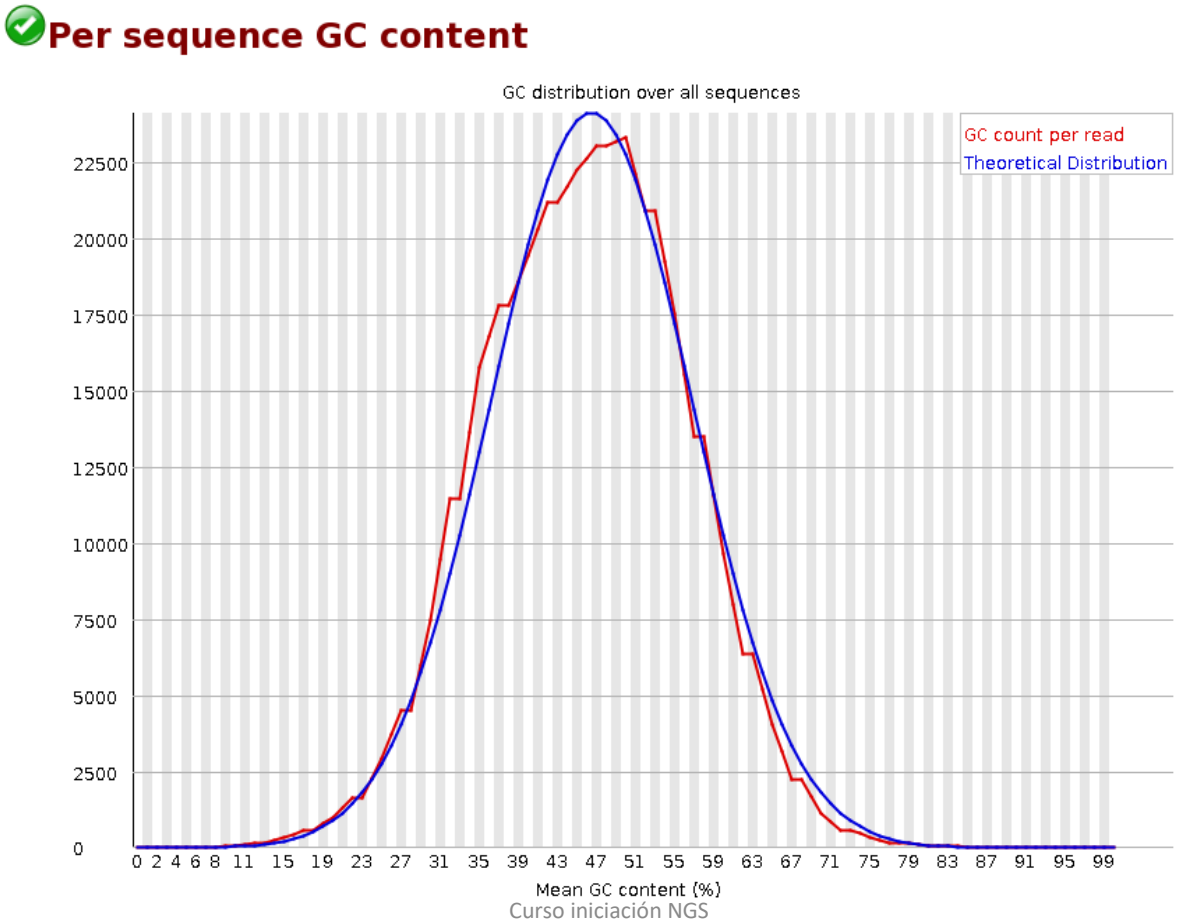
Control de Calidad

- Contenido por base



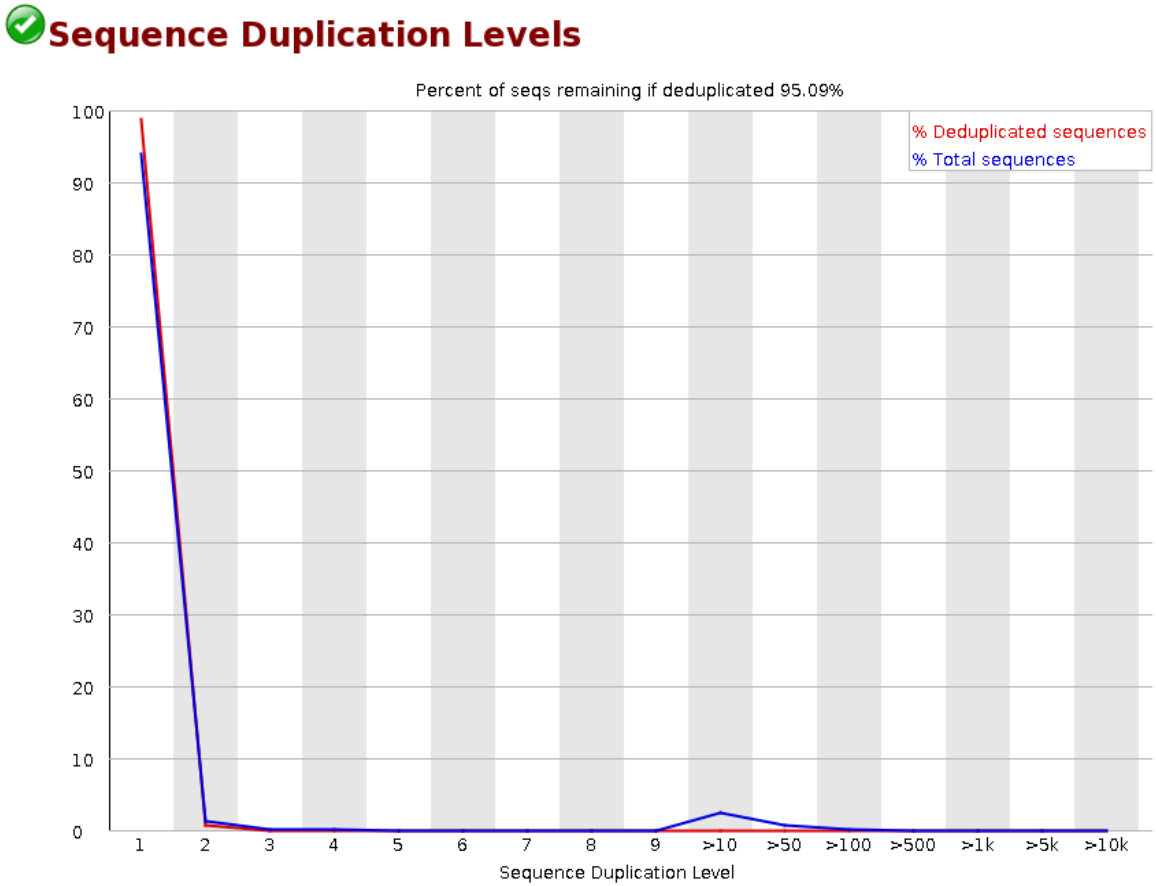
Control de Calidad

- Porcentaje de QC



Control de Calidad

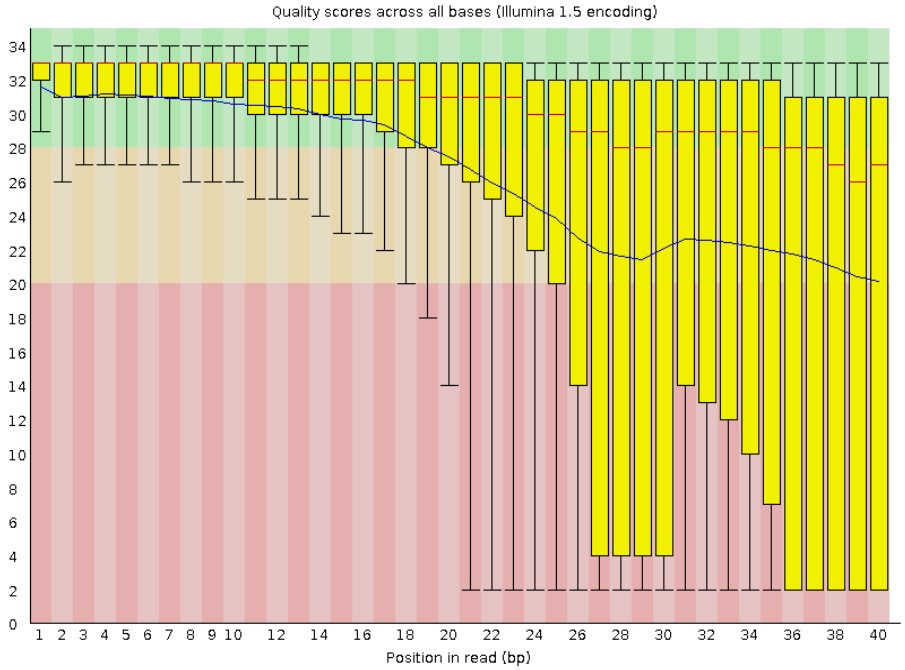
