Revista de I+D TECNOLÓGICO
Portal de revistas: http://revistas.utp.ac.pa

Análisis de microARN utilizando Galaxy: superando barreras bioinformáticas en la formación estudiantil

Analysis of microRNA using Galaxy: Overcoming bioinformatics barriers in student training

Abigail De Ávila[1,2], Grimaldo Ureña[1,3], Miryam Venegas-Anaya[1,4]*

[1]*Doctorado en Biociencias y Biotecnología, Facultad de Ciencias y Tecnología, Universidad Tecnológica de Panamá.*
[2]*Maestría en Docencia Superior, Facultad de Ciencias Educativas, Universidad Tecnológica OTEIMA.*
[3]*Grupo de Investigación en Biotecnología, Bioinformática y Biología de Sistemas (GIBBS), Dirección de Investigación, Universidad Tecnológica de Panamá*
[4]*Centro de Investigaciones Hidráulicas e Hidrotécnicas (CIHH), Universidad Tecnológica de Panamá.*

***Autor de correspondencia:** abigail.deavila@utp.ac.pa

RESUMEN. El análisis de datos ómicos es esencial en la biología moderna, pero la complejidad técnica de las herramientas bioinformáticas sigue siendo una barrera para quienes no poseen formación en programación. Este trabajo tuvo como objetivo demostrar que es posible realizar un análisis completo de microARNs utilizando únicamente Galaxy, como estrategia pedagógica para acercar a estudiantes de ciencias biológicas al análisis bioinformático. Se analizaron seis muestras de tejido cerebral humano, tres fetales y tres adultas, obtenidas de un repositorio público. El flujo de trabajo incluyó control de calidad con FastQC, eliminación de adaptadores con Cutadapt, alineamiento al genoma humano con HISAT2 y cuantificación de lecturas con featureCounts. El análisis de expresión diferencial se realizó con DESeq2. Se obtuvieron altos porcentajes de mapeo (87–93 %) y asignación confiable de lecturas a miRNAs conocidos. El análisis de componentes principales mostró una separación clara entre fetales y adultos, mientras que los mapas de calor confirmaron la consistencia de las réplicas y las diferencias entre regiones cerebrales. El histograma de valores *p* y las estimaciones de dispersión reflejaron patrones típicos de RNA-seq, y el MA-plot permitió identificar miRNAs diferencialmente expresados entre ambos grupos. El uso de Galaxy posibilitó completar el análisis sin necesidad de programación ni infraestructura avanzada, resaltando su valor como herramienta didáctica para la enseñanza de análisis de datos ómicos. En conclusión, este estudio evidencia que es posible implementar un flujo reproducible y accesible para la caracterización de perfiles de microARNs, ofreciendo un recurso pedagógico para la formación práctica en bioinformática.

Palabras clave. *Análisis de expresión diferencial, bioinformática, cerebro humano, Galaxy, microARN, sRNA-seq*

ABSTRACT. Omics data analysis has become a cornerstone of modern biology, yet the technical complexity of bioinformatics tools remains a significant barrier for students and researchers without programming expertise. This study aimed to demonstrate that a complete microRNA workflow can be carried out entirely within the Galaxy platform, as a pedagogical strategy to make bioinformatics more accessible in the life sciences. Six human brain tissue samples—three fetal and three adult—were obtained from a public repository and analyzed. The workflow included quality control with FastQC, adapter trimming with Cutadapt, alignment to the human genome using HISAT2, and read quantification with featureCounts. Differential expression analysis was conducted with DESeq2. The pipeline achieved high mapping rates (87–93%) and consistent assignment of reads to known miRNAs. Principal component analysis revealed clear separation between fetal and adult groups, while heatmaps confirmed the reproducibility of biological replicates and differences across brain regions. Additional outputs, including p-value distributions, dispersion estimates, and MA-plots, reflected typical RNA-seq patterns and highlighted sets of miRNAs with significant differential expression. By leveraging Galaxy, the entire analysis was completed without the need for programming skills or advanced computing infrastructure, underscoring its value as a teaching tool for omics data analysis. In conclusion, this study demonstrates that a reproducible and

Citación: A. De Ávila, G. Ureña y M. Venegas-Anaya, "Análisis de microARN utilizando Galaxy: superando barreras bioinformáticas en la formación estudiantil", *Revista de I+D Tecnológico*, vol. 22, no. 1, pp. (0), 2026.
Tipo de artículo: Original. **Recibido:** 8 de octubre de 2025. **Recibido con correcciones:** 29 de enero de 2026. **Aceptado:** 29 de enero de 2026.
DOI.

accessible workflow for microRNA profiling can be implemented in Galaxy, offering a practical educational resource for bioinformatics training.

Keywords. *Bioinformatics*, *Differential expression analysis, Galaxy, human brain, microRNA, sRNA-seq*

1. Introducción

El avance acelerado de las tecnologías de secuenciación de nueva generación ha generado una explosión en la producción de datos biológicos a gran escala, dando origen a lo que hoy se conoce como la era de los datos ómicos. En este nuevo panorama, disciplinas como la genómica, transcriptómica, epigenómica y metagenómica han incorporado herramientas computacionales como parte esencial de sus flujos de trabajo. El análisis de estos datos requiere habilidades específicas en bioinformática, un campo que combina conocimientos de biología molecular con competencias en programación, estadística y manejo de datos. Esta realidad ha creado una brecha significativa entre la generación de datos y su interpretación, especialmente en contextos académicos donde muchos estudiantes y docentes no cuentan con formación computacional formal. De hecho, se ha señalado que la mayoría de los estudiantes de ciencias biológicas completan sus estudios sin haber recibido formación alguna en bioinformática o biología computacional [1]. En consecuencia, existe un consenso creciente sobre la necesidad de integrar la bioinformática como un componente esencial en la educación en ciencias de la vida [2].

En este contexto, si bien la bioinformática ha sido reconocida como una competencia clave para la investigación moderna, su incorporación efectiva en entornos educativos aún presenta desafíos importantes. Muchos estudiantes de ciencias biológicas no cuentan con formación previa en programación, lo que limita su habilidad para interactuar con herramientas comunes en análisis ómicos (R, Python, entornos Linux). La integración de la bioinformática en los planes de estudio de pregrado continúa siendo insuficiente: en una encuesta nacional realizada en EE. UU., más del 70 % de los docentes reportaron enfrentar barreras para incorporar la bioinformática en sus cursos, identificando como principales desafíos la falta de formación técnica entre los estudiantes y la escasez de tiempo para reestructurar los contenidos curriculares [3]. Otro estudio descubrió que el obstáculo más frecuente informado por 1 260 profesores fue la falta de experiencia o formación docente en bioinformática, seguido por otros factores como la saturación del currículo y la preparación insuficiente de los estudiantes [4].

A estos retos se suma la complejidad técnica que implica instalar y configurar *softwares* bioinformáticos especializados, así como la necesidad de recursos computacionales que muchas universidades, especialmente en América Latina, no poseen debido a limitaciones presupuestarias persistentes [5]. La dispersión de la documentación técnica, a menudo redactada en un lenguaje poco accesible para principiantes, y la falta de acompañamiento docente capacitado agravan aún más el problema [6]. En consecuencia, muchos estudiantes perciben la bioinformática como un campo inaccesible o reservado exclusivamente a expertos, lo cual restringe su participación en proyectos reales de análisis de datos desde etapas tempranas de su formación.

