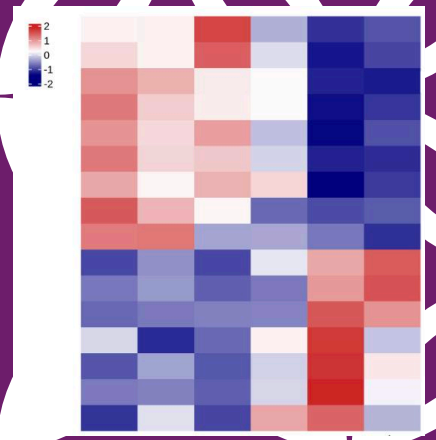
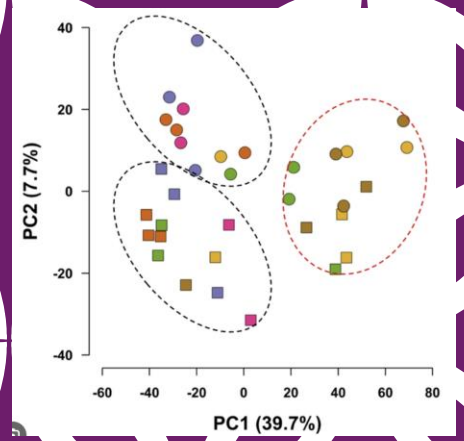
Instituto Nacional de Salud Pública

CISEI

Departamento de Bioinformática en Enfermades Infecciosas

Red Mexicana de Bioinformática





Cuernavaca

CONTROL DE CALIDAD EN TRANSCRIPTÓMICA: HEATMAP Y PCA

Dra. Ernestina Godoy Lozano
Instituto Nacional de Salud Pública
M. en C. Fernandina Nieves López
Universidad Autónoma de Nayarit

CONTENIDO

1. Introducción al control de calidad en transcriptómica
2. Tipos de gráficos usados en análisis del control de calidad
3. Heatmap
4. PCA

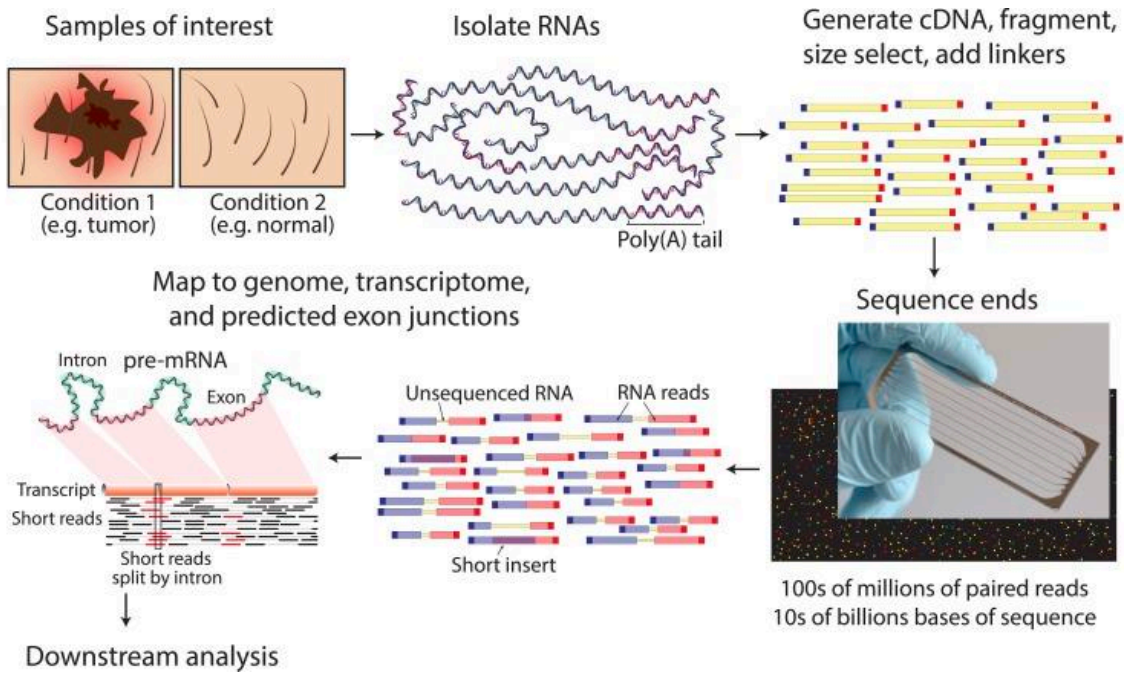
RNA-SEQ

La secuenciación de RNA (RNA-seq) se ha adoptado para la elaboración de perfiles de transcriptomas en muchas áreas de la biología, incluidos los estudios sobre la regulación génica, desarrollo y enfermedadades.

De particular interés es el descubrimiento de genes expresados diferencialmente en diferentes condiciones.

Los experimentos de transcriptómica se han utilizado ampliamente para medir los niveles de RNA expresados en tejidos o células de prácticamente cualquier organismo.

La transcriptómica se ha mejorado con la ayuda de tecnologías de secuenciación masiva y han reemplazando a los microarreglos mediante el uso de experimentos de RNA-Seq para evaluar la expresión génica a gran escala.



