

Calidad de los transcriptomas

2025-03-03

Descripcion del equipo

- **Equipo:** #1
- **Integrantes:**
 - Sofia Gamiño Estrada **sgamino**
 - Jorge Alfredo Suazo Victoria **jsuazo**
 - Emiliano Ferro Rodriguez **eferro**
- **Correo Electronico:**
 - ghobibohg@gmail.com
 - emiferro@comunidad.unam.mx
 - jasvpj@gmail.com

Descripcion de los datos

Descripción	Información
Bioproject	PRJNA494527
Especie	<i>Homo Sapiens</i>
Tipo de biblioteca	single-end
Método de Selecccion	RNA-Total
Número de transcriptomas	34
Número de réplicas biológicas	17 Replicas Biologicas por condicion (Control y Firma génica inducida por glucocorticoides en la piel humana)
Secuenciador Empleado	Illumina NextSeq 500
Profundidad de secuenciación de cada transcriptoma	12M a 40M
Tamaño de las lecturas	75 bp
Articulo Cientifico	Sarkar MK, Kaplan N, Tsoi LC, Xing X et al. Endogenous Glucocorticoid Deficiency in Psoriasis Promotes Inflammation and Abnormal Differentiation. J Invest Dermatol 2017 Jul;137(7):1474-1483. PMID: 28259685 Los datos se pueden descargar desde NCBI o usando ENA.

run_accession	sample_title	sample_alias	Individuals	Type
SRR7959189	aaF1	GSM3415442	1	Control
SRR7959190	aaF1_GC	GSM3415443	1	GC
SRR7959191	aaF2	GSM3415444	2	Control

run_accession	sample_title	sample_alias	Individuals	Type
SRR7959192	aaF2_GC	GSM3415445	2	GC
SRR7959193	aaF3	GSM3415446	3	Control
SRR7959194	aaF3_GC	GSM3415447	3	GC
SRR7959195	aaM1	GSM3415448	4	Control
SRR7959196	aaM1_GC	GSM3415449	4	GC
SRR7959197	aaM2	GSM3415450	5	Control
SRR7959198	aaM2_GC	GSM3415451	5	GC
SRR7959199	aaM3	GSM3415452	6	Control
SRR7959200	aaM3_GC	GSM3415453	6	GC
SRR7959201	HS15	GSM3415454	7	Control
SRR7959202	HS16	GSM3415455	7	GC
SRR7959203	HS17	GSM3415456	8	Control
SRR7959204	HS18	GSM3415457	8	GC
SRR7959205	HS19	GSM3415458	9	Control
SRR7959206	HS20	GSM3415459	9	GC
SRR7959207	HS1233_control_S10_R1_001	GSM3415460	10	Control
SRR7959208	HS1233_Gc_S11_R1_001	GSM3415461	10	GC
SRR7959209	HS1583_control_S1_R1_001	GSM3415462	12	Control
SRR7959210	HS1583_Gc_S2_R1_001	GSM3415463	12	GC
SRR7959211	HS1609_control_S12_R1_001	GSM3415464	13	Control
SRR7959212	HS1609_GC_S13_R1_001	GSM3415465	13	GC
SRR7959213	HS1625_control_S3_R1_001	GSM3415466	14	Control
SRR7959214	HS1625_Gc_S4_R1_001	GSM3415467	14	GC
SRR7959215	HS1659_control_S14_R1_001	GSM3415468	15	Control
SRR7959216	HS1659_GC_S15_R1_001	GSM3415469	15	GC
SRR7959217	HS1660_control_S16_R1_001	GSM3415470	16	Control
SRR7959218	HS1660_GC_S17_R1_001	GSM3415471	16	GC
SRR7959219	HS2100_control_S18_R1_001	GSM3415472	17	Control
SRR7959220	HS2100_GC_S19_R1_001	GSM3415473	17	GC
SRR7959221	HS2191_GC_S8_R1_001	GSM3415474	18	Control
SRR7959222	HS2191_control_S7_R1_001	GSM3415475	18	GC