Uno de los campos que ilustra claramente esta brecha es el estudio de los microARNs (miRNAs). Los miRNAs son pequeñas moléculas de ARN no codificante, con una longitud de aproximadamente 22 nucleótidos, que regulan la expresión génica a nivel postranscripcional [7], [8]. Actúan uniéndose a ARN mensajeros (mRNA) diana, lo que puede provocar su degradación o inhibir su traducción sin necesidad de degradarlos [9], [10]. Debido a su papel central en procesos celulares como proliferación, diferenciación, muerte celular y sinapsis neuronal, los miRNAs se han convertido en biomarcadores clave en cáncer, enfermedades neurodegenerativas, enfermedades cardiovasculares, entre otras [11], [12].

El análisis de miRNA implica un flujo de trabajo bioinformático multifactorial: desde control de calidad y recorte de adaptadores, hasta alineamiento de lecturas, cuantificación y análisis de expresión diferencial. Esto requiere ejecutar herramientas, normalmente, en entornos de línea de comandos [13].

Este nivel de complejidad convierte al análisis de miRNAs en un caso emblemático para mostrar la accesibilidad pedagógica de una plataforma como Galaxy, donde todos estos pasos pueden realizarse mediante interfaces visuales, sin escribir una sola línea de código [14].

Además de simplificar el proceso técnico, Galaxy favorece la colaboración y la transparencia científica al permitir que los flujos de trabajo sean compartidos, reutilizados y adaptados por otros usuarios [15]. Su interfaz gráfica, combinada con el acceso a recursos computacionales en la nube, elimina la necesidad de contar con equipos de alto rendimiento o conocimientos avanzados en administración de sistemas [14]. Esto es especialmente relevante en instituciones con recursos limitados, donde la adquisición y mantenimiento de infraestructura computacional avanzada no siempre es viable. En tales contextos, el uso de plataformas abiertas y alojadas en servidores externos, como Galaxy, permite que estudiantes y docentes accedan a herramientas bioinformáticas de alto nivel sin incurrir en costos adicionales, superando así una de las principales barreras para la enseñanza práctica en análisis ómicos [15]. De esta manera, Galaxy no solo reduce las barreras técnicas de entrada, sino que también promueve una enseñanza basada en proyectos reales, facilitando que los participantes se concentren en la interpretación biológica de los resultados más que en la resolución de problemas de infraestructura o configuración.

El presente artículo tiene como objetivo demostrar cómo puede desarrollarse un análisis completo de miRNAs utilizando exclusivamente Galaxy, desde datos públicos hasta la obtención de resultados interpretables. Se propone este enfoque como una estrategia práctica y pedagógica para introducir a estudiantes de ciencias biológicas en el mundo de la bioinformática, superando las barreras técnicas tradicionales. A través de un caso de estudio real y reproducible, se busca mostrar que el análisis de datos ómicos puede estar al alcance de estudiantes con conocimientos básicos, promoviendo una formación más integrada, moderna y participativa.

2. Materiales y Métodos

Para este estudio se utilizaron datos públicos de secuenciación de ARN pequeños (sRNA-seq) de tejido cerebral humano, correspondientes al proyecto PRJEB71709, disponible en el repositorio European Nucleotide Archive (ENA): https://www.ebi.ac.uk/ena/browser/view/PRJEB71709.

La elección de este conjunto de datos se basó en su disponibilidad pública, la inclusión de lecturas crudas y la relevancia para el análisis bioinformático de miRNAs en un contexto educativo.

Todo el procesamiento y análisis se llevó a cabo en la plataforma **Galaxy** (https://usegalaxy.org/), empleando únicamente herramientas disponibles en su interfaz gráfica. El flujo de trabajo incluyó el control de calidad de las lecturas con FastQC [16], cuyos reportes individuales fueron integrados mediante MultiQC [17]. Posteriormente, se realizó el recorte de adaptadores con Cutadapt [18], el alineamiento de lecturas con HISAT2 [19], y la cuantificación de miRNA con featureCounts [20] empleando anotaciones GFF3 de miRBase (versión 22.1) [21]. Finalmente, el análisis de expresión diferencial se llevó a cabo con DESeq2 [22].

2.1 Carga de archivos a la plataforma

Los archivos FASTQ correspondientes a las muestras seleccionadas del proyecto PRJEB71709 fueron obtenidos desde el repositorio ENA. Se seleccionaron tres muestras de tejido cerebral fetal (ERR12409245, ERR12409249, ERR12409251) y tres muestras de tejido cerebral adulto (ERR12409217, ERR12409229, ERR12409239), con el objetivo de ilustrar el análisis comparativo de perfiles de miRNA en dos grupos biológicos distintos.

Para la importación de los datos, se utilizaron los enlaces directos a los archivos FASTQ comprimidos (*formato fastq.gz*) disponibles en la página del proyecto en ENA. Dichos enlaces fueron copiados y pegados en la opción *Paste/Fetch data* de la herramienta **Upload Data** de Galaxy, lo que permitió cargar los datos directamente desde la fuente sin necesidad de descargarlos previamente al equipo local.

Una vez cargados, los archivos se renombraron y agruparon en una colección dentro de Galaxy, lo que facilitó su manejo como un único conjunto y permitió que todas las muestras fueran procesadas en bloque durante cada paso del flujo de trabajo. Los archivos se mantuvieron en su formato comprimido original (*gzip*), dado que Galaxy es compatible con su lectura directa, optimizando así tanto el uso de espacio de almacenamiento como el tiempo de transferencia.

2.2 Evaluación de calidad y limpieza de secuencias

El control de calidad inicial de las secuencias crudas, agrupadas en la colección generada en el paso anterior, se realizó con la herramienta **FastQC** (Galaxy versión

0.74) utilizando la configuración predeterminada. Este análisis permitió identificar la presencia de adaptadores remanentes en las lecturas.

Para la eliminación de adaptadores y el filtrado de secuencias se empleó **Cutadapt** (Galaxy versión 5.1), aplicando la herramienta sobre la colección de datos crudos. En la opción *Custom 3' adapter sequence*, se especificó manualmente la secuencia adaptadora estándar para bibliotecas de sRNA en plataformas Illumina (TGGAATTCTCGGGTGCCAAGG), con el fin de asegurar su remoción completa.

Se estableció un *quality cutoff (R1)* de 20, valor que corresponde a una probabilidad de error de 1 en 100 (calidad Phred Q20), considerado un umbral estándar para asegurar lecturas de alta calidad en estudios de expresión génica. Adicionalmente, se definió un rango de longitudes entre 18 y 26 nucleótidos (*minimum length (R1) = 18*; *maximum length (R1) = 26*), con el objetivo de conservar secuencias correspondientes a miRNAs maduros, los cuales suelen presentar longitudes de entre 18 y 24 nucleótidos [7].