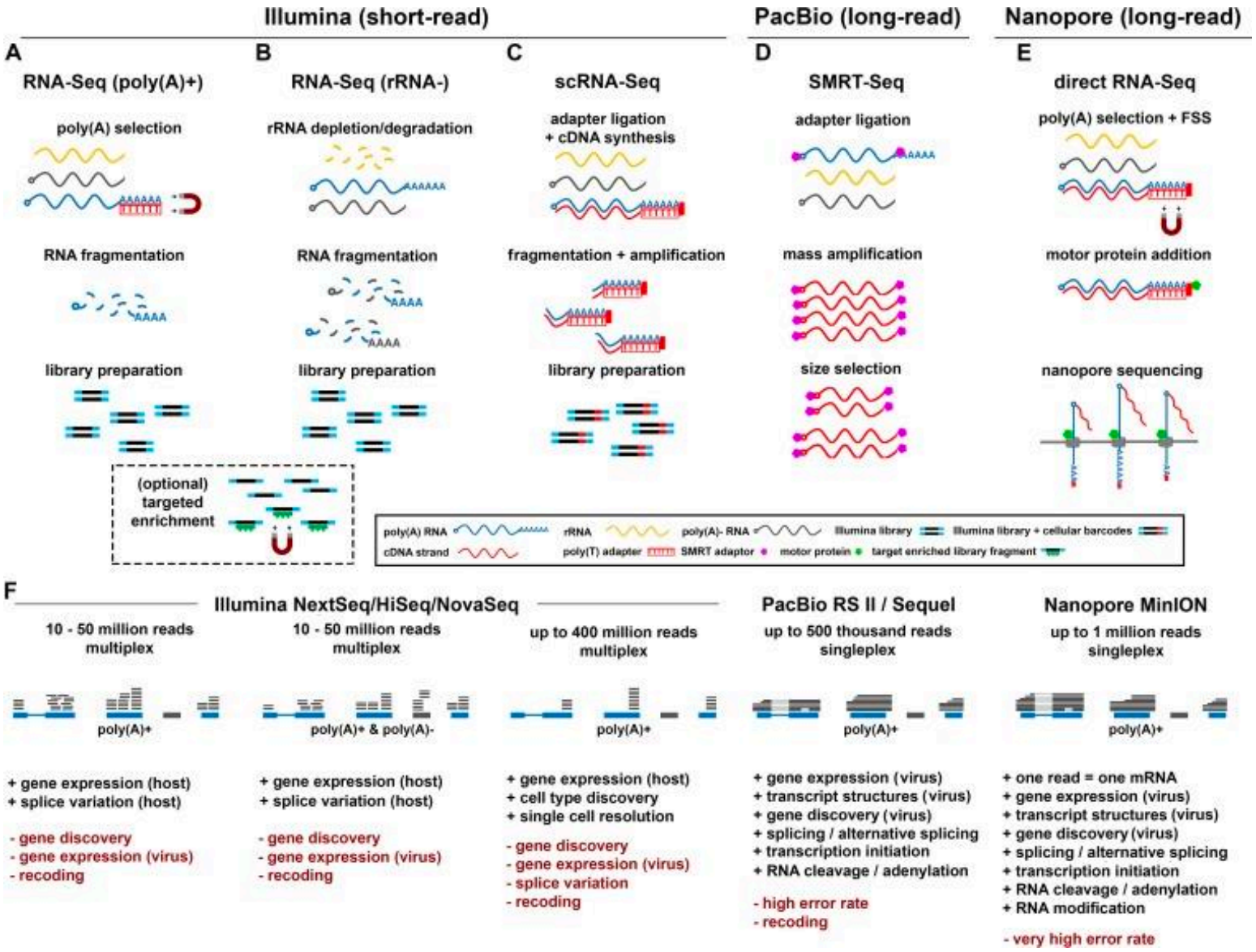
Griffith M, Walker JR, Spies NC, Ainscough BJ, Griffith OL. Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. *PLoS Comput Biol.* 2015;11(8):e1004393. Published 2015 Aug 6. doi:10.1371/journal.pcbi.1004393

RNA-SEQ

Dependera de nuestra pregunta de investigación que tipo de RNA-Seq elegiremos así como la plataforma de secuenciación masiva.

La elección de la plataforma de secuenciación viene dictada por la profundidad de secuenciación, ya que difieren notablemente en el número de secuencias y la longitud de las lecturas generadas.

Cada una tiene sus ventajas y desventajas.



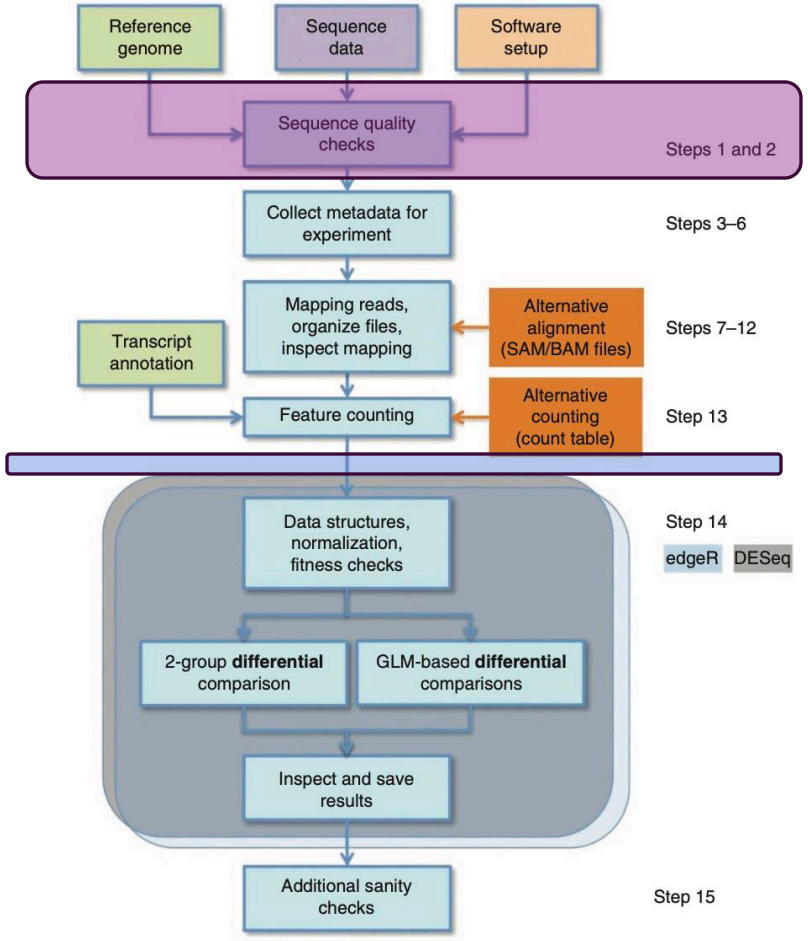
Depledge DP, Mohr I, Wilson AC. Going the Distance: Optimizing RNA-Seq Strategies for Transcriptomic Analysis of Complex Viral Genomes. *J Virol.* 2018;93(1):e01342-18. Published 2018 Dec 10. doi:10.1128/JVI.01342-18

