Universitat Oberta de Catalunya

uoc.edu

Estudio comparativo de *pipelines* de análisis de *small non-coding RNAs*

Álvaro Santacruz Roco
Máster en Bioinformática y Bioestadística
Área 4: Análisis de datos ómicos

Tutor: Mireia Ferrer Almirall
Profesor responsable de la asignatura: Antoni Pérez Navarro

02/06/2022

FICHA DEL TRABAJO FINAL

Título del trabajo:	Estudio comparativo de *pipelines* de análisis de *small non-coding RNAs*
Nombre del autor:	*Álvaro Santacruz Roco*
Nombre del consultor/a:	*Mireia Ferrer Almirall*
Nombre del PRA:	*Antoni Pérez Navarro*
Fecha de entrega (mm/aaaa):	06/2022
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*TFM – 4: Análisis de datos ómicos*
Idioma del trabajo:	Castellano
Número de créditos:	15 ECTS
Palabras clave	*small non-coding RNAs, microRNA, pipeline*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

El estudio de los *small non-coding RNA* (*sncRNAs*) ha adquirido una gran importancia en los últimos años, especialmente en el caso de los *microRNAs* (*miRNAs*). La introducción de protocolos de *high-throughput small RNA sequencing* (*sRNAseq*) ha hecho necesario el desarrollo de herramientas de análisis de los datos de secuenciación generados y de interpretación de los resultados. En los últimos años han sido publicados diferentes *pipelines* para llevar a cabo estas tareas.

Se ha realizado una búsqueda bibliográfica y revisión de *pipelines* publicados en los últimos 5 años que ha permitido conocer las características principales de los *pipelines* de análisis de *sncRNAs*, sus posibilidades de análisis y cómo realizar un análisis básico. Se han utilizado los *pipelines* COMPSRA, miRge3.0 y nf-core/smrnaseq sobre un *dataset* de muestras de tejido de pulmón en macho y hembra. De los tres *pipelines* se han obtenido las métricas de alineamiento de lecturas y el fichero de contajes de *miRNAs*. En el caso de miRge3.0 y nf-core smrnaseq también se han obtenido otros resultados como por ejemplo análisis de *isomirs* y *novel miRNAs*. A partir de los ficheros de contajes de cada uno de los *pipelines* se ha realizado un análisis de expresión diferencial y los resultados obtenidos han sido comparados.

Abstract (in English, 250 words or less):

The study of small non-coding RNAs (sncRNAs) has become very important in the last years, especially in the case of microRNAs (miRNAs). The introduction of high-throughput small RNA sequencing (sRNAseq) protocols has made it necessary to develop tools for analysing the sequencing data generated and interpreting the results. In recent years, different pipelines have been published to carry out these tasks.

A bibliographic search and review of pipelines published in the last 5 years has been carried out to learn about the main characteristics of the sncRNA analysis pipelines, their analysis possibilities and how to carry out a basic analysis. COMPSRA, miRge3.0 and nf-core/smrnaseq pipelines were used on a dataset of male and female lung tissue samples. The read alignment metrics and the file of miRNA counts were obtained from the three pipelines. In the case of miRge3.0 and nf-core smrnaseq, other results such as isomirs analysis and novel miRNAs were also obtained. From the count files of each of the pipelines, a differential expression analysis was performed and the results obtained were compared.

Índice

Lista de figuras

Figura 21. Exploración datos normalizados. **A:** miRge3.0; **B:** COMPSRA; **C:** nf-core/smrnaseq; **D:** Isakova *et al.* 2020.

Figura 22. *Heatmap* matriz de distancias. **A:** miRge3.0; **B:** COMPSRA; **C:** nf-core/smrnaseq; **D:** Isakova *et al.* 2020.

Figura 23. *Clúster* jerárquico. **A:** miRge3.0; **B:** COMPSRA; **C:** nf-core/smrnaseq; **D:** Isakova *et al.* 2020.

Figura 24. Visualización en dimensión reducida. **A:** miRge3.0; **B:** COMPSRA; **C:** nf-core/smrnaseq; **D:** Isakova *et al.* 2020.

Figura 25. Correlación de las expresiones de *miRNAs* entre los tres *pipelines* utilizados en las muestras de pulmón de hembra.

Figura 26. Correlación de las expresiones de *miRNAs* entre los tres *pipelines* utilizados en las muestras de pulmón de macho.

Figura 27. Volcano plot. **A:** miRge3.0; **B:** COMPSRA; **C:** nf-core/smrnaseq; **D:** Isakova *et al.* 2020.

Figura 28. *Heatmap* jerárquico. **A:** miRge3.0; **B:** COMPSRA; **C:** nf-core/smrnaseq; **D:** Isakova *et al.* 2020.

Figura 29. Diagramas de Venn de las listas de *miRNAs* diferencialmente expresados y filtrados como significativos. **A:** *pipelines* ejecutados en este trabajo; **B:** *pipelines* junto con la lista obtenida a partir de los datos de Isakova *et al.* 2020.

Lista de tablas

1 Resumen

El estudio de la expresión y función de los *small non-coding RNA* (*sncRNAs*) ha adquirido una gran importancia en los últimos años, especialmente en el caso de los *microRNAs* (*miRNAs*). La introducción de protocolos de *high-throughput small RNA sequencing* (*sRNAseq*) ha hecho necesario el desarrollo en paralelo de herramientas que permitan el análisis de los datos de secuenciación generados y la interpretación de los resultados. De esta forma, en los últimos años han sido publicados diferentes *pipelines* para llevar a cabo estas tareas.

En este trabajo se ha realizado una búsqueda bibliográfica y revisión de *pipelines* publicados en los últimos 5 años que ha permitido conocer en detalle las características principales de los *pipelines* de análisis de *sncRNAs*, las posibilidades de análisis que ofrecen y cómo se lleva a cabo un análisis básico. Posteriormente, se ha llevado a cabo la ejecución de algunos de estos *pipelines* sobre un *dataset* de muestras de tejido de pulmón en macho y hembra obtenido de la base de datos *Sequence Read Archive* (SRA). En concreto, se han ejecutado los *pipelines* COMPSRA, miRge3.0 y nf-core/smrnaseq. Estos *pipelines* se han seleccionado, fundamentalmente, por las diferencias que presentan en cuanto a la estrategia de alineamiento de lecturas que llevan a cabo. El primero de ellos realiza el alineamiento frente a un genoma de referencia, mientras que los otros dos lo realizan frente a bibliotecas específicas de anotación. De los tres *pipelines* se han obtenido las métricas de alineamiento de lecturas y el fichero de contajes de *miRNAs*. En el caso de miRge3.0 y nf-core smrnaseq también se han obtenido otros resultados como por ejemplo análisis de *isomirs* y *novel miRNAs*. A partir de los ficheros de contajes de cada uno de los *pipelines* se ha realizado un análisis de expresión diferencial mediante *R statistical software*.

El trabajo realizado ha permitido llevar a cabo un análisis de *sncRNA* mediante la utilización de tres *pipelines* distintos y la comparación de sus resultados, desde la lectura de los datos hasta la obtención de *miRNAs* diferencialmente expresados.

2 Introducción

2.1 Contexto y justificación del Trabajo

Las tecnologías de secuenciación masiva (*Next Generation Sequencing*, NGS) (1) han permitido en los últimos años aumentar y mejorar el conocimiento de los mecanismos y rutas biológicas. Una de las aplicaciones más recientes de estas tecnologías ha sido la identificación y detección de *small non-coding RNAs* (*sncRNA*). El estudio de la expresión y función de los *sncRNA* ha adquirido una gran importancia en los últimos años, especialmente en el caso de los *miRNAs* (2). Estas moléculas tienen un papel fundamental en la regulación post-transcripcional de la expresión de genes y participan en una gran variedad de procesos fisiológicos y/o patológicos (3-6). El interés en el estudio de este tipo de ARN ha traído consigo el desarrollo y la aplicación de protocolos de NGS como son los denominados *high-throughput small RNA sequencing* (*sRNAseq*).

La introducción de protocolos de *sRNAseq* ha hecho necesario el desarrollo en paralelo de herramientas que permitan el análisis de los datos de secuenciación generados y la interpretación de los resultados. De esta forma, en los últimos años han sido publicados diferentes *pipelines* y *workflows* para llevar a cabo estas tareas (7-11).

Todos ellos tienen puntos en común cuando se analiza su estructura general y los pasos que los componen. Por ejemplo, en aspectos como el preprocesado de datos o la necesidad de realizar un alineamiento de las lecturas. Sin embargo, difieren en otros aspectos como por ejemplo el tipo de *sncRNA* analizado, los programas y las bases de datos utilizadas para ejecutar cada uno de los pasos o el tipo de resultados obtenidos. Existe por tanto, una amplia variedad de *pipelines*, cada día mayor, que es necesario conocer en detalle en cuanto a su funcionamiento, las posibilidades que ofrecen y las limitaciones que pueden presentar; y así poder elegir en cada caso los más adecuados en función del tipo de datos y análisis que se quiera realizar.

2.2 Objetivos del Trabajo

Los objetivos principales de este trabajo son los siguientes:

1. Revisión de *pipelines* disponibles para el análisis de *sncRNA*.

2. Estudio comparativo del análisis de *sncRNA* realizado mediante diferentes *pipelines*.

2.3 Enfoque y método seguido

La estrategia a la hora de realizar este trabajo parte de una búsqueda bibliográfica que permita realizar una revisión de *pipelines* disponibles seguido de un análisis descriptivo de estos.

Seguidamente se llevará a cabo la búsqueda de un *dataset* con el cual ejecutar algunos de los *pipelines* analizados y que serán seleccionados de forma que se pueda hacer un análisis comparativo de ellos a partir de los resultados que permiten obtener.

2.4 Planificación del Trabajo

TAREAS/HITOS:

1. Revisión de *pipelines* disponibles para el análisis de *sncRNA*.

Se llevará a cabo una búsqueda de *pipelines* disponibles para el análisis de *sncRNA* a través de diferentes fuentes y recursos. No se pretende hacer una búsqueda de todos y cada uno de los *pipelines* existentes hasta la fecha, sino de aquellos más recientes, utilizados o actualizados.

TAREA 1.1. Búsqueda bibliográfica. La búsqueda se realizará a partir de bases de datos como *Pubmed*, *Web of Science* o *Scopus* de las publicaciones más recientes (últimos 5 años aproximadamente) en las cuales se presenten o se utilicen *pipelines* para el análisis de *sncRNA*. También se consultarán otros recursos como por ejemplo páginas web o repositorios de *Github* donde se encuentre disponibles estos *pipelines*.

TAREA 1.2. Estudio comparativo de *pipelines*. En cada uno de los *pipelines* se analizarán diferentes aspectos que permitan su comparación. Se realizará una descripción general de cada *pipeline* atendiendo a características como por ejemplo:

- Estructura general y descripción de los pasos que lo conforman.

- Tipos de *sncRNA* que permite analizar.

- Tipo de archivos o datos requeridos de inicio.

- Programas o herramientas utilizadas para ejecutar cada uno de los pasos.

- Bases de datos utilizadas para el alineamiento y/o la anotación.

- Tipo de archivos o resultados obtenidos.

- Otros aspectos generales como por ejemplo la flexibilidad o facilidad de uso, información y documentación disponible, etc…

HITOS: las tareas llevadas a cabo para este primer objetivo permitirán tener un conocimiento detallado de *pipelines* disponibles para llevar a cabo un análisis de *sncRNA*, así como las características y posibilidades que estos ofrecen. Esto será fundamental para llevar a cabo la selección de *pipelines* en etapas posteriores del trabajo.

2. Estudio comparativo del análisis de *sncRNA* realizado por diferentes *pipelines*.

A partir de la información obtenida a través de las tareas del objetivo 1 se podrá realizar la selección de algunos *pipelines* para ejecutarlos utilizando un *dataset* determinado y poder comparar los resultados obtenidos.

TAREA 2.1. Búsqueda de un *dataset* . Se realizará una búsqueda en bases de datos como *Gene Expression Omnibus* (GEO) o *Sequence Read Archive* (SRA) de un *dataset* que se pueda utilizar con los *pipelines* seleccionados. En principio, el área de investigación o temática del *dataset* seleccionado no es fundamental para los objetivos de este trabajo. Como idea inicial para la búsqueda se propone un *dataset* de ratón que permita el análisis de la expresión de *sncRNA*.

TAREA 2.2. Selección de *pipelines*. De todos los *pipelines* estudiados se realizará una selección de algunos de acuerdo con alguna característica diferencial que presenten entre ellos. Por ejemplo, se considerará la estrategia llevada a cabo por el *pipeline* a la hora de realizar el alineamiento de lecturas: frente a un genoma de referencia o frente a bases de datos de anotación de *sncRNA.* Otras diferencias que se puedan encontrar entre *pipelines* una vez se haya realizado su estudio comparativo, también podrían ser consideradas a la hora de realizar la selección.

También es importante tener en cuenta para realizar la selección de *pipelines* la posibilidad y/o complejidad a la hora de ejecutar el *pipeline* o parte de él o sus programas. Por lo que será necesario comprobar que es posible instalar y ejecutar correctamente todos los componentes del *pipeline* y que este se puede llevar hasta el final.

TAREA 2.3. Análisis del *dataset* con los *pipelines* seleccionados. Una vez seleccionados los *pipelines* de trabajo y se disponga del *dataset*, se llevará a cabo el análisis de este siguiendo las instrucciones de cada uno de los *pipelines*. Existen diferentes tipos de *sncRNA* que se pueden analizar, también dependiendo del *pipeline*; no obstante, el análisis que se lleve a cabo en este trabajo se centrará en *miRNAs* y durante la ejecución del trabajo se valorará la posibilidad de incluir otro tipo de *sncRNA*, si los *pipelines* seleccionadas permiten esa posibilidad y es factible llevarlo a cabo teniendo en cuenta el tiempo de ejecución de este trabajo y el progreso de este.

TAREA 2.4. Estudio comparativo de los resultados obtenidos con cada *pipeline*. Los resultados que proporcionen los *pipelines* ejecutados, que dependerán de las posibilidades que ofrezcan estos, permitirán hacer un estudio comparado y valorar su rendimiento. Para realizar este estudio, en este trabajo se tendrán en cuenta resultados como:

- Ficheros de resultados del alineamiento de lecturas.

- Ficheros de resultados del número y tipo de *sncRNAs* identificados.

- Obtención de los ficheros de contaje de cada uno de los *pipelines*.

- Análisis de expresión diferencial de *sncRNAs* de cada uno de los *pipelines*.

- Otros ficheros o resultados que pueda proporcionar cada uno de los *pipelines*.

La evaluación de estos resultados permitirá discutir sobre el rendimiento de los *pipelines* y cuáles proporcionan mejores resultados, más completos o adecuados.

HITOS: las tareas llevadas a cabo para este objetivo permitirán poder ejecutar *pipelines* completos correctamente, analizar los resultados obtenidos con cada *pipeline* y poder discutir el rendimiento o idoneidad de los *pipelines* en función de los resultados obtenidos con cada uno de ellos.

CALENDARIO:

Para elaborar un calendario de ejecución de este trabajo se ha utilizado el *software GanttProyect*, en el cual se han indicado las diferentes etapas del trabajo así como las tareas a realizar descritas en el plan de trabajo. Este calendario es orientativo, pudiendo variar algunas fechas tanto de inicio como de finalización de tareas de acuerdo al progreso de la ejecución de estas:

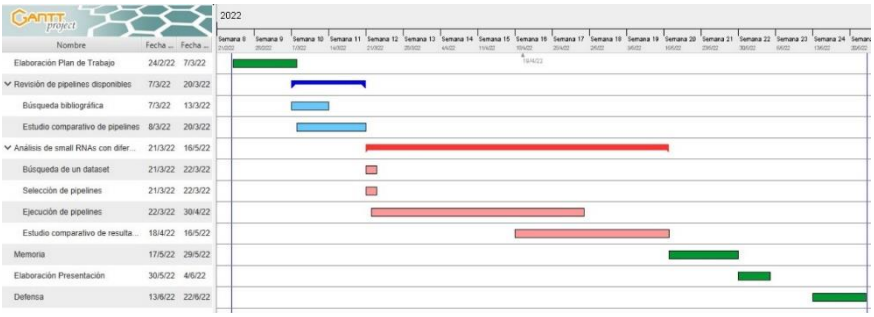

Figura 1. Calendario de trabajo.

ANÁLISIS DE RIESGOS:

En principio no se contemplan riesgos importantes que puedan poner en peligro la viabilidad de este trabajo.

- Existe una amplia variedad de bases datos y repositorios donde poder llevar a cabo una búsqueda bibliográfica sobre el tema de este trabajo.

- Los *pipelines* que se estudien en este trabajo serán de acceso libre a través de las publicaciones y repositorios y contendrán la información necesaria para su análisis y ejecución.

- El *dataset* que se utilice para ejecutar los *pipelines* será también de acceso libre al encontrarse publicado en bases de datos. Teniendo en cuenta el alcance del trabajo y el tiempo de ejecución disponible se seleccionará uno de un tamaño y número de muestras para el análisis adecuado para que los recursos de computación necesarios para ejecutar los *pipelines* no resulten excesivos, pero que permita realizar un análisis cuyos resultados sean valorables.

- La selección de *pipelines* para llevar a cabo un análisis completo será en función de la posibilidad de tener acceso a él y los programas necesarios y que dichos *pipelines* puedan ser

ejecutados correctamente. Para realizar la selección de *pipelines* en cuanto al número y características también se tendrá en cuenta el alcance del trabajo y el tiempo de ejecución disponible.

2.5 Breve sumario de contribuciones y productos obtenidos

Al finalizar este trabajo se habrá realizado una revisión de *pipelines* publicados en los últimos años y disponibles para el análisis de *sncRNA*, junto con la descripción de las características de estos y las posibilidades de análisis y resultados que pueden ofrecer.

Por otro lado, se obtendrán los resultados procedentes de la ejecución de los *pipelines* seleccionados para el análisis de *sncRNA*, así como el estudio comparativo de estos y una discusión sobre los resultados obtenidos con dichos *pipelines*.

3 Estado del arte

Una gran parte del genoma se transcribe en moléculas de ARN que no codifican para proteínas sino que directamente participan en numerosos procesos biológicos como por ejemplo la transcripción y la traducción, el silenciamiento de genes o el mantenimiento de la estructura de la cromatina y procesos relacionados con epigenética (2, 12). De esta forma están relacionadas con multitud de procesos fisiológicos y patológicos (3-6, 13).

Este tipo de ARN no codificante es denominado *non-coding RNA* (*ncRNA*) y es clasificado de acuerdo a la longitud en pares de bases de la molécula. Las moléculas mayores de 200 nucleótidos son denominadas *long non-coding RNA* (*lncRNA*). Y las moléculas menores de 200 nucleótidos *short non-coding RNA* o *small non-coding RNAs* (*sncRNA*).

Los *sncRNA* son un grupo heterogéneo de moléculas en el cual podemos encontrar diferentes tipos que incluyen (Figura 2): *microRNA* (*miRNA*)(14), *small interfering RNAs* (*siRNA*), *small nucleolar RNAs* (*snoRNA*) (15), *small nuclear RNA* (*snRNA*) (16), *PIWI-interacting RNA* (*piRNA*) (17) y *tRNA-derived small RNAs* (*tRFs*) (18). Para una revisión en detalle sobre los tipos de *sncRNA*, su biogénesis y funciones se pueden consultar por ejemplo las revisiones de Cech *et al.* 2014 (12) y Hombach et al. 2016 (19).

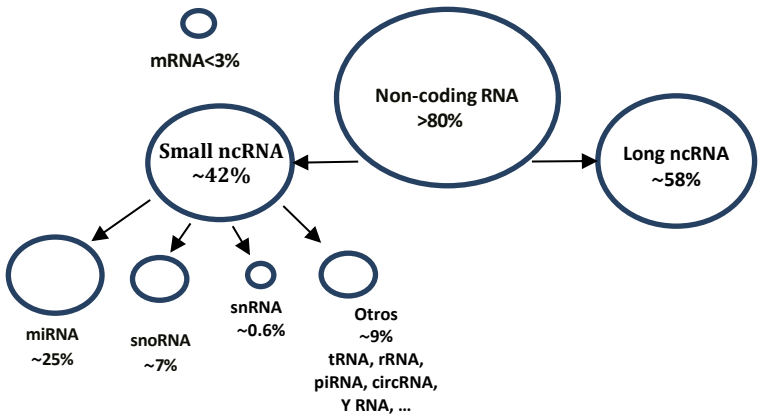

Figura 2. Tipos de ARN y su abundancia relativa en la célula con respecto al genoma. ***ncRNA***, *non-coding RNA*; ***miRNA***, *microRNA*; ***snoRNA***, *small nucleolar RNA*; ***snRNA***, *small nuclear RNA*; ***tRNA***, *transfer RNA*; ***rRNA***, *ribosomal RNA*; ***piRNA***, *PIWI-interacting RNA*; ***circRNA***, *circular RNA*; Y ***RNA***, *components of the Ro60 ribonucleoprotein particle*. Datos tomados de Seal et al. 2020 (20) y Uchida and Dimmeler et al. 2015 (21).

