Módulo:
Expresión diferencial

Bioinformática y Estadística 2

Dra. Evelia Coss

Dra. Alejandra Medina

21 al 24 de Febrero, 2023

Día 1

- Transcriptoma
- Variaciones en los transcriptomas
- Cuestiones experimentales en RNA-Seq
- Aspectos generales de Genética
- Tipos de bibliotecas (*paired-end y single-end*)
- Strand-specific
  - *Paired-end* y strand-specific
- Número de replicas
- Diseño de Secuenciación
- Pipeline bioinformática
  - Quality Check

Objetivo: Hacer de ustedes **bioinformáticos** aptos en sus nuevos laboratorios.

Transcriptoma

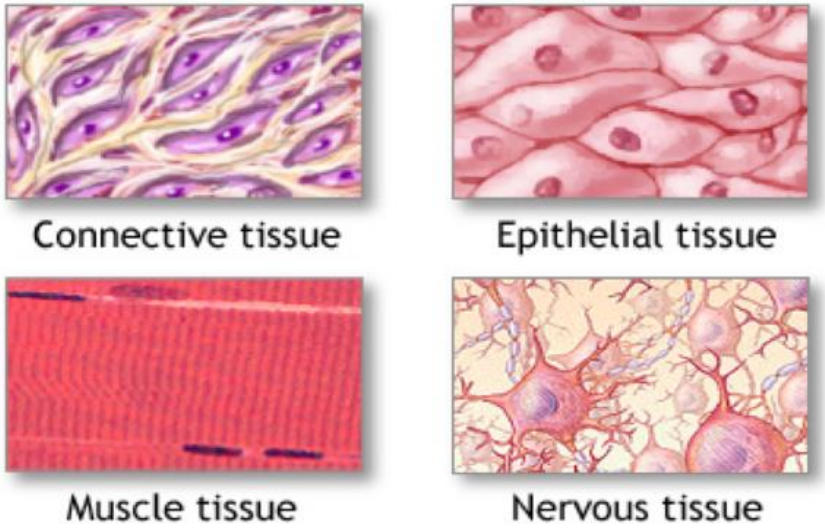
- Es el conjunto de todos las moléculas de RNA producidos por el genoma bajo **condiciones específicas** o en una **célula específica (scRNA-Seq)** o en una **población de celulas (bulk RNA-Seq)**.

Palabras claves:

- Genoma - Fijo
- Transcriptoma - Altamente variable

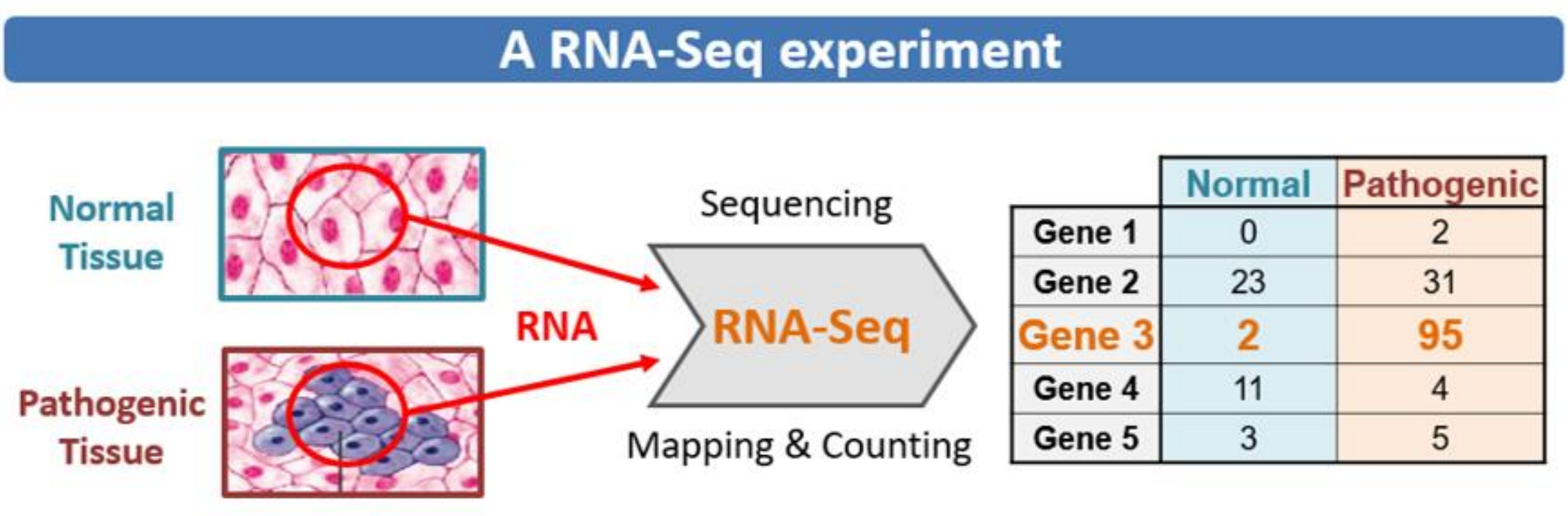
El transcriptoma varía según:

- Tejido / Órgano
- Célula
- Ambiente (estrés)
- Medicamentos (tratamientos)
- Salud
- Edad
- Etapa del desarrollo



Idea principal de RNA-Seq

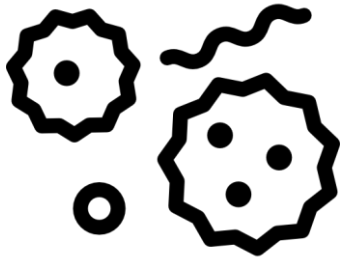
Relacionar un fenotipo con los cambios de expresión de los genes en una condición dada

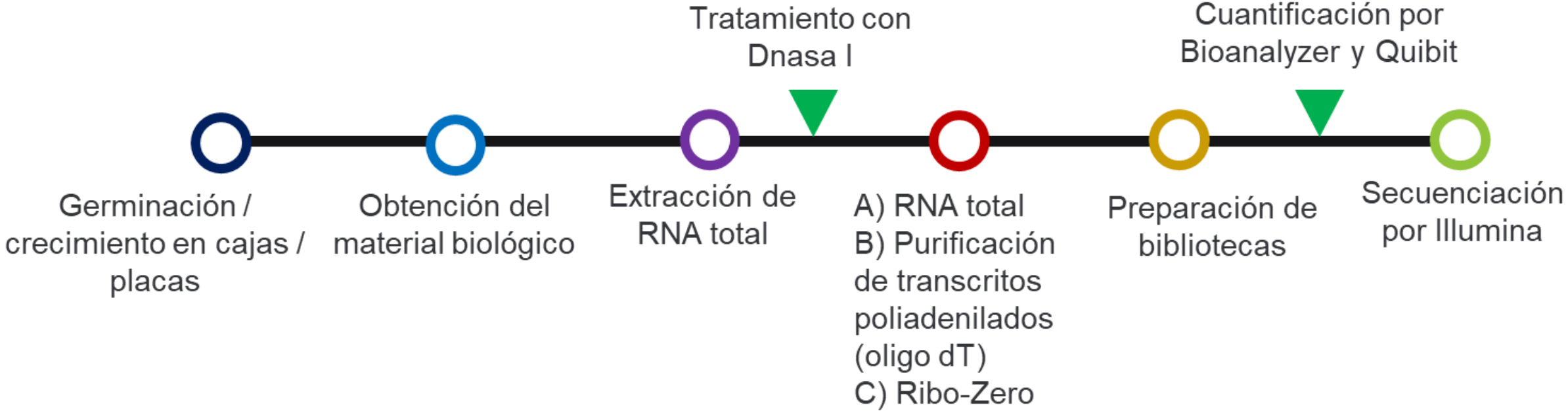


	Normal	Pathogenic
Gene 1	0	2
Gene 2	23	31
Gene 3	2	95
Gene 4	11	4
Gene 5	3	5

Imagen tomada de: https://www.acobiom.com/en/rna-seq-en/

Flujo experimental de RNA-Seq





Aproximadamente el 2 % del RNA es mRNA en células eucariotas

- 80 % rRNA
- 15 % tRNA
- 5 % otros (mRNA, non-coding RNA)

Lodish H, Berk A, Zipursky SL, et al. New York: W. H. Freeman 2000. Molecular Cell Biology. 4th edition

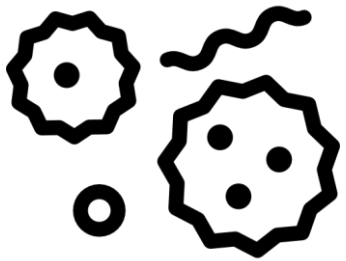
Aproximadamente el 2 % del RNA es mRNA en células eucariotas

- 80 % rRNA
- 15 % tRNA
- 5 % otros (mRNA, non-coding RNA)

Esperas tener ~ 10 % diferencialmente expresado

Lodish H, Berk A, Zipursky SL, et al. New York: W. H. Freeman 2000. Molecular Cell Biology. 4th edition

Flujo experimental de RNA-Seq



¿Qué pasa si seleccionamos alguna de estas opciones?

Germinación / crecimiento en cajas / placas

Obtención del material biológico

Extracción de RNA total

Tratamiento con Dnasa I

A) RNA total
B) Purificación de transcritos poliadenilados (oligo dT)
C) Ribo-Zero

Preparación de bibliotecas

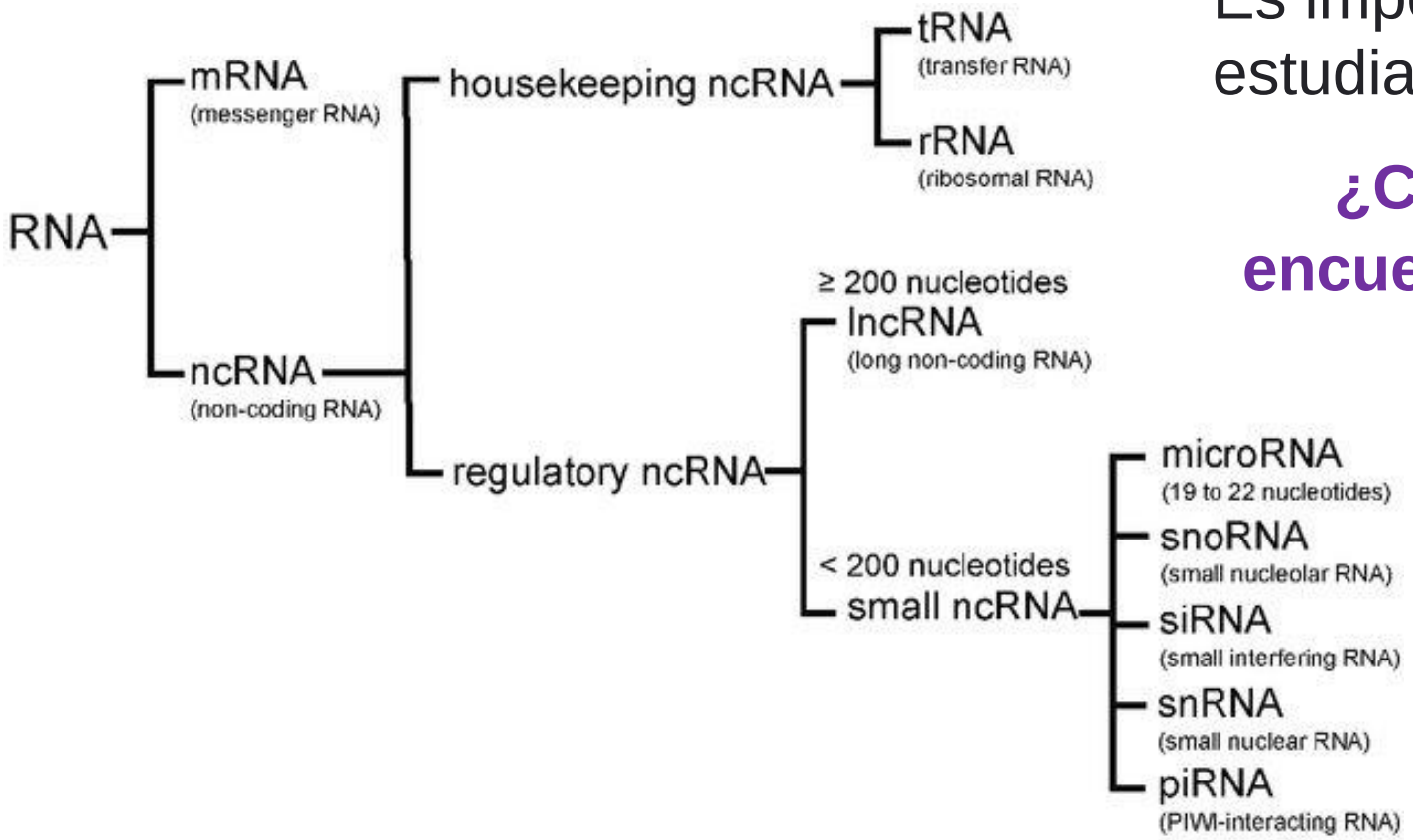
Cuantificación por Bioanalyzer y Quibit

Secuenciación por Illumina

Aspectos generales de Genética

- ¿Cuántos **tipos de RNA** existen? ¿Y en qué especies se encuentran?
- ¿**Un RNA** solo es transcrito por **una sola polimerasa**?
- ¿Cuántas **bandas** esperas encontrar en un gel de RNA (integridad)?
- ¿En qué **compartimientos celulares** podemos encontrar al rRNA en células eucariotas?