EXPRESIÓN DIFERENCIAL

Para cuantificar los niveles de transcripción e identificar genes expresados diferencialmente en diferentes condiciones, utilizando datos de RNA-Seq de tecnologías de secuenciación de alto rendimiento

Podemos describir un flujo de trabajo general:

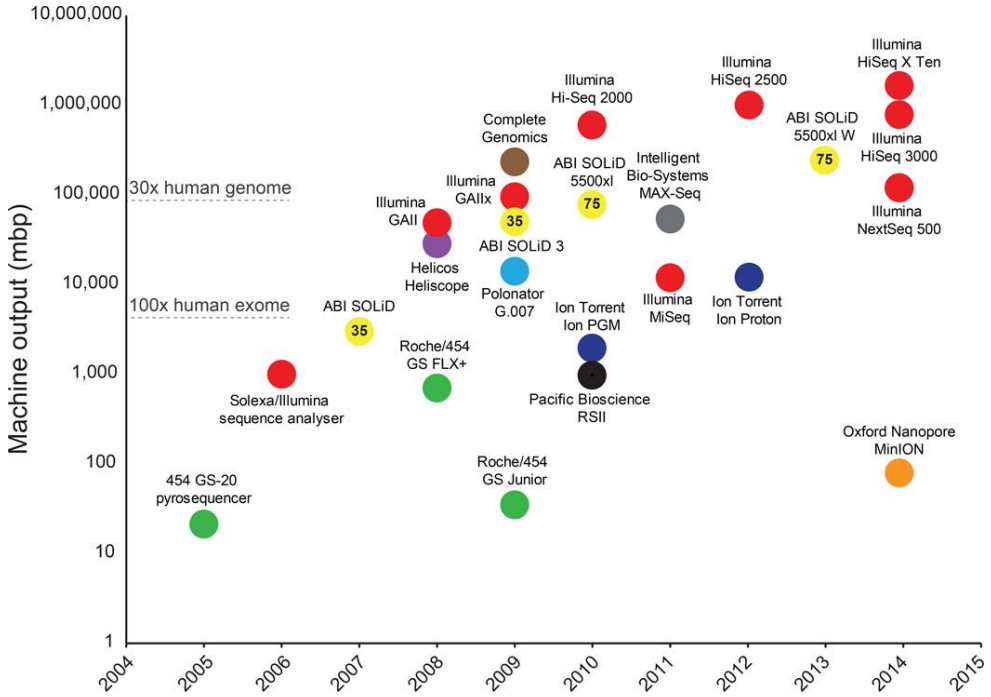
1. Control de calidad de las lecturas de RNA-Seq
2. Cortes en la lecturas (read trimming) y filtrado
3. Mapeo de lecturas trimeadas/filtradas a una referencia (genoma o transcriptoma) o ensamble de NOVO
   - Predicción de genes
4. Obtener el recuento de lecturas para cada gen.
5. Evaluación de la calidad de las replicas biológicas
6. Análisis de expresión diferencial



TECNOLOGÍAS DE SECUENCIACIÓN MASIVA

La secuenciación de alto rendimiento, también conocida como secuenciación de próxima generación (NGS), es el término integral utilizado para describir las tecnologías que secuencian el ADN y el ARN de manera rápida y rentable.

Actualmente los conocimientos sobre la respuesta inmune humana a la vacunación, los cánceres y las infecciones virales provienen de tecnologías "ómicas" de alto rendimiento que miden el comportamiento de genes, ARNm (transcriptómica de una sola célula), proteínas (proteómica), metabolitos (metabolómica), células (citometría de masas) y modificaciones epigenéticas (ATAC-seq), junto con enfoques computacionales.



Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015 May 21;58(4):586-97. doi: 10.1016/j.molcel.2015.05.004.

CARACTERÍSTICAS TECNOLOGÍAS DE SECUENCIACIÓN MASIVA

Plataforma de secuenciación	Longitud de secuencia (pb)	Precisión	Salida	Química de secuenciación	Tiempo de corrida	Ventajas
Sanger	400-900	99.999%	1.9-84 Kb	Terminación de la cadena didesoxi	20 min – 3 h	Lectura larga y de alta calidad
Illumina MiSeq	75-300	99.9%	12.2-20Gb	Secuenciación por Síntesis	21-56 h	Alto rendimiento, calidad de lectura
MinION	>200,000	~95%	~50 Gb	Secuenciación única lecturas largas en tiempo real	1-48 h	Alto rendimiento, gran longitud de lectura, portabilidad
PacBIO	10-15 Kb	99.999	5-10 Gb	Secuenciación única lecturas largas en tiempo real	4 h	Lectura larga y calidad

Santos A et al. Comput Struct Biotechnol J. 2020 Jan 31;18:296-305. doi: 10.1016/j.csbj.2020.01.005. PMID: 32071706; PMCID: PMC7013242

TECNOLOGÍAS DE SECUENCIACIÓN MASIVA

Es importante tomar en cuenta que tecnología de secuenciación se esta utilizando para poder determinar como analizar los datos ya que cada tecnología tiene características particulares:

- Profundidad
- Longitud de la lecturas
- Errores de secuenciación

TABLE 1 | Next-generation sequencing platforms for the analysis of display libraries.