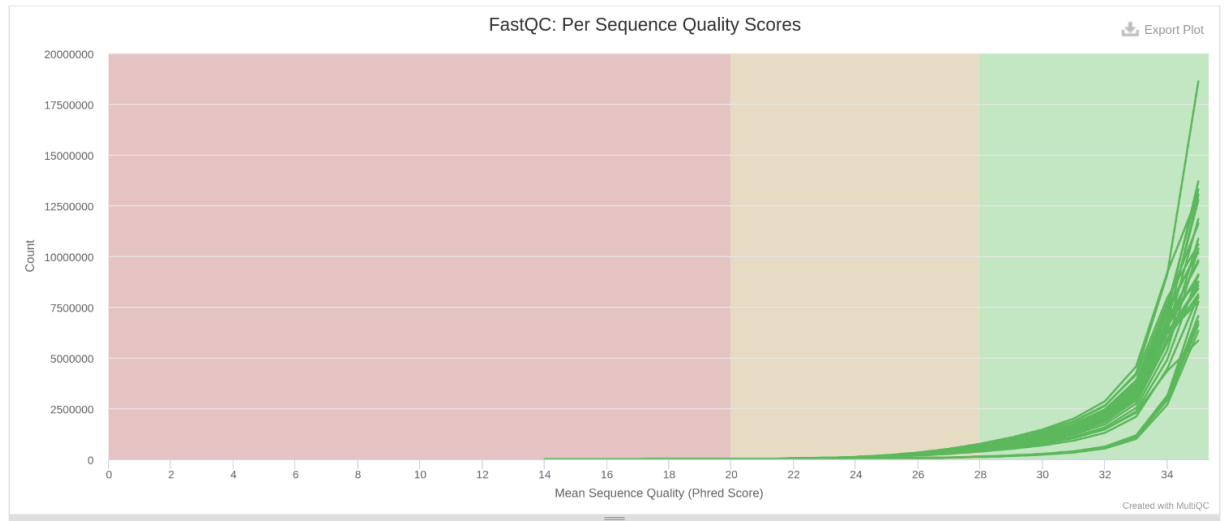
Calidad de las secuencias de los datos crudos

Per Sequence Quality Scores

32

The number of reads with average quality scores. Shows if a subset of reads has poor quality. See the [FastQC help](#).

Y-Limits: ☒ on



Per Base Sequence Content

32

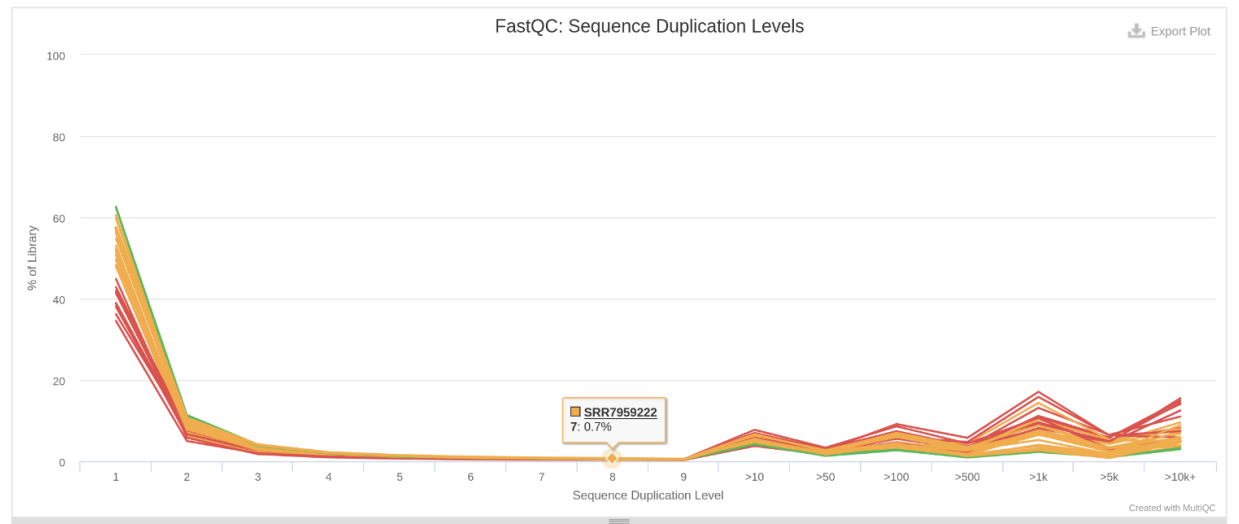
The proportion of each base (A, C, G, T) for each of the four bases (A, C, G, T) has been called. See the [FastQC help](#).

Sequence Duplication Levels

2 20 10

The relative level of duplication found for every sequence. See the [FastQC help](#).

Y-Limits: ☒ on



Conclusión sobre los datos

¿Son viables para continuar el análisis?

No, los datos como están no son viables para continuar con el análisis ya que hay muchas secuencias duplicadas (adaptadores) que pueden interferir al momento de alinear, ya que si los dejamos así pueden llegar a alinear en muchos sitios.