- Porcentaje de duplicados



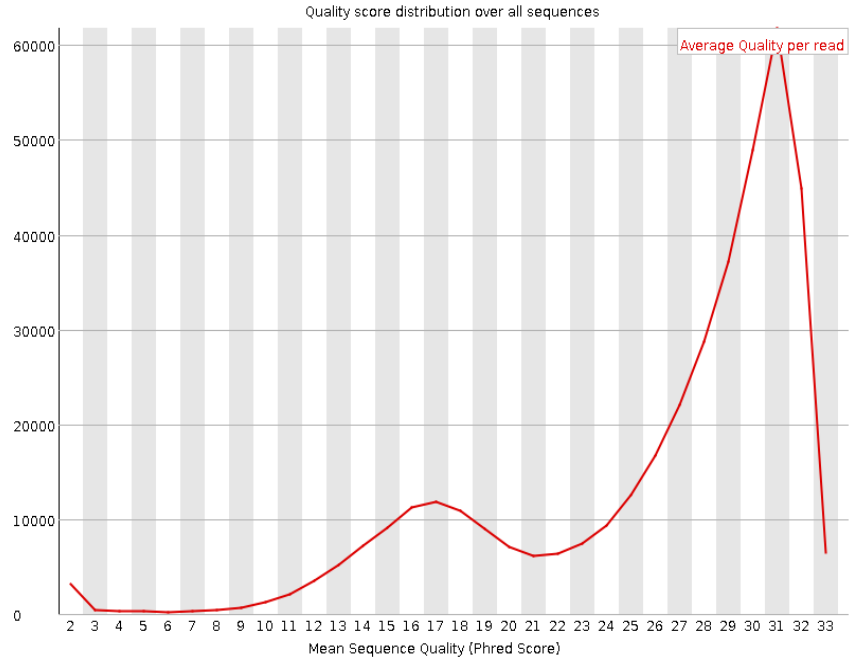
Control de Calidad

- Ejemplos
- Mala secuenciación de Illumina

Per base sequence quality



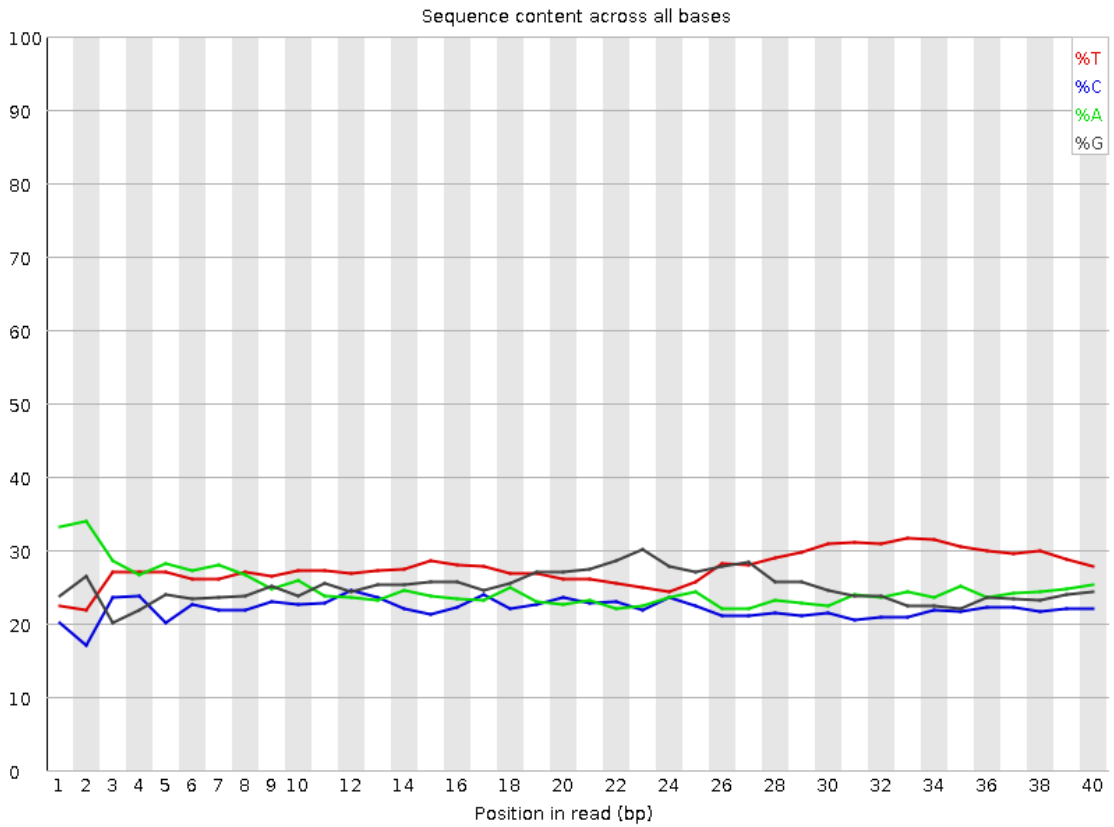