¿Cuántos tipos de RNA existen? ¿Y en qué especies se encuentran?

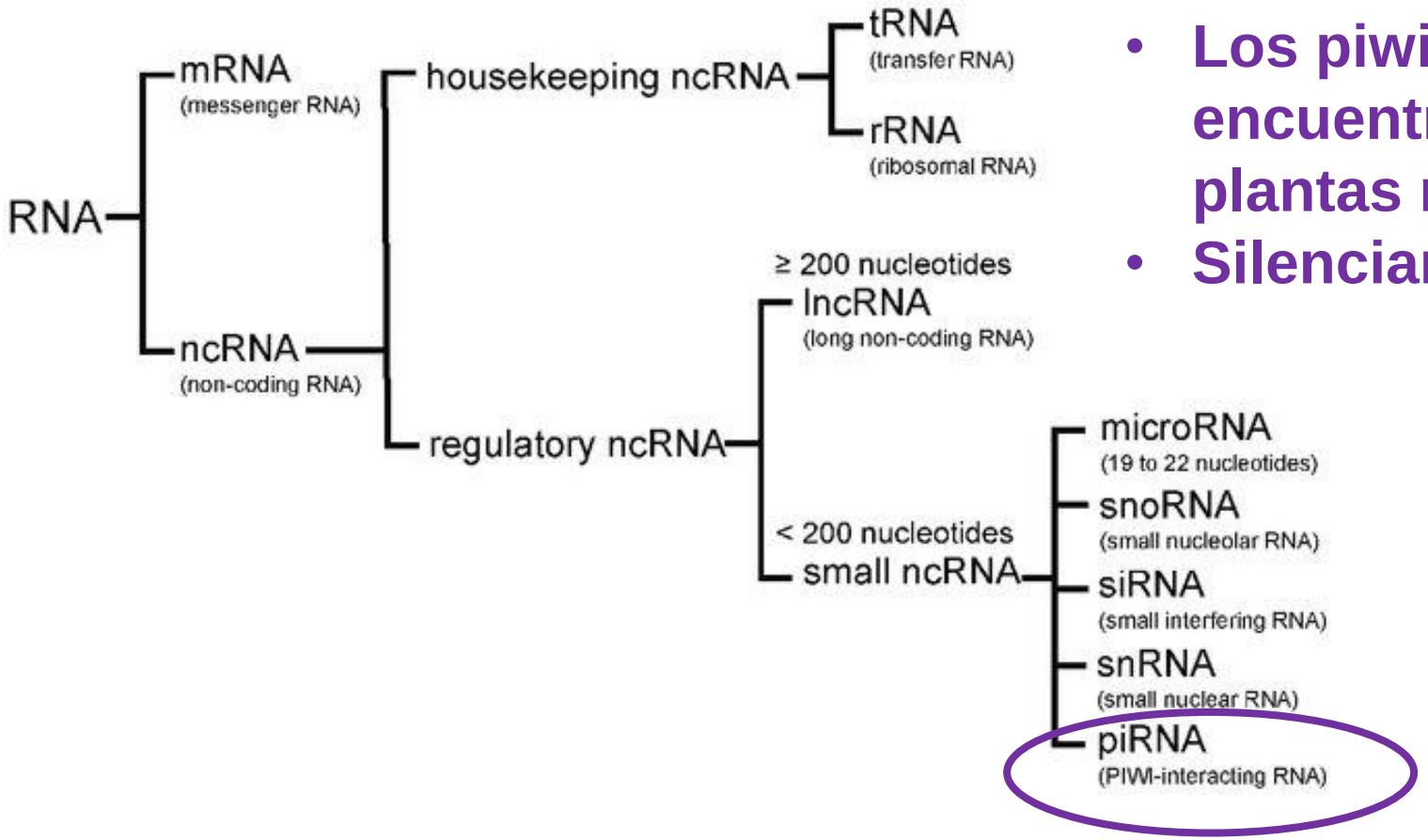


Es importante saber que quieren estudiar y el enfoque de su estudio.

¿Cuál de estos RNA no se encuentra presente en todas las células eucariotas?

Inamura K. 2017. *Cells*

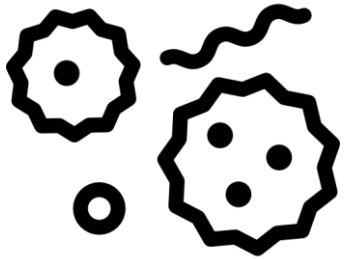
¿Cuántos tipos de RNA existen? ¿Y en qué especies se encuentran?



- **Los piwiRNAs o piRNAs no se encuentran presentes en plantas ni en hongos.**
- **Silenciamiento de transposons.**

Inamura K. 2017. *Cells*

Flujo experimental de RNA-Seq



¿Qué pasa si seleccionamos por la opción B?

Tratamiento con Dnasa I

Cuantificación por Bioanalyzer y Quibit

Germinación / crecimiento en cajas / placas

Obtención del material biológico

Extracción de RNA total

A) RNA total
B) Purificación de transcritos poliadenilados (oligo dT)
C) Ribo-Zero

Preparación de bibliotecas

Secuenciación por Illumina

¿Un RNA solo es transcrito por una sola polimerasa?

NO

- Oligo dT (18 T)
- Los componentes del transcrito del RNA dependerán de la RNA Pol que lo transcriba.

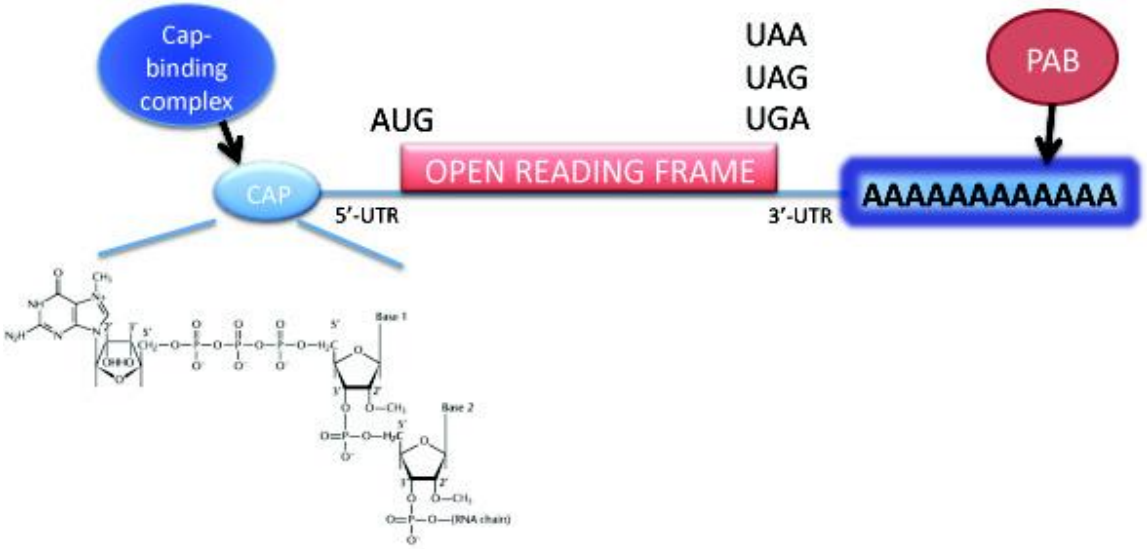
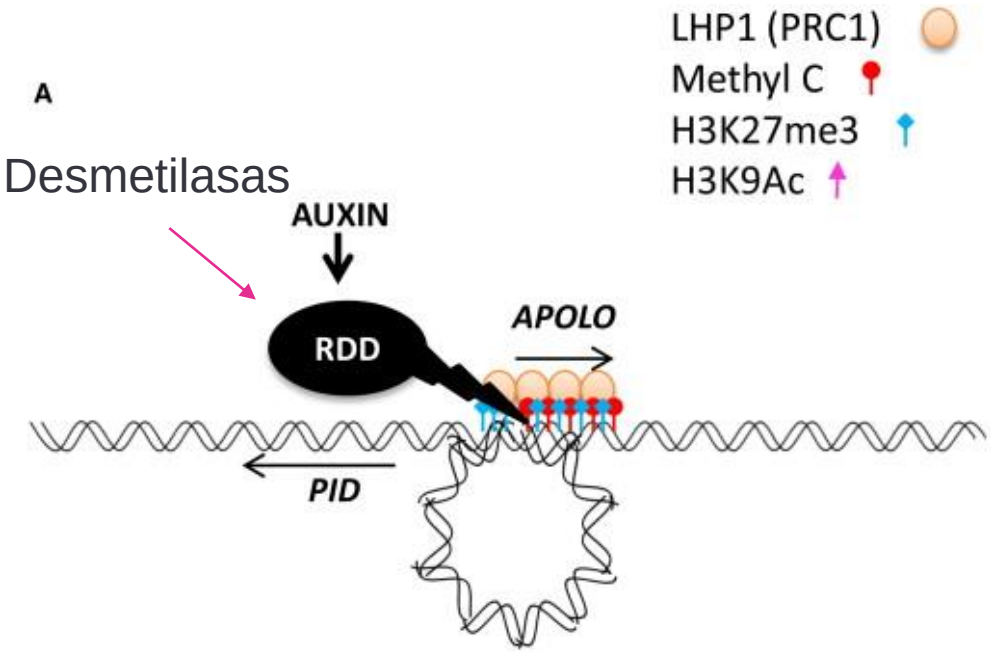
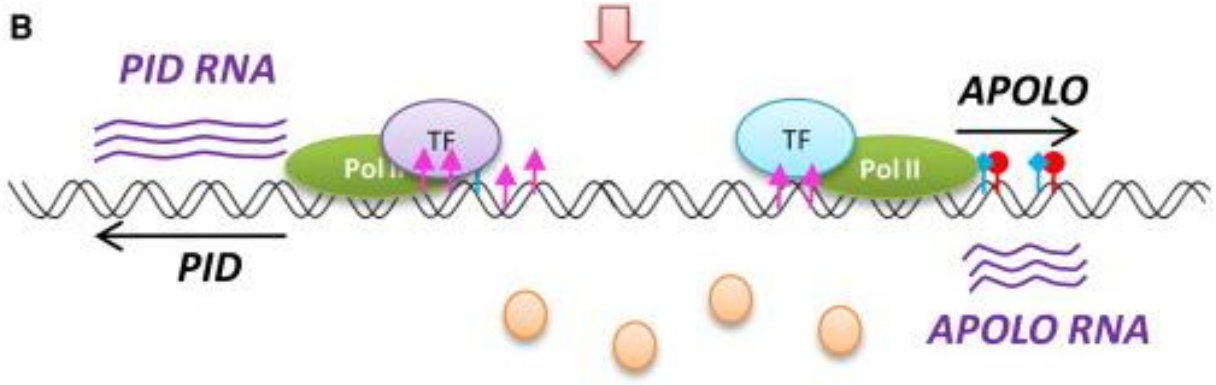