Tras el recorte, se generó un segundo reporte de **FastQC** para verificar la eliminación de adaptadores y la mejora en los perfiles de calidad, confirmando que las lecturas procesadas eran aptas para las etapas posteriores de alineamiento y cuantificación. Finalmente, los resultados individuales de FastQC fueron integrados mediante **MultiQC** (Galaxy versión 1.27), lo que permitió obtener un reporte consolidado y visualizar de forma conjunta las métricas de las seis muestras.

2.3 Mapeo y conteo de lecturas

El mapeo de la colección de secuencias obtenida tras el procesamiento con Cutadapt se realizó utilizando **HISAT2** (Galaxy versión 2.2.1) con la configuración predeterminada, empleando como referencia el genoma humano (GCF_000001405.40) disponible en la propia plataforma Galaxy. Durante este paso se habilitó la generación de un archivo resumen con las estadísticas de alineamiento para su posterior revisión.

La cuantificación de las lecturas alineadas se llevó a cabo con la herramienta **featureCounts** (Galaxy versión 2.1.1), empleando como archivo de anotación un GFF3 específico para miRNAs de *Homo sapiens*, importado desde miRBase mediante la opción *Paste/Fetch data* de Galaxy (https://www.mirbase.org/download/hsa.gff3).

En la opción *Gene annotation file* se seleccionó "A GFF/GTF file in your history", indicando el archivo .gff3 previamente importado. Se configuró el parámetro *GFF feature type filter* como "miRNA" y el *GFF gene identifier* como "Name". Para la asignación de lecturas se activó la opción "-M -O", que incluye tanto lecturas *multi-mapping* como *multi-overlapping*, dado que en el análisis de miRNAs estas situaciones son comunes y biológicamente relevantes [23].

Este procedimiento permitió generar, para cada muestra, un archivo de conteos por cada miRNA anotado en el archivo de referencia. Dichos archivos fueron utilizados directamente en el análisis de expresión diferencial posterior sin necesidad de combinarlos previamente.

2.4 Análisis de expresión diferencial

Previo al análisis, se editaron los nombres de los archivos de conteos generados en la etapa anterior, de manera que coincidieran con las etiquetas que se mostrarían en las representaciones gráficas posteriores. El análisis de expresión diferencial se llevó a cabo utilizando la herramienta **DESeq2** (Galaxy versión 2.11.40.8). En la opción *How* se seleccionó "*select datasets per level*", definiendo el factor experimental como "Edad". Para el *Factor level 1 (Fetal)* se seleccionaron los tres archivos de conteos correspondientes a las muestras fetales, mientras que para el *Factor level 2 (Adult)* se seleccionaron los tres archivos de conteos correspondientes a las muestras adultas. Todos los demás parámetros se mantuvieron en su configuración predeterminada.

3. Resultados y discusión

En esta sección se presentan los resultados obtenidos a lo largo del flujo de trabajo implementado en Galaxy, desde el control de calidad de las lecturas crudas hasta el análisis de expresión diferencial de miRNAs entre muestras de tejido cerebral fetal y adulto. Cada etapa se acompaña de su interpretación y discusión, destacando los aspectos técnicos relevantes observados en el procesamiento de los datos.

3.1 Control de calidad antes y después del recorte de adaptadores

El análisis inicial con FastQC mostró que todas las lecturas crudas presentaban alta calidad por base (PASS), descartando problemas de secuenciación como fuente de error. Sin embargo, el módulo *Adapter Content* indicó una presencia elevada de adaptadores de Illumina (FAIL en todas las muestras), con acumulación a partir de los 22 nt, reflejando el hecho de que las lecturas (51 nt) excedían la longitud típica de los miRNA maduros (18–

24 nt) y, por lo tanto, retenían fragmentos del adaptador (ver figura 1A).

Tras el recorte con Cutadapt, el módulo de Adapter Content pasó a PASS en todas las muestras, confirmando la remoción exitosa de adaptadores. De igual manera, las secuencias sobre-representadas, que inicialmente correspondían a adaptadores y *primers*, fueron reemplazadas por secuencias clasificadas como "No Hit", lo que refleja la alta abundancia relativa de unos pocos miRNA dominantes, un hallazgo esperado en bibliotecas de sRNA-seq [24].

La distribución de longitudes posterior al filtrado mostró un perfil uniforme y consistente entre todas las muestras, concentrado en el rango de 18–26 nt, con un pico principal alrededor de los 22 nt, característico de miRNA maduros (ver figura 1B), lo que asegura que las lecturas retenidas corresponden al tamaño biológico esperado y facilita la posterior anotación y cuantificación de miRNA [7].

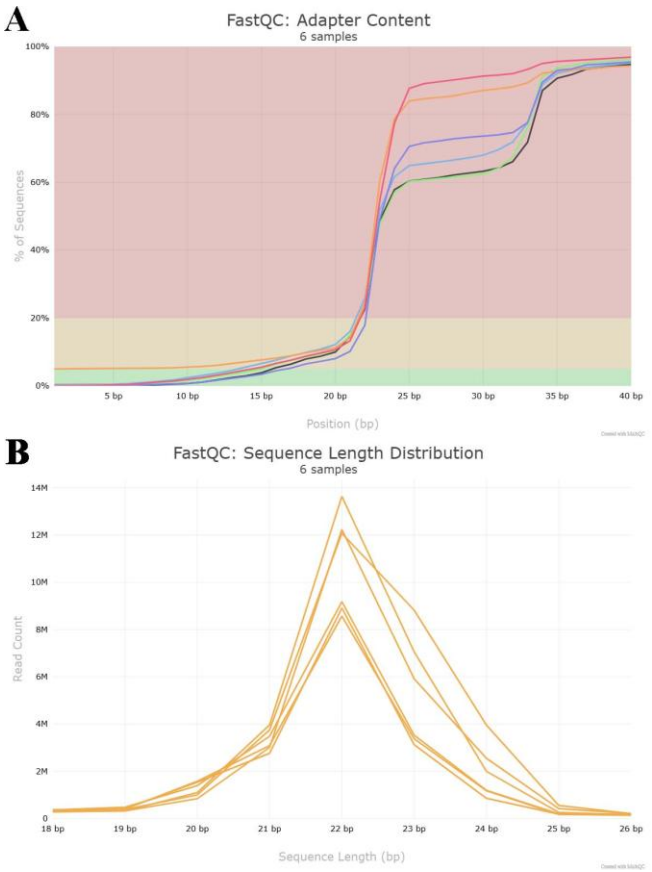

Figura 1. Resumen del control de calidad antes y después del recorte. **(A)**. Contenido de adaptadores en las lecturas crudas, mostrando acumulación a partir de los 22–24 nt (FAIL en todas las muestras). Tras el recorte con **Cutadapt**, el módulo pasó a **PASS**, confirmando la eliminación de adaptadores. **(B)** Distribución de longitudes de las lecturas después del filtrado, mostrando un perfil uniforme entre las seis muestras y concentrado en el rango esperado para miRNA maduros, con un pico principal alrededor de los 22 nt.