Platform	Read length	Max. depth	Error type (percentage)
Illumina Miseq	300 bp PE	40×10^6 reads	Substitutions (~0.1)
Illumina Hiseq 2500	250 bp PE	600×10^6 reads	Substitutions (~0.1)
Ion Torrent PMG	400 bp	5.5×10^6 reads	Indels (~1)
454 GS FLX	Up to 1 kb	1×10^6 reads	Indels (~1)
PacBio	250 bp–40 kb	0.4×10^6 reads	Indels (~1)

Front. Immunol., 02 February 2018 | https://doi.org/10.3389/fimmu.2018.00118

ANÁLISIS DE CALIDAD

- El análisis de calidad de los datos de secuenciación nos sirve para asegurarnos que los datos con los que vamos a trabajar sean de buena calidad y son suficientes para realizar los análisis subsecuentes.

- Problemas que pueden surgir:
  - Bases de calidad baja
  - Contaminación
  - Secuencias con adaptadores

- Métricas a evaluar en el análisis de calidad:
  - Calidad promedio
    - Porcentajes de Q>20 y Q>30
  - Longitud de las secuencias
  - Número de secuencias y bases
  - Porcentaje de lecturas con Ns

Si se tuvo problemas para la obtención de la muestra es probable que la secuenciación no sea de buena calidad

FUENTES DE ERROR EN DATOS DE RNA-SEQ

Rin Integrity

> Es un valor numérico que indica la integridad del ARN
> Se expresa en una escala de 1 a 10, donde 10 indica una integridad excelente del ARN y 1 indica una degradación grave del ARN

- **RT-qPCR:** RIN entre 5 a 6
- **qPCR:** RIN >7
- **Secuenciación de ARN:** RIN entre 8 y 10
- **Microarray:** RIN entre 7 y 10

Contaminación y sesgos de preparación de librerías

Una clase de errores se refiere a sesgos en la abundancia de secuencias de lectura debido a las preferencias de preparación del ARN, la selección del tamaño de los fragmentos y el contenido de GC.

Artefactos de secuenciación

Los errores de secuenciación, que son el resultado de errores del llamado de bases (mismatch) o la inserción o eliminación de una base (indeles)

doi:10.1186/1471-2199-7-3; https://doi.org/10.1093/nar/gkq224; doi: 10.1093/nar/gkt215

CALIDAD DE SECUENCIACIÓN

Phred Quality Score

- La precisión del llamado de bases es la métrica más común utilizada para determinar la precisión de una plataforma de secuenciación.
- Indica la posibilidad de que el secuenciador coloque incorrectamente a una determinada base.
- El Phred Quality Score se utiliza para indicar la medida de la calidad de la base en la secuenciación del ADN.
- La alta consistencia de una base secuenciada se indica mediante mayores valores de Phred.

$$Q = -10 \log 10 P$$

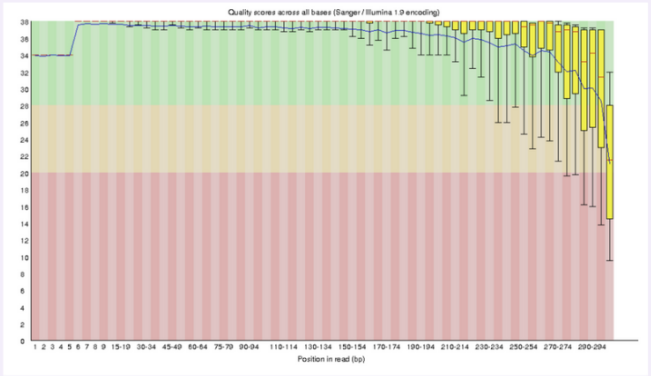
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

RNA-seq for revealing the function of the transcriptome. Kavita Goswami, Neeti Sanan-Mishra, in Bioinformatics, 2022

HERRAMIENTAS PARA EL CONTROL DE CALIDAD DE LA SECUENCIACIÓN

Herramientas que exploran los datos de calidad

- FastQC
- MultiQC



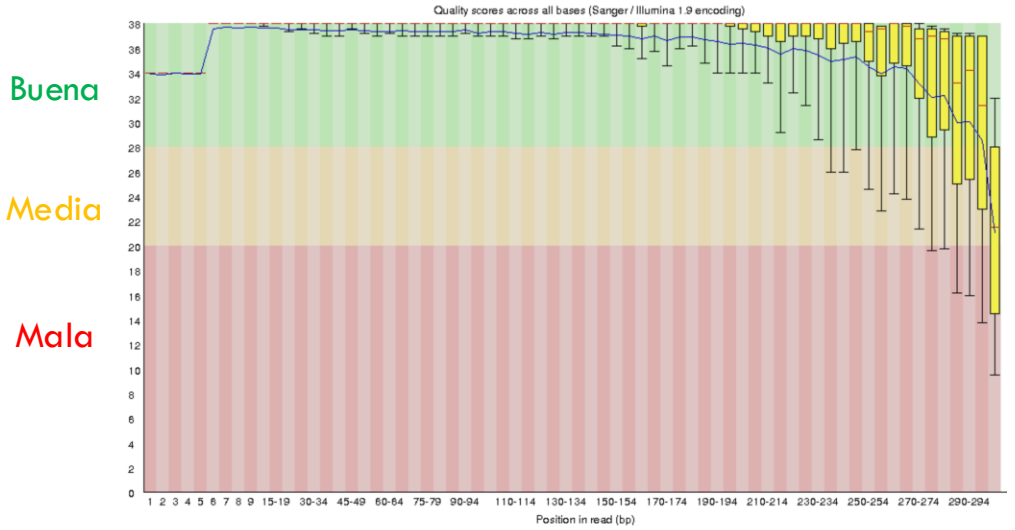
Herramientas que realizan el control de calidad