General Statistics

Copy table

Configure Columns

Plot

Showing 32/32 rows and 3/5 columns.

Sample Name	% Dups	% GC	M Seqs
SRR7959210	35.1%	46%	27.3
SRR7959211	43.0%	49%	24.7
SRR7959212	45.9%	49%	29.4
SRR7959213	31.9%	46%	31.4
SRR7959214	29.1%	45%	31.3
SRR7959215	32.5%	47%	27.0
SRR7959216	42.1%	49%	24.1
SRR7959217	61.1%	52%	27.2
SRR7959218	38.1%	50%	23.5
SRR7959219	36.3%	47%	28.9
SRR7959220	41.5%	48%	24.6
SRR7959221	29.3%	45%	23.5
SRR7959222	34.3%	45%	26.6

FastQC

FastQC is a quality control tool for high throughput sequence data, written by Simon Andrews at the Babraham Institute in Cambridge.

Sequence Quality Histograms 32

The mean quality value across each base position in the read. See the [FastQC help](#).

Y-Limits: on



Figure 1:

Per Base Sequence Content

32

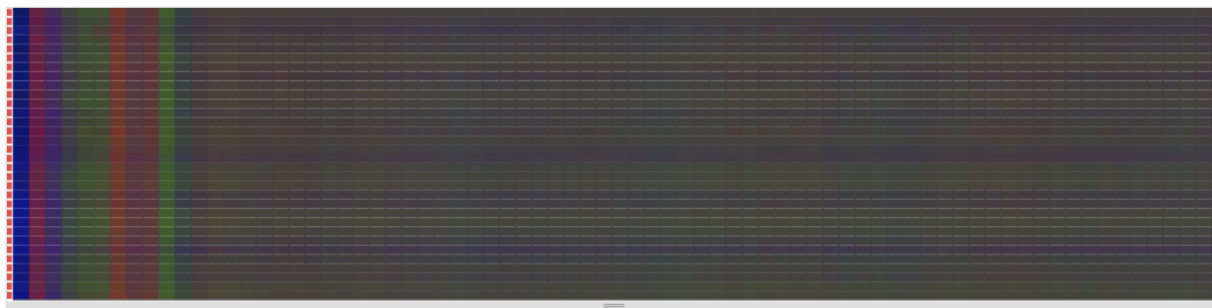
The proportion of each base position for which each of the four normal DNA bases has been called. See the [FastQC help](#).

Click a sample row to see a line plot for that dataset.

Rollover for sample name

Position: - %T: - %C: - %A: - %G: -

Export Plot



Per Sequence GC Content

4 24 4

The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content. See the [FastQC help](#).