Per sequence quality scores



Control de Calidad

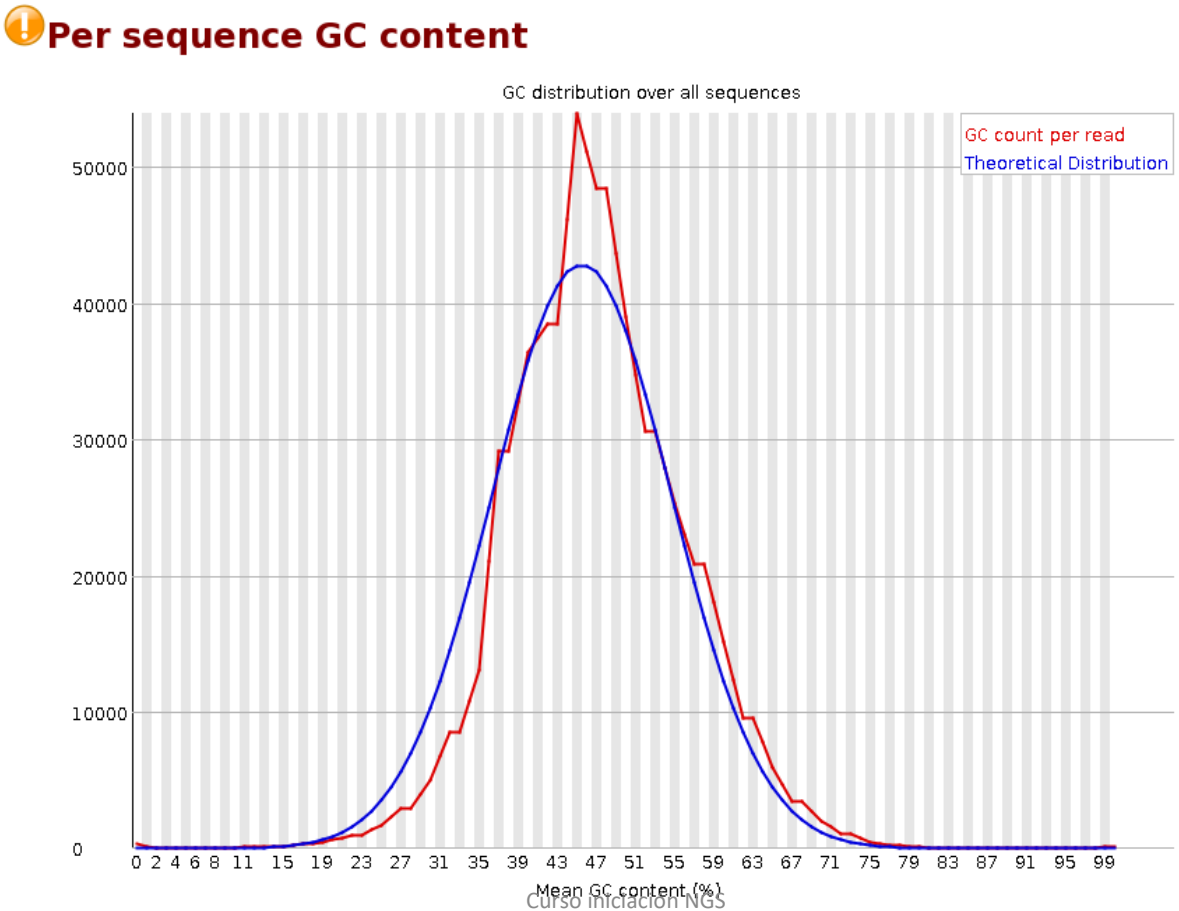
- Contenido por base

Per base sequence content



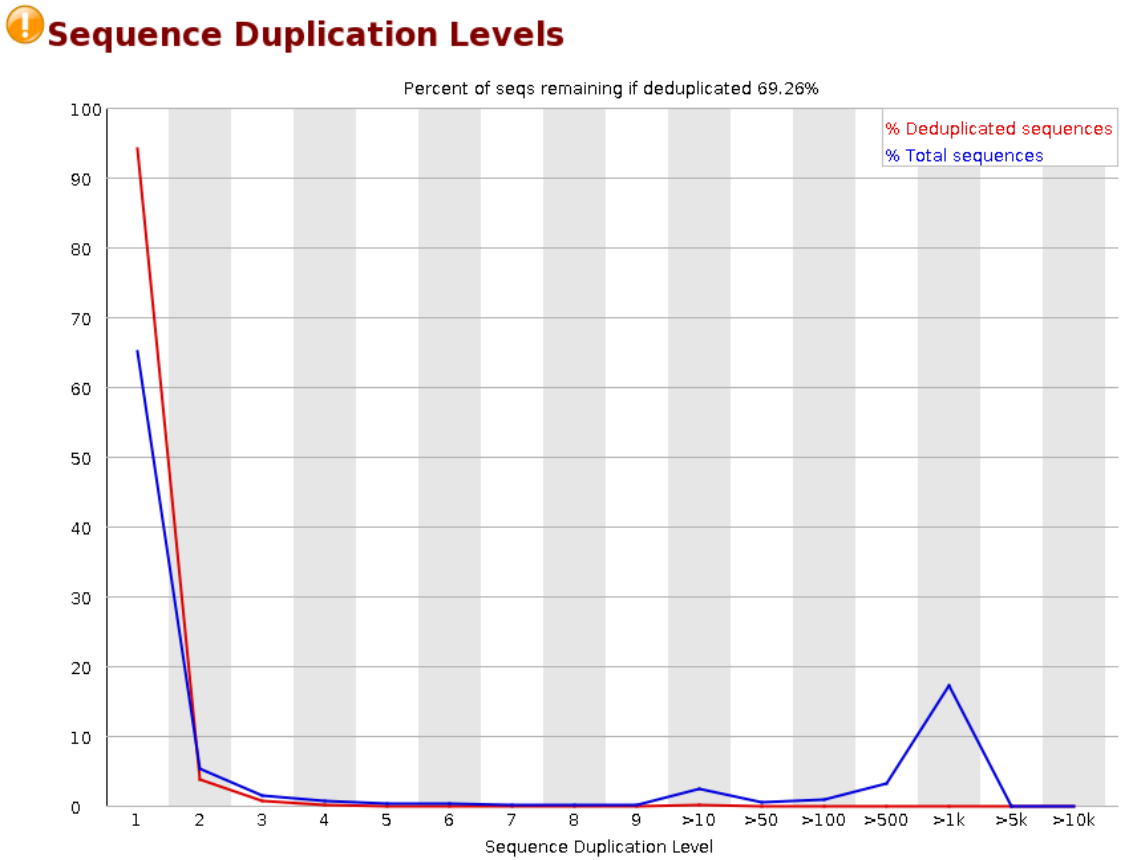
Control de Calidad