De todos los *sncRNAs*, los *miRNA* son hasta la fecha los más conocidos y estudiados, aunque cada vez más son los estudios que incluyen otros tipos como los *piRNAs* o los *siRNAs,* que en mamíferos parecen jugar, por ejemplo, un papel muy importante en el desarrollo de las células germinales y la fertilidad mediante el silenciamiento de transposones (17) (22). Los mecanismos implicados en la biogénesis de *miRNA* y su regulación están bien establecidos (14) (Figura 3). Los *miRNA* son moléculas de ARN de unos 21-22 nucleótidos. Una de sus funciones más conocidas es su papel como reguladores post-transcripcionales de la expresión génica a través de su unión a moléculas de ARN mensajero. Esta unión se produce por complementariedad entre la secuencia del ARN mensajero y la secuencia en el extremo 5' del *miRNA* y que es denominada *seed region*. Generalmente la *seed region* tiene una longitud entre 6 y 8 nucleótidos y se localiza entre las posiciones 2 y 10 del *miRNA*.

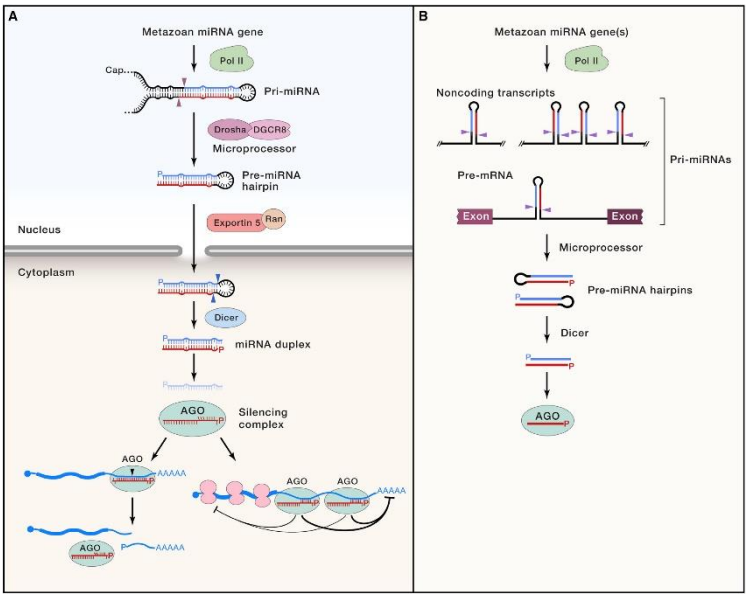

Figura 3. A: Biogénesis y función de *miRNAs* en animales. **B:** Origen de *miRNAs* en animales. Adaptado de Bartel *et al.* 2018 (2).

Los *miRNA* son transcritos inicialmente por una ARN polimerasa II en forma de precursores primarios que son procesados en el mismo núcleo por un complejo denominado Multiprocessor, en el que participan entre otras proteínas DROSHA y DGCR8 para generar los denominados *pre-miRNAs*. Estos son transportados al citoplasma mediante XPO5 y una vez allí son procesados por DICER (RNAse III-endocnuclease) para formar moléculas de ARN de doble cadena de unos 22 nucleótidos que se unen a la proteína AGO formando el complejo RISC (*RNA-induced silencing complex*). Durante este proceso, una de las hebras de ARN permanece en el complejo mientras que la otra es degradada, formándose así el *miRNA* maduro. La *seed region* de este actúa como guía que permite la unión a la región 3' UTR del ARN mensajero. De esta forma se impide la traducción del ARN mensajero, que en la mayoría de los casos además es degradado, y se regula la expresión del gen correspondiente.

En un principio se pensaba que cada pre-*miRNA* en forma de horquilla (*hairpin*) daba lugar a dos potenciales *miRNAs* maduros: uno a partir del extremo 5' (5p) y otro a partir del extremo 3' (3p). No obstante, un aspecto importante a tener en cuenta es que el procesamiento de los precursores de los *miRNA* maduros por parte de DROSHA y DICER no es completamente preciso por lo que para un mismo *miRNA* se generan diferentes variantes denominados *isomirs* que pueden diferir en longitud, composición de la secuencia o en ambas (23). Al tener diferente secuencia en la región 5' (y por tanto diferente *seed region*) o en la región 3', tienen diferentes

ARN mensajero dianas lo cual amplía la funcionalidad de los *miRNAs*. A diferencia de las variaciones en la región 5', el papel de las variaciones en la región 3' es menos conocida.

El estudio de los *sncRNA* se ha incrementado en los últimos años, adquiriendo cada vez mayor importancia en el análisis de la expresión génica (24). El interés en el estudio de este tipo de ARN ha traído consigo el desarrollo y la aplicación de protocolos de NGS como son los denominados high-throughput *small RNA sequencing* (*sRNAseq*). Y los datos de *sRNAseq* que son generados por estas plataformas han hecho necesario la utilización de nuevas herramientas que permitan un análisis eficaz y eficiente de estos y la interpretación de los resultados. De esta forma, diferentes *pipelines* y *workflows* se han ido desarrollando y publicando por parte de la comunidad científica para llevar a cabo estas tareas en la última década (7-11, 25-28).

Los *pipelines* de análisis de *sRNAseq* tienen una estructura general muy similar a los utilizados para el análisis de datos de *RNAseq*, con las particularidades que tienen este tipo de secuencias. En Potla *et al.* 2021 (29) incluyen un resumen bastante completo de la estructura general de un *pipeline* de análisis de *sncRNA* y las herramientas más habituales que se utilizan en cada uno de los pasos. Se puede observar en la Figura 4, adaptada de dicha publicación.

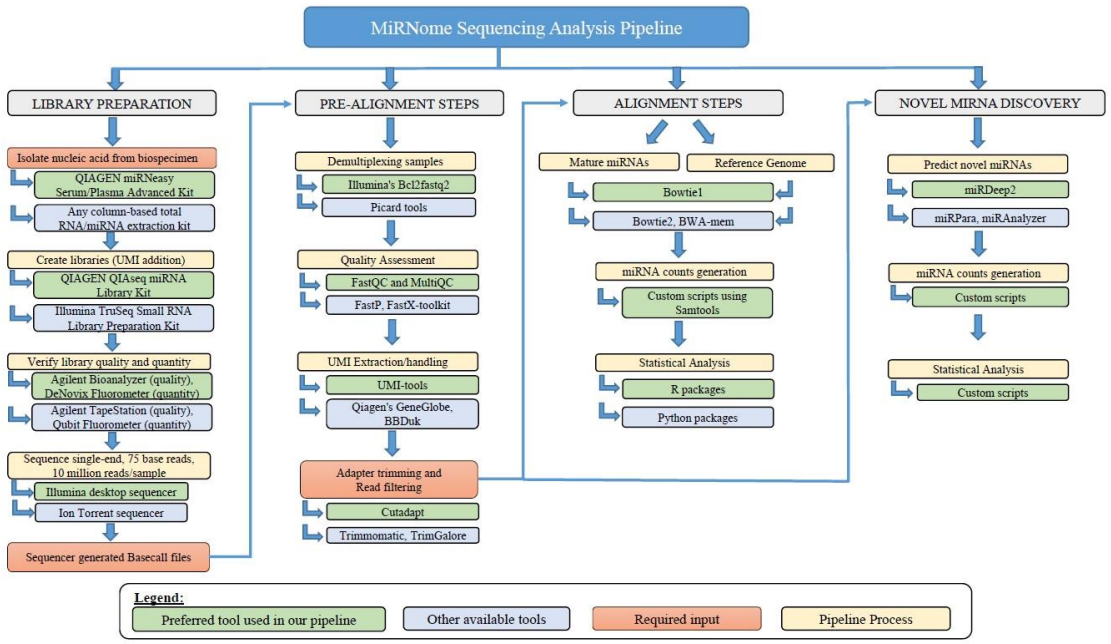

Figura 4. Esquema general de un *pipeline* de análisis de *sncRNA*. Adaptado de Potla *et al.* 2021 (29)

Es necesario realizar un análisis inicial de calidad de las lecturas de cada una de las muestras. Entre las herramientas utilizadas para realizar este paso están FASTQC (30), FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit) y MultiQC (31). Permiten visualizar métricas de calidad de las lecturas y las bases de nucleótidos, distribución de la longitud de las lecturas, contenido en GC o presencia de secuencias duplicadas o sobrerrepresentadas. A este respecto, a la hora de interpretar estos resultados es importante tener en cuenta que, dada la corta longitud de este tipo de lecturas, es muy probable encontrar en el control de calidad valores atípicos por ejemplo, en contenido en GC o por un alto número de secuencias duplicadas o sobrerrepresentadas. Por otro lado, antes de comenzar con el análisis de los datos es necesario eliminar la posible presencia de los adaptadores en posición 3' con los cuales ha sido construida la biblioteca para su secuenciación.

Uno de los aspectos que diferencian los distintos *pipelines* de análisis de *sRNAseq* disponibles es la estrategia de alineamiento de lecturas. Entre las herramientas disponibles utilizadas normalmente se encuentran Bowtie (32), BWA (33) o STAR (34). El alineamiento puede realizarse frente a un genoma de referencia de la especie de interés; frente a una base de datos de anotaciones, como por ejemplo miRBase (versión 22.1, con 38589 entradas) (35) o MirGeneDB (versión 2.1, con 16670 entradas) (36); o frente ambos. Algunos *pipelines* también disponen de bases de datos *pre-build* específicas tanto de *miRNAs* como de otros tipos que permiten ampliar el análisis de *sncRNA* (8, 9). Así, las estrategias de alineamiento pueden ser varias: alineamiento secuencial frente a bases de datos de cada uno de los tipos de *sncRNA*; alineamiento secuencial frente a una base de datos y posteriormente frente a un genoma, o viceversa; normalmente, en cada etapa de alineamiento se utilizan las lecturas que no han alineado en la etapa anterior.

Del resultado del alineamiento de las lecturas es conveniente llevar a cabo un control de calidad. En general los *pipelines* los llevan a cabo y proporcionan ficheros reporte con distintas métricas del porcentaje de lecturas alineadas y el rendimiento del alineamiento. Por ejemplo, el resultado del alineamiento frente a miRBase se puede considerar como bueno si la mayor parte de lecturas, entre un 60 y un 80%, han podido ser alineadas. Esto es también un indicador de que los pasos iniciales de procesado y filtraje de las muestras han sido adecuados y los datos son de alta calidad.

A partir del alineamiento se lleva a cabo la anotación de los *sncRNA* identificados y el recuento de lecturas que soportan dicha anotación. La mayoría de *pipelines* llevan a cabo estos pasos con herramientas como SAMtools (37) y featureCounts (38). El resultado final es un fichero de contajes que puede ser utilizado para posteriores análisis, fundamentalmente de expresión diferencial. Esta opción además es incluida por muchos *pipelines*.

De la misma forma que para el análisis de *RNAseq* se han desarrollado *pipelines* en los que se incluye la detección de polimorfismos (SNP), isoformas o variantes de un gen. Para el análisis de *sncRNA*, y en concreto de *miRNAs*, cada vez son más los *pipelines* que incluyen análisis adicionales como son la descripción de *isomirs* o la identificación de *novel miRNA* (9, 39).

Los *pipelines* también se pueden agrupar en función de los tipos de *sncRNA* analizados. Existen de esta forma dos grandes grupos: aquellos que únicamente analizan un tipo de *sncRNA*, *miRNA*, como por ejemplo miRDeep2 (39). Y aquellos que extienden el análisis en mayor o menor detalle a otros *sncRNAs*, como por ejemplo miRge3.0, COMPSRA o sRNAbench (8, 9, 40).

En el análisis de *sRNAseq* existe una particularidad que no se encuentra cuando se analizan datos *RNAseq*, cuyos protocolos además están mucho mejor desarrollados. Las lecturas de *sRNAseq* son cortas, de unos 18 a 30 nucleótidos una vez son procesadas en cuanto a calidad y se eliminan los adaptadores. Esto provoca la aparición de problemas de alineamiento múltiple de una lectura, es decir, una lectura puede alinear en múltiples localizaciones del genoma con la misma calidad de alineamiento. Además hay que tener en cuenta que muchos *sncRNAs* se transcriben de por sí a partir de diferentes *loci* del genoma (41). La estrategia de seleccionar únicamente las lecturas con alineamiento único, que es la llevada a cabo por la mayoría de *pipelines* de análisis de *RNAseq (42)*, no es aquí una opción ya que se pierde una gran cantidad de lecturas, lo cual afecta a la posterior cuantificación de los *sncRNAs* identificados. Otras estrategias posibles son la distribución de lecturas de manera equitativa o aleatoria o considerar todos los alineamientos múltiples. Cada *pipeline* opta por una u otra estrategia para solucionar este aspecto. En cualquier caso, cada una de estas estrategias puede llevar a obtener diferentes

resultados. Así pues, el alineamiento de lecturas de este tipo requiere de unos parámetros del programa de alineamiento más estrictos que reduzcan lo más posible los problemas de alineamiento múltiple. El alineamiento de lecturas frente a bases de datos de anotación en lugar de genomas de referencia puede ayudar a disminuir el problema del alineamiento múltiple (43). Sin embargo, de esta forma se limita la identificación y cuantificación de *sncRNAs* a aquellos que han sido descritos previamente.

Como se ha mencionado anteriormente, el estudio de los *sncRNAs* ha experimentado un gran avance en los últimos años, lo cual ha supuesto la aparición de numerosos *pipelines* para trabajar con estos datos. Si realizamos por ejemplo una búsqueda en Pubmed de publicaciones relacionadas con este campo de investigación con las palabras clave *non-coding RNA pipeline* o *miRNA pipeline* podemos ver como el número de publicaciones ha aumentado considerablemente últimamente (Figura 5).

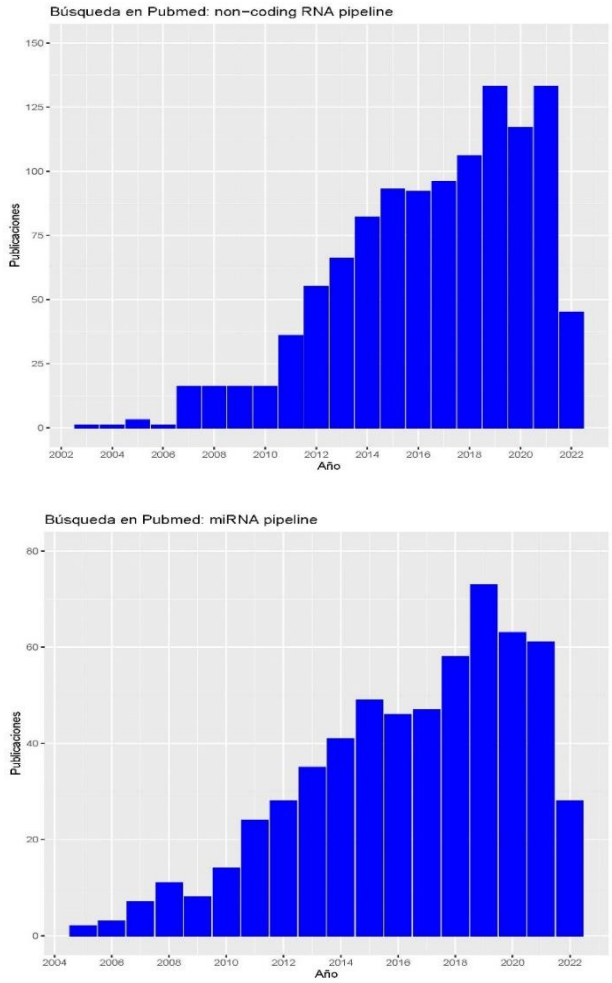

Figura 5. Número de publicaciones obtenidas en *Pubmed* para las búsquedas de palabras clave *non-coding RNA pipeline* y *miRNA pipeline*.

Por otro lado, y teniendo en cuenta todo lo anterior, la existencia de diferentes herramientas para el análisis en cada uno de los pasos, diferentes estrategias, diferentes tipos de análisis, etc… hacen que a día de hoy exista una gran heterogeneidad de *pipelines* para el análisis de *sncRNAs*. A modo de ejemplo, en el Anexo se adjunta una tabla resumen (Tabla 13) de los *pipelines* analizados para realizar este trabajo con las características principales de estructura, tipos de *sncRNA* analizados, tipos de ficheros y resultados que proporcionan y la información disponible consultada.

4 Metodología

4.1 Selección de un *dataset*

El *dataset* utilizado para ejecutar los *pipelines* forma parte del trabajo *"A mouse tissue atlas of small noncoding RNA"*, publicado por Isakova *et al.* en el año 2020 (44). En este trabajo llevan a cabo un estudio de los perfiles de expresión de diferentes clases de *sncRNAs* en 11 tejidos distintos de ratón obtenidos de 14 individuos. La descripción de este dataset se encuentra disponible en la base de datos *Gene Expression Onmibus* (GEO) https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119661. Y los ficheros de las muestras pueden obtenerse en la base de datos *Sequence Read Archive* (SRA) https://www.ncbi.nlm.nih.gov/sra?term=SRP160385.

Del total de 139 muestras disponibles se han seleccionado para realizar este trabajo muestras de tejido de pulmón de macho y de hembra. La selección de las muestras se ha realizado de acuerdo a los resultados obtenidos por los autores al comparar los perfiles de expresión de *miRNAs* entre sexos, teniendo en cuenta las condiciones en las cuales han observado diferencias evidentes. Se han seleccionado tres muestras para cada sexo, que se pueden considerar réplicas biológicas. La lista de ficheros utilizados en este trabajo aparece resumida en la Tabla 1.

4.2 Selección de *pipelines* de análisis de *sncRNA*

De todos los *pipelines* analizados que se incluyen en la Tabla 13 del Anexo se han seleccionado tres de ellos para realizar un análisis de *sncRNA*: miRge3.0 (9), COMPSRA (8) y nf-core/smrnaseq (28). Para seleccionarlos se ha tenido en cuenta, fundamentalmente, la estrategia de alineamiento de las lecturas que realiza el *pipeline*, frente a genoma o bibliotecas de anotación; el tipo de *sncRNA* analizado, al menos *miRNAs*; y si lleva a cabo otro tipo de análisis como *isomirs* o *novel miRNAs*.

El *pipeline* miRge3.0 (Figura 6) lleva a cabo el alineamiento de las lecturas frente a bibliotecas específicas de anotación de distintos *sncRNA*. Las lecturas alineadas y anotadas a partir de las bibliotecas de anotación específicas frente a *miRNAs* permiten obtener el fichero de contajes de aquellos que han sido identificados y que puede utilizarse posteriormente para un análisis de expresión diferencial. También permite llevar a cabo análisis de identificación de *isomirs*. Por otro lado, a partir de las lecturas que no alinean frente a ninguna biblioteca de *sncRNA* utilizada lleva a cabo la predicción de nuevos *miRNAs* mediante *Support Vector Machine* (SVM).

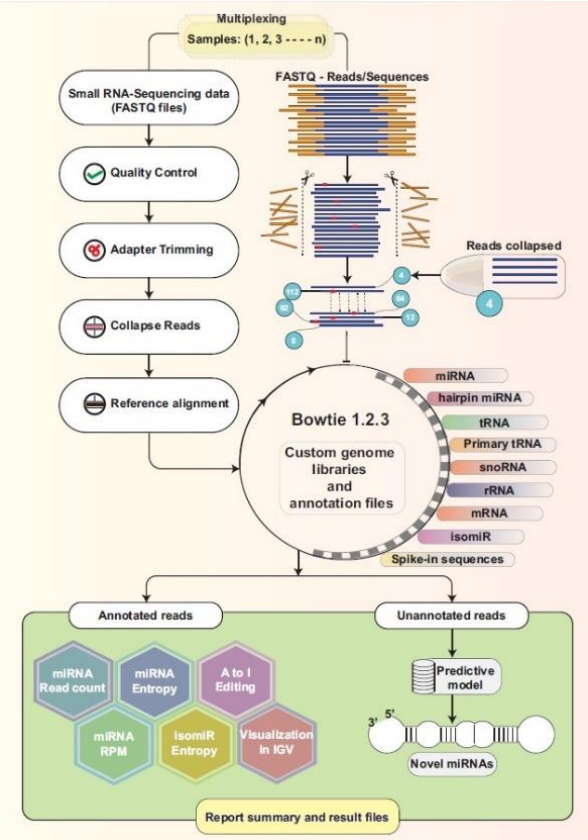

Figura 6. Esquema general de trabajo de miRge3.0 para el análisis de *sncRNAs* (adaptado de Patil *et al.* 2021) (9).