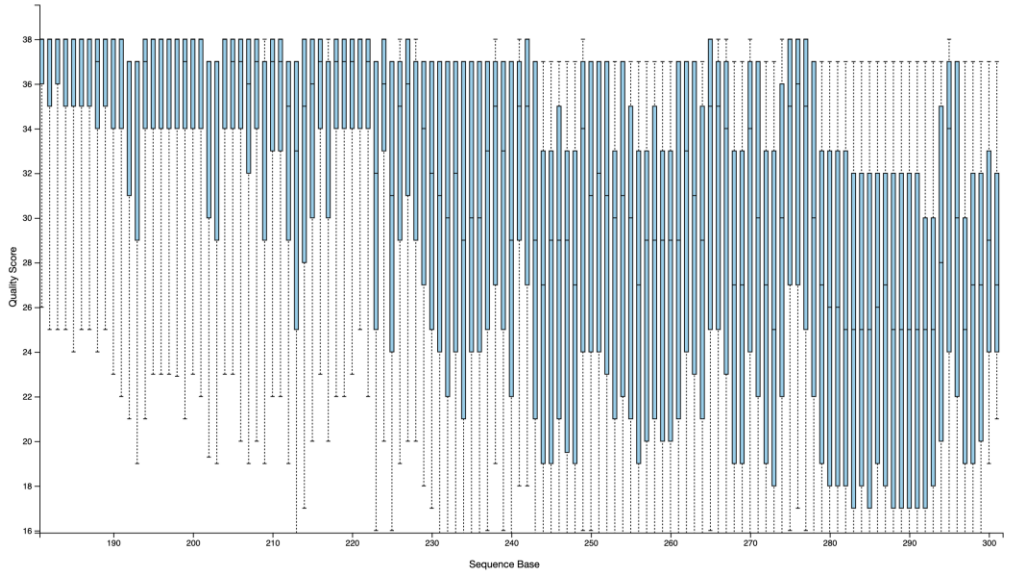
- Trimommatic
- Cutadapt

Herramientas que generan una modificación en los datos crudos de secuenciación para mejorar la calidad de secuenciación

TIPOS DE GRAFICOS USADOS EN EL CONTROL DE CALIDAD

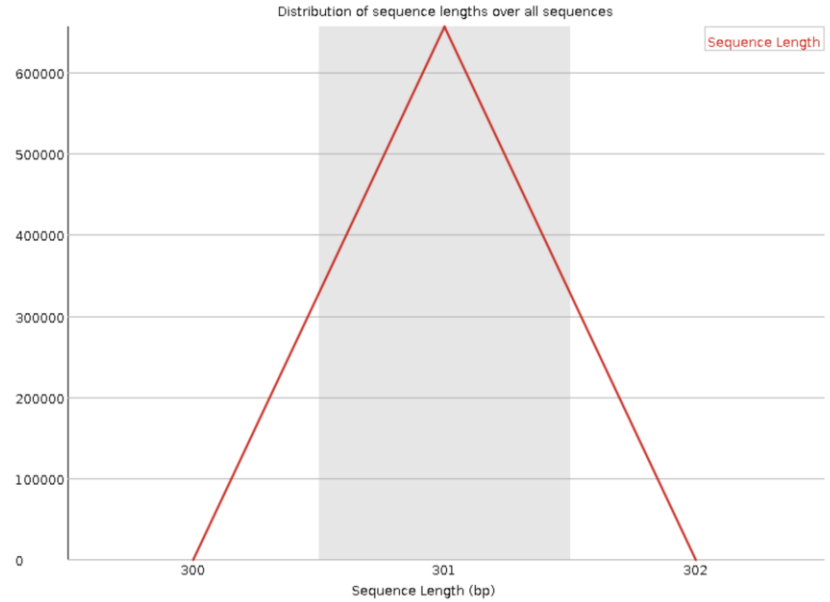
- Gráficos de calidad de bases
  - Boxplot de calidad a lo largo de las lecturas
- Distribución de longitud de lecturas
- Contenido de GC
- Gráficos de duplicación de secuencias