Los resultados consolidados en la tabla 1 reflejan este patrón de mejora: el *Adapter content* cambió de FAIL a PASS, la distribución de longitudes de WARN a PASS, mientras que la calidad por base se mantuvo en PASS en todo momento. La marca FAIL persistente en *Overrepresented sequences* corresponde a la sobreexpresión biológica de ciertos miRNA y no a contaminantes técnicos. En bibliotecas de sRNA es habitual que unos pocos miRNA muy abundantes superen el umbral definido por FastQC (≥0.1% del total de lecturas), lo que activa la alerta automática. Este resultado refleja la naturaleza regulatoria y la acumulación diferencial de determinados miRNA en el tejido analizado, más que un problema de calidad experimental o de preparación de la librería [24], [25].

Tabla 1. Estado de los principales módulos de FastQC en muestras crudas y procesadas

Muestra/ Estado	Per-b quality	Adapter content	Overrepresent seqs	Length dist
A. Total				
Antes	PASS	FAIL	FAIL	WARN
Después	PASS	PASS	FAIL (No Hit)	PASS
A. Parental Lobe				
Antes	PASS	FAIL	FAIL	WARN
Después	PASS	PASS	FAIL (No Hit)	PASS
A. Cerebellum				
Antes	PASS	FAIL	FAIL	WARN
Después	PASS	PASS	FAIL (No Hit)	PASS
F. Total				
Antes	PASS	FAIL	FAIL	WARN
Después	PASS	PASS	FAIL (No Hit)	PASS
F. Parental Lobe				
Antes	PASS	FAIL	FAIL	WARN
Después	PASS	PASS	FAIL (No Hit)	PASS
F. Cerebellum				
Antes	PASS	FAIL	FAIL	WARN
Después	PASS	PASS	FAIL (No Hit)	PASS

El control de calidad con FastQC se realiza tanto en las lecturas crudas como en las lecturas procesadas para responder a dos preguntas distintas pero complementarias: En las muestras crudas, permite diagnosticar posibles problemas técnicos derivados de la secuenciación, como la presencia de adaptadores, caídas

en la calidad de base, contaminantes o sesgos de composición. Tras el recorte con Cutadapt, un segundo análisis con FastQC confirma si las intervenciones aplicadas corrigieron efectivamente esas deficiencias, garantizando que las lecturas finales son aptas para el alineamiento y la cuantificación.

La herramienta MultiQC complementa este proceso al integrar los resultados de todas las muestras en un solo reporte, facilitando la comparación global. Mientras que FastQC muestra el detalle de cada muestra, MultiQC ofrece una visión panorámica que permite identificar patrones comunes o detectar muestras atípicas. Esta integración es especialmente útil en proyectos con múltiples réplicas biológicas, donde la consistencia entre muestras es clave para la validez estadística.

Este flujo metodológico refleja la lógica de diagnóstico → intervención → verificación: primero se identifican los problemas potenciales mediante FastQC en lecturas crudas, luego se aplican soluciones parametrizadas con Cutadapt, y finalmente se confirma la corrección con un segundo análisis de FastQC. Seguir esta secuencia garantiza la obtención de datos confiables para las etapas posteriores de mapeo y cuantificación, y constituye una práctica fundamental en el análisis bioinformático de datos de secuenciación.

3.2 Eficiencia de mapeo y asignación de lecturas

Los archivos resumen generados por HISAT2 y featureCounts resultaron fundamentales para la evaluación inicial del mapeo y la asignación de lecturas. En el caso de HISAT2, permitieron identificar fácilmente tanto el número total de lecturas como las lecturas mapeadas por muestra, lo que constituye un indicador directo de la calidad de la alineación y de la representatividad de los datos para los análisis posteriores de expresión diferencial [19]. Por su parte, los reportes de featureCounts proporcionaron información complementaria al cuantificar las lecturas asignadas específicamente al *feature* de interés, en este caso los miRNA, a la vez que informan sobre las lecturas no asignadas, lo que facilita el control y la veracidad de los conteos [20]. Gracias a esta información integrada, fue posible organizar rápidamente los datos en la tabla 2, que resume la calidad del alineamiento y la representatividad de cada muestra.

El alineamiento de las secuencias procesadas contra el genoma de *Homo sapiens* (GCF_000001405.40) utilizando HISAT2 mostró altos porcentajes de mapeo en todas las muestras, con valores entre 88,09 % y 91.54 %. Estos niveles de alineamiento son consistentes con lo reportado en experimentos de miRNA-seq cuando se emplean genomas de referencia de alta calidad y anotaciones específicas para la especie analizada [26], [27].

Tabla 2. Resumen de alineamiento y cuantificación de miRNAs

Muestra	*Lecturas totales*	*Lecturas mapeadas*	*% Mapeo*	*miRNA asignados*
Adult Total	18,811,270	17,900,223	91.54 %	25,731,752
Adult P_Lobe	18,266,856	16,491,388	88.09 %	23,998,199
Adult Cerebellum	20,009,572	17,872,828	88.94 %	27,771,945
Fetal Total	31,115,782	27,011,378	88.72 %	45,839,991
Fetal P_Lobe	25,803,146	22,777,547	88.17 %	37,090,512
Fetal Cerebellum	28,778,505	26,016,359	90.40 %	48,481,866

En todas las muestras, el número de lecturas asignadas superó al de lecturas mapeadas. Este fenómeno no implica un error de conteo, sino que refleja el comportamiento esperado en el análisis de miRNAs cuando se permite la inclusión de lecturas *multi-mapping* y *multi-overlapping* mediante el parámetro *-M -O* de featureCounts. En el caso de los miRNAs, es frecuente que una misma lectura pueda alinearse a más de una región del genoma (*multi-mapping*) o coincidir con más de una anotación de miRNA en el archivo de referencia (*multi-overlapping*). Esto se debe a que muchos miRNAs pertenecen a familias altamente conservadas y comparten secuencias idénticas o muy similares, además de que algunos genes de miRNA están presentes en múltiples copias genómicas [28], [29], [30].

Como consecuencia, una misma lectura puede contabilizarse en más de una entidad anotada, incrementando el total de lecturas asignadas. Lejos de ser un artefacto, esta característica es biológicamente relevante en el análisis de miRNAs, ya que permite retener información que se perdería si se excluyeran las lecturas *multi-mapping* o *multi-overlapping*. Sin embargo, es importante interpretarla en este contexto y no confundirla con una métrica de eficiencia de asignación comparable a la utilizada en análisis de transcritos largos [31].

3.3 Expresión diferencial

El análisis de expresión diferencial realizado con DESeq2 generó dos salidas principales: (1) una tabla que incluye todos los miRNAs detectados en las muestras

analizadas, junto con métricas de abundancia, magnitud de cambio y significancia estadística, y (2) un archivo en formato PDF con representaciones gráficas que permiten evaluar tanto la separación entre grupos como la magnitud de los cambios de expresión.