Illustration of the cap taken from Encyclopedia of Life Sciences
John Wiley & Sons, Ltd

El lncRNA *APOLO* es transcrito tanto por RNA Pol II como por RNA Pol IV/V



Los transcritos generados por la RNA Pol II reclutan a PRC2 (H3K27me3)

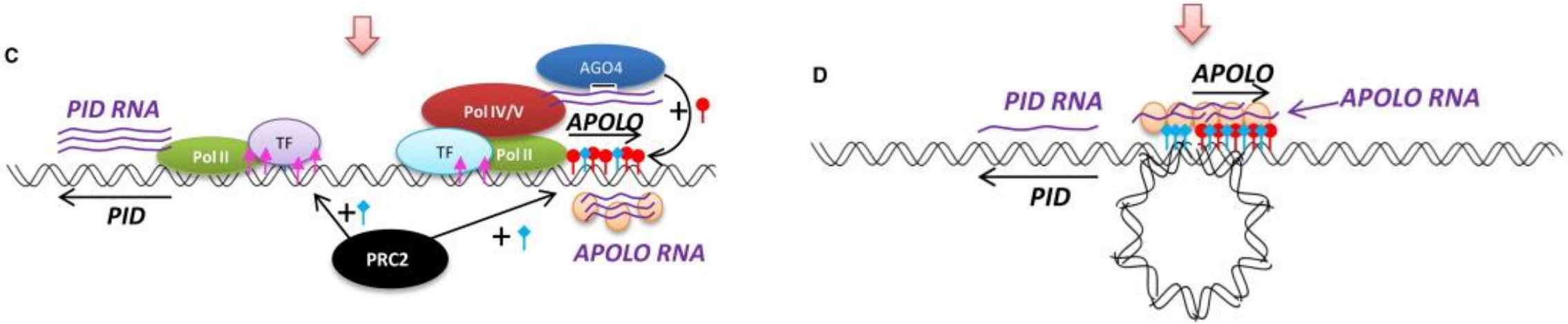


PID (PINOID) : Encargado del Transporte de Auxinas

Ariel, *et al.* 2014. *Molecular Cell*

El lncRNA *APOLO* es transcrito tanto por RNA Pol II como por RNA Pol IV/V

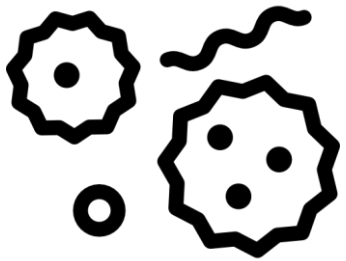
Los transcritos generados por la RNA Pol IV/V activan la maquinaria de silenciamiento por RNA (RdDM)



Pol II - CAP y PolyA
Pol V - CAP

Ariel, *et al.* 2014. *Molecular Cell*

Flujo experimental de RNA-Seq



¿Qué pasa si seleccionamos por la opción C?

Tratamiento con Dnasa I

Cuantificación por Bioanalyzer y Quibit

Germinación / crecimiento en cajas / placas

Obtención del material biológico

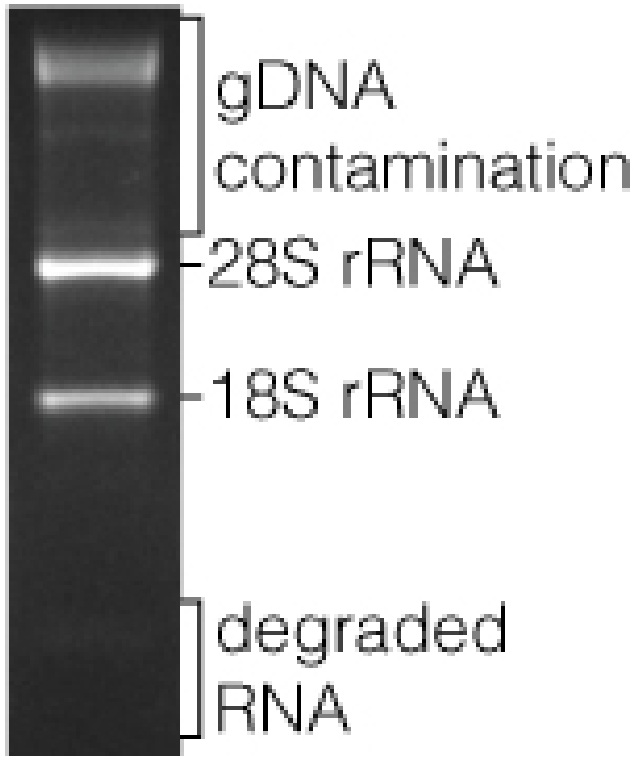
Extracción de RNA total

A) RNA total
B) Purificación de transcritos poliadenilados (oligo dT)
C) Ribo-Zero

Preparación de bibliotecas

Secuenciación por Illumina

¿Cuántas bandas esperas encontrar en un gel de RNA (integridad)?



Normalmente la respuesta es 2, correspondientes al 28S y 18S de rRNA.

Esto no es del todo cierto…

¿En qué compartimientos celulares podemos encontrar al rRNA en células eucariotas?

- **Citoplásmico** – 28S, 18S, 5.8S, 5S
- **Cloroplasto** – 23S, 16S, 5S, 4.5S
- **Mitocondrial** – 18S, 5S

Ribo-Zero y RiboMinus emplean perlas magnéticas para eliminar el rRNA, sin embargo estos beads son específicos de las especies.

- Si no cuentan con tu especie de interés no te conviene esta técnica de aislamiento.

Dependiendo del tejido u órgano analizado podemos localizar mas bandas integras

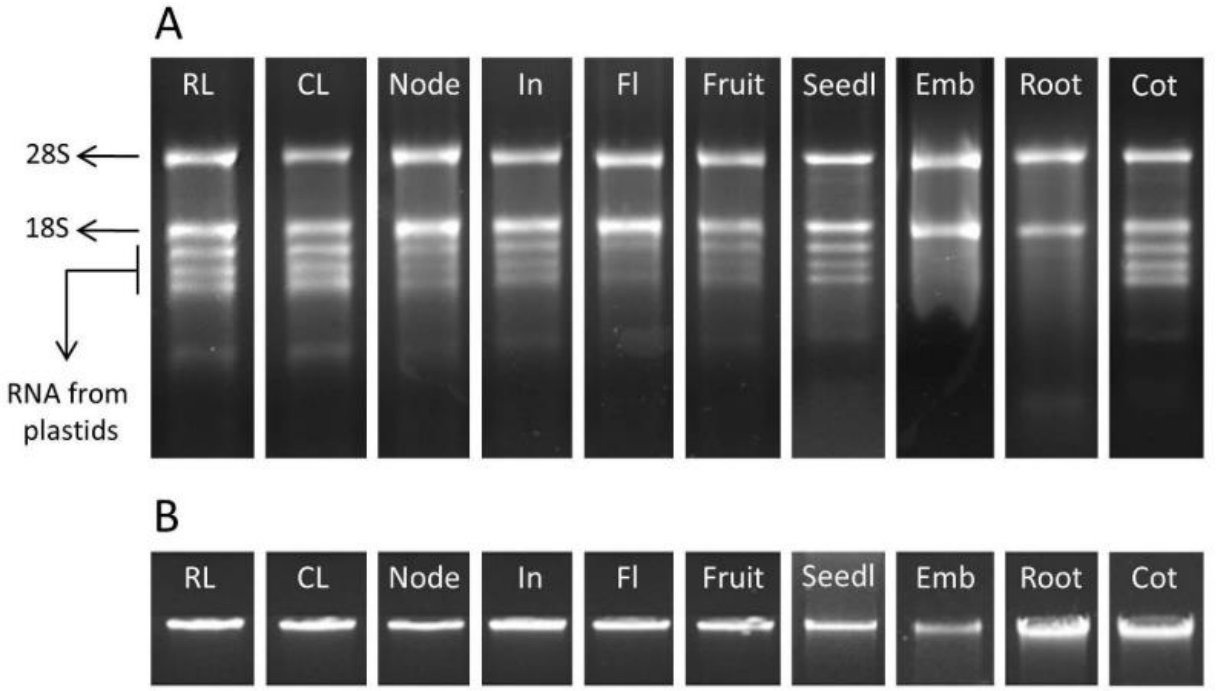
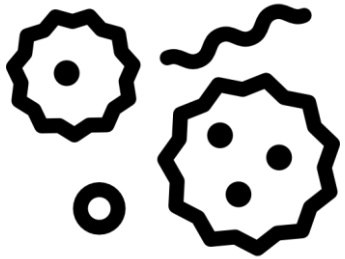
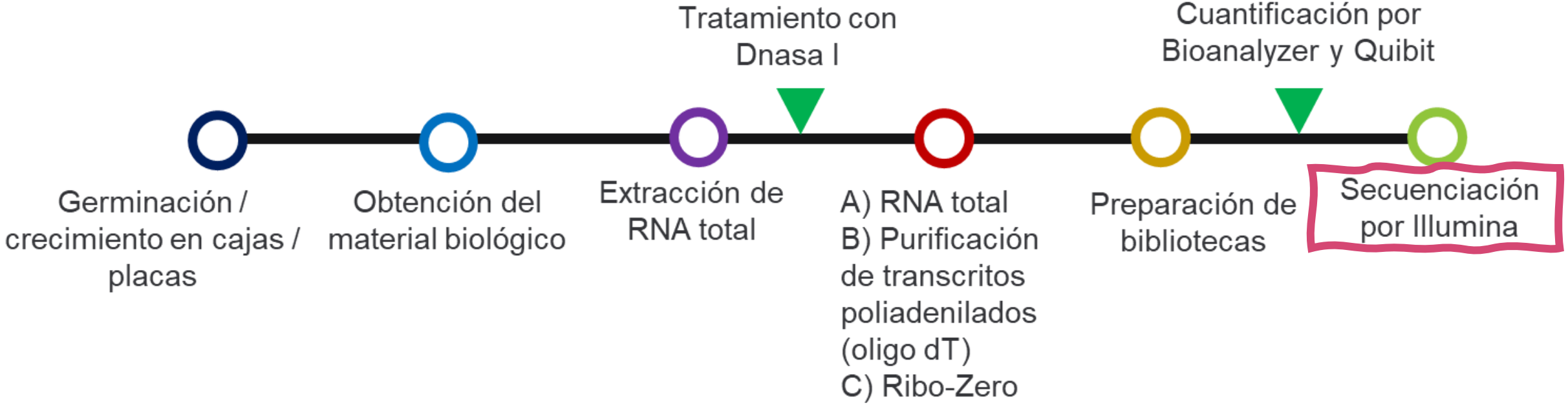


Figure 1. Integrity of RNA and DNA extracted from various *Arabidopsis thaliana* tissues. **A.** Total RNA (1 µg) extracted from various *Arabidopsis thaliana* tissues was fractioned on a denaturing 1% agarose gel. **B.** Fractionation of genomic DNA obtained from the same tissues as those shown in A on a 1% agarose gel. No apparent degradation (smear) is observed on either gel. RL = Rosette leaf; CL = Cauline leaf (bracts); In = Internode; Fl = Flower; Seedl = Seedlings nine days after germination, Emb = Embryo, and Cot = Cotyledon.

An efficient method for simultaneous extraction of high-quality RNA and DNA from various plant tissues. Oliveria et al. 2015. Genetics and molecular research: GMR.