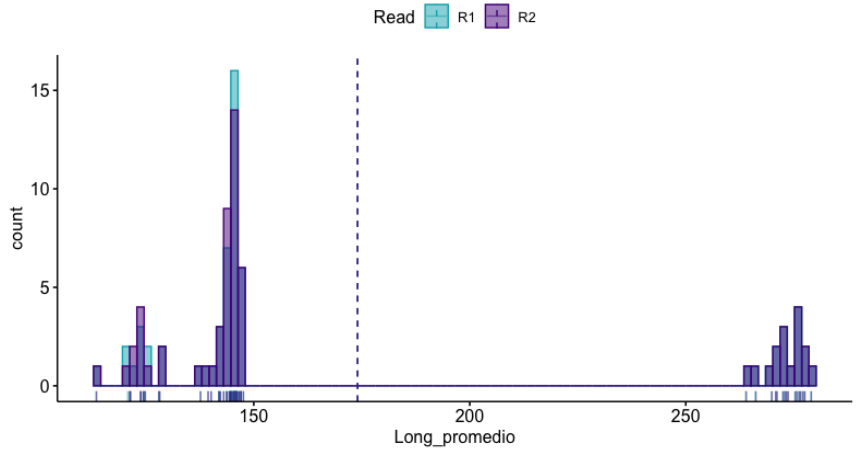




TIPOS DE GRAFICOS USADOS EN EL CONTROL DE CALIDAD

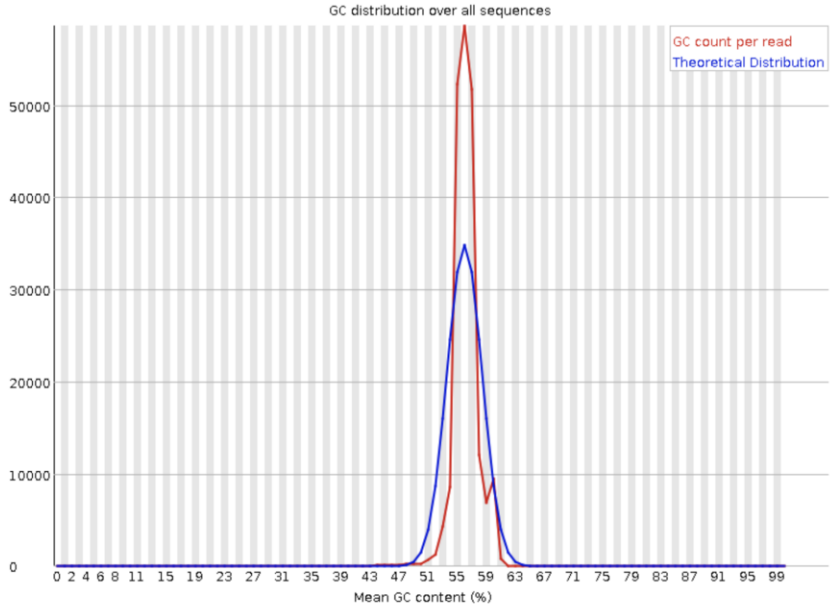
- Gráficos de calidad de bases
  - Boxplot de calidad a lo largo de las lecturas
- **Distribución de longitud de lecturas**
- Contenido de GC
- Gráficos de duplicación de secuencias

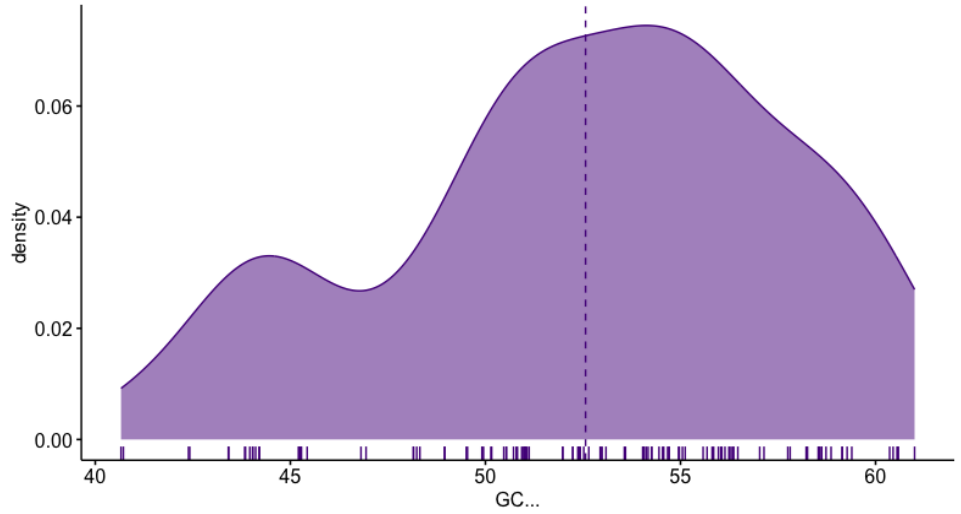




TIPOS DE GRAFICOS USADOS EN EL CONTROL DE CALIDAD

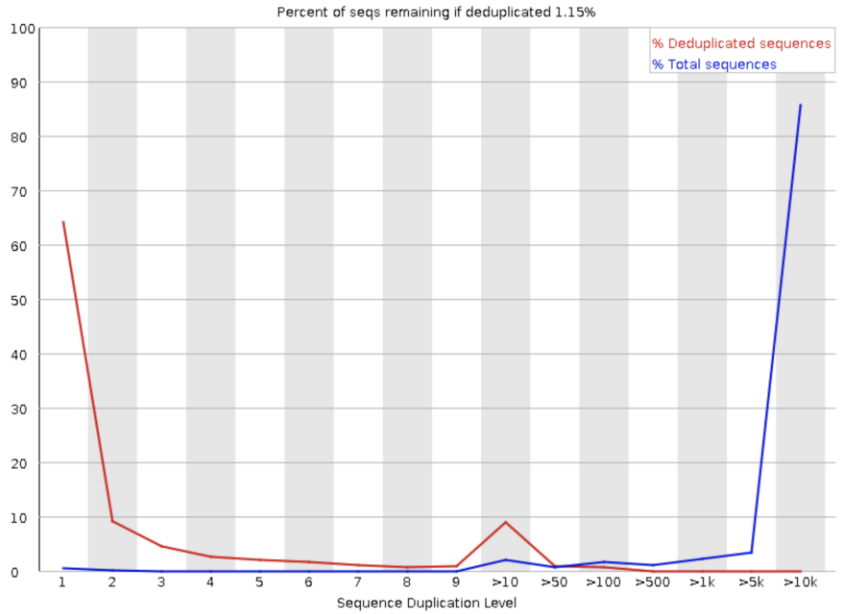
- Gráficos de calidad de bases
  - Boxplot de calidad a lo largo de las lecturas
- Distribución de longitud de lecturas
- Contenido de GC
- Gráficos de duplicación de secuencias