Para fines de presentación, en la tabla 3 se muestra un resumen de los resultados obtenidos con DESeq2, correspondiente a los 10 miRNAs más relevantes. Estos fueron seleccionados considerando tres criterios complementarios: i) la magnitud absoluta del cambio de expresión ($|\log_2 FC|$), que refleja la intensidad del efecto biológico y permite distinguir los miRNAs con diferencias más marcadas entre condiciones; ii) el valor de significancia ajustado (Padj < 0,05), que garantiza la solidez estadística de los hallazgos al controlar por comparaciones múltiples; y iii) un nivel de expresión mínimo (baseMean > 20), que asegura que los miRNAs seleccionados no solo presenten cambios significativos, sino que también se encuentren respaldados por una cantidad suficiente de lecturas, reduciendo así la posibilidad de artefactos derivados de baja cobertura. De esta manera, la tabla resume tanto miRNAs sobreexpresados como subexpresados entre los grupos comparados, proporcionando una visión sintética de los candidatos con mayor interés biológico y robustez estadística [22], [32].

Tabla 3. Diez miRNAs diferencialmente expresados con mayor relevancia según DESeq2 (criterios: $|\log_2 FC|$, Padj < 0,05 y baseMean > 20).

GeneID	*BaseMean*	*log2FC*	*Padj*
hsa-miR-4780	20.89	-6.93	0.00054
hsa-miR-31-3p	52.94	-6.31	0.00000
hsa-miR-561-5p	713.58	6.24	0.00000
hsa-miR-1343-3p	26.81	-5.92	0.00762
hsa-miR-219a-2-3p	24879.93	-5.78	0.00000
hsa-miR-219b-5p	24957.98	-5.78	0.00000
hsa-miR-135a-3p	189.05	5.72	0.00000
hsa-miR-874-3p	1608.11	-5.70	0.00054
hsa-miR-1224-5p	217.88	-5.30	0.00000
hsa-miR-31-5p	1944.63	-5.24	0.00000

La primera representación gráfica en el archivo PDF corresponde al análisis de componentes principales (PCA) generado a partir de la matriz de conteos normalizados de DESeq2 (ver figura 2). En este gráfico, cada punto representa una muestra en un espacio bidimensional definido por los dos primeros componentes principales, que concentran el mayor porcentaje de varianza en los datos de expresión [33]. El color distingue el grupo biológico (fetal o adulto), y las etiquetas identifican cada muestra según su tipo de tejido.

El PCA revela una clara separación entre los grupos adultos (naranja) y fetales (turquesa) a lo largo del primer componente principal (PC1), que explica el 52 % de la varianza total. El segundo componente (PC2), con un 22 % adicional, refleja variaciones intragrupo relacionadas con la sección cerebral de origen. Destaca que las muestras de cerebelo se distancian más del resto de las muestras dentro de cada grupo, lo que sugiere un perfil de expresión de miRNAs distintivo para esta región, independiente de la edad del donante.

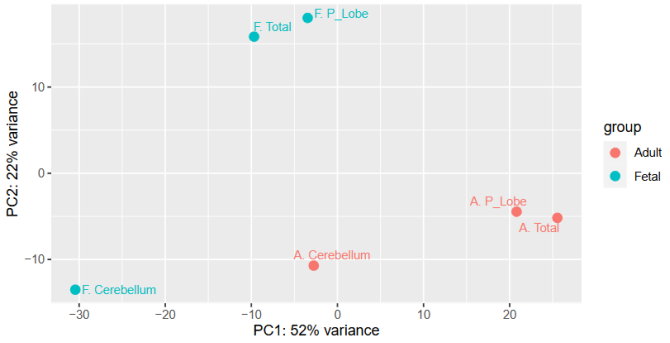

Figura 2. Análisis de componentes principales (PCA) de la expresión de miRNAs. Se observa separación clara entre muestras fetales y adultas (PC1: 52 % de varianza) y diferenciación adicional por sección cerebral (PC2: 22 %).

El PCA constituye un recurso visual de gran valor, ya que permite comprender de forma intuitiva cómo los patrones globales de expresión diferencian grupos biológicos sin necesidad de examinar cada miRNA de forma individual. Asimismo, la proporción de varianza explicada por los dos primeros componentes (52 % y 22 %) facilita comprender el concepto de varianza en el análisis multivariante y su relevancia para interpretar datos ómicos. Este tipo de visualización también fomenta la reflexión sobre la consistencia de las réplicas biológicas y la importancia de un diseño experimental balanceado [34].

La segunda representación gráfica (ver figura 3) corresponde al mapa de calor de distancias entre muestras, calculado a partir de la matriz de conteos normalizados de DESeq2. Cada celda del gráfico representa la distancia de expresión global entre un par de muestras, codificada en una escala de colores. La escala numérica (0–60) corresponde a la distancia euclidiana entre los perfiles de expresión: valores

cercanos a 0 indican alta similitud (tonos oscuros), mientras que valores más altos reflejan menor similitud (tonos claros). El dendrograma asociado muestra la agrupación jerárquica basada en estas distancias [35].

El patrón observado es coherente con el análisis de componentes principales, ya que las muestras tienden a agruparse primero por grupo biológico (fetal o adulto) y, dentro de estos, por sección cerebral. En particular, las muestras de cerebelo de ambos grupos presentan un perfil de expresión diferenciado que las separa del resto de las muestras de su mismo grupo, lo que coincide con lo evidenciado en el PCA.

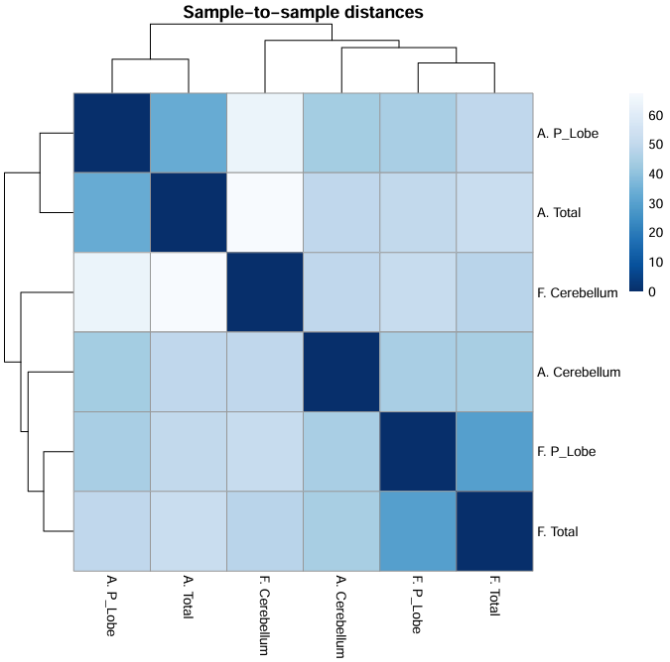

Figura 3. Mapa de calor de distancias entre muestras basado en perfiles de expresión de miRNAs normalizados. Las muestras se agrupan principalmente por grupo biológico y, dentro de estos, por región cerebral, destacando el perfil particular del cerebelo.

Este tipo de visualización refuerza la comprensión de cómo se pueden evaluar relaciones globales entre muestras en estudios de expresión génica. Permite identificar patrones de agrupamiento y evaluar la consistencia de las réplicas biológicas, así como detectar posibles muestras atípicas [36], [37]. Además, al complementarse con el PCA, este análisis contribuye a una interpretación más robusta de la estructura de los datos antes de proceder a examinar genes o miRNAs individuales.