- Porcentaje de QC



Control de Calidad

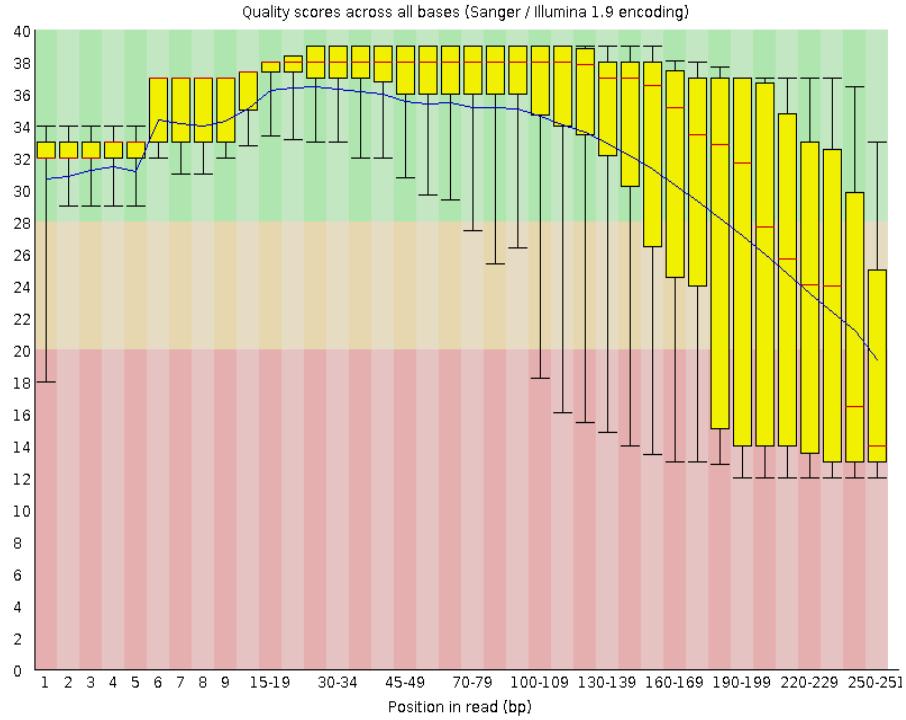
- Porcentaje de duplicados



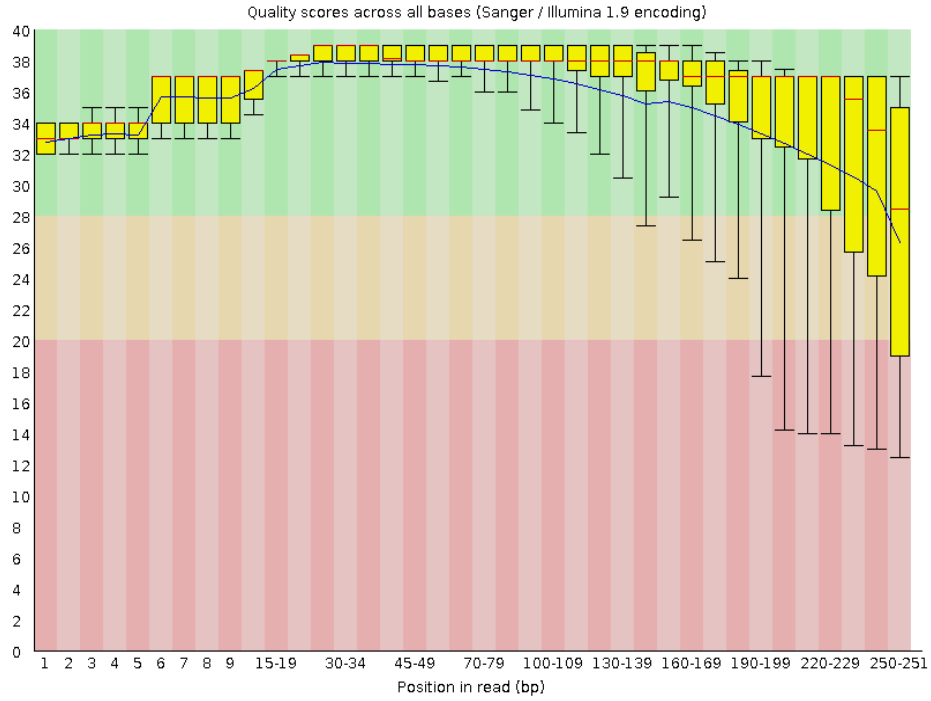
Control de Calidad

- Asimetría en MiSeq

Per base sequence quality