TIPOS DE GRAFICOS USADOS EN EL CONTROL DE CALIDAD

- Gráficos de calidad de bases
  - Boxplot de calidad a lo largo de las lecturas
- Distribución de longitud de lecturas
- Contenido de GC
- Gráficos de duplicación de secuencias

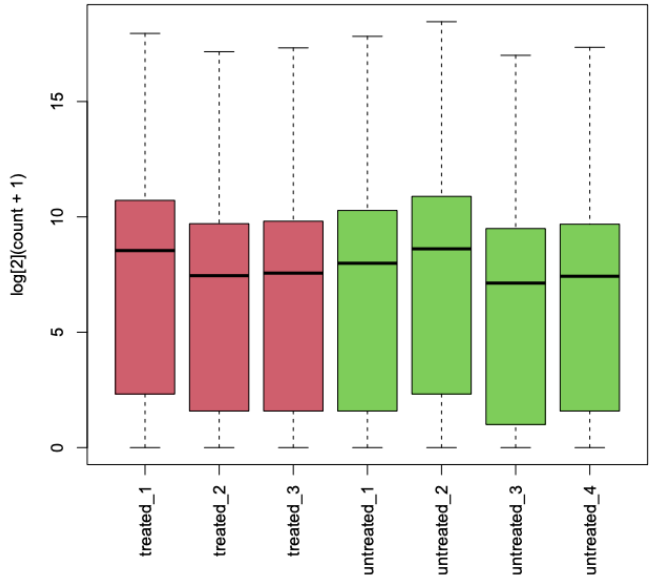


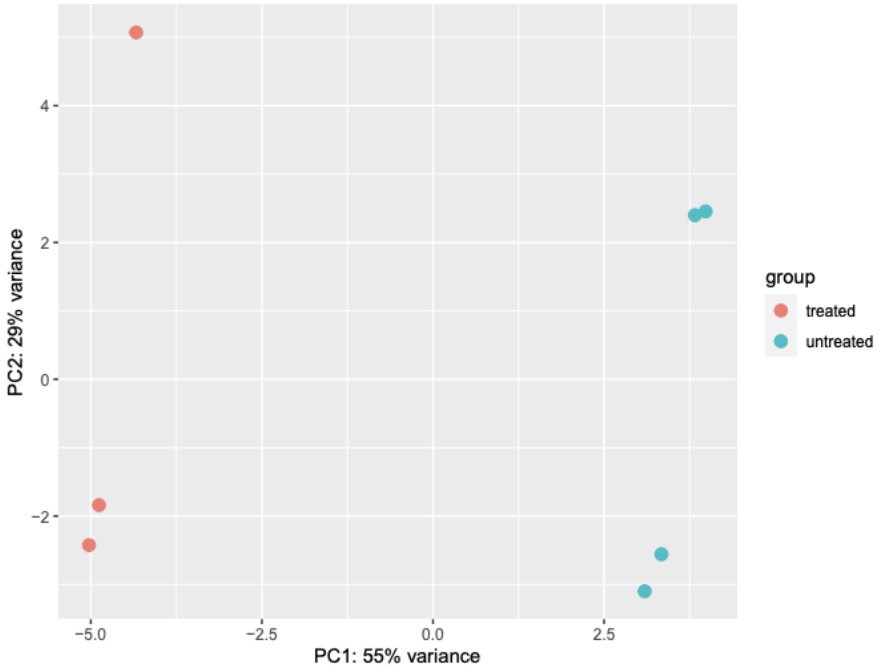
Overrepresented sequences

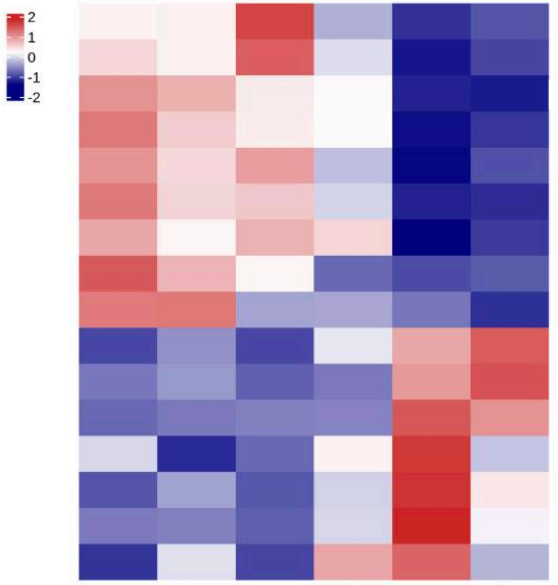
Sequence	Count	Percentage	Possible Source
GGGACCCAAAAACCACACCCCTCCTTGGGAGAATCCCCTAGATCACAGCT	117344	17.901176183429694	No Hit
GGGAACCACATCCCTCCTCAGAAGCCCCCAGAGCACAACTCCTTACCATG	110605	16.873121691507375	No Hit
GGGATCATCCAACAACCACATCCCTTCTCTACAGAAGCCTCTGAGAGGAA	53268	8.126191820109534	No Hit
GGGAGCTCTGGGAGAGGAGCCCCAGCCCTGAGATTCCCACGTGTTTCCAT	42904	6.545132797363884	No Hit
GGGCTTTCTGAGAGTCCTGGACCTCCTGTGCAAGAACATGAAACATCTGT	42726	6.517978367988284	No Hit
GGGATCACATAACAACCACATTCCTCCTCTAAAGAAGCCCCTGGGAGCAC	41340	6.306539945996247	No Hit

TIPOS DE GRAFICOS USADOS EN EL CONTROL DE CALIDAD

- Gráficos de abundancia y distribución de reads
  - Violin/box plots, PCA, heatmaps para exploración de datos







IMPORTANCIA DE LAS REPLICAS EN TRANSCRIPTÓMICA

- Las réplicas en los experimentos de RNA-seq y expresión diferencial son fundamentales para garantizar la precisión, reproducibilidad y robustez de los resultados.