Y-Limits: on

Percentages Counts

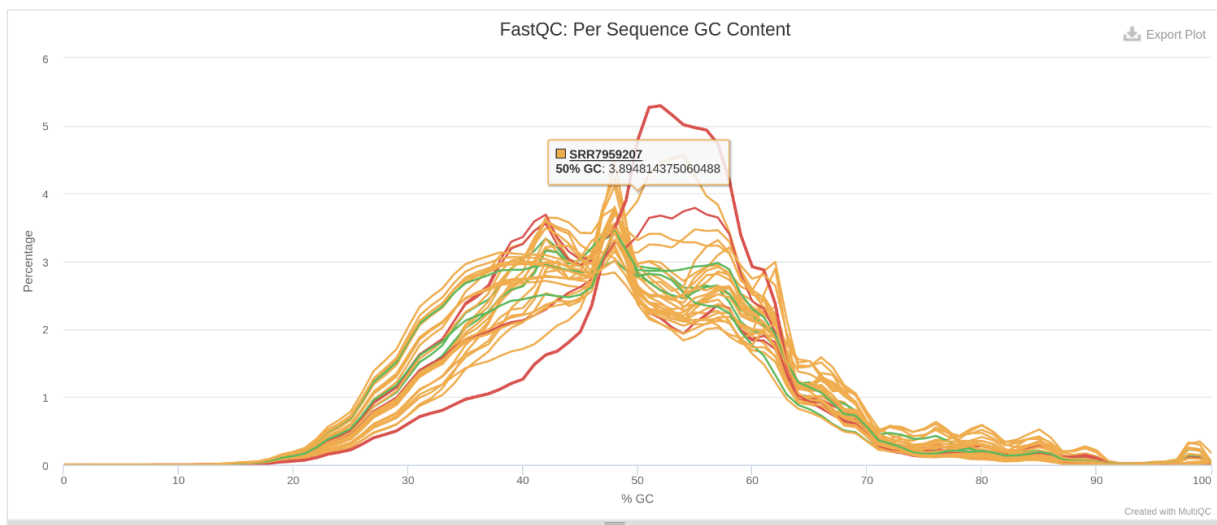


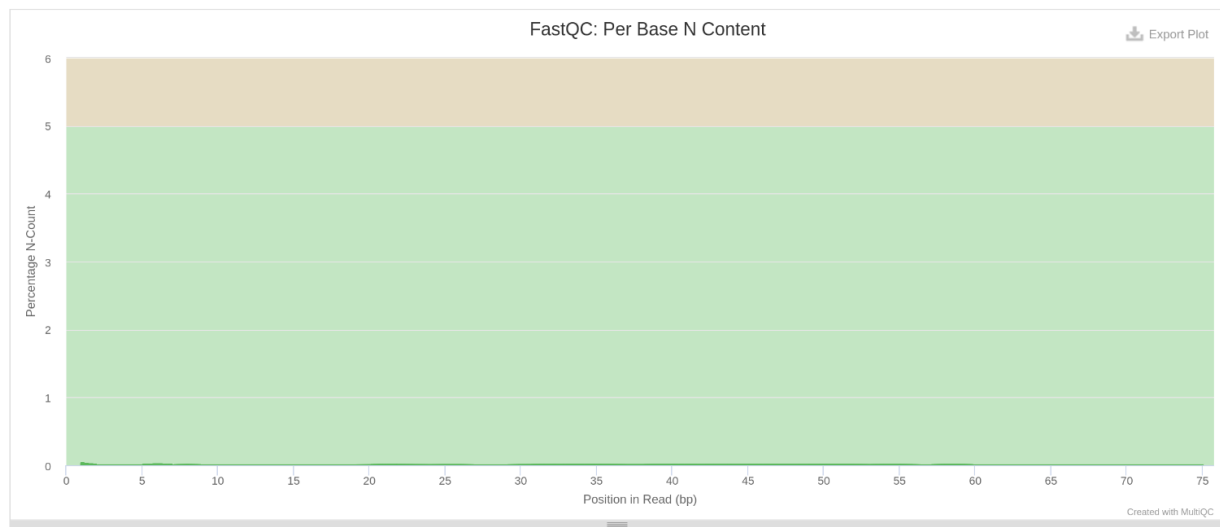
Figure 2:

Per Base N Content

32

The percentage of base calls at each position for which an N was called. See the [FastQC help](#).

Y-Limits: ☒ On



Sequence Length Distribution

32

All samples have sequences of a single length (75bp).

Figure 3:

Esto es evidente gracias a la alta cantidad de transcriptomas con un contenido de GC no adecuado, el hecho de que todos los fastq muestran una cantidad desbalanceada de bases por secuencia, las secuencias sobrerrepresentadas en una cantidad alta y los niveles altos de duplicación. Todos estos son signos de que los archivos fastq están crudos.

Algo importante a notar es el contenido de GC. Errores en este test pueden deberse a contaminaciones del tejido epitelial por virus o bacterias, lo cual no es nada raro en la piel humana. Es importante considerar esto, ya que en caso de que no mejore se tendría que volver a mandar a secuenciar pero ahora en condiciones estériles y más estrictas. H

¿Qué pasos deben seguirse para mejorar la calidad de los datos?

Aparte de los adaptadores en las reads, tenemos una calidad de transcriptomas muy buena y con buena cobertura, por lo cual la secuenciación fue sin problemas.

El principal problema que veo aparte de los adaptadores es checar que tanta contaminación tienen nuestras muestras con microorganismos, y que tanto pueden afectar la calidad de nuestras lecturas. En caso de que sea muy prevalente, sería bueno separar las lecturas entre las células epiteliales y los micro-organismos.