Flujo experimental de RNA-Seq





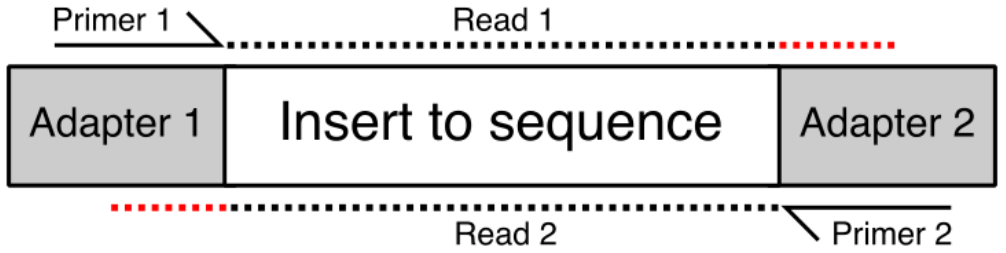
Tipos de bibliotecas

Single-end (SE)

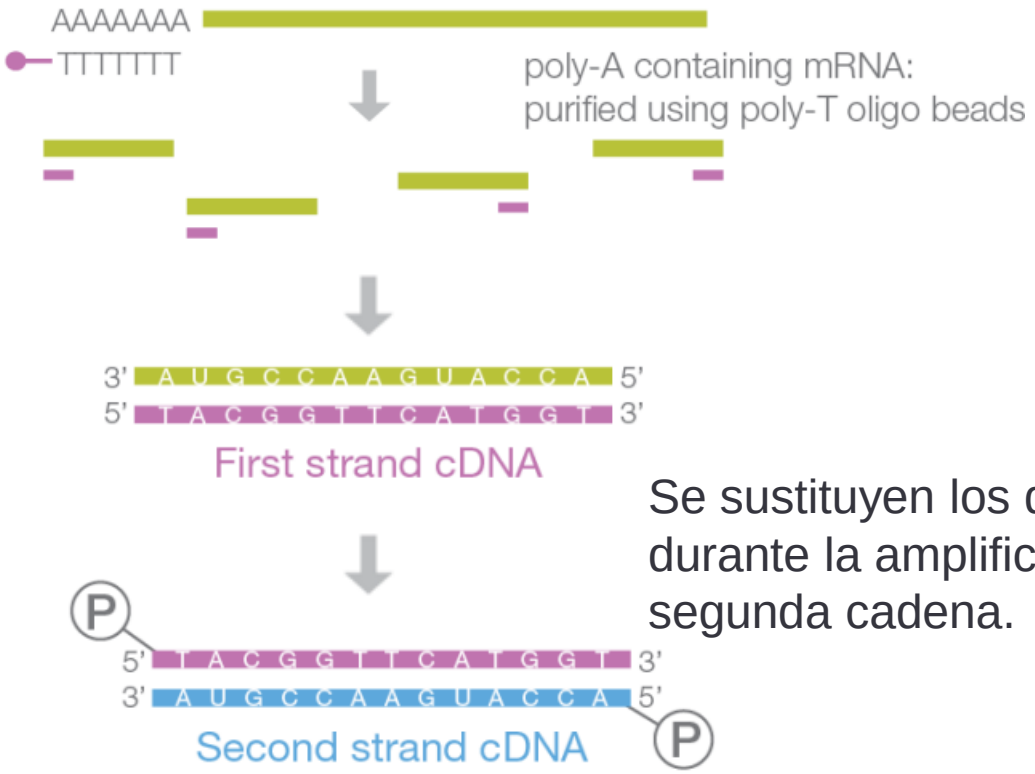
- Organismo bien anotado.
- Bajo costo.
- Solo un sentido en la lectura

Paired-end (PE)

- Anotación de nuevos genes
- Análisis de expresión de isoformas
- Análisis de expresión de genes antisenido



Strand-specific

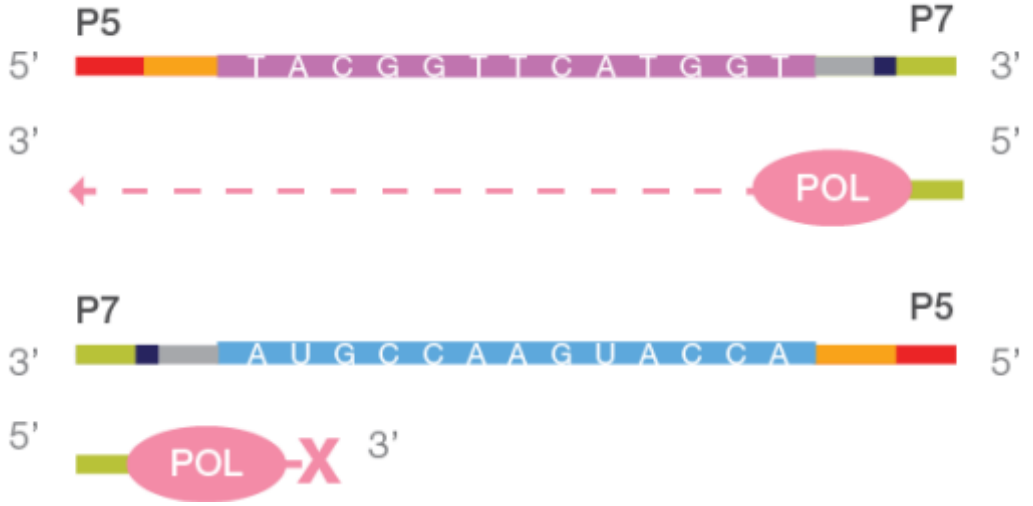


Adenylate 3' ends

Ligate Adapters

Se sustituyen los dTTp por dUTP durante la amplificación de la segunda cadena.

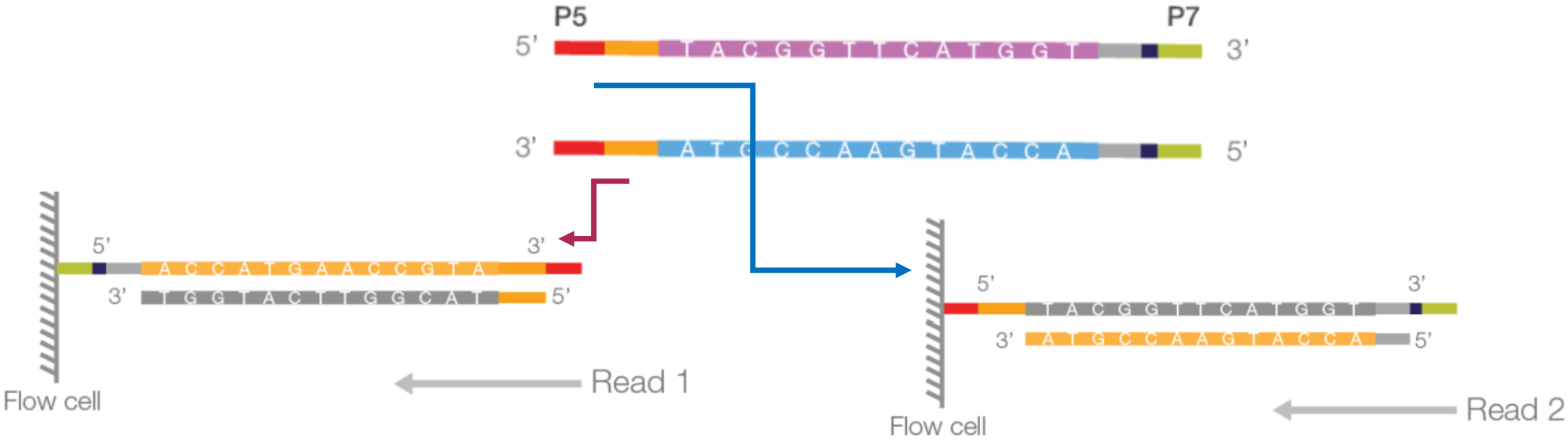
Enrich DNA Fragments (PCR)



Permiten darle dirección a lo transcritos generados.

Manual TruSeq Stranded mRNA

Paired-end y Strand-specific



Read 1 , se genera a partir de la hebra antisentido (antisense strand), para obtener la sentido.

Read 2 , se genera a partir de la hebra sentido (sense strand), para obtener la antisentido.

Manual TruSeq Stranded mRNA

Profundidad de secuenciación

Para lncRNAs 30 M en plantas

Sequencing applications	Recommended Coverage
Whole genome sequencing (WGS)	15X to 60X
Whole exome sequencing (WES)	100X
RNA sequencing (RNA-seq)	5 to 100 M reads per sample depending on target study
ChIP-Seq	100X
Whole genome sequencing (WGS) for *de novo* assembly (PacBio HiFi reads)	10X-15X per haplotype
Whole genome sequencing (WGS) for variant detection (PacBio HiFi reads)	≥ 15X (for human genome)

https://www.reneshbedre.com/blog/sequencing-coverage.html

Número de replicas

Depende de la **variabilidad técnica** y la **variabilidad biológica** del **objeto de estudio**, así como del **poder estadístico** deseado.

Variabilidad en mediciones

Extracción o preparación de bibliotecas

Variabilidad biológica

Inferencias poblacionales (mínimo 3)

Poder estadístico

Depende del método estadístico elegido

Table 1 Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

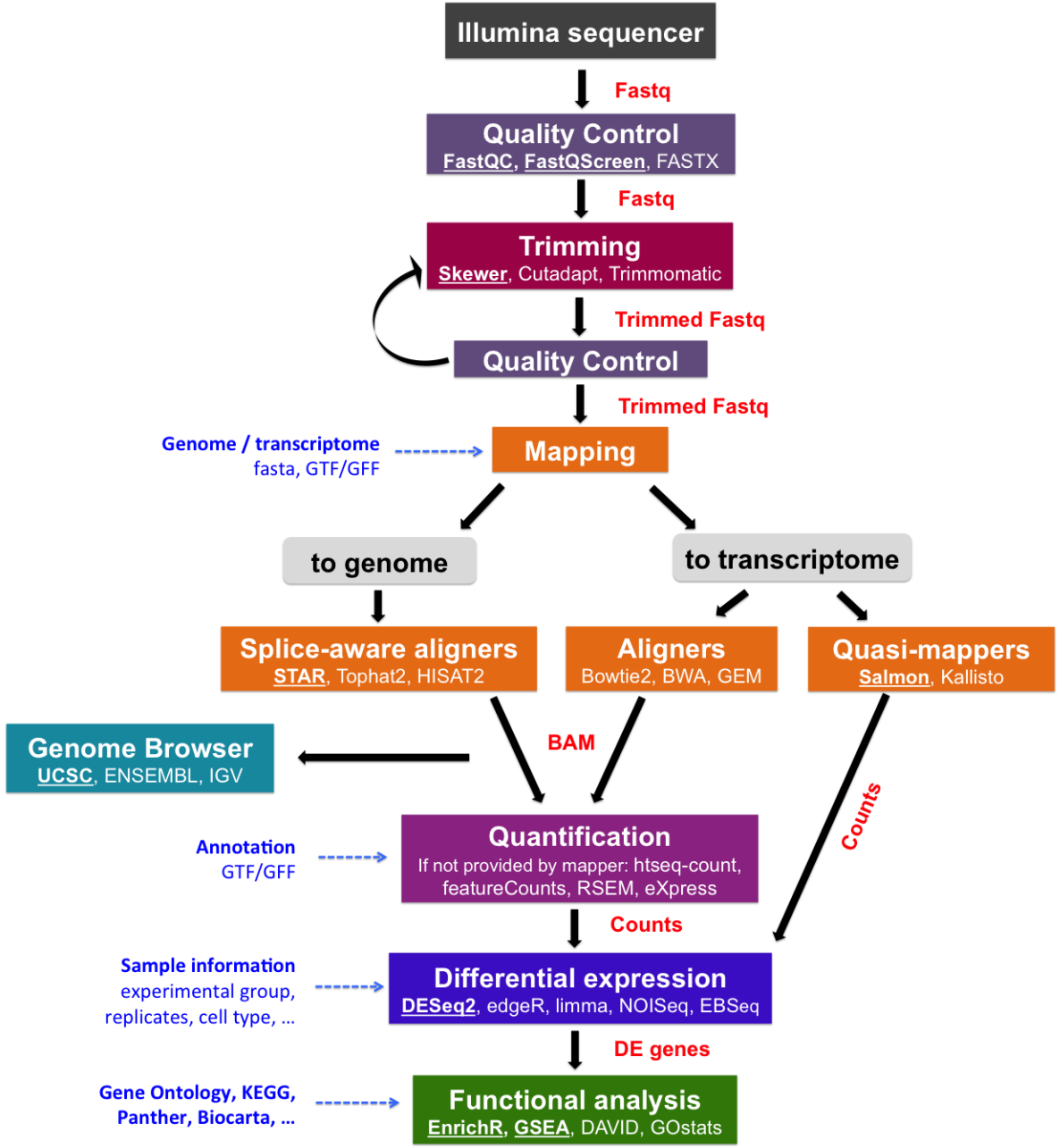
Example of calculations for the probability of detecting differential expression in a single test at a significance level of 5 %, for a two-group comparison using a Negative Binomial model, as computed by the RNASeqPower package of