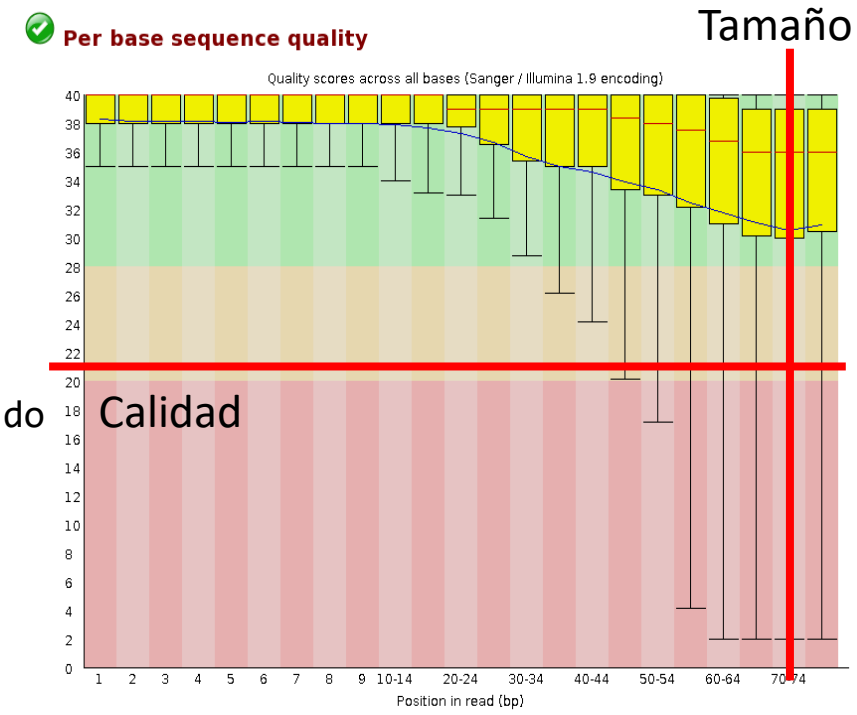
Per base sequence quality



Preprocesamiento: filtrado

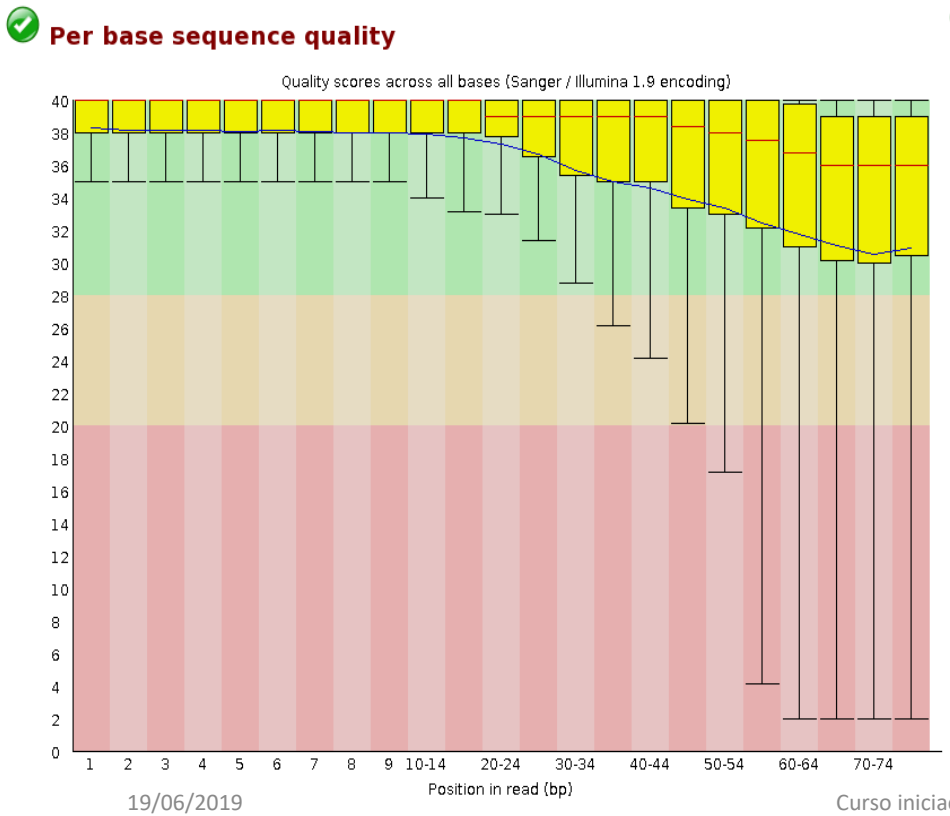
- Quitar adaptadores residuales
  - Según librería utilizada