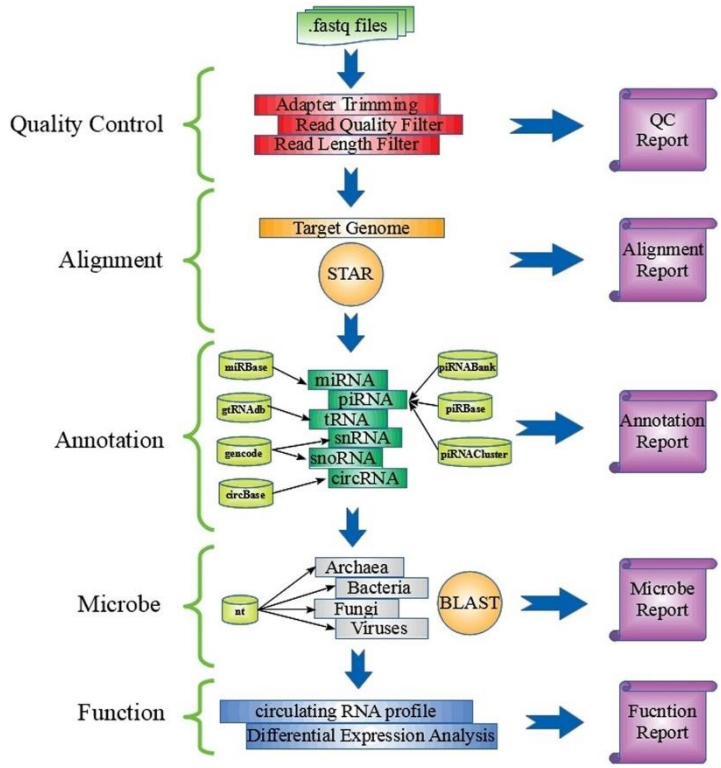

Figura 7. Esquema general de trabajo de COMPSRA para el análisis de *sncRNAs* (adaptado de Li *et al.* 2020)(8).

En cuanto a COMPSRA (Figura7), este *pipeline* consta de cinco módulos que pueden ejecutarse de manera independiente o todos en conjunto: *Quality Control* (QC), *Alignment*, *Annotation*, *Microbe* y *Function*. El alineamiento de las lecturas (*Alignment*) se lleva a cabo

frente a un genoma de referencia. Y la identificación y cuantificación (Annotation) es realizada a partir de bibliotecas de anotación específicas del *pipeline*. En este caso, la identificación y cuantificación se realiza para *miRNAs* y también para otros tipos de *sncRNAs*. Para todos ellos, se puede obtener un fichero de contajes con el cual realizar análisis de expresión diferencial (Function). El módulo Microbe, que es opcional, permite identificar posibles contaminaciones de las muestras a partir de bibliotecas de secuencias de virus y bacterias.

Finalmente, el *pipeline* de nf-core/smrnaseq (Figura 8) lleva a cabo diferentes tipos de alineamiento. En primer lugar realiza un alineamiento frente bibliotecas de anotación de miRBase (35) de *miRNAs* maduros y posteriormente *miRNAs* precursores. Este alineamiento es el que permite la identificación y cuantificación de *miRNAs* y la obtención de un fichero de contajes a partir del cual realizar un análisis de expresión diferencial. También permite la identificación y el análisis de *isomirs*. Posteriormente, un primer alineamiento de las lecturas frente a un genoma de referencia es utilizado como control de calidad del alineamiento de las lecturas. Finalmente, un segundo alineamiento de las lecturas frente a un genoma de referencia es llevado a cabo mediante el *script* de mirDeep2 (39) para la identificación y cuantificación *de novo* de *miRNAs*.

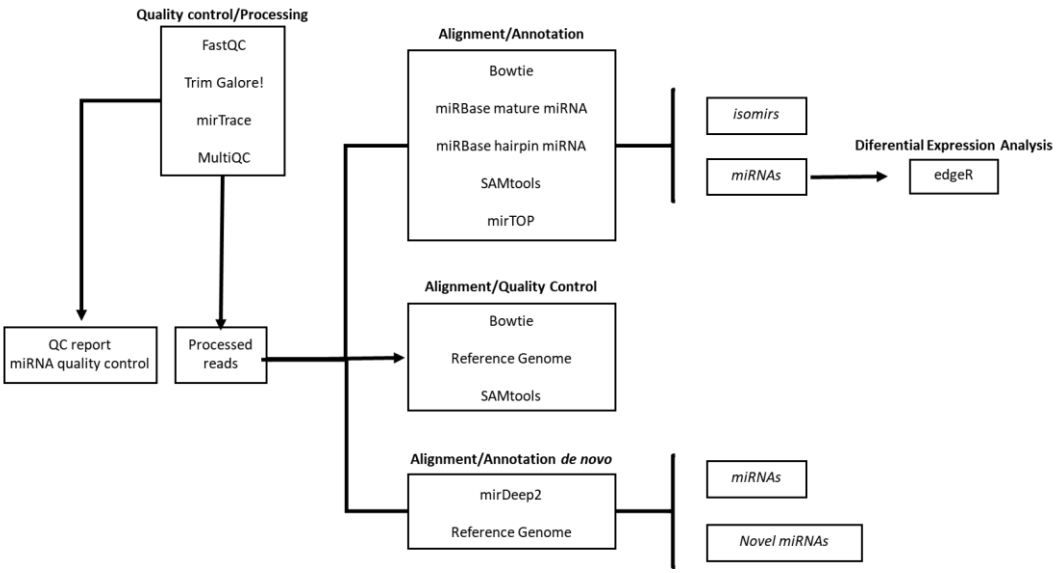

Figura 8. Esquema general de trabajo de nf-core/smrnaseq para el análisis de *sncRNAs*.

4.3 Métodos

El conjunto de comandos necesarios para obtener los ficheros de lecturas, procesarlos e instalar y ejecutar los *pipelines* de análisis de *sncRNAs* utilizados en este trabajo se encuentran en el *script "Analisis sncRNA.txt"* guardado en la carpeta "*Pipelines*" y que puede consultarse en la siguiente dirección: https://github.com/asroco/TFM. Se han ejecutado en una máquina virtual proporcionada por la *Universitat Oberta de Catalunya* (UOC) equipada con sistema operativo Ubuntu 20.04.4 LTS. A continuación, se describen brevemente.

4.3.1 Obtención de los ficheros de la base de datos SRA

Los ficheros de las muestras utilizadas se han descargados de la base de datos SRA mediante *sratoolkit* 3.0.0 (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software) utilizando el comando *fastq-dump*. Este comando permite descargar los datos en ficheros en formato *.fastq*. Para cada muestra hay dos ficheros de lecturas, por lo que una vez descargados

los ficheros de cada muestra se unen mediante el comando cat en un único fichero de trabajo (Tabla 1).

Ficheros de SRA	Muestras	Sexo
SRR7807267 SRR10695926	Lung_F1	Hembra
SRR7807270 SRR10695929	Lung_F2	Hembra
SRR7807271 SRR10695930	Lung_F3	Hembra
SRR7807259 SRR10695918	Lung_M1	Macho
SRR7807260 SRR10695919	Lung_M2	Macho
SRR7807261 SRR10695920	Lung_M3	Macho

Tabla 1. Muestras utilizadas en este trabajo como dataset.

4.3.2 Procesado de los ficheros

Algunos de los *pipelines* que se ejecutan en este trabajo disponen de sus propios pasos de control de calidad y procesado de los ficheros *.fastq* para eliminar adaptadores y lecturas de baja calidad. No obstante, este proceso se ha realizado previamente para que los ficheros de entrada, a partir de los cuales cada *pipeline* realiza el alineamiento de las lecturas, sean lo más similares posible en todos los casos.

Los ficheros de lecturas de cada una de las muestras son analizadas mediante FastQC (versión 0.11.7, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (30) y MultiQC (versión 1.12, https://multiqc.info/) (31). De acuerdo con la información obtenida, para la preparación de las bibliotecas se ha utilizado como adaptador *Illumina Small RNA 3' Adapter*.

Para eliminar los adaptadores en posición 3' y las lecturas de baja calidad se utiliza Cutadapt (versión 4.0, https://cutadapt.readthedocs.io/en/stable/) (45) con los argumentos *-a TGGAATTCTCGG*, para indicar el adaptador a eliminar; y *-m 15* para eliminar las lecturas con una longitud menor de 15 nucleótidos. El resto de argumentos se mantienen por defecto. Posteriormente, los ficheros *.fastq* generados son analizados mediante FASTQC y MultiQC. Los ficheros resultantes son los que se utilizan como ficheros de entrada en la ejecución de los *pipelines*.

4.3.3 Ejecución de *pipelines*

Se han ejecutado los siguientes *pipelines*: miRge3.0, COMPSRA y nf-core/smrnaseq. A continuación se describe brevemente los pasos necesarios para instalar los programas necesarios para cada *pipeline* y cómo éstos han sido ejecutados.

- miRge3.0

La instalación de miRge3.0 se ha llevado a cabo siguiendo las instrucciones de los autores que se pueden encontrar en la siguiente dirección: https://mirge3.readthedocs.io/en/latest/. Para poder instalar y ejecutar miRge3.0 es necesario tener instalado previamente Python 3.8.0 o superior (https://www.python.org/). Además de la instalación de miRge3.0, para poder

ejecutar el *pipeline* es necesario instalar Bowtie (versión 1.3.0), SAMtools (versión 1.10) y ViennaRNAfold (versión 2.4.16). Y también una serie de bibliotecas de anotación de *sncRNAs* específicas de la especie con la que se está trabajando (en este caso ratón), y que se obtienen a través de SourceForge (https://sourceforge.net/projects/mirge3/files/miRge3_Lib/).

Para ejecutar miRge3.0 a través de la línea de comandos se utiliza la instrucción *miRge3.0*. El *pipeline* tiene una serie de argumentos, algunos de ellos desactivados por defecto, que se pueden configurar y permiten realizar diferentes análisis y obtener diferentes resultados (https://mirge3.readthedocs.io/en/latest/quick_start.html#parameters). En este caso se han utilizado los siguientes argumentos: *-s* con el nombre de los ficheros *.fastq*; *-lib miRge3_Lib*, con la ruta a las bibliotecas de anotación de *sncRNAs* de miRge3.0; *-on mouse*, indicando el organismo (ratón); *-db miRBase*, indicando miRBase como la base de datos de *miRNAs* de referencia para la anotación; *-o* con la ruta de escritura de los ficheros de resultados; *-gff* para obtener los resultados de *miRNAs* e *isomiRs* en formato *.gff*; *-bam* para obtener los resultados del alineamiento en formato *.bam*; *-nmir* para incluir los resultados de predicción de nuevos *miRNAs*; -ai para incluir el análisis de edición A to I; *-pbwt* con la ruta del programa Bowtie; *-prf* con la ruta del programa ViennaRNAfold.

- COMPSRA

La instalación de COMPSRA se ha llevado a cabo siguiendo las instrucciones de los autores que se pueden encontrar en la siguiente dirección: https://github.com/cougarlj/COMPSRA. Para poder instalar y ejecutar COMPSRA es necesario instalar previamente Java Runtime Environment (JRE) version 8 o superior:

(http://www.oracle.com/technetwork/java/javase/downloads/index.html)

Una vez instalado JAVA y COMPSRA correctamente se puede utilizar el sistema de instalación COMPSRA *Toolkit* (tk) para poder descargar el programa de alineamiento STAR (34), las bibliotecas de anotación de *sncRNAs* específicas y el genoma de referencia de ratón mediante las siguientes instrucciones:

java -jar COMPSRA.jar -tk -dr -ck star

java -jar COMPSRA.jar -tk -dr -ck miRNA_mm10,piRNA_mm10,tRNA_mm10,snoRNA_mm10,snRNA_mm10,circRNA_mm10

java -jar COMPSRA.jar -tk -dr -ck star_mm10

El *pipeline* de COMPSRA está formado por cinco módulos que se pueden ejecutar de manera conjunta o independiente. En este trabajo se han ejecutado los módulos de control de calidad (QC), alineamiento, anotación y función de forma conjunta a través de la instrucción *java -jar COMPSRA.jar* y con los siguientes argumentos: *-ref mm10*, para indicar el genoma de referencia; *-qc*, para ejecutar el módulo de control de calidad; *-ra TGGAATTCTCGG,* para indicar el adaptador en posición 3'; *-aln*, para ejecutar el módulo de alineamiento; *-mt star*, para indicar el programa de alineamiento; *-ann*, para ejecutar el módulo de anotación; *-ac 1,3,4,5,6*, para indicar las clases de *sncRNAs* anotadas; *-fun*, para ejecutar el módulo función; *fdclass 1,3,4,5,6*, para indicar las clases de sncRNAs de las que obtener los ficheros de contaje; *-inf*, para indicar la ruta con el fichero que contiene la lista de ficheros *.fastq* a analizar; *-out* con la ruta de escritura de los ficheros de resultados.

En este *pipeline* el módulo de control de calidad es necesario ejecutarlo, independientemente del resto de módulos que se ejecuten. Como se ha indicado anteriormente, el control de calidad y procesado de los ficheros se ha realizado previamente. Por lo que los adaptadores y lecturas de baja calidad ya han sido en principio eliminados. El módulo del *pipeline* se ha ejecutado con los parámetros por defecto excepto para el argumento *-ra* en el cual se ha indicado el adaptador en 3' de la biblioteca que ya ha sido eliminado.

El resto de argumentos del *pipeline* se utilizaron con sus valores por defecto. El listado completo de argumentos disponibles se puede consultar en la siguiente dirección: https://github.com/cougarlj/COMPSRA. El fichero con la biblioteca de anotación específica de *piRNAs*, aunque era posible descargarlo, no ha sido posible utilizarlo sin que se produjera un error en la ejecución del *pipeline*. Por ello el análisis se realizó excluyendo este tipo de *sncRNA*.

- Nf-core/smrnaseq

La instalación de nf-core/smrnaseq se ha llevado a cabo siguiendo las instrucciones de los autores que se pueden encontrar en la siguiente dirección: https://nf-co.re/smrnaseq.

Para instalar y ejecutar correctamente el *pipeline* en primer lugar es necesario instalar Nextflow (versión 20.04.0 o superior) https://nf-co.re/usage/installation. Y para ello es necesario tener instalado Java Runtime Environment (JRE) version 8 o superior. Nextflow puede instalarse utilizando Bioconda mediante la instrucción *conda install nextflow*.

Para asegurar la reproducibilidad de los resultados y facilitar la ejecución del *pipeline* los autores indican que es necesario instalar alguno de los *docker containers* disponibles. Estos contienen una serie de instrucciones y configuraciones de ejecución necesarias. Esta configuración se indica en el momento de ejecutar el *pipeline* mediante el argumento *profile*. En este trabajo se ha utilizado como argumento *profile conda*.

Para ejecutar el *pipeline* se utiliza la instrucción *nextflow run nf-core/smrnaseq*. El *pipeline* tiene una serie de argumentos configurables que se pueden consultar en esta dirección: https://nf-co.re/smrnaseq/1.1.0/parameters. En este caso se han utilizado los siguientes parámetros: *-profile conda*; --input, con la ruta de los ficheros a analizar; *--genome mm10*, para indicar el genoma de referencia; *--fasta*, con la ruta al fichero con el genoma de referencia en formato *.fasta*; *--mirna_gft*, con la ruta al fichero *.gff3* con las anotaciones de precursores y *miRNAs* maduros; *--mature*, con la ruta al fichero .fasta de *miRNAs* maduros; *--hairpin*, con la ruta al fichero de *miRNAs* precursores; *-bt_index*, con ruta al índice del genoma de referencia; *--min_length 15*; *--three_prime_adapter 'TGGAATTCTCGG'*; *--trim_galore_max_length '76'*; *--skip_mirdeep*; *--outdir*, con la ruta de escritura de los ficheros de resultados.

El *pipeline* lleva a cabo un paso de procesado de los ficheros para eliminar los adaptadores y lecturas de baja calidad que no es posible omitir durante la ejecución. Como se ha indicado anteriormente, el control de calidad y procesado de los ficheros se ha realizado previamente. Por lo que los adaptadores y lecturas de baja calidad ya han sido en principio eliminados. Por ello, para los argumentos de este paso se han utilizado los mismos valores que en el procesado previo para que los cambios en las lecturas de los ficheros sean mínimos.

El *pipeline* incluye un paso de identificación de *miRNAs* conocidos y no conocidos en datos de secuenciación profunda mediante mirDeep2 (39). No obstante, no ha sido posible ejecutar el *pipeline* con este paso sin obtener errores en la ejecución. Por lo que ha sido necesario omitir esta análisis.

4.3.4 Análisis de expresión diferencial

Para el análisis de expresión diferencial se ha utilizado *R statistical software* (versión R 4.2.0, https://cran.r-project.org/index.html), mediante el software RStudio (versión 2022.02.2+485 "Prairie Trillium" Release for Windows) (https://www.rstudio.com/).

El *script* utilizado para realizar el análisis de expresión diferencial puede consultarse en el fichero *"Differential Expression Analysis.rmd"* que se encuentra guardado en la carpeta "Análisis de expresión diferencial" en la siguiente dirección: https://github.com/asroco/TFM. Para poder ejecutarlo es necesario instalar una serie de bibliotecas, algunas estándar de CRAN, y otras más específicas para un análisis de datos de secuenciación del proyecto Bioconductor: (https://www.bioconductor.org/ versión 3.15).

El análisis se realiza para cada fichero de contaje de *miRNAs* maduros obtenido con cada uno de los *pipelines* ejecutados y el fichero de contaje publicado por los autores del trabajo original (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE119661). De acuerdo con la descripción realizada en la publicación por los autores el *pipeline* utilizado para obtener este fichero de contajes está basado en el *pipeline ENCODE small RNA-seq pipeline* (46). Este es un *pipeline* diferente a los que se han seleccionado para ejecutar en este trabajo y que han sido descritos anteriormente. Este *pipeline* permite la identificación y cuantificación de distintos *sncRNAs*. Y en cuanto a los pasos para obtener el fichero de contajes, las lecturas procesadas mediante BaseSpace (Illumina) son alineadas frente a un genoma de referencia mediante STAR (34). Y la anotación y cuantificación de los diferentes *sncRNAs* se lleva a cabo utilizando las bibliotecas de GENCODE (47) y miRBase (35) mediante featureCounts (38).

En primer lugar se realiza un filtraje de *miRNAs* que se encuentran poco expresados. Para poder realizar el filtraje, los valores de contaje se estandarizan en función del tamaño de la biblioteca de cada muestra utilizando la función *cpm* del paquete edgeR. Los valores de contaje se expresan como *counts per million* (CPMs). Para realizar el filtraje de los datos se eliminan todos aquellos en los que para las muestras de macho o las muestras de hembra alguna de sus tres réplicas tenga un valor de contaje igual a 0. Para trabajar de forma más eficiente, los valores de contaje filtrados son almacenados en un objeto de la clase S4 de tipo *DGEList*.

Para realizar una primera exploración de los datos del análisis se representa su distribución mediante *boxplot*. Como los valores de contaje no siguen una distribución normal, en primer lugar es necesario realizar una transformación de los datos utilizando logaritmos. Los valores de contaje son transformados en recuentos log_2 por millón utilizando la función *cpm* del paquete *edgeR*.

La normalización de los datos se lleva a cabo mediante la función calcNormFactors. Una vez normalizados los datos, se lleva a cabo un análisis exploratorio de estos. De nuevo, los valores no siguen una distribución normal, por lo que es necesario realizar una transformación de estos utilizando logaritmos (log_2 *counts per million*) como la indicada anteriormente. Los datos transformados y normalizados se representan gráficamente mediante *boxplot*.

Para comprobar la similitud entre muestras y que estas se agrupan correctamente de acuerdo a los grupos establecidos (macho y hembra) se lleva a cabo un análisis no supervisado de similitud a partir del cálculo de la matriz de distancias. Esta matriz se calcula mediante la comparación dos a dos de todas las muestras utilizando la función *dist* a partir de los valores transformados y normalizados. Se representa en un *heatmap* mediante la función *fviz_dist* del paquete *FactorExtra* y un *clúster* jerárquico mediante la función *hclust*. Finalmente, también se

lleva a cabo un análisis en dimensión reducida a partir de la matriz de distancias calculadas mediante la función *plotMDS* del paquete *limma*.

Los perfiles de expresión de *miRNAs* comunes de los *pipelines* ejecutados (miRge3.0, COMPSRA, nf-core/smrnaseq) se comparan a partir de su correlación. La comparación se realiza mediante la función *ggscatter* del paquete *ggpubr*, para cada una de las réplicas, y comparando los *pipelines* dos a dos. Se utilizan los valores de contaje transformados y normalizados de los *miRNAs* comunes en los ficheros de contaje, y se calcula la correlación entre ellos.

El análisis de expresión diferencial se lleva a cabo mediante el paquete *DESEQ2* a partir de los valores de contaje filtrados sin normalizar. Para realizar el análisis se crea un objeto de tipo *DESeqDataSet* que contiene los datos filtrados y la información sobre los grupos que se quieren comparar (macho y hembra). El análisis se realiza mediane la función *Deseq*, indicando previamente mediante la función *relevel* el grupo de referencia a la hora de realizar la comparación (en este caso el grupo referencia es "Hembra").