Conesa, A., *et al*. 2017. RNA-seq methods. Journal of Cellular Biochemistry, 8(1), 1–24. https://doi.org/10.1002/wrna.1364.RNA-Seq

Diseño de Secuenciación

- Propósito: Evitar introducir sesgos técnicos que afecten el procesamiento de los datos.
- Se propone la aleatorización de muestras:
  - Durante la preparación de las bibliotecas.
  - Rondas de secuenciación (**batch**).
- Lo ideal es incluir todas las muestras en una misma línea para minimizar el *lane effect.*
- Tener cuidado de no mezclar el mismo **adaptador** en la línea de secuenciación (~24 adaptadores).

Pipeline bioinformática



mRNA-Seq data analysis workflow "https://biocorecrg.github.io/RNAseq_course_2019/workflow.html"

Antes que nada! Checa si los datos se descargaron bien… (Después de secuenciar)

KO84_At_shoot
- KO84_At_shootR1_L7_1.fq.gz
- KO84_At_shootR1 _L7_2.fq.gz
- **MD5.txt**

KO85_At_shoot
- KO85_At_shootR2 _L7_1.fq.gz
- KO85_At_shootR2 _L7_2.fq.gz
- **MD5.txt**

Checar los numeros md5 contenido en los archivos:

```
$ md5sum KO*/*gz
```

Numero de referencia:

```
$ cat KO*/MD5.txt
```

Antes que nada! Checa si los datos se descargaron bien… (Después de secuenciar)

KO84
- KO84_At_shootR1_L7_1.fq.gz
- KO84_At_shootR1 _L7_2.fq.gz
- **MD5.txt**

Todo debe coincidir, sino es así debes descargarlos de nuevo

Checar los numeros md5 contenido en los archivos:

$ md5sum KO*/*gz

```
57ee9c814da4494ee597bcaa3518b5e2  KO84/KO84_At_shootR1_L7_1.fq.gz
fc4794b93b566506788b3e52b299c2b4  KO84/KO84_At_shootR1 _L7_2.fq.gz
```

Numero de referencia:

$ cat KO*/MD5.txt

```
57ee9c814da4494ee597bcaa3518b5e2 KO84/KO84_At_shootR1_L7_1.fq.gz
fc4794b93b566506788b3e52b299c2b4 KO84/KO84_At_shootR1 _L7_2.fq.gz
```

Cuando los subas a NCBI checa de nueva cuenta el MD5

BioProject: PRJNA765039 Arabidopsis thaliana Col-0, wild type, shoot and roots from seedlings at 8 day-old

BioSample: SAMN21582179 A. thaliana shoot, biological replicate 1

SRR16093081 **Transcriptome of Arabidopsis thaliana Col-0, under normal growing condition, shoot from seedlings (biological replicate 1) (8 day-old)**

✔ **Released**

2022-05-19

2021-09-27 19:33

2022-05-19 17:03

Experiment	Library ID	Library strategy	Library source	Library selection	Library layout	Platform	Instrument
SRX12379381	A.thaliana_shoot_Biological_replica e_1	RNA-Seq	TRANSCRIPTOMIC	Oligo-dT	PAIRED	ILLUMINA	Illumina HiSeq X

Transcriptome of aerial part (shoot) of Arabidopsis t aliana Col-0 at 8 days after germination.

File name	File type	MD5SUM
At_shootR1_1.fq.gz	fastq	57ee9c814da4494ee597bcaa3518b5e2
At_shootR1_2.fq.gz	fastq	fc4794b93b566506788b3e52b299c2b4

Total number of spots	Total number of bases	GC percentage
36075409	10822622700	46.82

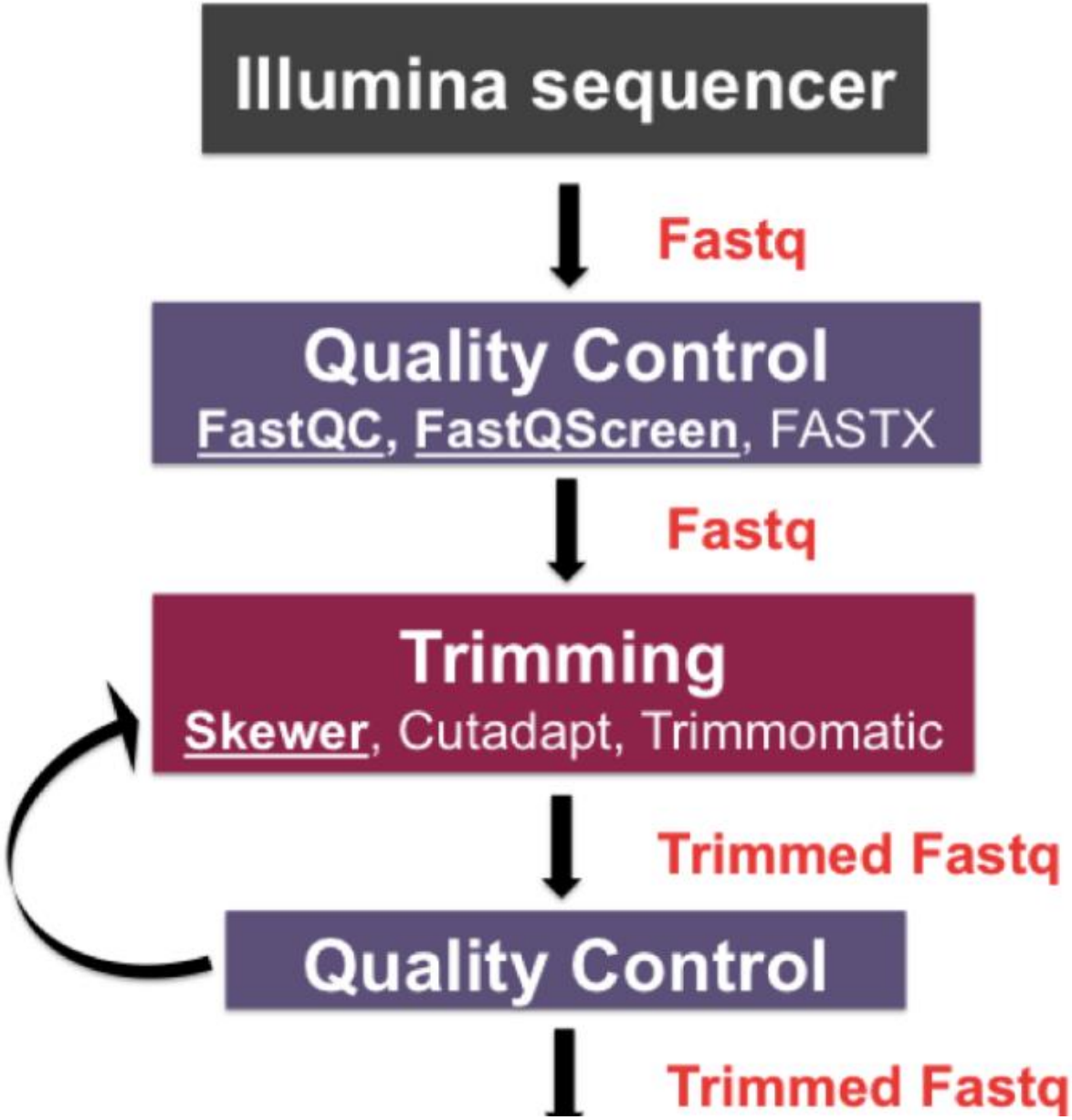
57ee9c814da4494ee597bcaa3518b5e2

fc4794b93b566506788b3e52b299c2b4

Quality Check

FastQC → MultiQC

- Uno de los pasos mas importantes. Dedícale tiempo.
- La calidad de tus datos importa, bibliotecas mal secuenciadas genera datos desconfiables.
- Debes analizar su calidad para poder reclamar en la secuenciación (~1 semana).



Archivos fasq (fasq.gz / fq.gz)

- Derivan del formato FASTA.
- Muestra la calidad de cada nucleótido.
- Cada secuenta esta representada por 4 líneas:
  - @ ID del read + información de la corrida
  - Secuencia
  - Símbolo "+“
  - Valor de Calidad codificado (Escala **Phred** y código **ASCII**)

Ejemplo:


1. ID
2. Seq
4. Q
@SRR12038075.1 D00635:426:CD19YANXX:5:2301:1410:2130/1
NCACTAGCCAGCTGCTTCAGGAAAACCACCCTCTTGCCCCTGTGGCGTCCA
+
#<<@BGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG
@SRR12038075.2 D00635:426:CD19YANXX:5:2301:1262:2153/1
CCTGCCCAAAGTAGCTGAGTTCGCTGCCGTCCAGGACGGCACTGGCCGTGT
+
33B>BGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGEGGGGGG
@SRR12038075.3 D00635:426:CD19YANXX:5:2301:1338:2197/1
GGATCAGCCAAGAAGGCCTTGACCTTTTCAGCAAGTGGGAAGGTGTAATCC
+
3A>BBGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGEFGGGGGGGGG

Lo ideal

FastQC Report

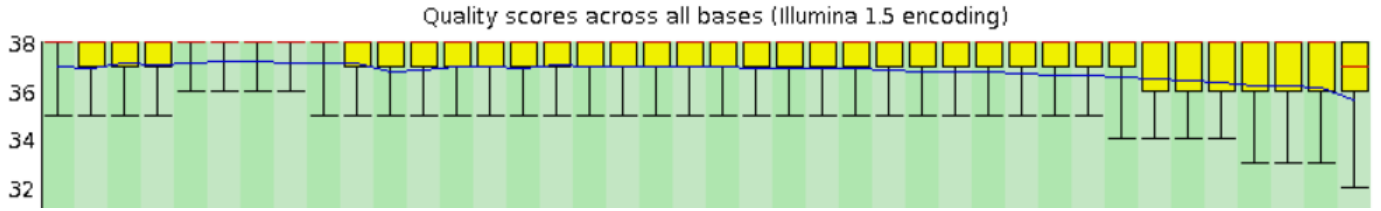
Summary

- 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

Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Per base sequence quality



Buena calidad pero con adaptadores

Buena calidad

Dimeros de adaptadores

FastQC Report

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

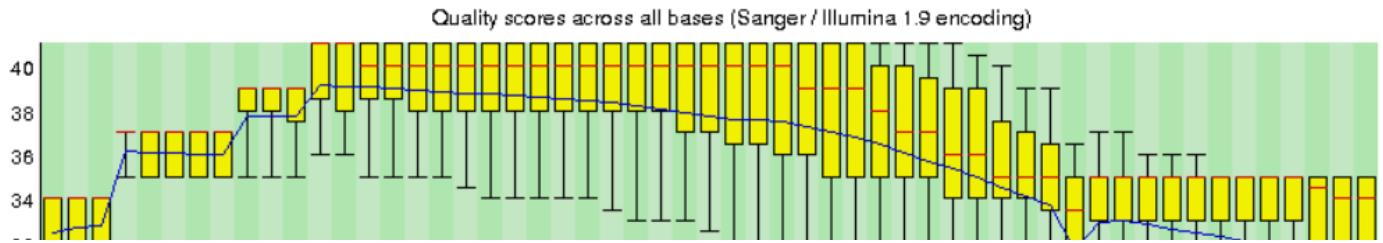
Basic Statistics

Measure	Value
Filename	SRR7947071_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	60349690
Sequences flagged as poor quality	0
Sequence length	100
%GC	47