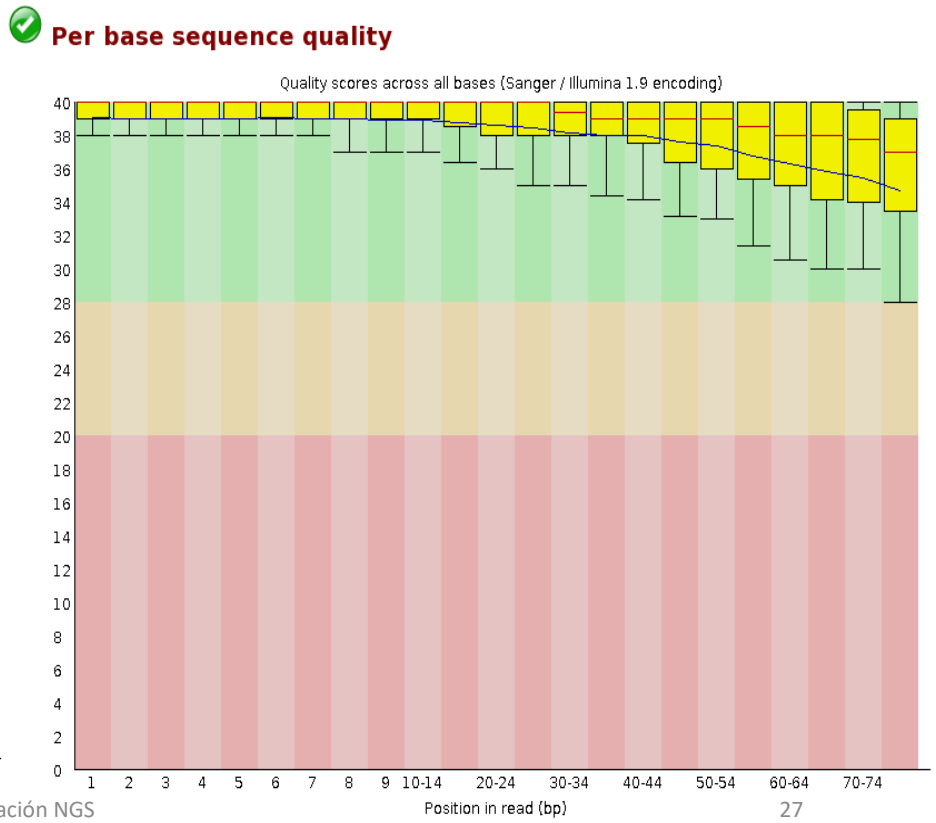
- Distintos parámetros de filtrado
  - Filtro por calidad
    - Media de calidad del read
    - Porcentaje de calidad
      - Extremo del read
      - Ventana deslizante
  - Filtro por tamaño del read.
    - Tamaño fijo del read
    - Tamaño restante después de filtrado



Preprocesamiento: filtrado

- Ejemplo filtro por porcentaje de bases por debajo de Q





Preprocesamiento: filtrado

- Ejemplo de estadísticas finales que se obtienen

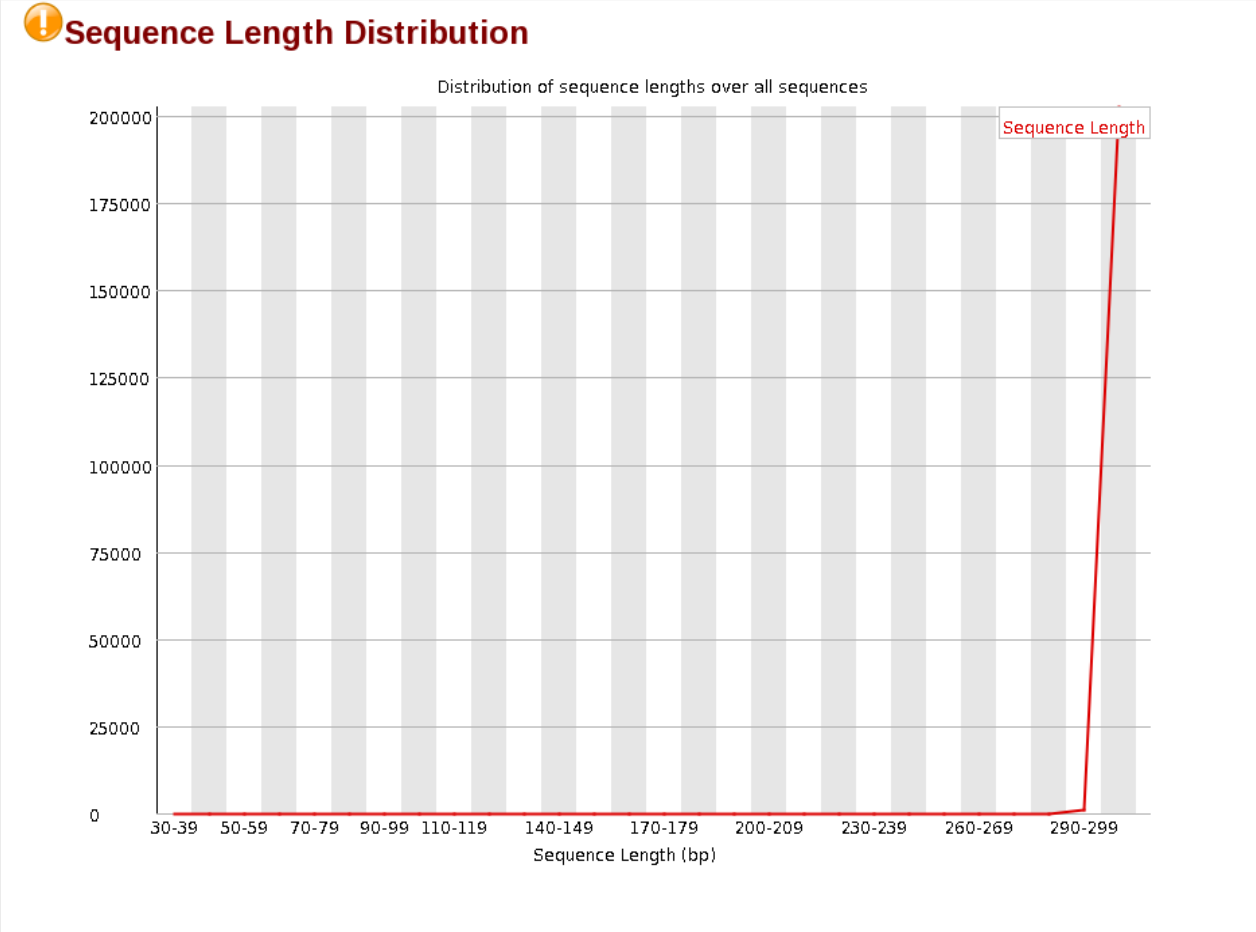
1. Preprocessing: Filter and Quality Control