La lista de *miRNAs* diferencialmente expresados se visualiza de manera gráfica en un *Volcano plot* mediante el paquete *EnhancedVolcano*. En la gráfica aparecen indicados en rojo aquellos con un valor absoluto de *log2-Fold-Change* (*log_2FC*) mayor de 1 y un *p* valor ajustado por el método de Benjamini-Hochberg menor de 0.05.

Se selecciona un subconjunto de *miRNAs* diferencialmente expresados con un valor absoluto de *log_2FC* mayor de 1 y un *p* valor ajustado menor de 0.05. El perfil de expresión de este conjunto de *miRNAs* se visualiza en un *heatmap* mediante el paquete *pheatmap*.

Las listas de *miRNAs* diferencialmente expresados sin filtrar y filtradas, obtenidas a partir de los datos de cada *pipeline*, se comparan utilizando diagramas de Venn mediante el paquete *VennDetail*.

5 Resultados

5.1 Procesado de las muestras

El análisis mediante FASTQC de los ficheros de lecturas procesados con Cutadapt 4.0 permite comprobar que la calidad de las muestras es la adecuada para poder utilizarlas en un *pipeline* de análisis. De manera general, en la Tabla 2 observamos que la longitud media de las lecturas es de entre 28 y 38 nucleótidos, con un contenido en GC aproximado de entre 45 y 48%. El número de lecturas por muestra está en torno a unos 20 millones o incluso superior como es el caso de las muestras de macho. Por otro lado, es de destacar que el número de secuencias duplicadas es muy elevado, superior al 90%. No obstante, tratándose de lecturas cortas de unos 20 ó 30 nucleótidos como es el caso de estas muestras de *sRNAseq*, estas cifras elevadas de duplicaciones en principio son de esperar. Las métricas generales de estas muestras son las que cabría esperar de este tipo de lecturas.

Sample Name	% Dups	% GC	Read Length	M Seqs
Lung_F1	93.3%	46%	28 bp	20.7
Lung_F2	92.6%	47%	28 bp	17.9
Lung_F3	97.0%	48%	30 bp	20.1
Lung_M1	97.6%	45%	38 bp	25.5
Lung_M2	97.2%	46%	37 bp	28.7
Lung_M3	95.7%	45%	36 bp	34.2

Tabla 2. Métricas generales de los ficheros de lecturas procesadas con Cutadapt 4.0.

Cuando se analiza los valores medios de calidad de las secuencias a lo largo de todas las bases (Figura 9), teniendo en cuenta todas las secuencias para cada una de las muestras, se observa que para todas ellas estos valores se encuentran dentro del rango óptimo. La calidad media disminuye ligeramente, especialmente en las muestras de hembra, en la parte final. Pero en general, los valores de calidad son adecuados.

En cuanto al valor medio de calidad de cada una de las secuencias (*Phred score*) contenidas por cada muestra Figura 9, en todos los casos el valor es superior a 30, llegando en la mayoría de los casos a alcanzar un valor de 35. Lo cual indica que para todas las muestras el valor medio de calidad es óptimo.

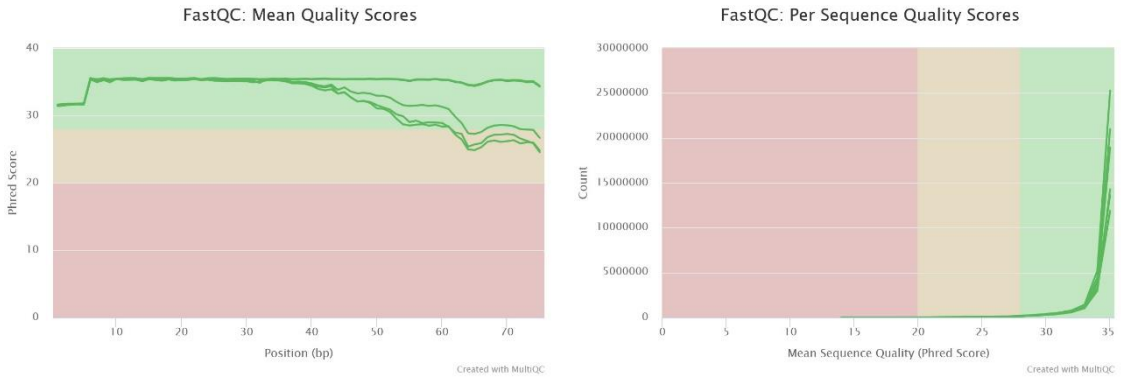

Figura 9. MultiQC. Valores medios de calidad de las secuencias a lo largo de todas las bases (izquierda) y valor medio de calidad de cada una de las secuencias (Phred score) (derecha).

En otros aspectos sí que aparecen *a priori* valores de calidad poco frecuentes o inadecuados. Es el caso de los valores de las secuencias por bases, en cuanto al contenido relativo de cada nucleótido a lo largo de cada una de las posiciones (Figura 10A). En una biblioteca aleatoria, esperamos que exista poca diferencia (o ninguna) entre las diferentes bases que componen las secuencias. Las líneas en este gráfico deberían ir paralelas a lo largo de las diferentes posiciones de las secuencias. Esto no ocurre en ninguna de las muestras, donde el perfil de cada línea es muy heterogéneo.

El contenido en GC por secuencia tampoco tiene un perfil adecuado en ninguna de las muestras (Figura 10B), al no seguir una distribución normal estándar como cabría esperar de una biblioteca aleatoria. No existen problemas de contenido de *Ns* en ninguna de las muestras (Figura 10C). Pero sí que la distribución de la longitud de la secuencias es irregular en todas las muestras (Figura 3D). La mayoría de las secuencias de todas las muestras tienen una longitud en torno a 20-25 nucleótidos. Pero también hay algunas de en torno a 30 y 40 nucleótidos. Especialmente significativo es el caso de la muestra 3 de hembra (Lung_F3), cuyo pico en torno a 30 nucleótidos es bastante alto. También existen secuencias en todas las muestras con una longitud entre 70 y 76 nucleótidos (que en principio es el tamaño de las lecturas sin procesar).

Finalmente, otros dos aspectos en los que los valores de calidad de las muestras son poco frecuentes o inadecuados son el número de secuencias duplicadas y el número de secuencias sobrerrepresentadas (Figura 11 A y B). En todas las muestras estos números son muy elevados.

Para poder interpretar los resultados anteriores hay que tener en cuenta que las muestras están obtenidas a partir de bibliotecas de *sRNAseq*. Estas bibliotecas están formadas por conjuntos de secuencias cortas. En este caso la longitud máxima de las lecturas de la biblioteca es de 76 nucleótidos y tras el procesado de las muestras hay lecturas de menor tamaño. Cuando se preparan estas bibliotecas, los fragmentos de ARN no son cortados al azar antes de añadir los adaptadores. Por lo que las lecturas para *sncRNAs* específicos serán idénticas. Por ello, es de esperar en el análisis de calidad de este tipo de muestras los resultados poco frecuentes mencionados anteriormente.

Si se tiene en cuenta que los valores medios de calidad de las secuencias para cada una de las posiciones y por *Phred score* han resultado adecuados. Y que los resultados poco frecuentes obtenidos son los que cabría esperar y consecuencia del tipo de biblioteca de las muestras (*sRNAseq*) podemos concluir que la calidad de las todas las muestras es adecuada para poder ser utilizadas en la ejecución de los *pipelines*. Únicamente el resultado obtenido para la muestra 3 de hembra con un pico muy elevado en torno a 30 nucleótidos para la longitud de las secuencias debe de ser tenido en cuenta. Este resultado, sólo en esta muestra, podría ser consecuencia de algún tipo de contaminación o error en la construcción de la biblioteca que podría afectar a los resultados obtenidos o posteriores análisis.

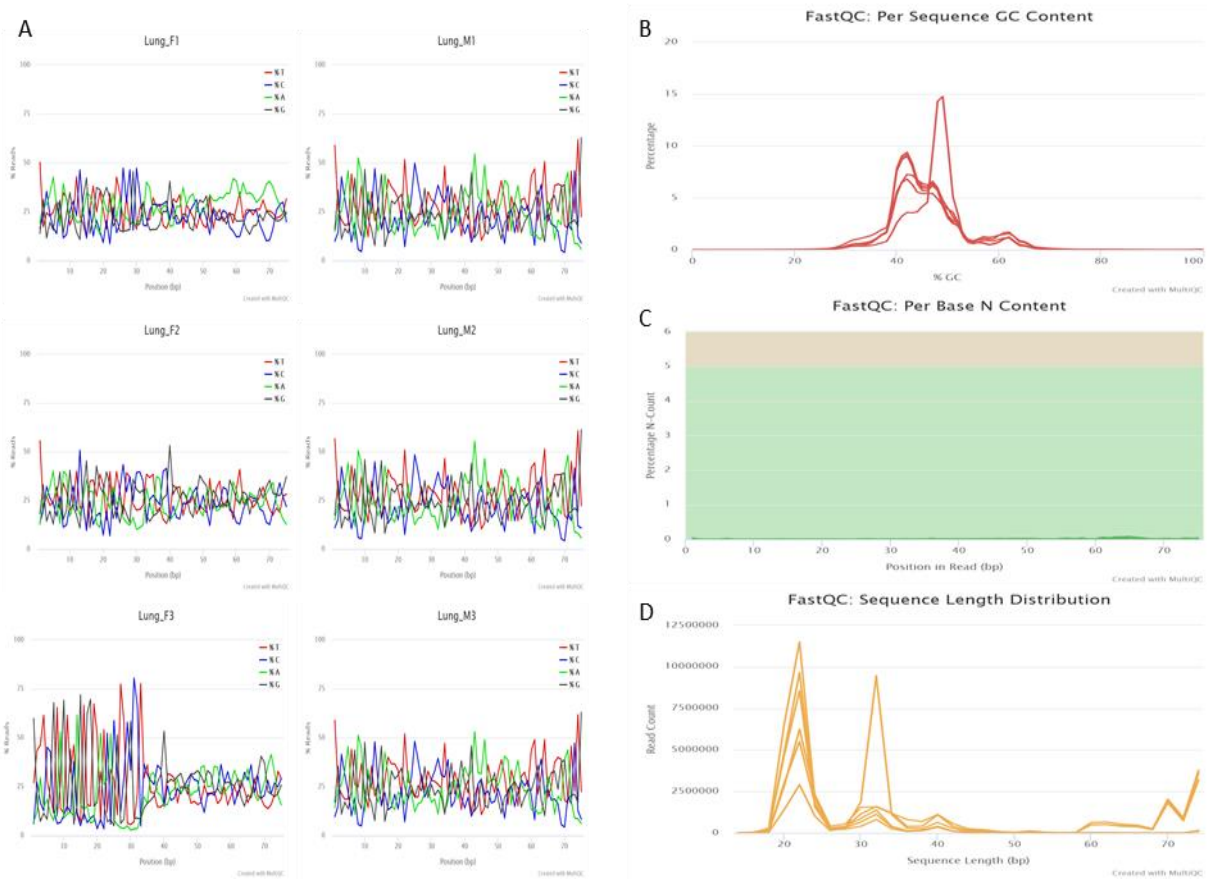

Figura 10. MultiQC. **A:** contenido relativo de cada nucleótido por posición; **B:** contenido en GC; **C:** contenido en Ns; **D:** Distribución de la longitud de las secuencias.

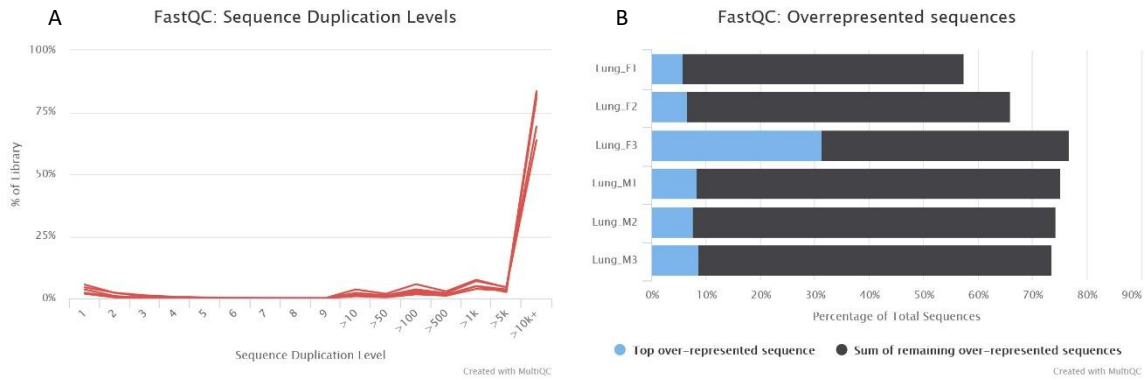

Figura 11. MultiQC. **A:** Número de secuencias duplicadas; **B:** Número de secuencias sobrerrepresentadas.

Otro aspecto importante que podemos observar en los resultados de calidad de MultiQC es que los adaptadores se han eliminado correctamente. La ejecución de FASTQC sobre los ficheros originales con las lecturas ha permitido identificar que los adaptadores utilizados para construir la biblioteca son del tipo *Illumina small RNA 3' adapater*. Mediante Cutadapt 4.0 este adaptador ha sido eliminado correctamente en todas las muestras (Figura 12).

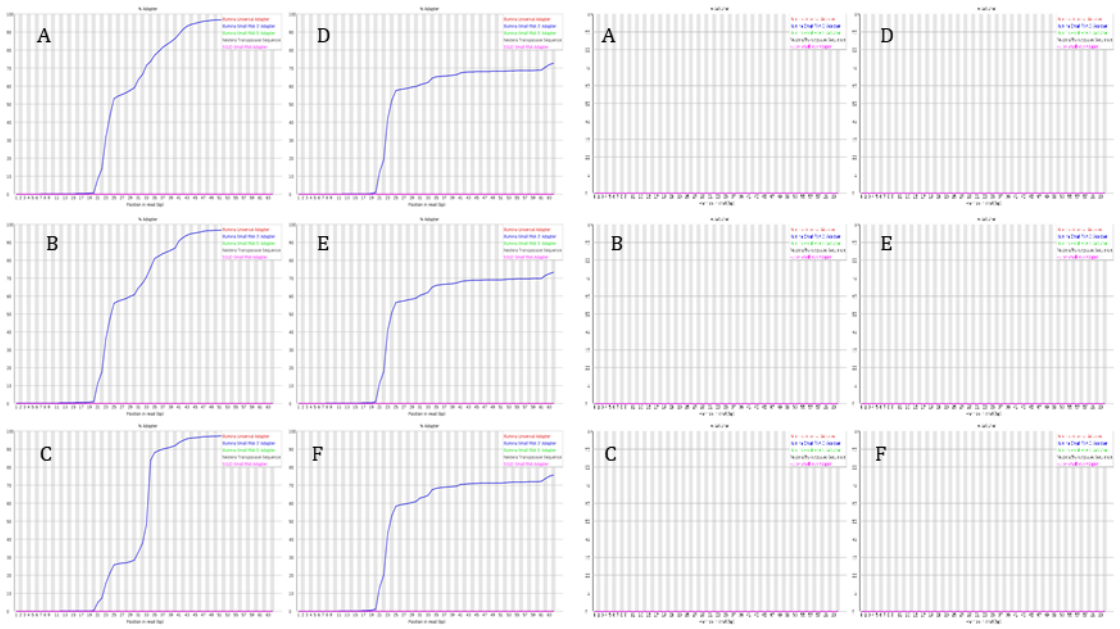

Figura 12. Detección (izquierda) y eliminación (derecha) del adaptador mediante Cutadapt4.0. **A:** Lung_F1; **B:** Lung_F2; **C:** Lung_F3; **D:** Lung_M1; **E:** Lung_M2; **F:** Lung_M3.

5.2 Análisis de *sncRNAseq* mediante diferentes *pipelines*

5.2.1 Análisis mediante miRge3.0

Los ficheros de resultados obtenidos de la ejecución del *pipeline* miRge3.0 se pueden consultar al completo en la carpeta "*Pipelines*", en la dirección: https://github.com/asroco/TFM. Los resultados obtenidos a través de este *pipeline* pueden variar en función de los parámetros indicados mediante los diferentes argumentos que permite el *pipeline*.

El *pipeline* miRge 3.0 lleva a cabo el alineamiento de las lecturas frente bibliotecas de anotación específicas. Las métricas del resultado del alineamiento de lecturas de cada una de las muestras se puede consultar en las Tablas 3 y 4, que el propio *pipeline* proporciona como resultado. En la Tabla 3 se incluyen las métricas de lecturas totales, filtradas y alineadas frente

a la base de datos de *miRNAs* (en este trabajo se ha utilizado miRBase v22.1) (35), así como el número de *miRNAs* identificados. Los valores de *Trimmed Read all* superiores al 99% en todas las muestras nos indican que los archivos procesados que se han utilizado como entrada apenas han sido modificados en cuanto a la eliminación de secuencias o bases de baja calidad por parte del *pipeline*. Los valores de *Trimmed Read unique* por debajo de un 10% en todas las muestras indican el alto número de duplicaciones, tal como el análisis previo que se ha realizado también lo indicaba. El porcentaje de lecturas alineadas frente a la base de datos de *miRNAs* maduros se encuentra de manera general por encima del 50% del total de lecturas (52-57%). Sin embargo, para la muestra 3 de hembra (Lung_F3) este porcentaje es mucho menor (25-26%).

Sample name(s)	Total Input Reads	Trimmed Reads (all)	Trimmed Reads (unique)	All miRNA Reads	Filtered miRNA Reads	Unique miRNAs
Lung_F1	20739536	20728333 99.9%	1338516 6.45%	11088664 53.47%	10821633 52.18%	555
Lung_F2	17931049	17919074 99.9%	1269486 7.08%	10141292 56.56%	10016198 55.86%	564
Lung_F3	20137854	20131926 99.9%	569910 2.83%	5270501 26.17%	5167849 25.66%	471
Lung_M1	25479710	25469227 99.9%	686192 2.69%	14745377 57.87%	14629699 57.42%	604
Lung_M2	28669675	28655390 99.9%	984313 3.44%	16146363 56.32%	16022545 55.89%	639
Lung_M3	34230890	34212487 99.9%	1261741 3.69%	19949919 58.28	19804088 57.85	659

Tabla 3. Métricas generales del alineamiento de lecturas en miRge 3.0

En la Tabla 4 se muestran las métricas de lecturas que han alineado frente las bases de datos de otros tipos de *sncRNAs*, así como el número de lecturas que han quedado sin alinear frente a ninguna de las bases de datos utilizadas.

De manera gráfica, los resultados de las Tablas 3 y 4 se pueden observar en la Figura 13, donde para cada muestra se ha representado el porcentaje de lecturas que corresponde a cada uno de los tipos de *sncRNAs*. La mayoría de las muestras tienen un perfil similar, donde el porcentaje mayoritario (52-57%) se corresponde con *miRNAs* maduros, y en torno a un 35-40% se corresponde con otros tipos de *sncRNAs*. Sin embargo, la muestra 3 de hembra (Lung_F3) tiene un perfil diferente y el mayor porcentaje se corresponde con *tRNA* maduro. En la tabla 4 se puede observar que para esa muestra, el número de lecturas alineadas para esa categoría es muy superior al resto (11056388). Este resultado está relacionado con lo que ya se ha observado en el análisis de calidad previo en cuanto a la distribución de la longitud de las secuencias y una posible contaminación o fallo en la preparación o secuenciación de la biblioteca. Finalmente, el número de lecturas no alineadas es en todos las muestras inferior a un 10%.

Sample name(s)	Hairpin miRNAs	mature tRNA Reads	primary tRNA Reads	snoRNA Reads	rRNA Reads	ncRNA others	mRNA Reads	Remaining Reads
Lung_F1	86362	2025418	26890	759389	2666510	1574963	877737	1622400 7.82%
Lung_F2	84607	2263985	21538	1259650	1518384	282121	667520	1679977 9.36%
Lung_F3	51424	11056388	14828	647420	1366104	124736	379381	1221144 6.06%
Lung_M1	39145	1114897	35843	6391052	1236845	165997	325088	1414983 5.55%
Lung_M2	41722	1656528	33490	6684429	1703325	217043	380034	1792456 6.25%
Lung_M3	48856	2088887	27259	7708025	1346355	232511	505762	2304913 6.73%

Tabla 4. Métricas del alineamiento de lecturas en miRge3.0 frente a diferentes tipos de *sncRNAs*.