Datos crudos (raw data)

Paired-end read1

Per base sequence quality



Muy mala calidad …

FastQC Report

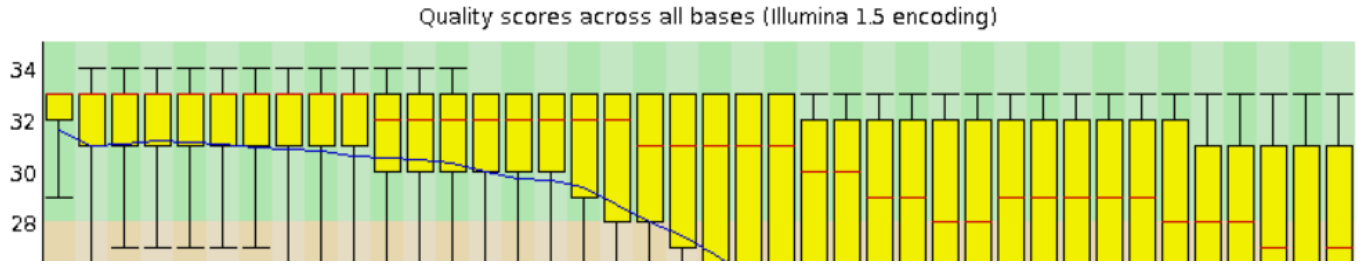
Summary

- 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

Basic Statistics

Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

Per base sequence quality



Datos crudos
(raw data)

Per base sequence quality

Distribución de la calidad de los datos en cada posición (bp).

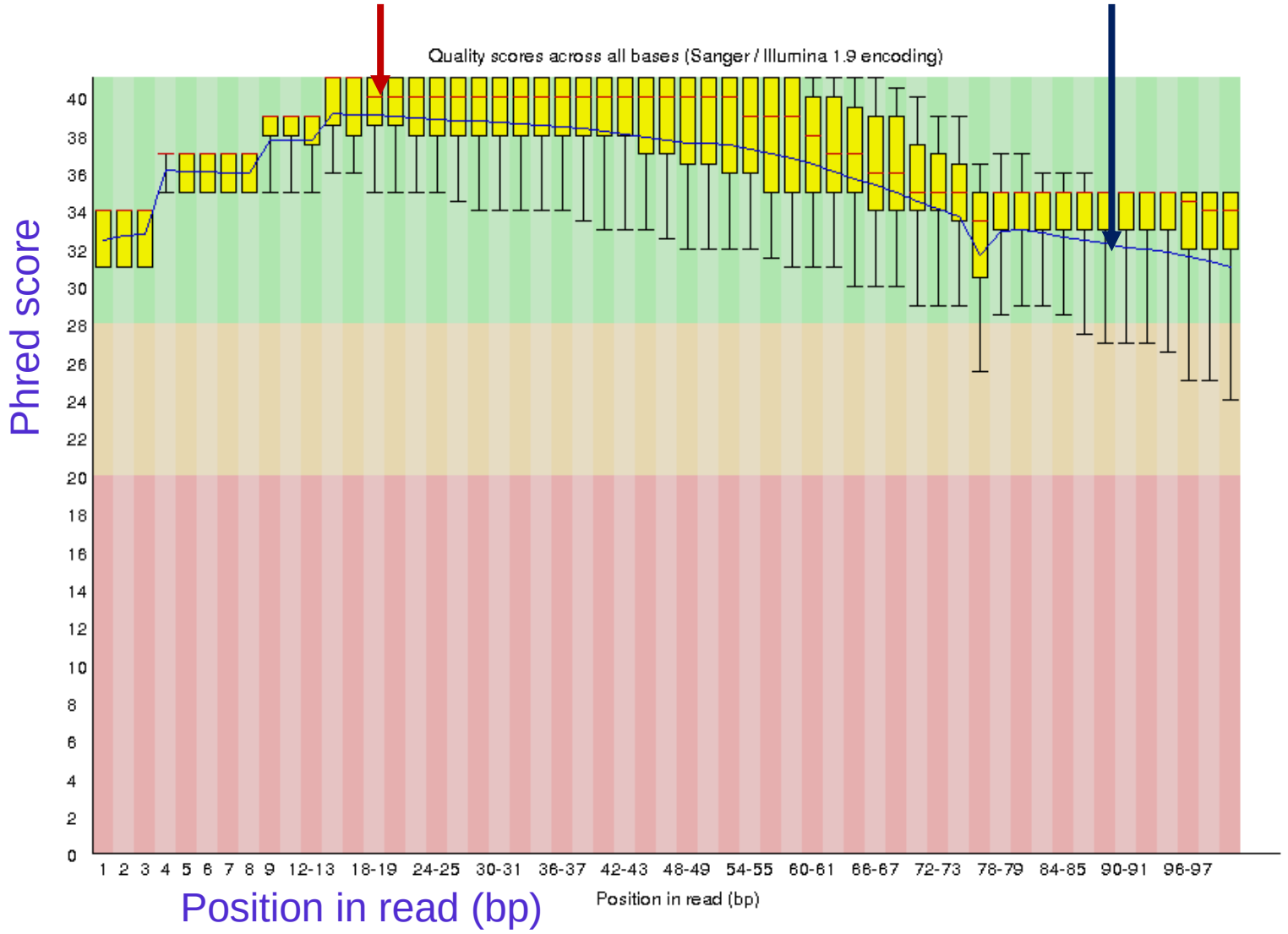
Profundidad de secuenciación ~60 M read

Basic Statistics

Measure	Value
Filename	SRR7947071_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	60349690
Sequences flagged as poor quality	0
Sequence length	100
%GC	47

Mediana de la calidad

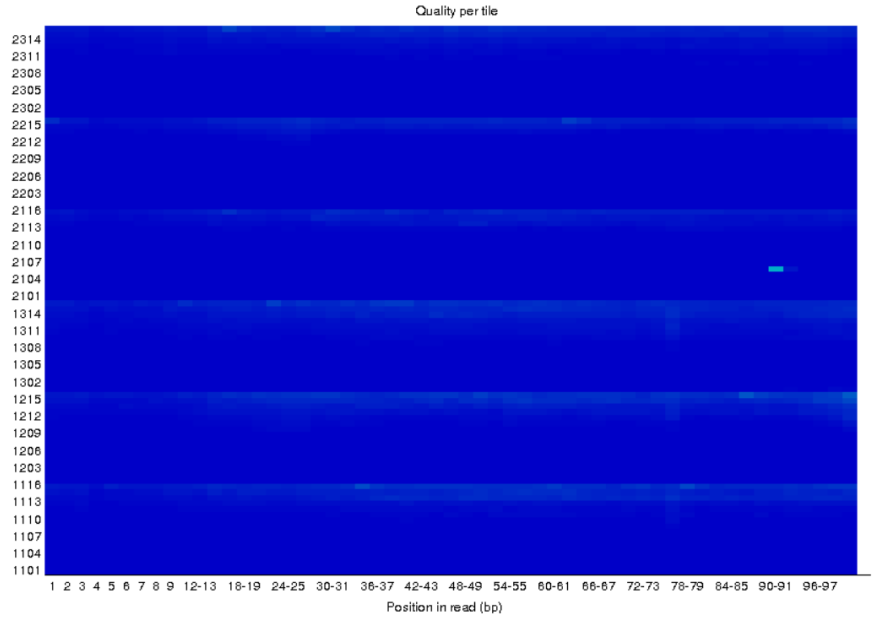
Promedio de la calidad



Per tile sequence quality

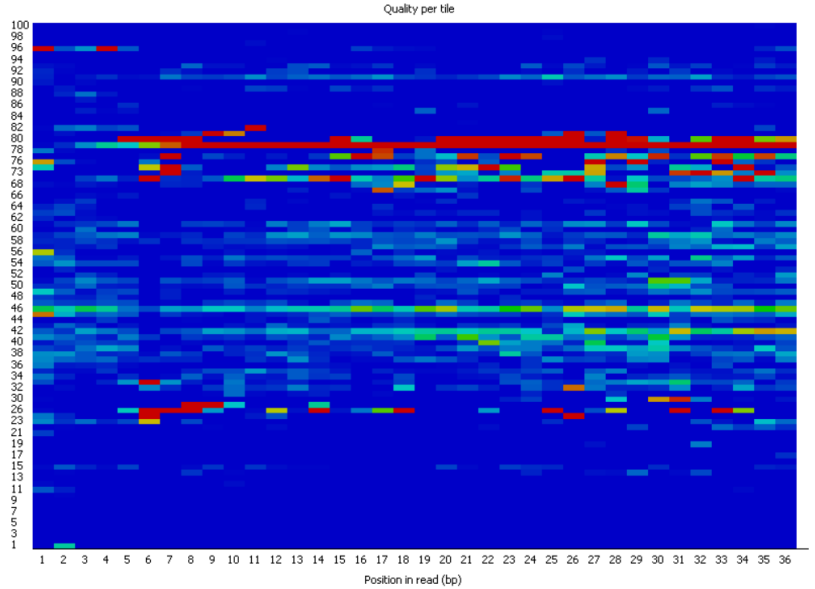
- **Perdida de la calidad** de las secuencias que se encuentran asociadas a una sola parte o a varias partes de la secuencia.
- **Desviación del promedio de la calidad.**
- Escala de colores de **azul** a **rojo**.
- Lo idóneo es encontrar el análisis en **azul**.
- Problemas con la secuenciación.

Buena calidad



Position in read (bp)

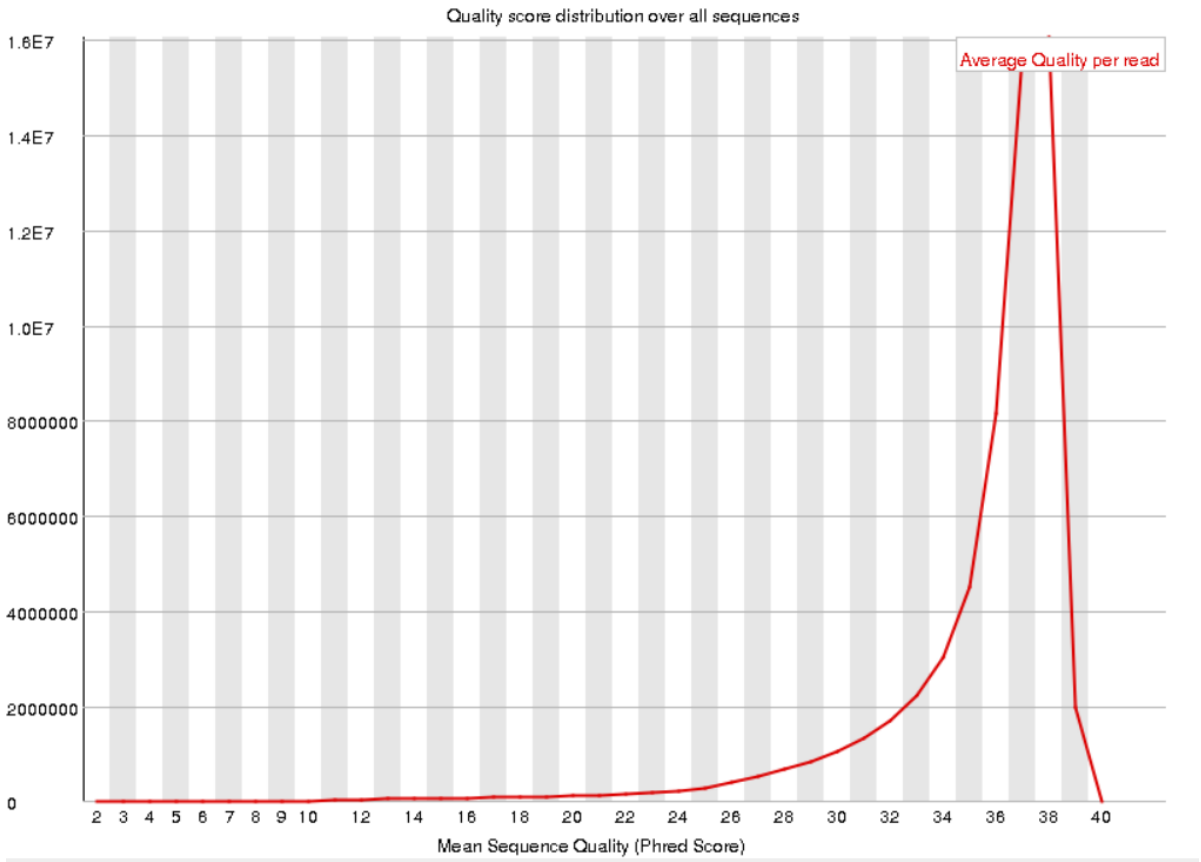
Mala calidad



Position in read (bp)