La figura 4 muestra la relación entre la media de los conteos normalizados y las estimaciones de dispersión obtenidas por DESeq2 para cada miRNA incluido en el análisis. Cada punto negro representa la variabilidad calculada para un miRNA de manera individual (gene–est), mientras que la línea roja indica la tendencia general, es decir, cómo debería variar la dispersión según el nivel de expresión promedio (fitted) y la línea azul corresponde a los valores finales que usa el modelo estadístico, después de ajustar y estabilizar las estimaciones (final) [22].

En términos simples, la dispersión nos dice qué tan consistente es la expresión de un miRNA entre las réplicas biológicas, más allá de las fluctuaciones esperadas por azar. La tendencia observada —con mayor dispersión en miRNAs de baja abundancia y una estabilización progresiva a medida que aumenta la media de conteos— es un patrón típico en datos de RNA-seq y está directamente relacionada con la precisión de las estimaciones de cambio de expresión [22], [37], [38].

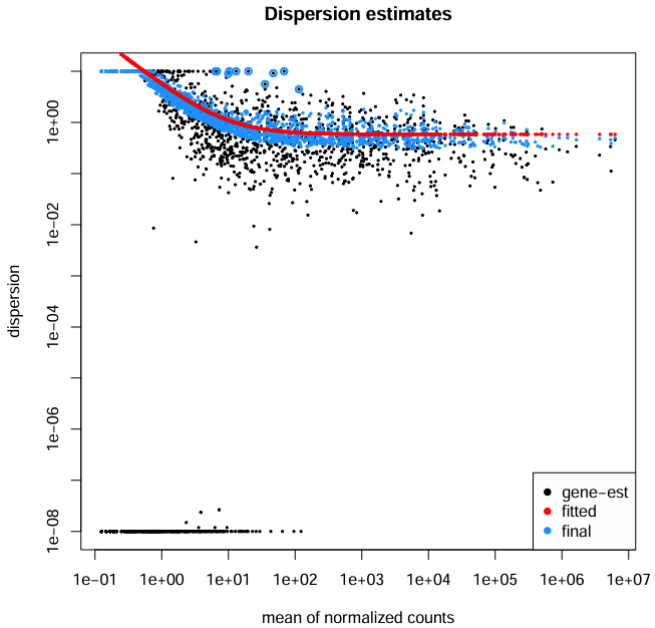

Figura 4. Estimaciones de dispersión para los miRNAs analizados mediante DESeq2. Se observa mayor variabilidad en miRNAs de baja abundancia y estabilización en los más expresados, patrón típico en RNA-seq.

Esta visualización es clave para entender uno de los pasos menos visibles, pero más importantes del flujo de análisis: el modelado de la variabilidad biológica. Este gráfico permite reforzar conceptos de estadística aplicada en ómicas, como el uso de ajustes paramétricos para estabilizar estimaciones y mejorar la detección de diferencias reales en la expresión génica.

La figura 5 presenta el histograma de valores *p* obtenidos en la comparación de expresión diferencial entre los grupos fetal y adulto. En este gráfico, cada barra representa la frecuencia de miRNAs que presentan un valor *p* dentro de un intervalo específico. Se observa una clara acumulación de valores *p* cercanos a cero, lo que indica la presencia de un conjunto importante de miRNAs con diferencias de expresión estadísticamente significativas entre los grupos comparados.

La distribución relativamente uniforme de los valores *p* intermedios sugiere que, más allá de los miRNAs con cambios claros, existe un amplio conjunto de transcritos cuya variación podría atribuirse al azar o a efectos biológicos menores. Desde un punto de vista estadístico, este patrón es consistente con experimentos bien diseñados, donde los genes diferencialmente expresados representan una fracción del total analizado, mientras que la mayoría presenta variaciones aleatorias [39], [40].

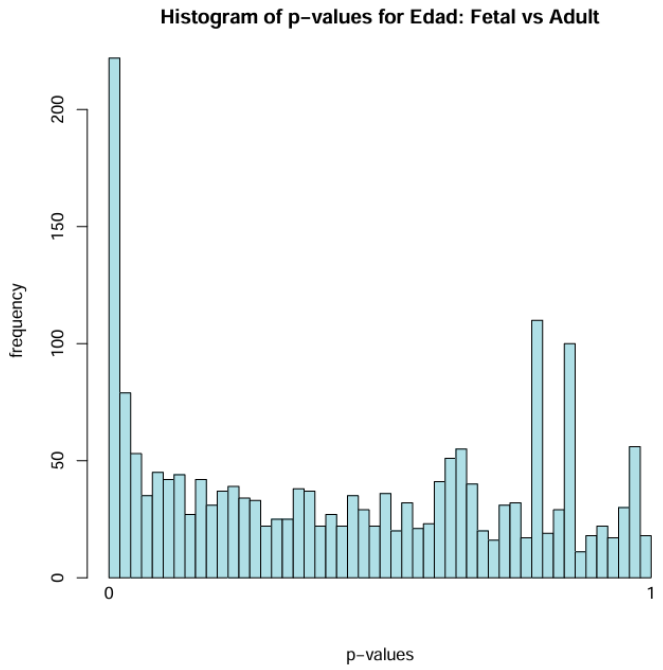

Figura 5. Distribución de valores *p* en la comparación fetal vs. adulto. El exceso de valores cercanos a cero indica un número importante de miRNAs diferencialmente expresados.

Este tipo de visualización es particularmente útil para la interpretación de los valores *p* en estudios ómicos. Permite reconocer que un exceso de valores *p* bajos señala diferencias biológicas reales, mientras que una distribución uniforme o acumulaciones inesperadas en valores altos pueden indicar sesgos, problemas en el diseño experimental o artefactos técnicos [41], [42], [43].

La figura 6 muestra el MA-plot resultante de la comparación de expresión diferencial entre las muestras fetales y adultas. En este gráfico, cada punto representa un miRNA, donde el eje *x* indica la media de los conteos normalizados (*mean of normalized counts*) y el eje *y* muestra el cambio de expresión como $\log_2$ de la razón de cambio (*$\log_2$ fold change*). Los puntos en color azul corresponden a miRNAs con diferencias de expresión estadísticamente significativas tras la corrección por pruebas múltiples (*Padj* < 0,05), mientras que los puntos en gris representan miRNAs sin significancia estadística.

El patrón característico en forma de "embudo" refleja la relación entre la magnitud del cambio de expresión y la abundancia media: los miRNAs con baja abundancia presentan una mayor dispersión en sus valores de $\log_2$FC, mientras que los más abundantes tienden a mostrar cambios más consistentes [38], [44]. Este tipo de representación permite identificar rápidamente miRNAs sobreexpresados (valores positivos de $\log_2$FC) y subexpresados (valores negativos de $\log_2$FC) entre los grupos comparados [40].