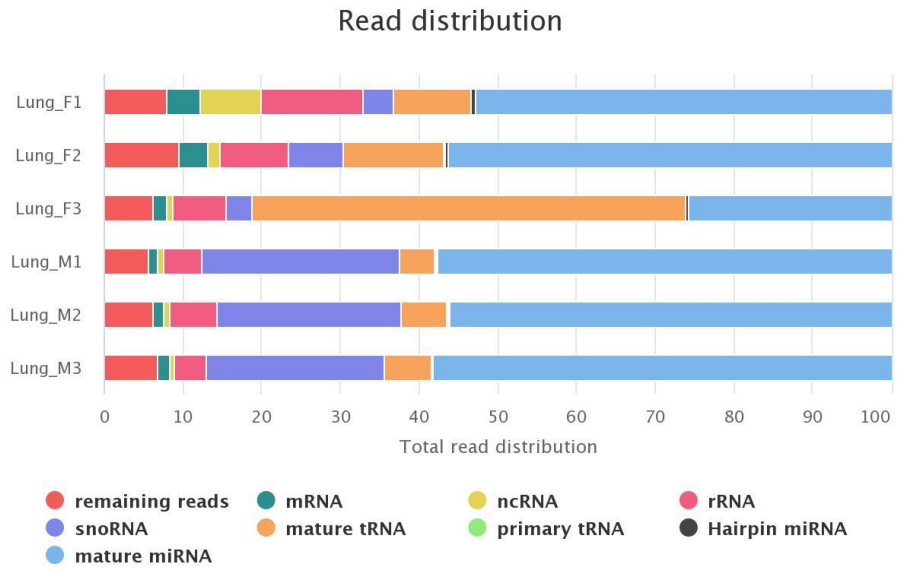

Figura 13. Distribución de lecturas alineadas por miRge3.0 de acuerdo al tipo de *sncRNA*.

Como resultado del alineamiento de lecturas el *pipeline* también proporciona para cada muestra un fichero en formato *.bam* que permite visualizar los resultados mediante un navegador como por ejemplo IGV (https://software.broadinstitute.org/software/igv/) (48).

Por otro lado, del alineamiento y cuantificación de lecturas se obtienen los ficheros de contajes normalizados y sin normalizar que pueden ser utilizados en posteriores análisis de expresión diferencial. El número de elementos cuantificados en estos ficheros es de 1927. En este trabajo no se ha utilizado el paso de análisis de expresión diferencial durante la ejecución del *pipeline*. Este análisis se ha realizado posteriormente mediante un *script* en R a partir del fichero de contajes sin normalizar.

Como resultado de la ejecución del *pipeline* también se han obtenido los gráficos de la Figura 14. Cada gráfico en forma de panel se corresponde a una de las muestras y representa los 40 *miRNAs* que se expresan en mayor abundancia de acuerdo a los valores de contaje en *reads per million* (RPM). Los *miRNAs* más abundantes son en su mayoría los mismos para todas las muestras, destacando por ejemplo *miR10a-5p*, *miR143-3p*, *miR181a-5p*, *miR26a-5p* o *miR30a-5p* entre otros.

Figura 14. Top 40 *miRNAs* de mayor expresión en cada una de las muestras obtenido por miRge3.0.

El análisis de *isomirs* llevado a cabo por miRge3.0 da como resultado un fichero en formato *.gff3* con los *isomirs* identificados. Además representa gráficamente la proporción de *isomirs* en con respecto a la región de la secuencia en la que se produce las variantes. Y para el top 20 de *miRNAs* más abundantes en cada una de las muestras representa gráficamente la distribución de las lecturas en las regiones de la secuencia donde se producen las variantes (Figura 15). Se observa que en las muestras analizadas el mayor número de variantes se encuentra acumulado en la región 3p.

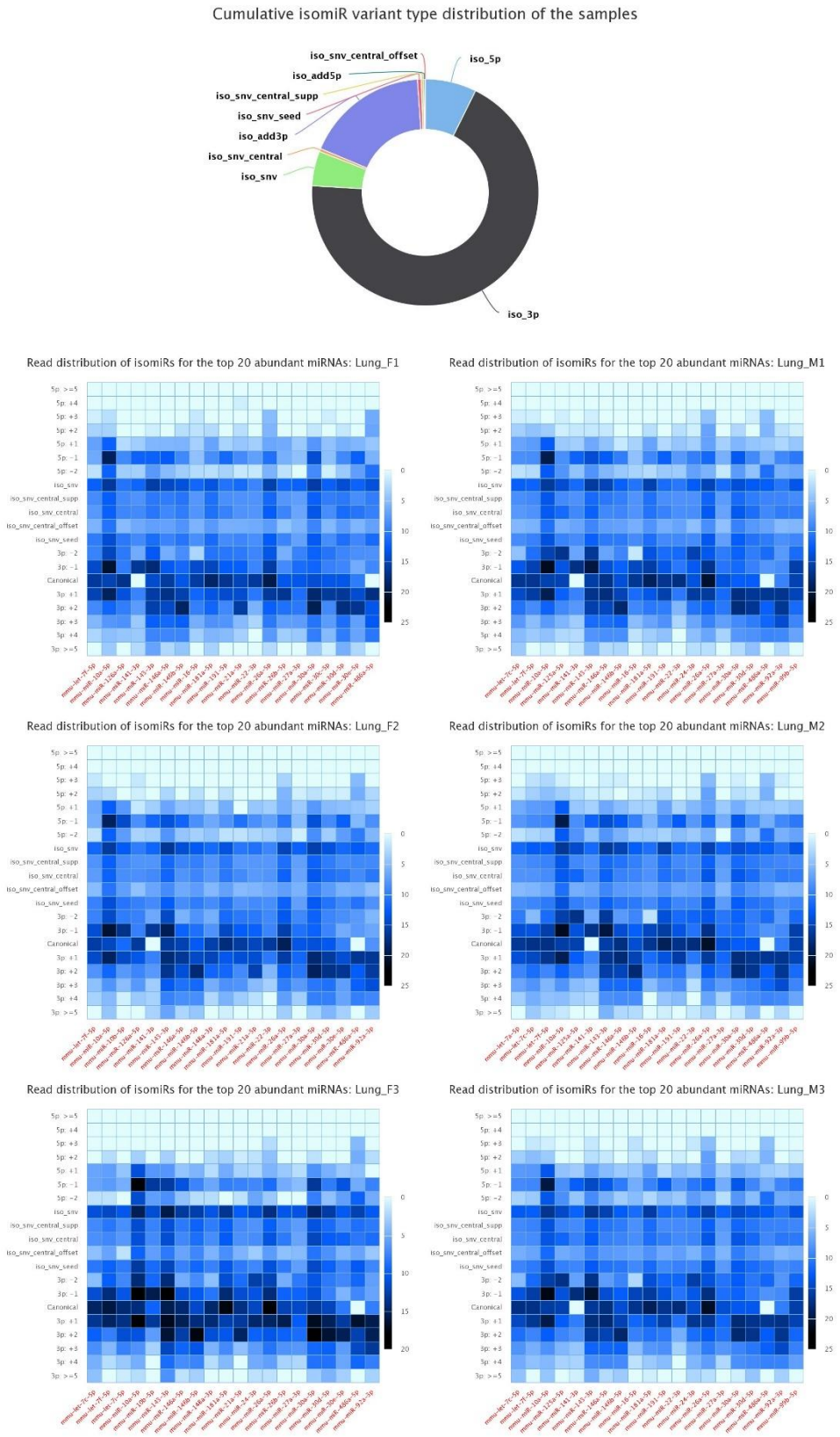

Figura 15. Proporción de *los distintos tipos de variantes isomirs y distribución de las lecturas a lo largo de* las regiones donde se producen las variantes.

Finalmente, a partir de las lecturas que no han alineado frente a ninguna base de *sncRNAs* en ninguna de las etapas de alineamiento, el *pipeline* realiza un análisis basado en *supporting vector machine* (SVM) para identificar posibles nuevos *miRNAs* no descritos (*novel miRNAs*). Los

resultados de este análisis aparecen en la Tabla 5, donde se indica la secuencia del *miRNA*, su posición en el genoma así como un valor de probabilidad y de contajes. De todos los identificados se puede destacar el localizado en el cromosoma 14. Este posible *miRNA* se ha identificado en todas las muestras excepto en una de las muestras de hembra (Lung_F1).

id	Name	Probability	Chr	Start pos.	End Pos.	Mature *miRNA* sequence	*miRNA* read Count
1	Lung_F1_novel_miRNA_1	0.88	chr7	19327284	19327305	ACCGAUCCCGGGUUAGUCUCCU	14
2	Lung_F2_novel_miRNA_1	0.82	chr14	31128290	31128309	CUUAACCUGAAUUUCUGAGC	13
3	Lung_F3_novel_miRNA_1	0.99	chr14	31128290	31128309	CUUAACCUGAAUUUCUGAGC	16
4	Lung_M1_novel_miRNA_1	0.99	chr12	110663149	110663169	AUUCCAAAUGUCCUGCUUUCU	14
5	Lung_M1_novel_miRNA_2	0.82	chr14	31128290	31128309	CUUAACCUGAAUUUCUGAGC	10
6	Lung_M2_novel_miRNA_1	0.99	chr14	31128290	31128309	CUUAACCUGAAUUUCUGAGC	11
7	Lung_M3_novel_miRNA_1	0.99	chr1	55449415	55449434	UUGGUACUGAGGGAAUUAGA	13
8	Lung_M3_novel_miRNA_2	0.97	chr6	90772958	90772978	ACCCUGGACUGUCUACAAAUA	11
9	Lung_M3_novel_miRNA_3	0.91	chr7	19327284	19327305	ACCGAUCCCGGGUUAGUCUCCU	12
10	Lung_M3_novel_miRNA_4	0.82	chr14	31128290	31128309	CUUAACCUGAAUUUCUGAGC	11

Tabla 5. Listado de *novel miRNAs* identificados por miRge3.0.

5.2.2 Análisis mediante COMPSRA

En el análisis de los datos mediante la ejecución del pipeline COMPSRA se han ejecutado los módulos de control de calidad, alineamiento y anotación. El módulo de función no se ha ejecutado ya que en este trabajo se ha optado por realizar el análisis de expresión diferencial de manera independiente con un mismo *script* para todos los *pipelines* ejecutados. Los ficheros de resultados obtenidos de la ejecución del *pipeline* COMPSRA se pueden consultar al completo en la carpeta *"Pipelines"*, en la dirección: https://github.com/asroco/TFM.

El *pipeline* COMPSRA lleva a cabo el alineamiento de las lecturas frente a un genoma de referencia. Las métricas del alineamiento de lecturas de cada una de las muestras se recogen en la Tabla 6. Los valores de *Total input reads* superiores al 99% nos indican que los ficheros de lectura de entrada, que previamente han sido procesados, apenas han sido modificados en cuanto a la eliminación de secuencias o bases de baja calidad por parte del *pipeline*. En este caso el alineamiento de las lecturas se lleva a cabo frente a un genoma de referencia. El porcentaje total de lecturas alineadas, sumando aquellas con una única localización (*unique*) o varias (*multiple*) es en todas las muestras superior al 90%, siendo el número de lecturas no alineadas en todas las muestras por tanto inferior al 10%. No obstante, hay que destacar las métricas obtenidas para la muestra 3 de hembra (Lung_F3) donde el porcentaje de lecturas alineadas en una única localización es muy bajo respecto al resto de muestras (27.56%) y el de múltiples muy alto (67.85%). Al contrario que en el resto de muestras. El resultado para esta muestra está relacionado con el resultado obtenido en el análisis de calidad inicial de las muestras y con el obtenido en las métricas de alineamiento de miRge3.0, donde hay un alto porcentaje de lecturas alineadas frente a *tRNAs* maduros.

Sample	Total processed reads	Total input reads	Uniquely mapped reads	% Uniquely mapped reads	Multiple mapped reads	% Multiple mapped reads	% of reads unmapped
Lung_F1	20821100	20779213 99.79%	11439149	55.05	7558106	36.37	8.57
Lung_F2	18028815	17971780 99.68%	11125510	61.91	5625506	31.30	6.79
Lung_F3	20189824	20167472 99.88%	5558061	27.56	13683405	67.85	4.59
Lung_M1	25545228	25518199 99.89%	16737106	65.59	7411893	29.05	5.37
Lung_M2	28758079	28723260 98.87%	18031315	62.78	8839093	30.77	6.45
Lung_M3	34333318	34281857 99.85%	22665167	66.11	10019907	29.23	4.65

Tabla 6. Métricas del alineamiento de lecturas en COMPSRA.

En la tabla 7 aparecen las métricas resultado de la anotación llevada a cabo por COMPSRA a partir de sus bases de anotación específicas. Con el número de elementos únicos identificados para cada uno de los tipos de *sncRNAs* incluidos en el análisis y el número total de lecturas que soportan la anotación en el fichero *.bam*.

	miRNA		tRNA		snoRNA		snRNA		circRNA	
Sample	**Items**	**Reads**	**Items**	**Reads**	**Items**	**Reads**	**Items**	**Reads**	**Items**	**Reads**
Lung_F1	1343	13934762	409	21776408	4605	1254	498	183849	11533	1206038
Lung_F2	1344	12151488	406	22120317	1290	4196795	636	143595	11547	1226477
Lung_F3	1123	6445759	416	97889723	1095	2167630	408	47793	9449	703931
Lung_M1	1419	18671603	413	10124545	1251	25198489	525	51051	9977	1508108
Lung_M2	1443	20667609	415	15652509	1305	15652509	546	66459	11104	1763871
Lung_M3	1481	25037018	418	19533899	1458	29885496	636	100236	11575	2056392

Tabla 7. Métricas de la anotación en COMPSRA.

Para cada una de las muestras analizadas COMPSRA proporciona los ficheros de reporte en formato *.txt* de control de calidad, alineamiento y anotación a partir de los cuales obtener la información con la que se han construido las tablas de métricas (Tabla 6 y 7).

El *pipeline* también proporciona el resultado del alineamiento de lecturas de cada muestra en forma de fichero *.bam* para su visualización.

Finalmente, para cada uno de los tipos de *sncRNAs* analizados proporciona un fichero de contajes. Estos ficheros pueden ser utilizados para llevar a cabo análisis de expresión diferencial. Se ha utilizado el fichero de contajes de *miRNAs*, con un total de 1830 elementos, para llevar a cabo el análisis de expresión diferencial mediante un *script* de R del mismo modo que para los otros *pipelines*.

5.2.3 Análisis mediante Nf-core/smrnaseq

Los ficheros de resultados obtenidos de la ejecución del *pipeline* nf-core/smrnaseq se pueden consultar al completo en la carpeta *"Pipelines"* que puede encontrarse en la dirección: https://github.com/asroco/TFM.

Este *pipeline* ejecuta un paso de control de calidad y procesado de muestras que no se puede omitir. Los resultados del análisis se presentan en un fichero de reporte MultiQC. Sus resultados indican que apenas se han producido modificaciones en cuanto a la eliminación de secuencias o bases de baja calidad por parte del *pipeline* con respecto a los ficheros procesados que se han utilizado para ejecutar los *pipelines*. Las métricas de calidad que se obtienen son muy similares a las obtenidas previamente en el análisis inicial (Tabla 8). Y el porcentaje de bases de baja calidad recortadas (*% BP Trimmed*) se sitúa en torno a un 3% (Tabla 8, Figura 16B). Además, no se ha eliminado ninguna de las lecturas en ninguna de las muestras (Figura 16A)

Sample Name	% BP Trimmed	% Dups	% GC	Length	M Seqs
Lung_F1_*	3.3%	93.3%	46%	28 bp	20.7
Lung_F2_*	3.6%	92.6%	47%	28 bp	17.9
Lung_F3_*	3.0%	97.0%	48%	30 bp	20.1
Lung_M1_*	2.8%	97.6%	45%	38 bp	25.5
Lung_M2_*	2.7%	97.2%	46%	37 bp	28.7
Lung_M3_*	2.9%	95.7%	45%	36 bp	34.2

Tabla 8. Métricas calidad de nf-core/smrnaseq.

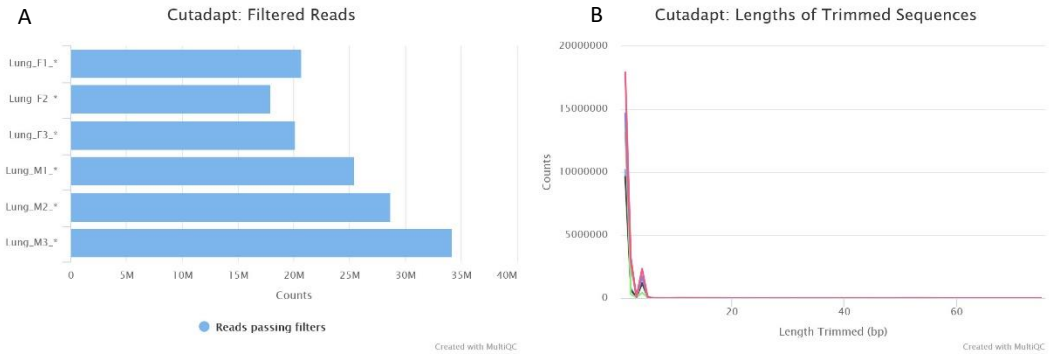

Figura 16. Procesado de secuencias mediante Cutadapt en el *pipeline* nf-core/smrnaseq. A: Número de lecturas filtradas; **B:** Número de bases recortadas.

Las métricas del alineamiento de lecturas aparecen recogidas en la Tabla 9. En este *pipeline* el alineamiento de las lecturas se realiza en primer lugar frente bases de datos de *miRNAs* maduros y precursores de manera secuencial. Estos alineamientos permiten la identificación y cuantificación de *miRNAs*. El porcentaje de lecturas alineadas se sitúa en torno a un 70% del total de lecturas si se suman ambas etapas. Esto ocurre para todas las muestras excepto para la muestra 3 de hembra cuyo porcentaje es de tan solo un 31%. Al igual que con los otros *pipelines* se obtienen valores anómalos para esta muestra respecto al resto de ellas.

Posteriormente se realiza un alineamiento de las lecturas frente a un genoma de referencia, aunque en este caso no para la identificación y cuantificación de *miRNAs* si no como control de calidad de las secuencias. En ese sentido, se observa que para todas las muestras el porcentaje de lecturas alineadas es superior a un 97% del total de secuencias utilizadas.

Sample Name	M Reads Mapped	Error rate	% Mapped	M Total seqs
Lung_F1_*.genome	58.0	1.21%	97.1%	59.8
Lung_F1_*.hairpin	3.9	2.75%	28.9%	13.5
Lung_F1_*.mature	8.9	0.34%	41.7%	21.5
Lung_F2_*.genome	67.6	0.83%	98.2%	68.8
Lung_F2_*.hairpin	3.1	2.71%	28.9%	10.8
Lung_F2_*.mature	8.2	0.33%	44.6%	18.4
Lung_F3_*.genome	126.2	0.89%	98.9%	127.6
Lung_F3_*.hairpin	1.9	2.88%	11.3%	16.7
Lung_F3_*.mature	4.2	0.34%	20.4%	20.4
Lung_M1_*.genome	51.7	0.95%	97.4%	53.1
Lung_M1_*.hairpin	2.8	3.04%	20.6%	13.5
Lung_M1_*.mature	13.2	0.21%	50.6%	26.0
Lung_M2_*.genome	62.4	0.98%	97.1%	64.3
Lung_M2_*.hairpin	3.1	3.01%	19.8%	15.6
Lung_M2_*.mature	14.4	0.21%	49.1%	29.3
Lung_M3_*.genome	75.8	0.97%	97.9%	77.5
Lung_M3_*.hairpin	3.6	3.03%	20.2%	17.8
Lung_M3_*.mature	17.8	0.21%	51.1%	34.9

Tabla 9. Métricas de alineamiento de nf-core/smrnaseq.

Los resultados recogidos en la Tabla 9 también se muestran de manera gráfica en la Figura 17, donde se representan las proporciones de lecturas alineadas y no alineadas en cada una de las tres etapas de alineamiento del *pipeline*.

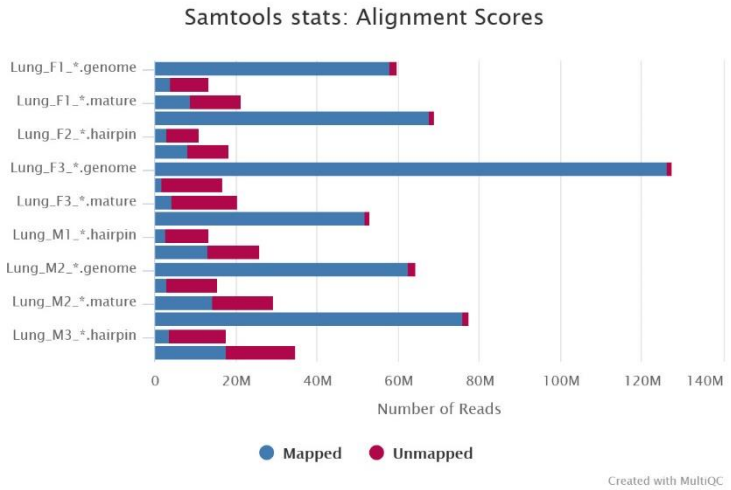

Figura 17. Proporción de lecturas alineadas en cada una de las etapas de alineamiento del *pipeline* de nf-core/smrnaseq.