Filtering Options:

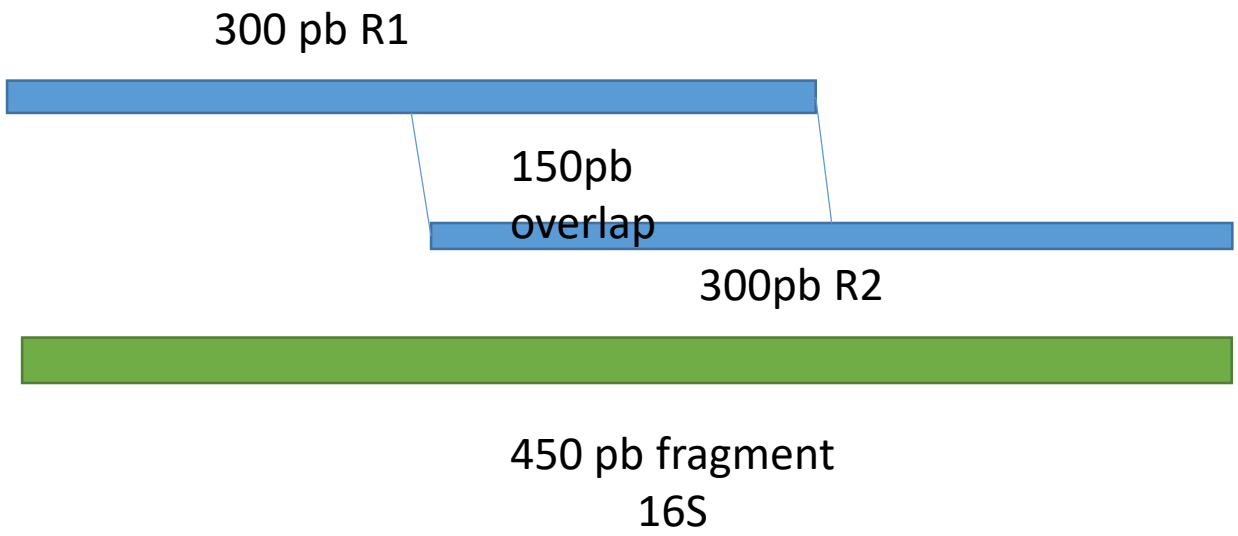
- Single-End
  - 70 % of the bases with more than 30 phred quality
  - Trimming: trim bases from the rigth with less than 30 phred quality without allowing less than length 70.
- Paired-End
  - 70 % of the bases with more than 30 phred quality
  - No trimming

Sample	3233-S.fastq	3233-T.fastq	3353-S.fastq	3353-T.fastq
Pre-Filter				
Sequence length	75-76	75-76	75-76	75-76
Total Sequences	101204363	128694353	134984248	127907012
%GC	44	45	46	46
Post-Filter				
Sequence length	70-76	70-76	70-76	70-76
Total Sequences	83027834	104199275	108632895	102735796
%GC	43	44	44	44