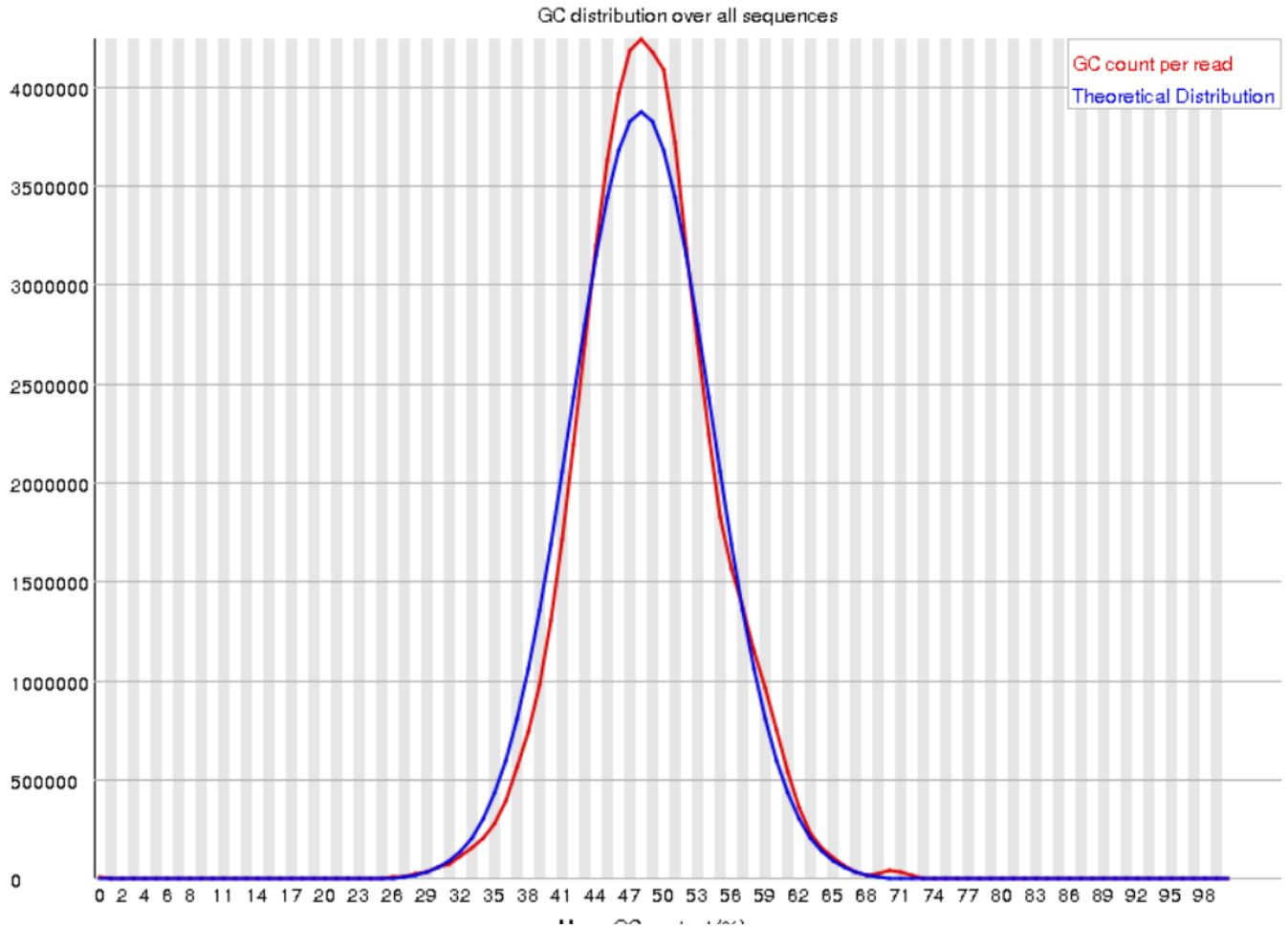
Per sequence quality scores

- Subgrupo de secuencias con baja calidad universal.
- Normalmente los datos de baja calidad se encuentran relacionados con un abaja representación de secuencias, por lo que, suele darse solo en un pequeño porcentaje de las secuencias totales.
- **Picos altos = la mayoría de los datos buena calidad.**

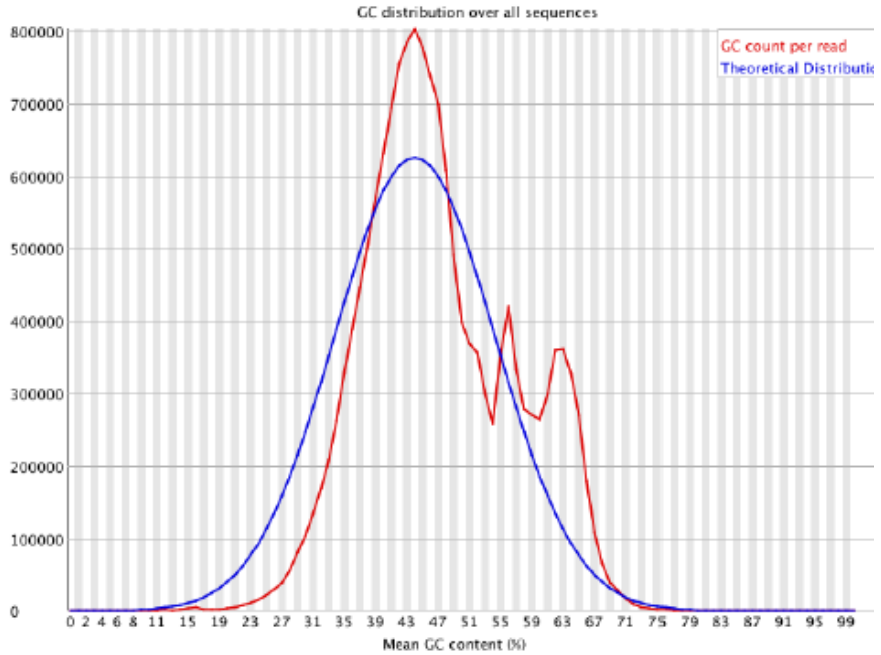


Per sequence GC content

- Porcentaje de G/C.
- **~50% en una secuencia.**
- Distribución normal.
- En caso de encontrar mas picos se relaciona con **contaminación** o dímeros de adaptadores.



Variaciones en el contenido de GC se relacionan con contaminaciones

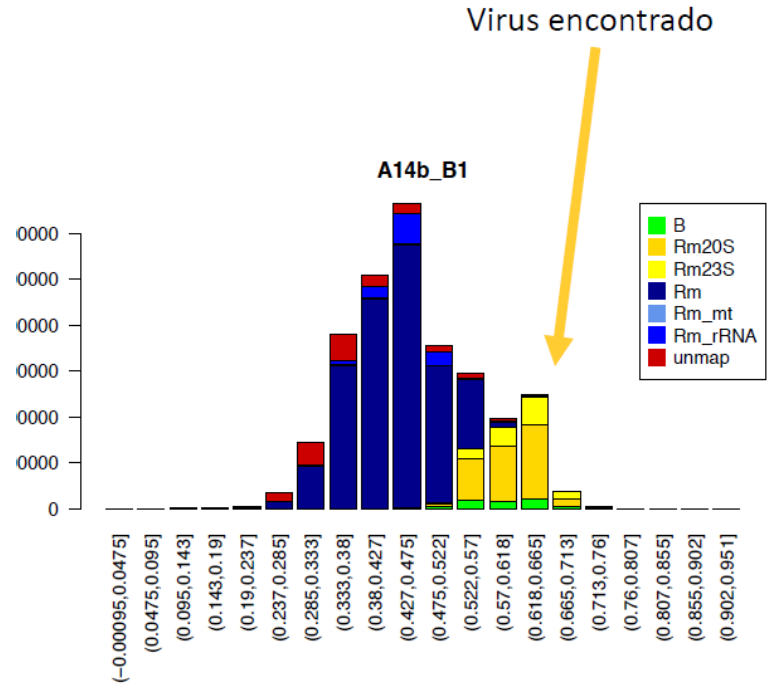


Nucleic Acids Research

Oxford University Pre

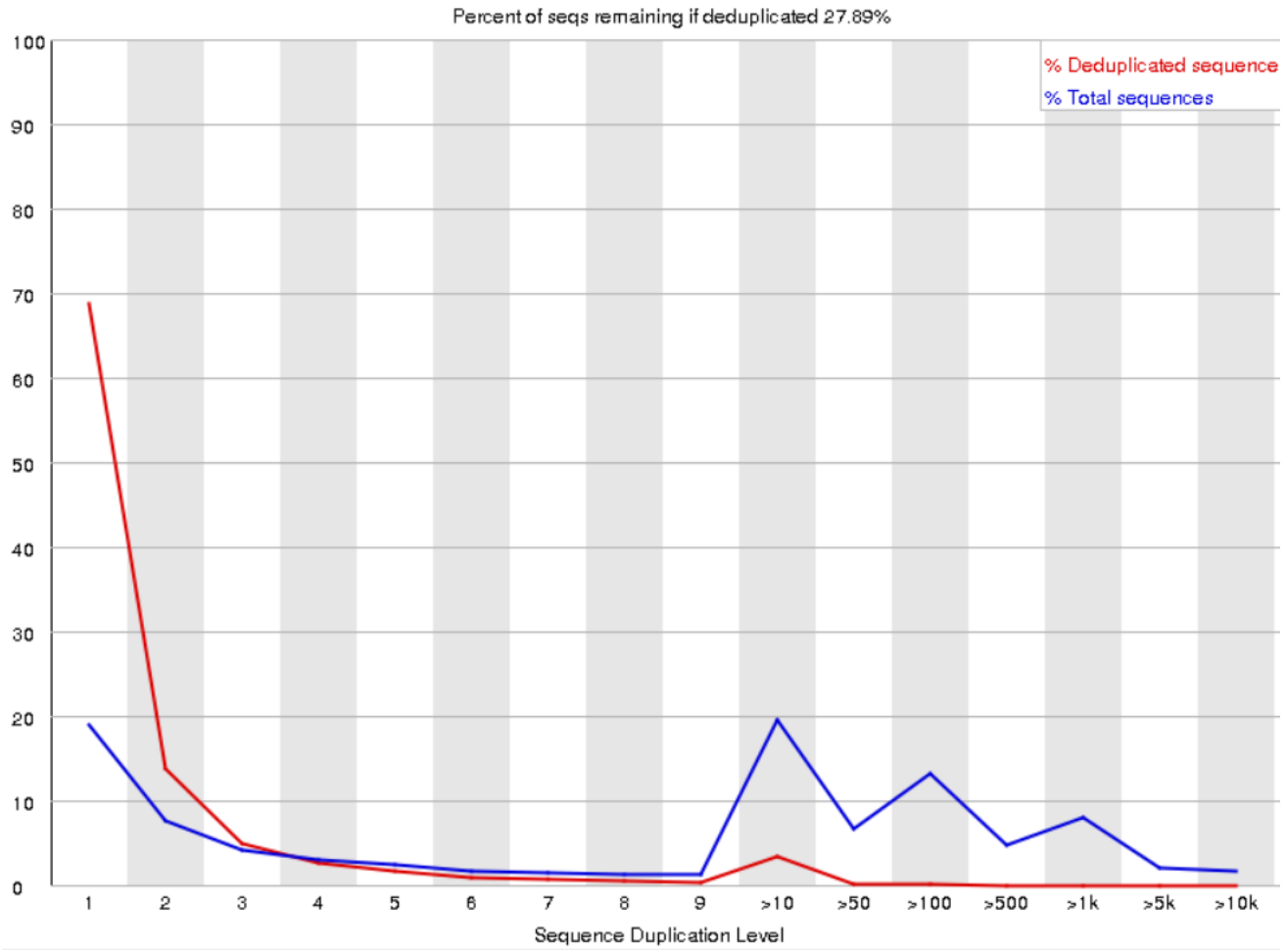
Disentangling sRNA-Seq data to study RNA communication between species