Los resultados del alineamiento de las lecturas en cada una de las etapas son proporcionados también por el pipeline en ficheros en formato *.bam* que permiten su visualización. Estos ficheros además son utilizados para generar ficheros *.gff3* con las anotaciones de los *miRNAs* e *isomirs* identificados mediante mirTOP (https://github.com/miRTop/mirtop).

El *pipeline* de nf-core/smrnaseq lleva a cabo también un análisis de calidad específico para datos de *sRNAseq* mediante miRTrace (49). En este análisis se evalúa por ejemplo la calidad de las secuencias, la cantidad y tipos de *sncRNAs* detectados o la presencia de secuencias no deseadas como pueden ser los casos de contaminación con otras especies o artefactos. Los resultados de este análisis permiten comprobar que los adaptadores de las secuencias de todas las muestras se han eliminado correctamente (Figura 18A), la proporción de lecturas que corresponden a *miRNAs* detectado en las muestras está en torno al 48% (Figura 18B) y no parece existir problemas de contaminación ya que el 100% de los *miRNAs* detectados pertenece al orden Rodentia, al cual pertenece la especie ratón (Figura 18C).

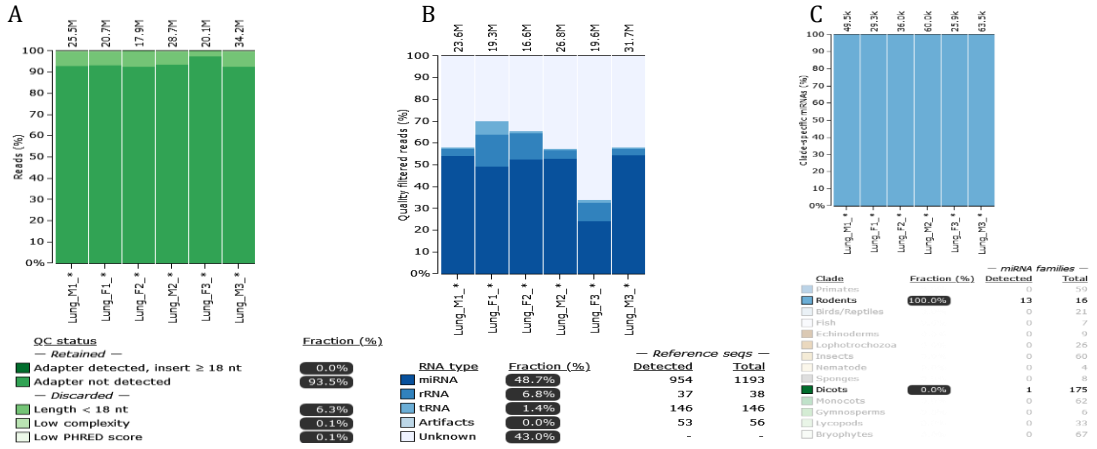

Figura 18. Resultados de miRTrace. **A:** Control de calidad de las secuencias; **B:** Tipos de *sncRNA* detectados; **C:** Porcentaje de lecturas de *miRNAs* específicas de cada clado.

El *pipeline* puede ejecutar un análisis de expresión diferencial mediante *edgeR*. No obstante, al igual que en los otros *pipelines*, en este trabajo el análisis de expresión diferencial a partir del fichero de contajes que genera el *pipeline* se ha llevado a cabo mediante el *script* de R utilizado para los otros *pipelines*.

5.3 Análisis de expresión diferencial

Si se tienen en cuenta los ficheros de contaje para *miRNAs* obtenidos con cada *pipeline* existen diferencias en el número total de *miRNAs* identificados: 1927 por miRge 3.0, 1830 por COMPSRA, y 1389 por nf-core/smrnaseq. Las diferencias no son muy grandes en cuanto al número, siendo el de nf-core/smrnaseq sensiblemente inferior. Pero se puede comprobar que sólo 1090, un 41% del total son comunes a los tres *pipelines*, mientras que hasta un 45% del total son *miRNAs* que sólo aparecen identificados por uno de los *pipelines* (643 en COMPSRA, 556 en miRge 3.0 y 9 en nf/core smrnaseq) (Figura 19A). Estos números varían si añadimos el fichero de contajes publicado por Isakova *et al.* 2020 (Figura 19B). El número de elementos en común se reduce a 715. No obstante, el número de elementos que aparecen en un solo *pipeline* se reduce en el caso de miRge3.0 a sólo 25 y en el caso de nf-core/smrnaseq no aparece ninguno. Sólo COMPSRA se mantiene con un número de elementos exclusivos elevado, con 551.

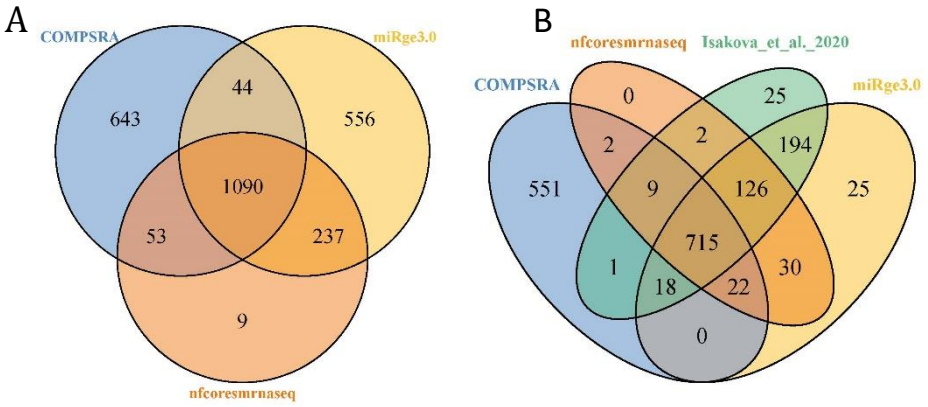

Figura 19. Diagrama de Venn de *miRNAs* identificados y cuantificados de acuerdo con los ficheros de contaje. **A:** *pipelines* ejecutados en este trabajo; **B:** *pipelines* junto con el fichero obtenido a partir de Isakova *et al.* 2020

El análisis de expresión diferencial se ha realizado para *miRNAs*, ya que son el tipo de *sncRNAs* para el cual los tres *pipelines* y el artículo original proporcionan una matriz de contajes. Si bien es cierto que algunos *pipelines* como COMPSRA o el utilizado por Isakova *et al.* 2020 proporcionan matriz de contajes para otros tipos de *sncRNAs* y se podría realizar un análisis similar. El análisis se ha realizado utilizando el mismo *script* en R para todos los ficheros, que puede consultarse en el fichero *"Differential Expression Analysis.rmd"* que se encuentra guardado en la carpeta "Análisis de expresión diferencial" en la siguiente dirección: https://github.com/asroco/TFM.

La distribución en el *boxplot* de los valores sin normalizar de los datos obtenidos con cada *pipeline* muestra un perfil similar para todas las muestras (Figura 20), sin ningún problema aparente que impida normalizar los datos y utilizarlos en el análisis. Una vez normalizados los valores, de nuevo se puede observar mediante *boxplot* (Figura 21) que la normalización ha sido correcta y la distribución de los valores en todas las muestras tiene un perfil similar para los cuatro *pipelines*.

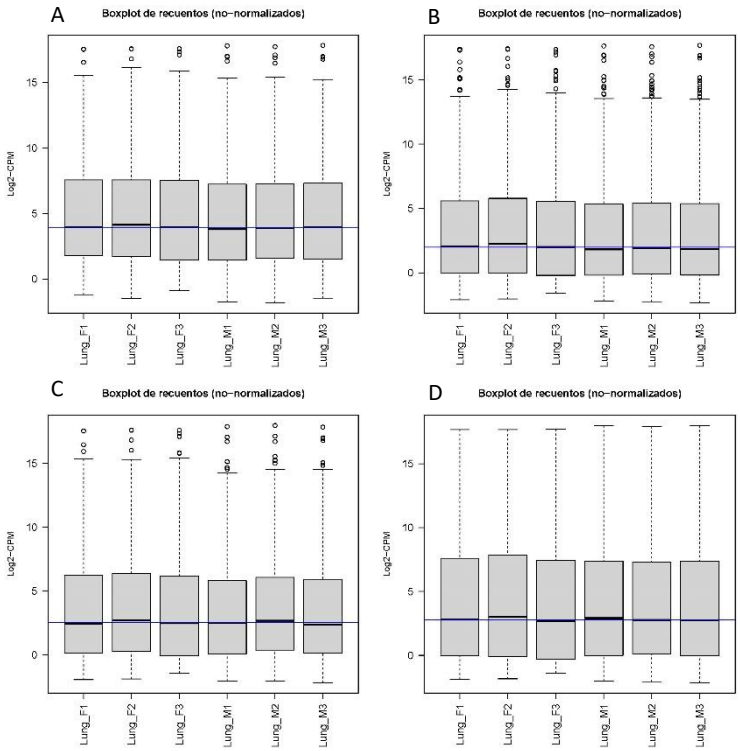

Figura 20. Exploración datos no normalizados. **A:** miRge3.0; **B:** COMPSRA; **C:** nf-core/smrnaseq; **D:** Isakova *et al.* 2020.

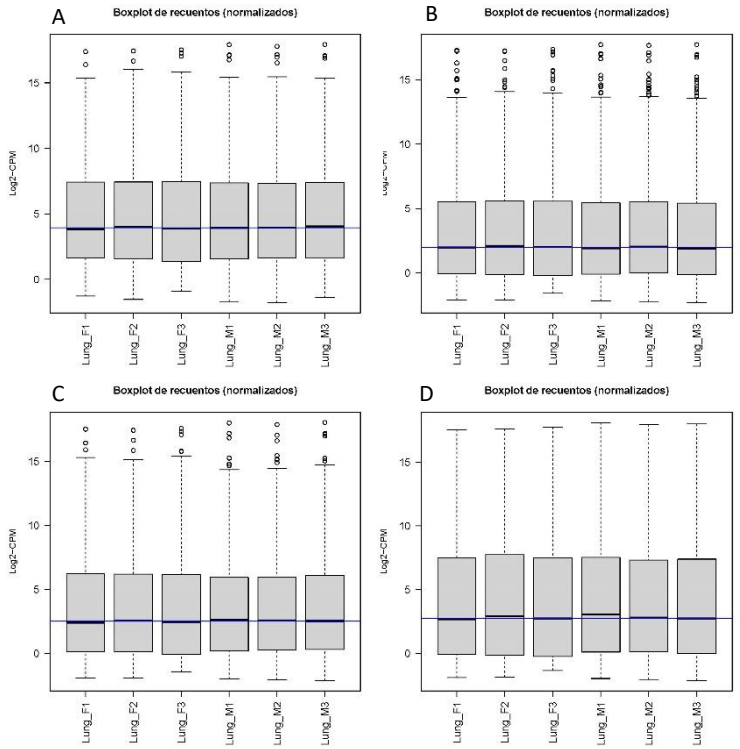

Figura 21. Exploración datos normalizados. **A:** miRge3.0; **B:** COMPSRA; **C:** nf-core/smrnaseq; **D:** Isakova *et al.* 2020.

Del análisis no supervisado de similitud entre las muestras mediante el cálculo de la matriz de distancias podemos deducir que el agrupamiento de las muestras utilizadas en el análisis es correcto para los cuatro *pipelines*. Como podemos observar tanto en los *heatmap* de la matriz

de distancias (Figura 22) como en los *clúster* jerárquicos (Figura 23), las muestras se agrupan de acuerdo al sexo en los cuatro casos tal y como se esperaba.

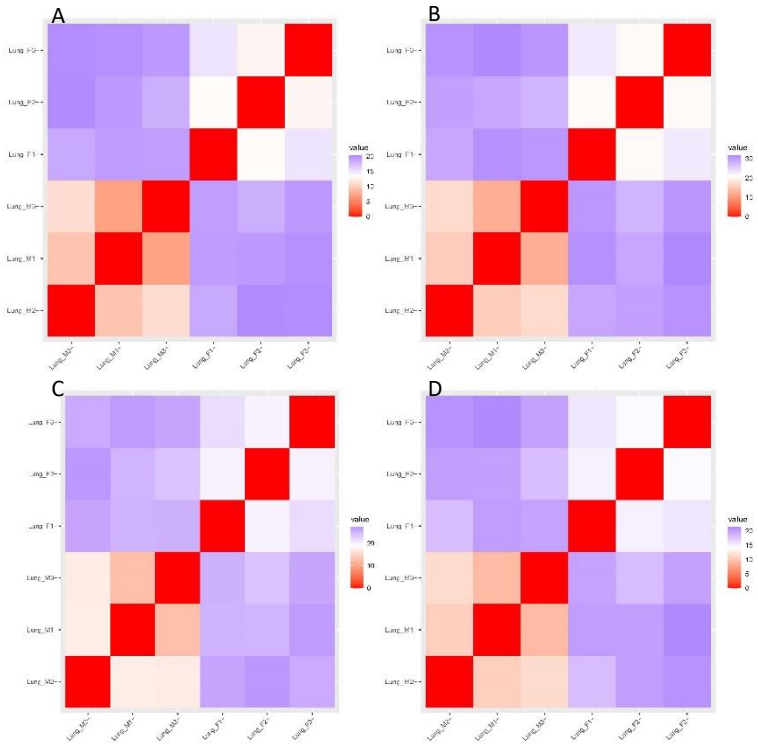

Figura 22. *Heatmap* matriz de distancias. **A:** miRge3.0; **B:** COMPSRA; **C:** nf-core/smrnaseq; **D:** Isakova *et al.* 2020.

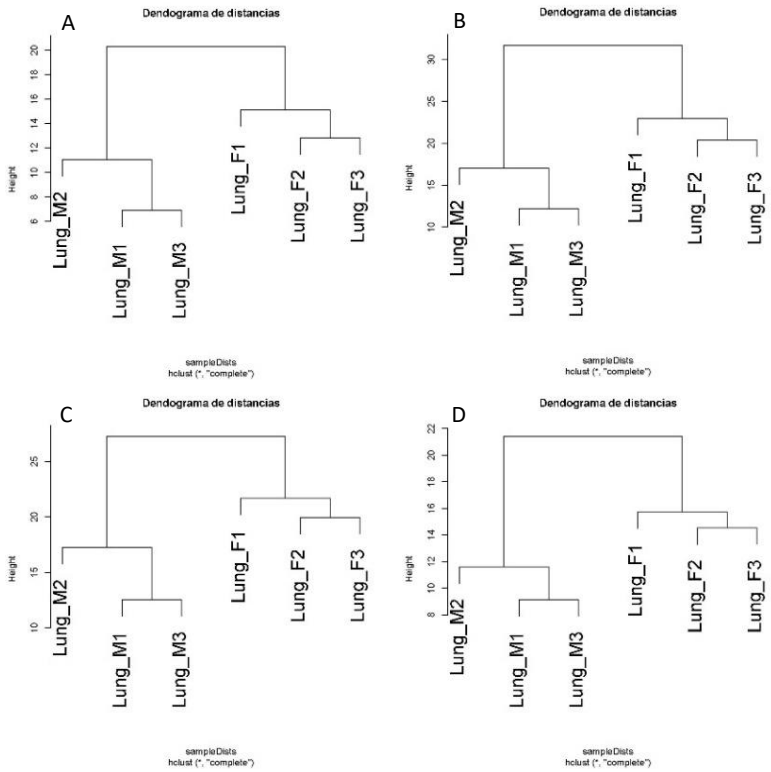

Figura 23. *Clúster* jerárquico. **A:** miRge3.0; **B:** COMPSRA; **C:** nf-core/smrnaseq; **D:** Isakova *et al.* 2020.

Las gráficas de visualización en dimensión reducida (Figura 24) confirman los resultados anteriores de similitud y agrupamiento. En los cuatro gráficos se observa cómo en la primera componente se produce una buena separación en función de los grupos de muestras que consideramos (macho y hembra). También podemos observar en todos los casos que las réplicas de macho son más similares entre sí que las de hembra.

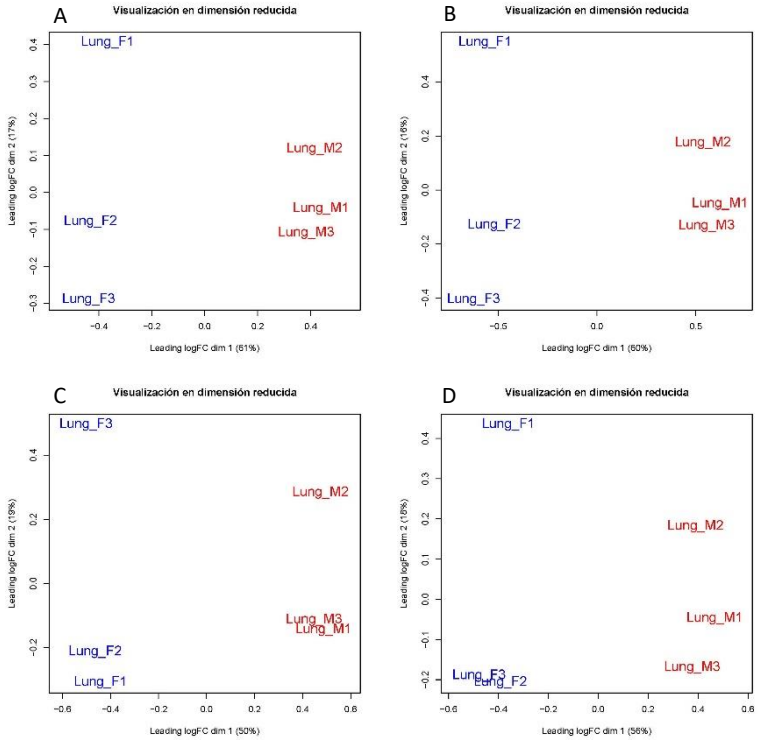

Figura 24. Visualización en dimensión reducida. **A:** miRge3.0; **B:** COMPSRA; **C:** nf-core/smrnaseq; **D:** Isakova *et al.* 2020.

Así pues, del análisis exploratorio de los datos realizado podemos concluir en los cuatro casos que la calidad de estos, de acuerdo con las distribuciones de los *boxplot*, es la correcta para realizar un análisis de expresión diferencial. La normalización de los datos también ha sido correcta. Y del análisis de similitud y agrupamiento concluimos que las muestras se agrupan correctamente en función del sexo, siendo las muestras de macho más parecidas entre sí que las muestras de hembra.

5.3.1 Correlación de los perfiles de expresión de *miRNAs*

Los perfiles de expresión de *miRNAs* que se obtienen con cada uno de los *pipelines* es muy parecido para todas las muestras (macho y hembra) cuando se comparan los pipelines entre sí (Figuras 25 y 26). En todas las comparaciones y para todas las muestras los datos se ajustan bastante bien a una recta. Y como podemos observar, los valores de correlación en todas y cada una de las comparaciones son muy altos, entre 0.98 y 0.99. Las mejores correlaciones se obtienen cuando se comparan COMPSRA con miRge3.0 (0.99) y éste con nf-core/smrnaseq (0.99). No obstante, la comparación COMPSRA con nf-core/smrnaseq también es muy alta (0.98). De ello podemos deducir que, independientemente de que podamos esperar que la lista de *miRNAs* identificados y cuantificados por cada *pipeline* sea diferente, la comparación de *miRNAs* comunes nos indica que la cuantificación llevada a cabo por los tres *pipelines* a partir del mapeo de lecturas es muy similar.