- Su importancia radica en varios aspectos clave:

1. Reducción de la Variabilidad Técnica y Biológica

Variabilidad biológica: Las diferencias entre organismos o condiciones pueden ser significativas.

Variabilidad técnica: Factores como la eficiencia de la extracción de ARN, la preparación de librerías o el secuenciamiento pueden introducir ruido en los datos.

2. Mejora de la Confianza Estadística
3. Mayor Poder Estadístico
4. Mejor Control de Errores Tipo I y II
5. Reproducibilidad y Generalización

TABLE 2. A summary of the recommendations of this paper

	Agreement with other tools[a]	WT vs. WT FPR[b]	Fold-change threshold (T)[c]	Tool recommended for: (# good replicates per condition)[d] ≤3	≤12	>12
DESeq	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
DESeq2	Consistent	Pass	0	-	-	Yes
			0.5	Yes	Yes	Yes
			2.0	Yes	Yes	Yes
EBSeq	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
edgeR (exact)	Consistent	Pass	0	-	-	Yes
			0.5	Yes	Yes	Yes
			2.0	Yes	Yes	Yes
Limma	Consistent	Pass	0	-	-	Yes
			0.5	-	Yes	Yes
			2.0	Yes	Yes	Yes
cuffdiff	Consistent	Fail				
BaySeq	Inconsistent	Pass				
edgeR (GLM)	Inconsistent	Pass				
DEGSeq	Inconsistent	Fail				
NOISeq	Inconsistent	Fail				
PoissonSeq	Inconsistent	Fail				
SAMSeq	Inconsistent	Fail				

[a]Full clean replicate data set, see section "Tool Consistency with High Replicate Data" and Figure 3.
[b]See section "Testing Tool False Positive Rates" and Figure 4.
[c]See section "Differential Expression Tool Performance as a Function of Replicate Number."
[d]See Figure 2.

más réplicas = resultados más confiables, robustos y reproducibles

DOI:10.1261/rna.053959.115

BATCH EFFECT

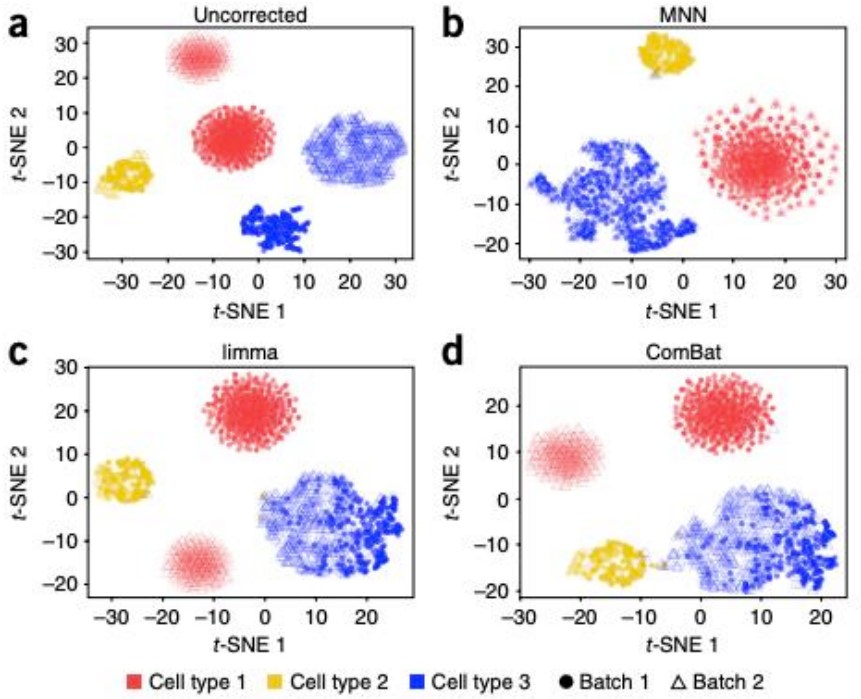
La corrección del efecto de batch o correción por lotes es el procedimiento de eliminar la variabilidad de los datos que no se debe a las variables de interés (por ejemplo, tipo de cáncer).

Los efectos de lote se deben a diferencias técnicas entre las muestras, como el tipo de equipo de secuenciación o incluso el técnico que analizó la muestra.

Ocurre cuando las muestras se procesan y miden en diferentes lotes y que no están relacionadas con ninguna variación biológica registrada durante el experimento.

Eliminar esta variabilidad significa cambiar los datos de las muestras individuales.

Para los datos de expresión, esto significa que nunca debe tomar muestras individuales de un conjunto corregido por lotes para un análisis separado, como la expresión diferencial.



doi:10.1038/nbt.4091

CONTACTO

Dra. Elizabeth Ernestina Godoy Lozano

elizabeth.godoy@insp.mx

@Tina_Godoy @tinagodoy.bsky.social @tgodoy

M. en C. . Fernandina Nieves López

zulia.nieves@uan.edu.mx

@ZuliaFer