José Roberto Bermúdez-Barrientos, Obed Ramírez-Sánchez, [...], and Cei Abreu-Goodger



Sequence duplication levels

- Secuencias que se repiten varias veces en el análisis.
- Dímeros de adaptadores.
- rRNA



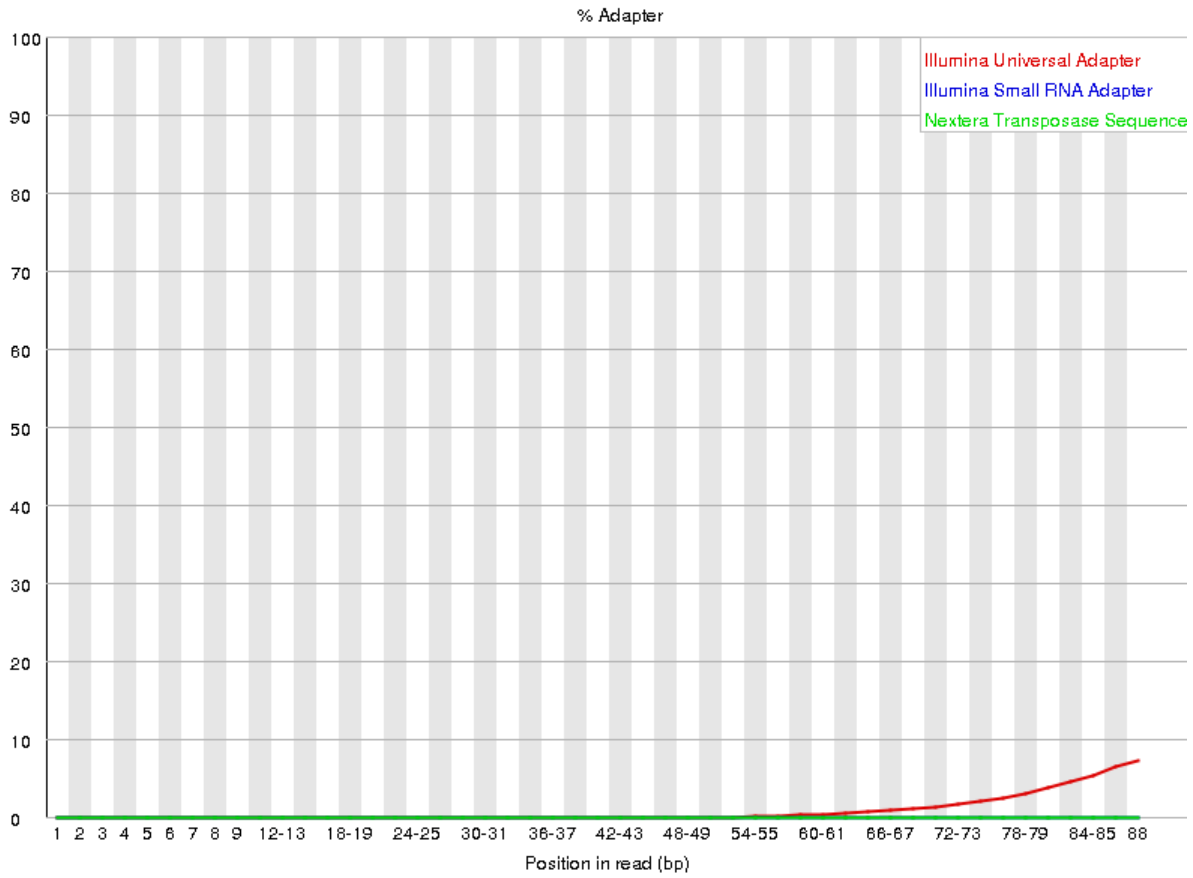
Overrepresented sequences

- Secuencias dentro del 0.1% del total de las secuencias.
- Secuencias representadas en una alta proporción o repetidas.

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACTAGCTTATCTCGTATGC	189760	0.31443409237064845	TruSeq Adapter, Index 10 (100% over 50bp)

Adapter Content



No emplees datos donde alguno de los reads esta mal *(paired-end)*

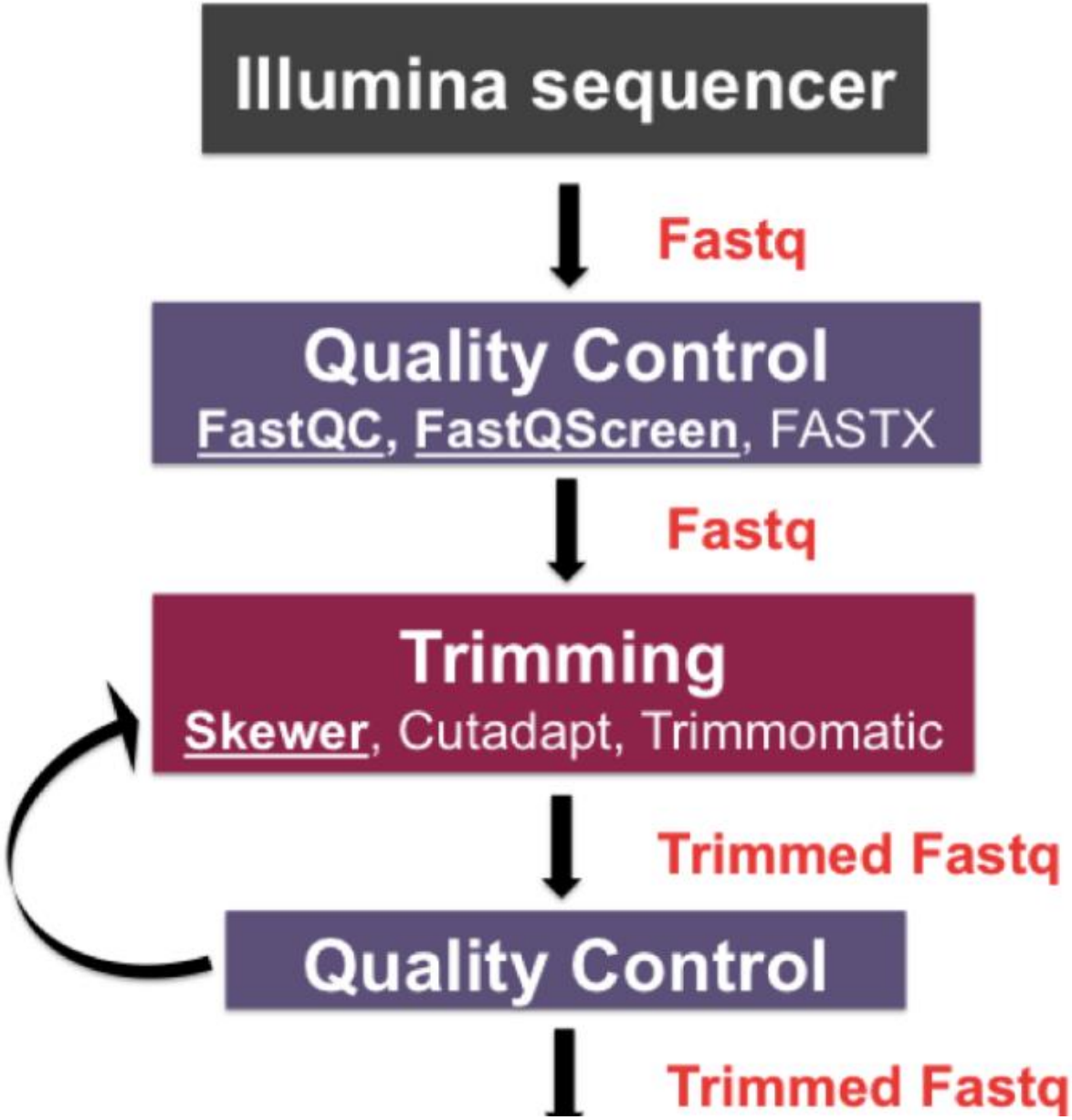
General Statistics

Copy table | Configure Columns | Plot | Showing 8/8 rows and 4/5 columns.

Sample Name	% Dups	% GC	Length	M Seqs
SRR12414267_1	87.4%	51%	68 bp	13.7
SRR12414267_2	99.5%	48%	8 bp	13.7
SRR12414268_1	70.9%	50%	68 bp	15.2
SRR12414268_2	99.5%	48%	8 bp	15.2
SRR12414269_1	74.2%	49%	68 bp	26.7
SRR12414269_2	99.7%	48%	8 bp	26.7
SRR12414270_1	87.1%	53%	68 bp	19.6
SRR12414270_2	99.6%	47%	8 bp	19.6

Trimming

- Quitar lecturas de mala calidad.
- Quitar bases con baja calidad.
- Cortar secuencias de adaptadores.



Empecemos con sus proyectos

Proyectos

Proyecto	GEO	Titulo del registro en GEO	Referencia
PRJNA826506	GSE200762	Regulation of human trophoblast gene expression by endogenous retroviruses	https://www.biorxiv.org/content/10.1101/2022.04.26.489485v2
PRJNA842067	GSE204739	Translation and Natural Selection of long non-canonical RNA micropeptides	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9622821/
PRJNA821620	GSE199834	SHORT-ROOT Stabilizes PHOSPHATE1 to Regulate Phosphate Allocation in Arabidopsis	https://pubmed.ncbi.nlm.nih.gov/36050464/
PRJNA858106	GSE208076	Genome wide expression from COVID-19 patients and controls' lung	https://pubmed.ncbi.nlm.nih.gov/36768969/

Ejercicio 1: Análisis con FastQC

/mnt/Timina/bioinfoII/rnaseq/BioProject_2023/rawData

- 1.- Arabidopsis_thaliana (PRJNA821620)
- 2.-COVID_virus (PRJNA858106)
- 3.-Drosophila_melanogaster (PRJNA842067)
- 4.-Homo_sapiens (PRJNA826506)

```
[ecoss@chromatin rawData]$ ls
Arabidopsis_thaliana  COVID_virus  Drosophila_melanogaster  Homo_sapiens  SRAData_dow.sh  SRA_run.sge
```

Solo descargue **2 SRA de cada Proyecto**, así que es solo la calidad de 2 archivos, debes analizar la calidad para los demás.

Ejercicio 2: Descarga los demás fastq.gz contenidos en el BioProject

Necesitaras dos scripts:

SRAData_dow.sh	Descarga de SRA
SRA_run.sge	Mandar como job al cluster

Ya que esto demora tiempo, de **tarea vuelvan a correr el análisis de calidad.**