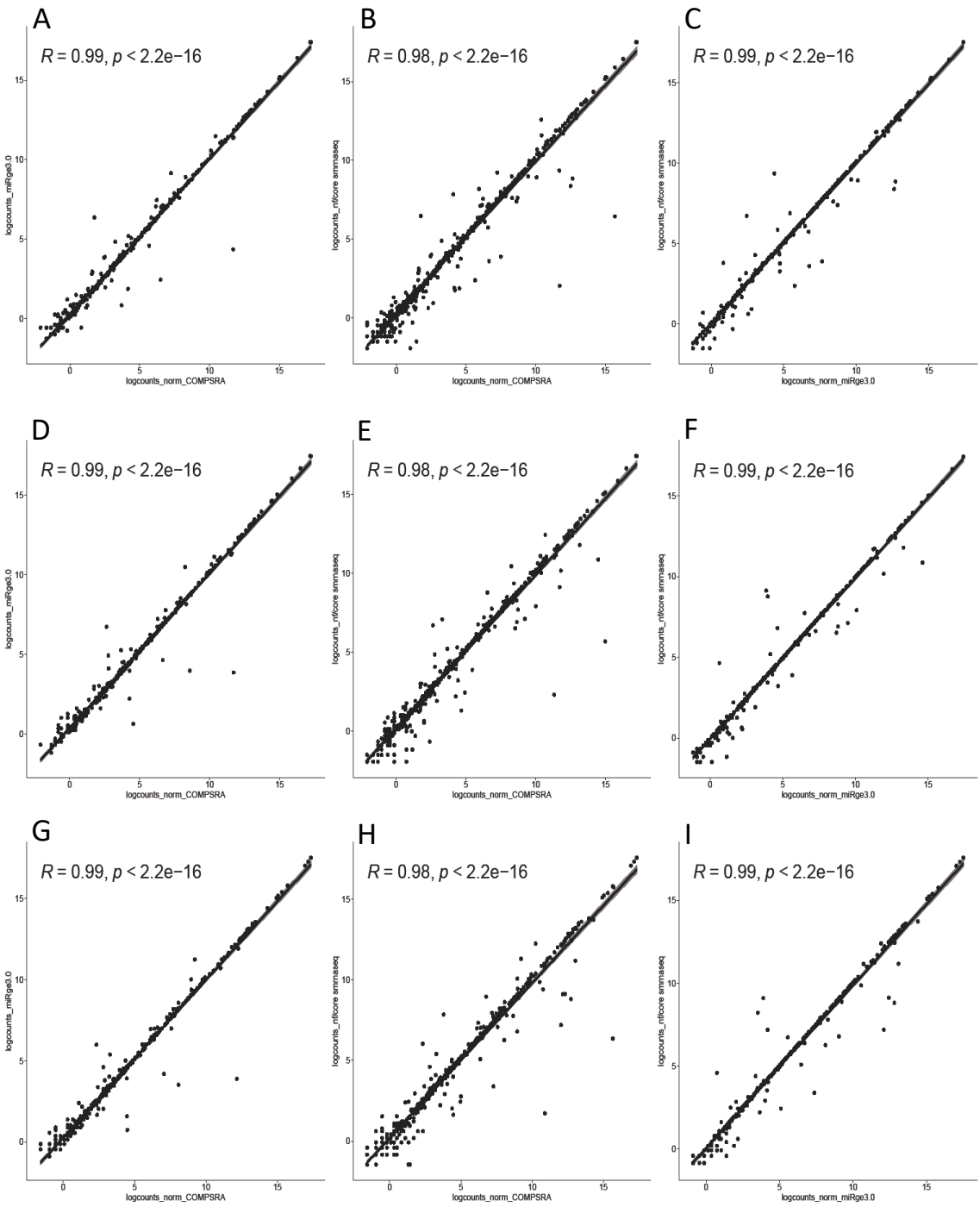

Figura 25. Correlación de las expresiones de *miRNAs* entre los tres *pipelines* utilizados en las muestras de pulmón de hembra.
A-B-C: LungF1, **D-E-F:** LungF2, **G-H-I:** LungF3.
A-D-G: COMPSRA frente a miRge3.0; **B-E-H:** COMPSRA frente a nf-core/smrnaseq; **C-F-I:** miRge3.0 frente a nf-core/smrnaseq.

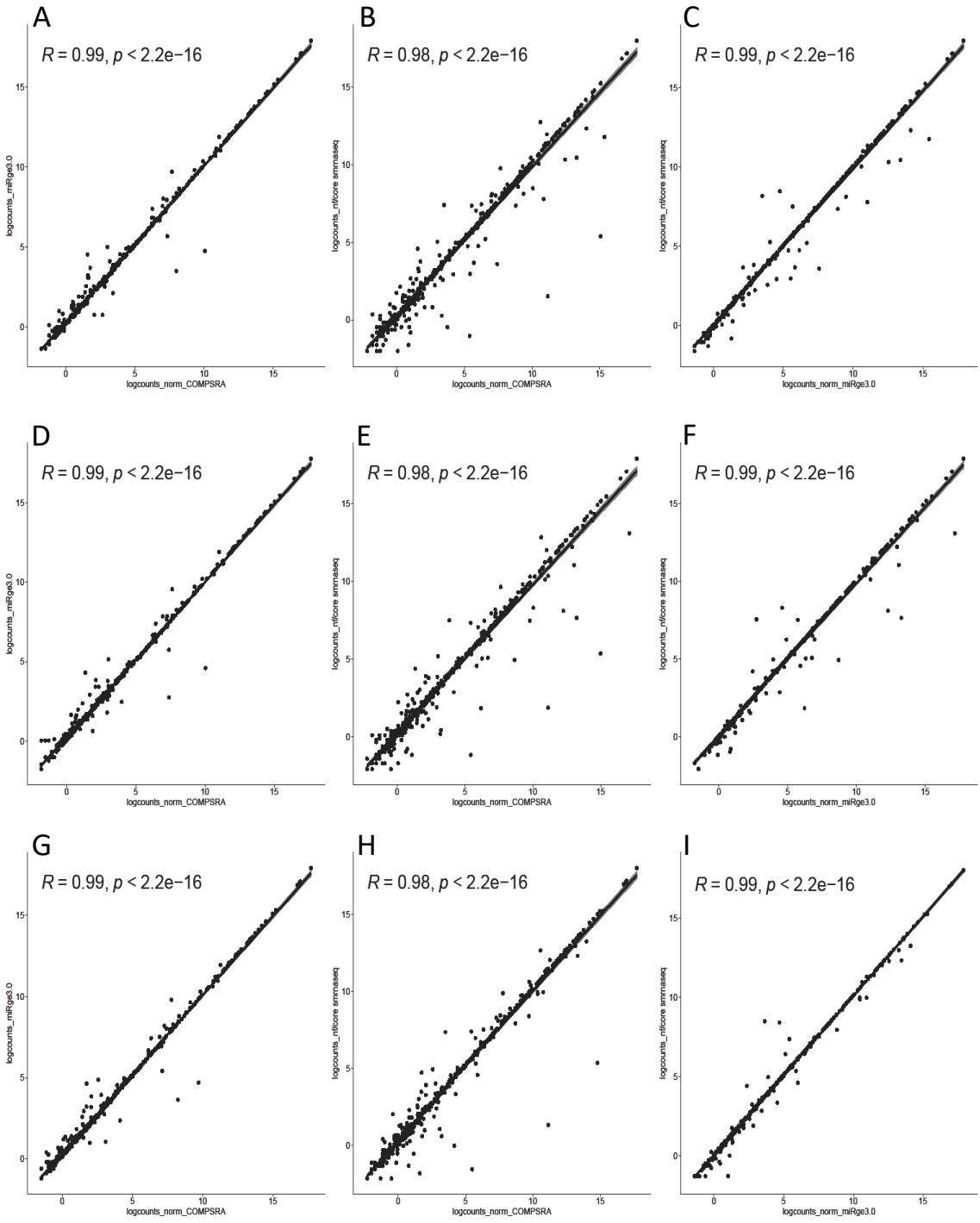

Figura 26. Correlación de las expresiones de *miRNAs* entre los tres *pipelines* utilizados en las muestras de pulmón de macho.
A-B-C: LungM1, **D-E-F**: LungM2, **G-H-I:** LungM3.
A-D-G: COMPSRA frente a miRge3.0; **B-E-H:** COMPSRA frente a nf-core/smrnaseq; **C-F-I:** miRge3.0 frente a nf-core/smrnaseq.

5.3.2 Análisis de *miRNAs* diferencialmente expresados

El análisis de expresión diferencial se ha realizado mediante el paquete *DESEQ2*. Se han identificado un total de 417 *miRNAs* diferencialmente expresados a partir de los datos proporcionados por miRge3.0, 733 por COMPSRA, 528 por nf-core/smrnaseq y 355 por los datos de Isakova *et al.* 2020 (Tabla 10). Las cuatro listas de *miRNAs* diferencialmente expresados se han representado gráficamente mediante *Volcano plots* (Figura 27). Aquellos *miRNAs* con expresión diferencial con un valor absoluto de log_2FC mayor de 1 y un *p* valor adjunto menor de 0.05 aparecen identificados como los puntos marcados en rojo. A la vista de los cuatro gráficos *a priori* no se podría decir que el número de éstos sea mayor en uno u otro caso.

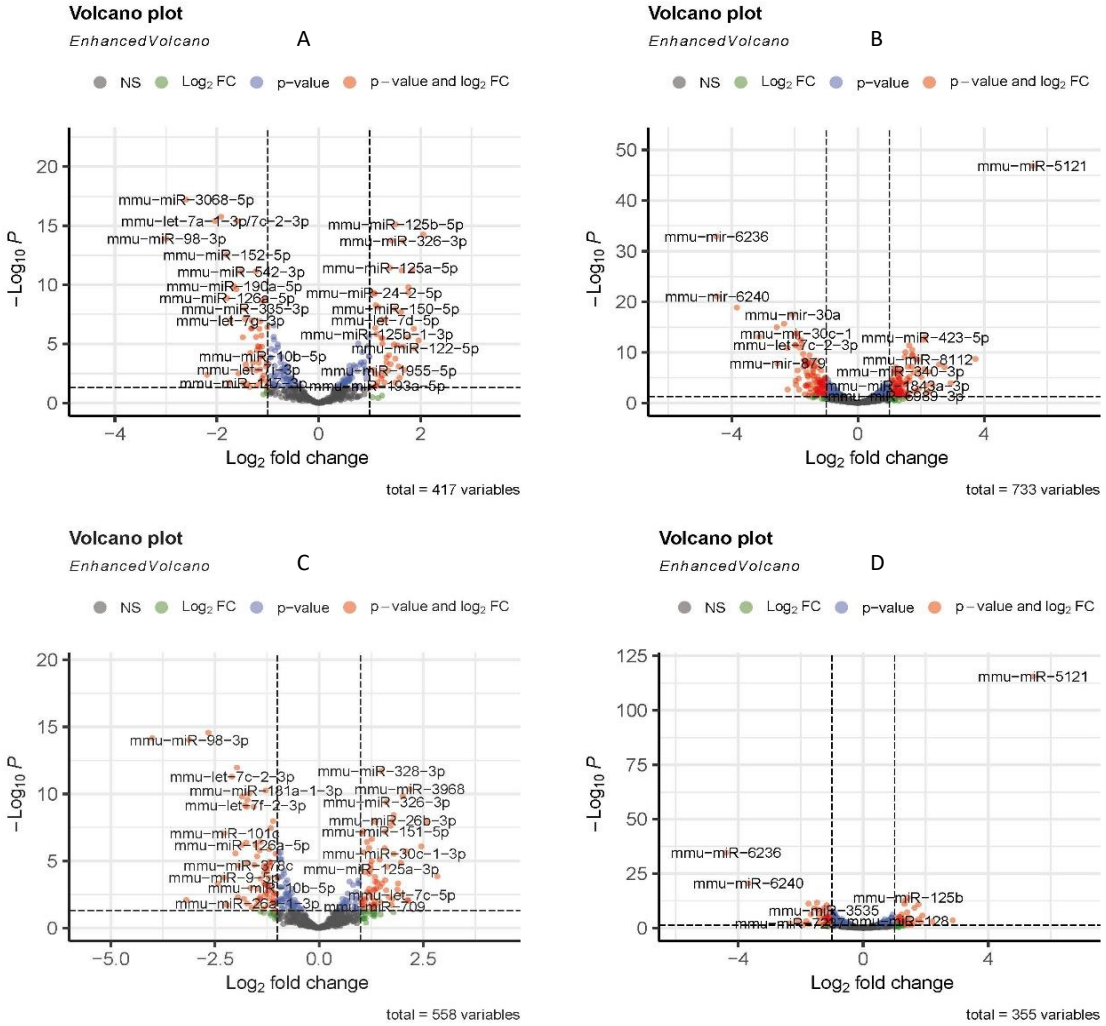

Figura 27. Volcano plot. **A:** miRge3.0; **B:** COMPSRA; **C:** nf-core/smrnaseq; **D:** Isakova *et al.* 2020.

A modo de resumen, el número de *miRNAs* diferencialmente expresados con un valor absoluto de log_2FC mayor de 1 y un *padj* menor de 0.05 que se obtienen a partir de los datos proporcionados por cada uno de los *pipelines* son los siguientes (Tabla 10):

Pipeline	**Total**	***Upregulated***	***Downregulated***	**No significativos**
miRge3.0	417	46	47	324
COMPSRA	733	77	93	563
nf-core/smrnaseq	558	56	65	437
Isakova *et al.* 2020	355	26	33	296

Tabla 10. *miRNAs* diferencialmente expresados *DESEQ2*: log_2FC absoluto mayor de 1 y *padj* menor de 0.05.

El *pipeline* con el cual se identifican un mayor número de *miRNAs* diferencialmente expresados mediante *DESEQ2* es COMPSRA con 733, seguido de nf-core/smrnaseq con 528 y miRge3.0 con 417. A partir de los datos de Isakova *et al.* 2020 se identifica el menor número con 355. En cuanto a aquellos que se han considerado significativos (log_2FC absoluto mayor de 1 y *padj* menor de 0.05), el número de *upregulated* y *downregulated* es bastante similar para un mismo *pipeline*. Y de nuevo el número total es mayor en COMPSRA con 170, seguido de nf-core/smrnaseq con 111, miRge3.0 con 93 y finalmente Isakova *et al.* 2020 con 59.

Los perfiles de expresión de estos *miRNAs* seleccionados se han representado mediante su correspondiente *heatmap* jerárquico (Figura 28). En estos *heatmaps*, aunque las listas de *miRNAs* no son exactamente iguales en cuanto a número y miembros, se pueden identificar perfiles de expresión claramente específicos de sexo de acuerdo con el agrupamiento jerárquico de las muestras de acuerdo a los grupos comparados (macho y hembra) en todos los casos.

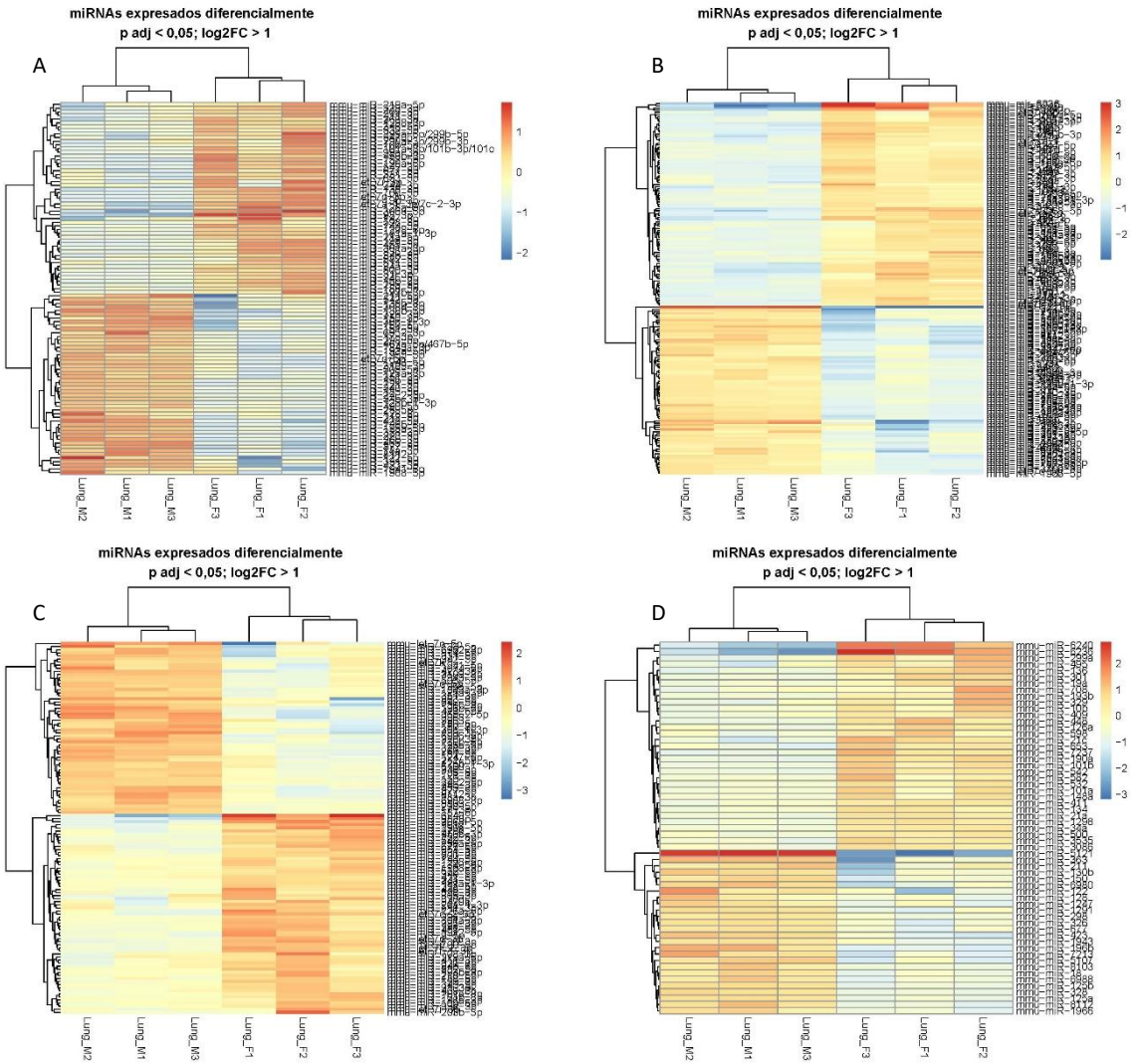

Figura 28. *Heatmap* jerárquico. **A:** miRge3.0; **B:** COMPSRA; **C:** nf-core/smrnaseq; **D:** Isakova *et al.* 2020.

Como se ha mencionado y se puede deducir también a partir de la (Figura 29A), las listas de *miRNAs* diferencialmente expresados y filtrados como significativos que se han obtenido a partir de cada uno de los *pipelines* es diferente. Cuando se comparan los tres *pipelines* ejecutados en este trabajo se observa que el número de *miRNAs* en común entre los tres *pipelines* es de 72. La lista de *miRNAs* puede consultarse en la Tabla 11. COMPSRA es el que tiene

más *miRNAs* exclusivos, con un total de 69, más incluso que en común con los otros dos *pipelines*. Mientras que el número de *miRNAs* exclusivos de los otros dos *pipelines* es mucho menor, con 7 para miRge3.0 y 14 para nf-core/smrnaseq.

Cuando incluimos la lista de *miRNAs* obtenida a partir de los datos de la publicación de Isakova *et al.* 2020, el número de elementos en común a los cuatro *pipelines* se reduce con un total de 27 (Figura 29B). Esta lista se puede consultar en la Tabla 12**.** Y de nuevo, COMPSRA sigue siendo el *pipeline* con el mayor número de *miRNAs* exclusivo con 54, mientras que para los otros *pipelines* el número es mucho menor.