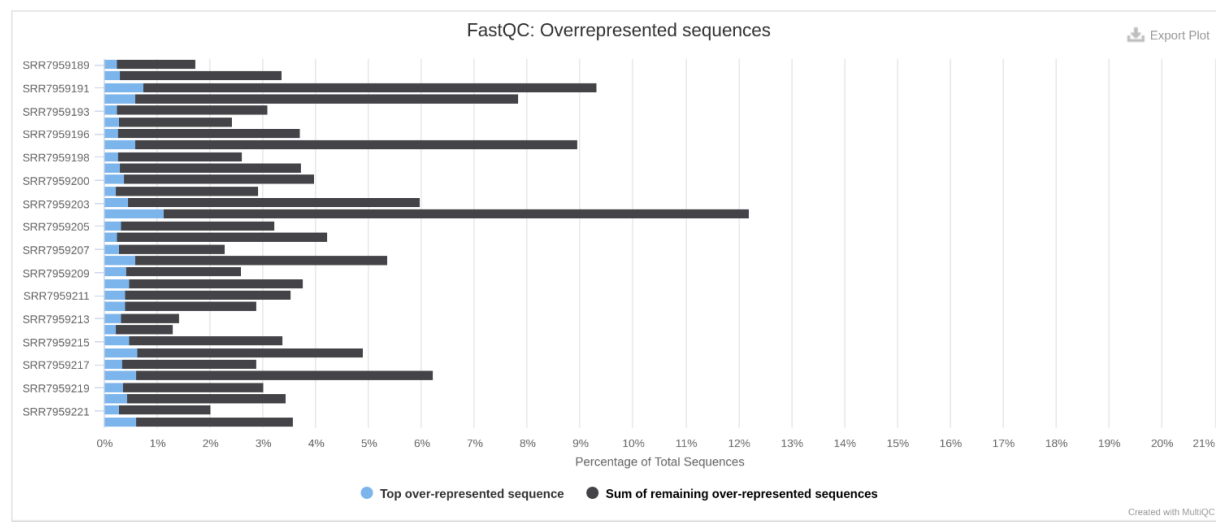
Entonces:

1. Hacer trimming de los adaptadores, cuidando que la longitud de nuestras reads se mantenga entre 50 y 60.
2. Volver a hacer un multiqc como este y volver a evaluar las calidades.
3. Checar si ha cambiado el contenido de GC. En caso de que no, seguir con los siguientes pasos.
4. Identificar la especie a la que pertenecen los transcritos contaminantes, y agruparlos para dejarlos fuera de nuestras lecturas.

Overrepresented sequences

31

The total amount of overrepresented sequences found in each library. See the [FastQC help](#) for further information.



Adapter Content

32

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position. See the [FastQC help](#). Only samples with $\geq 0.1\%$ adapter contamination are shown.

Y-Limits: ☒ on

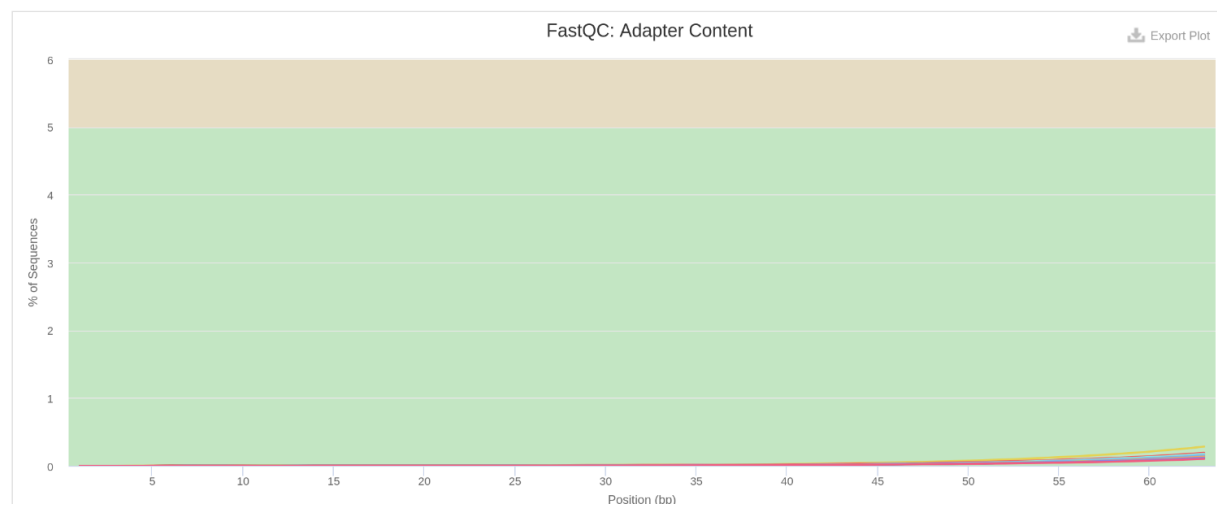


Figure 4:

5. Volver a hacer un multiqc, y repetir los pasos a partir del 3 en caso de ser necesario.