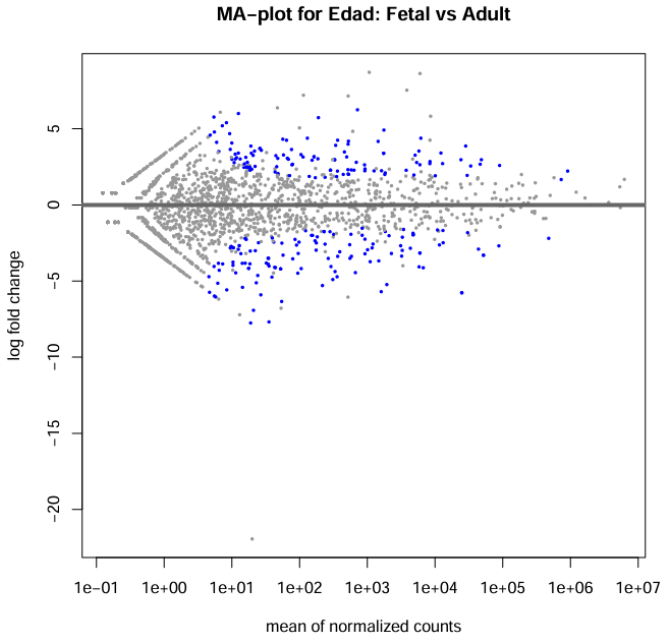

Figura 6. MA-plot del análisis de expresión diferencial entre grupos fetales y adultos. Los puntos azules marcan miRNAs con expresión diferencial significativa (Padj < 0,05).

El MA-plot es una herramienta visual efectiva para interpretar resultados de análisis de expresión diferencial,

ya que facilita comprender cómo la abundancia influye en la variabilidad de los cambios de expresión y resalta la importancia de aplicar criterios estadísticos para discernir cambios biológicamente relevantes de variaciones aleatorias [22], [38].

Los resultados obtenidos evidencian la solidez del flujo de trabajo implementado, desde la adecuada calidad de las lecturas procesadas hasta la alta eficiencia de mapeo y asignación de lecturas a miRNAs. Las representaciones gráficas derivadas del análisis con DESeq2 confirmaron la consistencia de las réplicas biológicas y permitieron identificar patrones claros de agrupamiento según la edad y la región cerebral, reforzando la validez del diseño experimental. Asimismo, la detección de un conjunto definido de miRNAs diferencialmente expresados entre muestras fetales y adultas aporta información valiosa para la comprensión de los procesos reguladores asociados al desarrollo cerebral.

4. Conclusiones

Este trabajo demostró que es posible realizar un análisis completo de miRNAs, desde datos públicos hasta la obtención de resultados biológicamente interpretables, utilizando exclusivamente la plataforma Galaxy. La estrategia propuesta, desarrollada a partir de un caso de estudio real y reproducible, constituye una herramienta pedagógica valiosa para introducir a estudiantes y profesionales de las ciencias biológicas en el análisis de datos ómicos, superando las barreras técnicas asociadas al uso de herramientas de línea de comando. Del mismo modo, al basarse en un flujo reproducible y accesible, esta propuesta también abre un espacio para que estudiantes de ingeniería en sistemas se acerquen al campo de la bioinformática, contribuyendo con sus competencias en programación, modelado y gestión de datos a la interpretación de resultados biológicos.

Entre las principales ventajas de este enfoque se encuentra su accesibilidad, ya que Galaxy no requiere instalaciones complejas ni conocimientos avanzados de programación, permitiendo que el análisis bioinformático sea más inclusivo y adaptable a distintos entornos académicos y de investigación.

Adicionalmente, la integración de herramientas como Cutadapt, HISAT2, featureCounts y DESeq2 en un flujo de trabajo reproducible facilita la comprensión de los pasos clave en el procesamiento y análisis de datos de miRNA-seq. No obstante, el trabajo también presenta limitaciones, como la dependencia de la disponibilidad de herramientas en la instancia de Galaxy utilizada y el rendimiento computacional, que puede verse afectado en análisis con grandes volúmenes de datos.

Los resultados obtenidos, que incluyen la identificación de miRNAs diferencialmente expresados entre grupos fetales y adultos y la caracterización de patrones de expresión asociados a distintas regiones cerebrales, pueden servir como punto de partida para estudios de validación experimental y exploración funcional. Asimismo, este flujo de trabajo puede ser adaptado para el análisis de miRNAs en otros modelos biológicos o condiciones experimentales, ampliando su aplicabilidad.

En conjunto, este estudio no solo contribuye al conocimiento sobre el perfil de expresión de miRNAs en distintas etapas del desarrollo cerebral humano, sino que también ofrece una propuesta concreta para democratizar el acceso a herramientas de bioinformática, con potencial de impacto en la formación de la próxima generación de investigadores en biología molecular y genómica.

AGRADECIMIENTOS

AD agradece a la Secretaría Nacional de Ciencia, Tecnología e Innovación (SENACYT) por su apoyo mediante la beca para el estudio de Doctorado en Biociencias y Biotecnología.

CONFLICTO DE INTERESES

Los autores declaran no tener algún conflicto de interés.

CONTRIBUCIÓN Y APROBACIÓN DE LOS AUTORES

AD: Conceptualización, curación de datos, análisis formal, investigación, metodología, visualización, redacción (borrador original).

GU: Conceptualización, metodología, validación, redacción (revisión y edición).

MV: Conceptualización, supervisión, redacción (revisión y edición).

Todos los autores afirmamos que hemos leído y aprobado la versión final de este artículo.

REFERENCIAS

[1] Rustici G, Larcombe L, Hendricusdottir R, Attwood TK, Bacall F, Beard N, et al. ELIXIR-UK role in bioinformatics training at the national level and across ELIXIR. F1000Res 2017;6. https://doi.org/10.12688/f1000research.11837.1.

[2] Mulder N, Schwartz R, Brazas MD, Brooksbank C, Gaeta B, Morgan SL, et al. The development and application of bioinformatics core competencies to improve bioinformatics training and education. PLoS Comput Biol 2018;14. https://doi.org/10.1371/journal.pcbi.1005772.

[3] Drew J, Morgan W, Galindo S, Kleinschmit AJ, McWilliams M, Pauley M, et al. Revisiting barriers to implementation of bioinformatics into life sciences education. Front Educ (Lausanne) 2023;8. https://doi.org/10.3389/feduc.2023.1317191.

[4] Williams JJ, Drew JC, Galindo-Gonzalez S, Robic S, Dinsdale E, Morgan WR, et al. Barriers to integration of bioinformatics into undergraduate life sciences education: A national study of US life sciences faculty uncover significant barriers to integrating bioinformatics into undergraduate instruction. PLoS One 2019;14. https://doi.org/10.1371/journal.pone.0224288.

[5] Gómez D, Villalobos R, Lobo S, Pirela V. Restricciones presupuestarias contra las universidades en Las Américas. Aula Abierta Latinoamérica 2019.

[6] Vivian Jiménez. Falta de capacitación docente afecta desarrollo tecnológico. Panamá América 2024.

[7] Bartel DP. Review MicroRNAs: Genomics, Biogenesis, Mechanism, and Function ulation of hematopoietic lineage differentiation in mam-mals (Chen et al., 2004), and control of leaf and flower development in plants (Aukerman and Sakai, 2003. vol. 116. 2004.