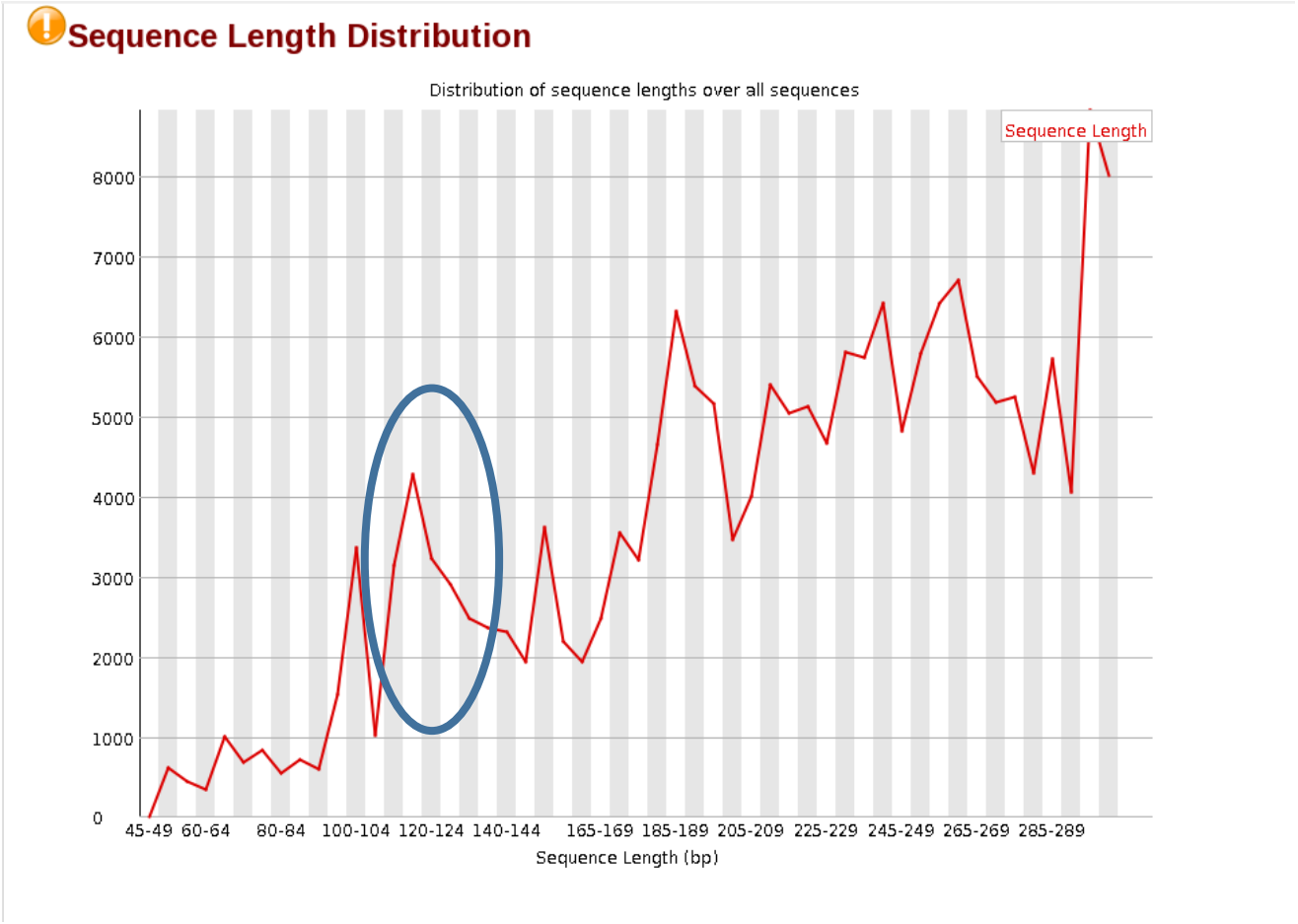
Ejemplo de filtrado en metagenómica



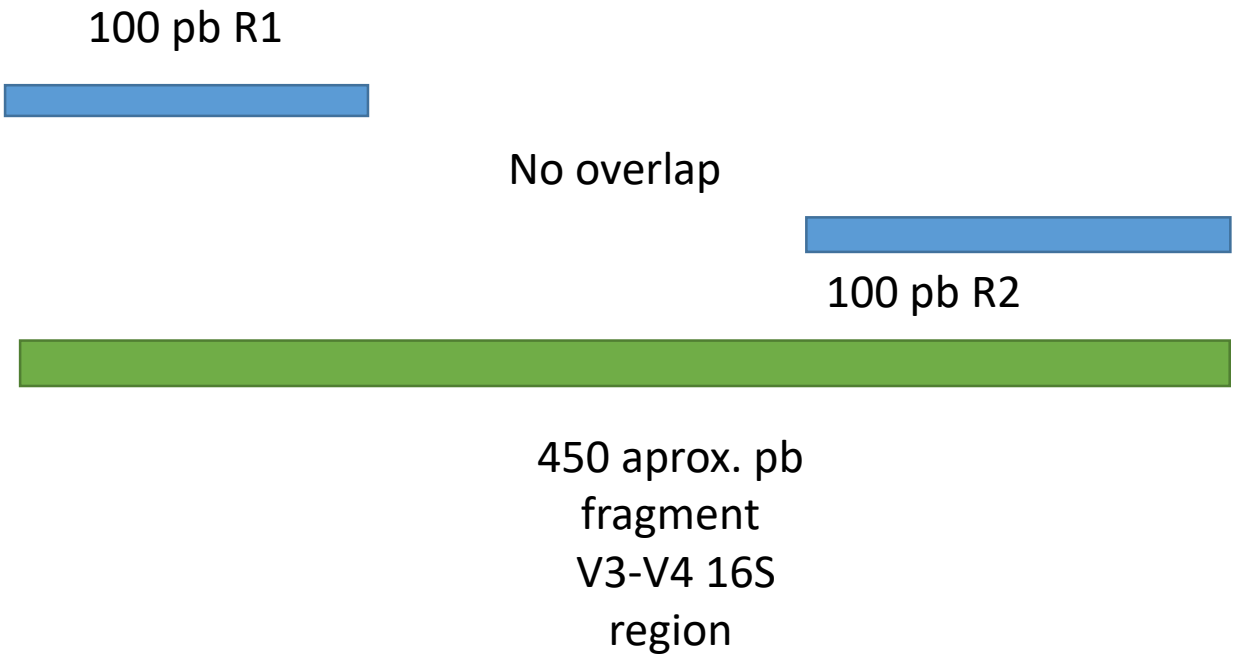
Ejemplo de filtrado en metagenómica



Ejemplo de filtrado en metagenómica



Ejemplo de filtrado en metagenómica



Objetivos de la práctica

- Visualización de ficheros fastq reales.
- Contabilización de número de reads.
- Pasar de fastq a fasta fácilmente.
- Ejecución de fastQC en muestras *dummy*
- Ejecución de NGSQCToolkit y Trimmomatic de muestas dummy

¿Preguntas?