[8] Ha M, Kim VN. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol 2014;15:509–24. https://doi.org/10.1038/nrm3838.

[9] O'Brien J, Hayder H, Zayed Y, Peng C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. Front Endocrinol (Lausanne) 2018;9. https://doi.org/10.3389/fendo.2018.00402.

[10] Naeli P, Winter T, Hackett AP, Alboushi L, Jafarnejad SM. The intricate balance between microRNA-induced mRNA decay and translational repression. FEBS Journal 2023;290:2508–24. https://doi.org/10.1111/febs.16422.

[11] Backes C, Meese E, Keller A. Specific miRNA Disease Biomarkers in Blood, Serum and Plasma: Challenges and Prospects. Mol Diagn Ther 2016;20:509–18. https://doi.org/10.1007/s40291-016-0221-4.

[12] Mori MA, Ludwig RG, Garcia-Martin R, Brandão BB, Kahn CR. Extracellular miRNAs: From Biomarkers to Mediators of Physiology and Disease. Cell Metab 2019;30:656–73. https://doi.org/10.1016/j.cmet.2019.07.011.

[13] Deshpande D, Chhugani K, Chang Y, Karlsberg A, Loeffler C, Zhang J, et al. RNA-seq data science: From raw data to effective interpretation. Front Genet 2023;14. https://doi.org/10.3389/fgene.2023.997383.

[14] Afgan E, Nekrutenko A, Grüning BA, Blankenberg D, Goecks J, Schatz MC, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Res 2022;50:W345–51. https://doi.org/10.1093/nar/gkac247.

[15] Hiltemann S, Rasche H, Gladman S, Hotz HR, Larivière D, Blankenberg D, et al. Galaxy Training: A powerful framework for teaching! PLoS Comput Biol 2023;19. https://doi.org/10.1371/journal.pcbi.1010752.

[16] Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data. 2010.

[17] Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 2016;32:3047–8. https://doi.org/10.1093/bioinformatics/btw354.

[18] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 2011;17:10. https://doi.org/10.14806/ej.17.1.200.

[19] Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 2019;37:907–15. https://doi.org/10.1038/s41587-019-0201-4.

[20] Liao Y, Smyth GK, Shi W. featureCounts: An efficient general-purpose program for assigning sequence reads to genomic features 2013. https://doi.org/10.1093/bioinformatics/btt656.

[21] Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res 2019;47:D155–62. https://doi.org/10.1093/nar/gky1141.

[22] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15. https://doi.org/10.1186/s13059-014-0550-8.

[23] Hackenberg M, Sturm M, Langenberger D, Falcón-Pérez JM, Aransay AM. miRanalyzer: A microRNA detection and analysis tool for next-generation sequencing experiments. Nucleic Acids Res 2009;37. https://doi.org/10.1093/nar/gkp347.

[24] Babraham Bioinformatics. Overrepresented Sequences. FastQC Help: Overrepresented Sequences Module 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/9%20Overrepresented%20Sequences.html? (accessed September 4, 2025).

[25] Seco-Cervera M, González-Rodríguez D, Ibáñez-Cabellos JS, Peiró-Chova L, Pallardó F V., García-Giménez JL. Small RNA-seq analysis of circulating miRNAs to identify phenotypic variability in Friedreich's ataxia patients. Sci Data 2018;5. https://doi.org/10.1038/sdata.2018.21.

[26] Kukurba KR, Montgomery SB. RNA sequencing and analysis. Cold Spring Harb Protoc 2015;2015:951–69. https://doi.org/10.1101/pdb.top084970.

[27] Baras AS, Mitchell CJ, Myers JR, Gupta S, Weng LC, Ashton JM, et al. MiRge - A multiplexed method of processing small RNA-seq data to determine MicroRNA entropy. PLoS One 2015;10. https://doi.org/10.1371/journal.pone.0143066.

[28] Johnson NR, Yeoh JM, Coruh C, Axtell MJ. Improved placement of multi-mapping small RNAs. G3: Genes, Genomes, Genetics 2016;6:2103–11. https://doi.org/10.1534/g3.116.030452.

[29] Guo L, Liang T, Gu W, Xu Y, Bai Y, Lu Z. Cross-mapping events in miRNAs reveal potential miRNA-Mimics and evolutionary implications. PLoS One 2011;6. https://doi.org/10.1371/journal.pone.0020517.

[30] Cuperus JT, Fahlgren N, Carrington JC. Evolution and functional diversification of MIRNA genes. Plant Cell 2011;23:431–42. https://doi.org/10.1105/tpc.110.082784.

[31] Deschamps-Francoeur G, Simoneau J, Scott MS. Handling multi-mapped reads in RNA-seq. Comput Struct Biotechnol J 2020;18:1569–76. https://doi.org/10.1016/j.csbj.2020.06.014.

[32] Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 2016;22:839–51. https://doi.org/10.1261/rna.053959.115.

[33] Jolliffe IT. Principal Component Analysis. Second Edition. New York: Springer; 2002. https://doi.org/10.1007/b98835.

[34] Ringnér M. What is principal component analysis? Nat Biotechnol 2008;26:303–4. https://doi.org/10.1038/nbt0308-303.

[35] Zhao S, Guo Y, Sheng Q, Shyr Y. Advanced Heat Map and Clustering Analysis Using Heatmap3. Biomed Res Int 2014;2014. https://doi.org/10.1155/2014/986048.

[36] National Cancer Institute, Center for Cancer Research. Data visualization with R: Lesson 5 – Intro to ggplot (version 4). CCR Bioinformatics Training and Education Program 2023. https://bioinformatics.ccr.cancer.gov/docs/data-visualization-with-r/Lesson5_intro_to_ggplot_version4/ (accessed September 5, 2025).

[37] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol 2016;17. https://doi.org/10.1186/s13059-016-0881-8.

[38] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11. https://doi.org/10.1186/gb-2010-11-10-r106.

[39] Storey JD, Tibshirani R, Green PP. Statistical significance for genomewide studies. n.d.

[40] Koch CM, Chiu SF, Akbarpour M, Bharat A, Ridge KM, Bartom ET, et al. A beginner's guide to analysis of RNA sequencing data. Am J Respir Cell Mol Biol 2018;59:145–57. https://doi.org/10.1165/rcmb.2017-0430TR.

[41] Breheny P, Stromberg A, Lambert J. P-Value histograms: Inference and diagnostics. High Throughput 2018;7. https://doi.org/10.3390/HT7030023.

[42] Päll T, Luidalepp H, Tenson T, Maiväli Ü. A field-wide assessment of differential expression profiling by high-throughput sequencing reveals widespread bias. PLoS Biol 2023;21. https://doi.org/10.1371/journal.pbio.3002007.

[43] STAT 555. Using the Histogram of p-values. Penn State Eberly College of Science, Department of Statistics 2018. https://online.stat.psu.edu/stat555/node/61/ (accessed September 5, 2025).

[44] Eilers PHC, Goeman JJ. Enhancing scatterplots with smoothed densities. Bioinformatics 2004;20:623–8. https://doi.org/10.1093/bioinformatics/btg454.