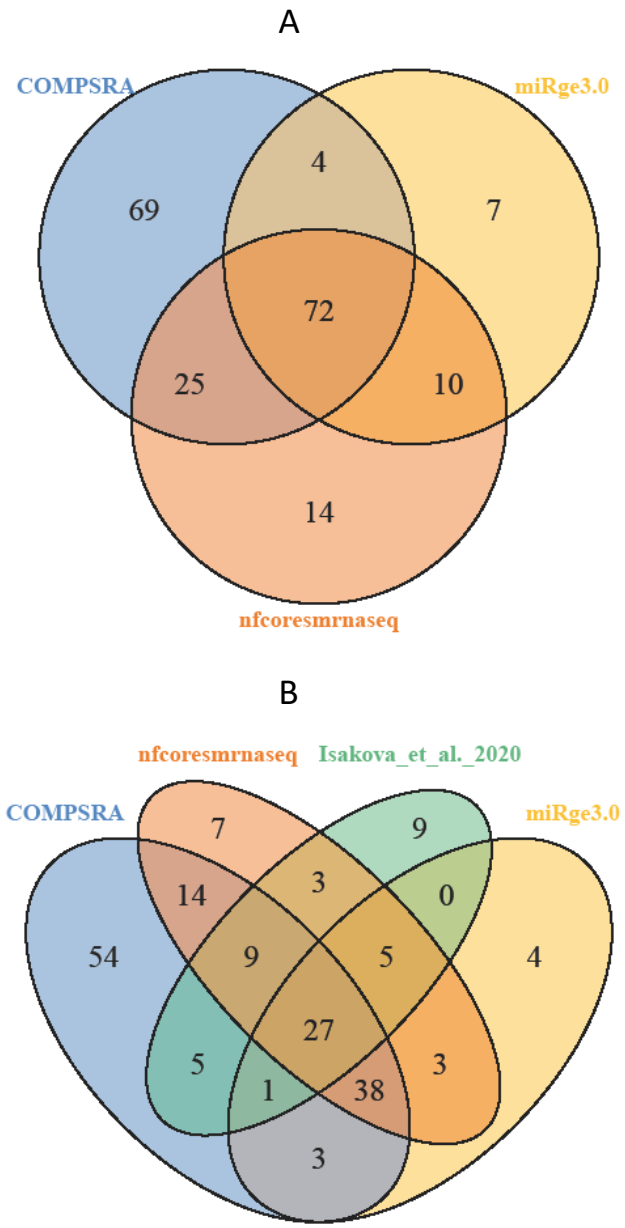

Figura 21. Diagramas de Venn de las listas de *miRNAs* diferencialmente expresados y filtrados como significativos. **A:** *pipelines* ejecutados en este trabajo; **B:** *pipelines* junto con la lista obtenida a partir de los datos de Isakova *et al.* 2020.

miRNAs			
mmu-let-7d-5p	mmu-miR-15b-3p	mmu-miR-29a-5p	mmu-miR-423-5p
mmu-let-7g-3p	mmu-miR-15b-5p	mmu-miR-301a-3p	mmu-miR-450b-3p
mmu-let-7i-3p	mmu-miR-17-3p	mmu-miR-3068-5p	mmu-miR-455-3p
mmu-miR-122-5p	mmu-miR-181a-1-3p	mmu-miR-30c-1-3p	mmu-miR-455-5p
mmu-miR-125a-3p	mmu-miR-1843a-3p	mmu-miR-3109-3p	mmu-miR-500-3p
mmu-miR-125a-5p	mmu-miR-185-5p	mmu-miR-31-3p	mmu-miR-505-5p
mmu-miR-125b-1-3p	mmu-miR-188-5p	mmu-miR-326-3p	mmu-miR-511-3p
mmu-miR-126a-5p	mmu-miR-18a-3p	mmu-miR-328-3p	mmu-miR-532-5p
mmu-miR-128-3p	mmu-miR-190a-5p	mmu-miR-331-3p	mmu-miR-542-3p
mmu-miR-1298-3p	mmu-miR-193b-3p	mmu-miR-331-5p	mmu-miR-542-5p
mmu-miR-1298-5p	mmu-miR-1943-5p	mmu-miR-335-3p	mmu-miR-582-5p
mmu-miR-130b-3p	mmu-miR-1955-5p	mmu-miR-340-3p	mmu-miR-598-3p
mmu-miR-130b-5p	mmu-miR-19a-3p	mmu-miR-345-5p	mmu-miR-6952-3p
mmu-miR-144-5p	mmu-miR-212-5p	mmu-miR-34b-3p	mmu-miR-700-5p
mmu-miR-150-3p	mmu-miR-21a-5p	mmu-miR-34b-5p	mmu-miR-8112
mmu-miR-150-5p	mmu-miR-24-2-5p	mmu-miR-351-3p	mmu-miR-877-5p
mmu-miR-151-5p	mmu-miR-26b-3p	mmu-miR-409-3p	mmu-miR-98-3p
mmu-miR-152-5p	mmu-miR-298-5p	mmu-miR-423-3p	mmu-miR-99a-3p

Tabla 11. *miRNAs* diferencialmente expresados (*log2FC* absoluto mayor de 1 y un *padj* menor de 0.05) comunes a los tres *pipelines* (COMPSRA, miRge3.0 y nf-core/smrnaseq).

miRNAs	
mmu-miR-101a	mmu-miR-298
mmu-miR-122	mmu-miR-299a
mmu-miR-125a	mmu-miR-326
mmu-miR-125b	mmu-miR-328
mmu-miR-126a	mmu-miR-409
mmu-miR-128	mmu-miR-423
mmu-miR-1298	mmu-miR-500
mmu-miR-130b	mmu-miR-532
mmu-miR-150	mmu-miR-542
mmu-miR-190a	mmu-miR-582
mmu-miR-193b	mmu-miR-598
mmu-miR-1943	mmu-miR-677
mmu-miR-19a	mmu-miR-8112
mmu-miR-21a	

Tabla 12. *miRNAs* diferencialmente expresados (*log2FC* absoluto mayor de 1 y un *padj* menor de 0.05) comunes a los tres *pipelines* (COMPSRA, miRge3.0 y nf-core/smrnaseq) e Isakova *et al.* 2020.

6 Discusión

En este trabajo se ha llevado a cabo un análisis de *sncRNAs*, fundamentalmente *miRNAs*, utilizando diferentes *pipelines* y a partir de un mismo *dataset* de *sRNAseq*. Los tres *pipelines* utilizados llevan a cabo una estrategia de alineamiento de lecturas distintas: mientras que COMPSRA lo realiza frente a un genoma de referencia, miRge3.0 y nf-core/smrnaseq siguen la estrategia de alineamiento frente a bases de datos de anotaciones, como por ejemplo miRBase u otras específicas. Por otro lado, los tres *pipelines* tienen una estructura general común (control de calidad, alineamiento y cuantificación) pero difieren en cuanto al tipo de *sncRNA* analizado. COMPSRA permite la identificación y cuantificación de distintos tipos mientras que los otros dos *pipelines* están más enfocados al análisis en exclusiva de *miRNAs*. Y también difieren en el tipo o variedad de análisis llevados a cabo. COMPSRA realiza el más simple, con un control de calidad, alineamiento y cuantificación para obtener el fichero de contajes que puede ser utilizado para análisis de expresión diferencial dentro del propio *pipeline* o posteriormente por el usuario. Por su parte, miRge3.0 y nf-core/smrnaseq además de los anteriores análisis incluyen otros como la descripción de *isomirs* o la identificación de nuevos *miRNAs*. Con respecto a la forma de presentar los resultados, COMPSRA los proporciona fundamentalmente en ficheros *.bam* para visualización y ficheros *.txt* o *.csv* a partir de los cuales extraer la información del análisis. Los otros dos *pipelines* incluyen además ficheros de reporte en formato *.html* donde se presentan diferentes tablas y gráficos de descripción de los resultados.

Los tres *pipelines* ejecutados tienen un paso inicial de procesado de los ficheros con el fin de eliminar adaptadores y secuencias y bases de baja calidad. No obstante, con el objetivo de que los ficheros de entrada fueran lo más similares posible y las modificaciones que hiciera cada *pipeline* fueran mínimas, estos fueron procesados previamente. Las métricas de calidad y alineamiento de cada uno de los *pipelines* indican que la pérdida de secuencias o modificaciones que han realizado han sido mínimas.

Los resultados obtenidos en el control de calidad de las muestras, así como del alineamiento de lecturas en los tres *pipelines*, han permitido identificar que una de las muestras del *dataset*, en concreto la muestra 3 de hembra (Lung_F3), es atípica respecto al resto de muestras. Tiene un alto contenido en lecturas de 30-35 nucleótidos y lecturas sobrerrepresentadas, un alto porcentaje alinean frente a *tRNA* maduros y sus métricas de alineamiento son totalmente diferentes al resto en los tres *pipelines*. El análisis llevado a cabo por el *pipeline* nf-core/smrnaseq mediante miRTrace, en principio descarta una posible contaminación. No obstante, se podría realizar un análisis mediante BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) y comprobar si esas lecturas corresponden a un ARN de ratón o de otra especie. También se podría optar por filtrar esas lecturas para que los resultados posteriores no se vean afectados. Lo que sí parece claro es que esta muestra, de alguna manera, está enriquecida en lecturas que se corresponden con *tRNA* maduros.

El porcentaje de lecturas alineadas respecto al total de los tres *pipelines* es muy similar. En los tres, más de un 90% de las lecturas son alineadas, bien frente al genoma de referencia en el caso de COMPSRA o bien frente a bibliotecas específicas como miRge3.0. En el caso de nf-core/smrnaseq el alineamiento se realiza frente a bibliotecas específicas pero la medida del porcentaje la estima a partir del alineamiento de las lecturas frente a un genoma de referencia. En cualquier caso, en este aspecto el resultado de los tres *pipelines* es muy parecido y ha funcionado correctamente. Y como consecuencia, el porcentaje de lecturas no alineadas también es muy similar en los tres *pipelines*, siempre menor a un 10% para todas las muestras,

incluida la muestra Lung_F3 mencionada anteriormente. Este último resultado también podría indicar que esas lecturas atípicas de la muestra no son consecuencia de una contaminación procedente de otra especie, sino que tan sólo, de alguna manera la muestra está enriquecida en *tRNA* maduro.

Los tres *pipelines* permiten identificar y cuantificar *miRNAs* y obtener un fichero de contajes para análisis posteriores. El porcentaje de lecturas alineadas correspondientes a *miRNAs* de los *pipeline*s, varía entre 52-57% de miRge3.0 y el 70% que se obtiene de media para COMPSRA y nf-core/smrnaseq. Estos porcentajes son en conjunto lo que cabría esperar. En este aspecto es significativo el diferente resultado obtenido entre *pipelines* si se tiene en cuenta que tanto miRge3.0 como nf-core/smrnaseq utilizan como programa de alineamiento Bowtie frente a miRBase, mientras que COMPSRA utiliza el programa STAR y frente a un genoma de referencia. Es evidente que las diferencias en los parámetros de los programas de alineamiento utilizados por cada *pipeline* son los que dan lugar a las diferencias de resultado.

Con respecto a los *sncRNAs* identificados por cada *pipeline* también existen diferencias entre ellos. COMPSRA permite identificar y cuantificar diferentes tipos de *sncRNAs*. En el caso de miRge 3.0 únicamente lo permite para *miRNAs*, aunque es cierto que proporciona datos sobre la distribución del alineamiento de lecturas frente a otros tipos de *sncRNAs*. Por su parte, nf-core/smrnaseq únicamente lo lleva a cabo para *miRNAs*. Así pues, diferencias en los tipos de *sncRNAs* analizados y cómo presentan los resultados de estos, no sólo estos tres *pipelines* sino en general, es un aspecto importante a tener en cuenta a la hora de seleccionar un *pipeline* de trabajo.

Respecto a los *miRNAs* identificados, a partir de los ficheros de contaje se ha observado que existen diferencias tanto en el número total como en el número de elementos comunes y no comunes. A este respecto, hay que tener en cuenta que cada *pipeline* es distinto en cuanto a las herramientas utilizadas. Por ejemplo el programa de alineamiento (Bowtie o STAR). Y que incluso utilizando la misma herramienta, como es el caso de miRge3.0 y nf-core/smrnaseq para Bowtie, los parámetros de distintas opciones como pueden ser el número de *mismatches* permitidos o el *overlapping* son diferentes en cada caso. Lo cual puede dar lugar a diferencias en los resultados como es este caso. Tratar de reanalizar los datos unificando herramientas y/o parámetros podría ser una opción para obtener resultados más similares. En caso contrario, la opción de analizar los datos con más de un *pipeline* y consensuar resultados puede hacer que estos sean más fiables o completos. Otro aspecto a tener en cuenta por el que se pueden obtener diferencias en este caso, pero también en general, entre *pipelines* es el *dataset* analizado. En este trabajo, se ha utilizado un *dataset* sencillo en el que además se ha buscado, de acuerdo con la publicación original, aquellas muestras donde hay diferencias evidentes entre sexos. Un *dataset* más complejo en cuanto a número de muestras y/o donde las diferencias entre grupos no fueran tan marcadas podría dar lugar a la aparición de más o menos diferencias en los resultados de cada *pipeline*.

El análisis de expresión diferencial de *miRNAs* se ha realizado con el mismo *script* para todos los *pipelines* en lugar de utilizar la opción de análisis que algunos de ellos permitían. De esta forma se ha asegurado que todos los pasos del análisis sean iguales. Y que por tanto las diferencias que han aparecido entre *pipelines* se deban tan sólo a los ficheros de contajes. No existen grandes diferencias en los perfiles de expresión de *miRNAs* que se obtienen con cada *pipeline*. Independientemente de las diferencias que hay en las listas de *miRNAs* identificados por cada uno, la comparación de *miRNAs* comunes indica que la cuantificación llevada a cabo

por los tres *pipelines* a partir de las lecturas alineadas, aún siendo llevada a cabo por métodos distintos, proporciona resultados similares.

En el caso de los *miRNAs* diferencialmente expresados identificados por cada *pipeline* la situación es similar a lo obtenido para el número total de *miRNAs* identificados y cuantificados. Tanto en el número total de *miRNAs* diferencialmente expresados como en los filtrados como significativos (*log2FC* absoluto mayor de 1 y *padj* menor de 0.05), existen diferencias. No obstante, la mayoría de estas diferencias se deben a que se obtienen un mayor número de *miRNAs* diferencialmente expresados a partir de los datos proporcionados por COMPSRA. En el resto de comparaciones, incluidos los datos del paper original, el solapamiento es bastante elevado, por lo que el número de *miRNAs* en común es elevado en estos casos. Las razones de estas diferencias, de nuevo, son las mencionadas anteriormente, puesto que el *script* de análisis utilizado es el mismo para todos los *pipelines* y la única diferencia son los ficheros de contajes obtenidos a partir de los pasos anteriores: diferencias en las herramientas utilizadas, en los parámetros de los argumentos o las características del *dataset* utilizado.

Los tres *pipelines* utilizados en este trabajo disponen de una documentación detallada para poder realizar su instalación y llevar a cabo su ejecución, describiendo los distintos argumentos que se pueden configurar en cada uno de ellos. De todos ellos existen publicaciones que describen el *pipeline*, así como repositorios en Github para poder acceder a ellos. Los tres *pipelines* especifican todas las herramientas y bibliotecas necesarias para realizar el análisis. Tanto COMPSRA como miRge3.0 proporcionan las instrucciones de acceso a las bibliotecas específicas de anotación que utilizan. Y en el caso de COMPSRA también al genoma de referencia. Respecto a nf-core/smrnaseq, es recomendable disponer de los ficheros de *miRNAs* maduros y precursores de miRBase y del fichero con el genoma de referencia y su índice y proporcionarlos directamente.

En cualquier caso, determinados análisis de los *pipelines* ha sido imposible realizarlos. La biblioteca de *piRNAs* de COMPSRA no está disponible o está dañada, por lo que ha sido necesario omitir ese paso. Y en el caso de nf-core/smrnaseq, el análisis con mirDeep2 no ha sido posible ejecutarlo, probablemente por el perfil de configuración utilizado (*profile* conda) pero que era el único disponible en el momento de realizar este trabajo. Es posible que con otros perfiles de configuración pueda ser ejecutado.

Finalmente, el tiempo de ejecución de los tres *pipelines* en la máquina virtual utilizada no ha sido excesivo en ningún caso para el dataset utilizado en este trabajo (6 muestras en total). Tanto para COMPSRA como miRge3.0 el tiempo total de análisis ha sido de unos a 40 minutos. Mientras que para nf-core/smrnaseq, teniendo en cuenta que no se ha ejecutado mirDeep2, el tiempo total ha sido una hora.

7 Conclusiones

7.1 Conclusiones

Las conclusiones principales que se han obtenido de este trabajo son las siguientes:

- El estudio de *sncRNAs* ha adquirido una gran importancia y desarrollo a lo largo de los últimos diez años. Una prueba de ello es la cantidad de *pipelines* publicados para el análisis de este tipo de datos.

- En general, todos los *pipelines* tienen una estructura similar, con una serie de pasos básicos, como por ejemplo, control de calidad y procesado, alineamiento de lecturas o anotación y cuantificación de *sncRNAs*. No obstante, se diferencian en cuanto a las herramientas y/o estrategias llevadas a cabo para realizar cada uno de los pasos, tipos de *sncRNA* analizados, y la posibilidad de incluir análisis más específicos como *isomirs* o *novel miRNAs*.

- El análisis de la calidad de las lecturas de las muestras utilizadas, así como su correcto procesamiento, son fundamentales para obtener buenos resultados. Un *dataset* con muestras con perfiles atípicos como la muestra Lung_F3 del *dataset* utilizado pueden afectar negativamente en los resultados obtenidos.

- Los tres *pipelines* utilizados han tenido un rendimiento en el alineamiento de lecturas similar, y superior al 90%, independientemente de si la estrategia de alineamiento está basada en un genoma de referencia o en bibliotecas de anotación específicas. En cualquier caso, sería necesario evaluar otros *pipelines* o utilizar otros dataset más complejos para confirmar si existen o no diferencias significativas entre ambas estrategias de alineamiento.

- Las diferencias en los programas utilizados por cada *pipeline* en cada uno de sus pasos, o incluso para un mismo programa, la configuración de los argumentos disponibles, da lugar a diferencias en el número de *sncRNAs* identificados, su cuantificación, etc… y afecta directamente a análisis posteriores como por ejemplo de expresión diferencial.

- La gran variedad de *pipelines* disponibles hace que el tipo de *sncRNAs* que se quieren analizar, cómo y qué resultados y ficheros se proporcionan, así como los análisis posteriores que se quieran realizar a partir de esos ficheros, sean aspectos importantes a tener en cuenta a la hora de seleccionar un *pipeline* de trabajo.

- Alternativamente, se puede analizar los datos con más de un *pipeline* y consensuar resultados, lo que puede hacer que estos sean más fiables o completos.

Se han alcanzado los objetivos iniciales que se plantearon en este trabajo. En primer lugar se ha realizado una búsqueda bibliográfica y revisión de *pipelines* que ha permitido conocer en detalle las características principales que poseen los *pipelines* de análisis de *sncRNAs*, las posibilidades de análisis que ofrecen y cómo se lleva a cabo un análisis básico. En la Tabla 13 incluida en el Anexo aparece resumida esta parte del trabajo. Posteriormente se han ejecutado algunos de los *pipelines* estudiados. Esto ha permitido llevar a cabo un análisis básico de *sncRNAs*, desde la lectura de los datos hasta la obtención de una lista de *miRNAs* diferencialmente expresados, comparando los resultados obtenidos con cada uno de ellos.

7.2 Líneas de futuro

El trabajo realizado puede ser continuado o ampliado incluyendo otros *pipelines* diferentes a los utilizados aquí y siguiendo un análisis similar. También puede ser interesante utilizar otros *dataset* de mayor complejidad en cuanto al número de muestras, grupos y/o con mayor o menor número de diferencias en los datos. Y comprobar si las diferencias entre *pipelines* que se obtienen son más o menos significativas.

Con respecto a los *pipelines* utilizados en este trabajo, podrían volver a ser ejecutados modificando alguno de los argumentos disponibles. Por ejemplo, miRge3.0 puede utilizar

MirGene DB en lugar de miRBase como biblioteca de anotación. Por otro lado, sería interesante volver a ejecutar nf-core/smrnaseq de nuevo si fuera posible realizar el análisis con mirDeep2. Esto permitiría, por ejemplo, obtener un análisis de *novel miRNAs* que podría ser comparado con el obtenido mediante miRge3.0.

7.3 Seguimiento de la planificación

El trabajo ha podido ser llevado a cabo siguiendo la metodología propuesta al inicio, en el plan de trabajo. Todas las tareas incluidas para cumplir los objetivos propuestos se han llevado a cabo y han permitido que estos se hayan alcanzado. Y la planificación temporal propuesta se ha cumplido sin que se haya producido ningún retraso o modificación en la ejecución de las tareas.

8 Bibliografía

1. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9):418-26.
2. Bartel DP. Metazoan MicroRNAs. Cell. 2018;173(1):20-51.
3. Burgos M, Hurtado A, Jimenez R, Barrionuevo FJ. Non-Coding RNAs: lncRNAs, miRNAs, and piRNAs in Sexual Development. Sex Dev. 2021;15(5-6):335-50.
4. He C, Wang K, Gao Y, Wang C, Li L, Liao Y, et al. Roles of Noncoding RNA in Reproduction. Front Genet. 2021;12:777510.
5. Kuo MC, Liu SC, Hsu YF, Wu RM. The role of noncoding RNAs in Parkinson's disease: biomarkers and associations with pathogenic pathways. J Biomed Sci. 2021;28(1):78.
6. Xiao L, Wang J, Ju S, Cui M, Jing R. Disorders and roles of tsRNA, snoRNA, snRNA and piRNA in cancer. J Med Genet. 2022.
7. Alexiou A, Zisis D, Kavakiotis I, Miliotis M, Koussounadis A, Karagkouni D, et al. DIANA-mAP: Analyzing miRNA from Raw NGS Data to Quantification. Genes (Basel). 2020;12(1).
8. Li J, Kho AT, Chase RP, Pantano L, Farnam L, Amr SS, et al. COMPSRA: a COMprehensive Platform for Small RNA-Seq data Analysis. Sci Rep. 2020;10(1):4552.
9. Patil AH, Halushka MK. miRge3.0: a comprehensive microRNA and tRF sequencing analysis pipeline. NAR Genom Bioinform. 2021;3(3):lqab068.
10. Pogorelcnik R, Vaury C, Pouchin P, Jensen S, Brasset E. sRNAPipe: a Galaxy-based pipeline for bioinformatic in-depth exploration of small RNAseq data. Mob DNA. 2018;9:25.
11. Zhong X, Pla A, Rayner S. Jasmine: a Java pipeline for isomiR characterization in miRNA-Seq Data. Bioinformatics. 2019.
12. Cech TR, Steitz JA. The noncoding RNA revolution-trashing old rules to forge new ones. Cell. 2014;157(1):77-94.
13. Esteller M. Non-coding RNAs in human disease. Nat Rev Genet. 2011;12(12):861-74.
14. Ha M, Kim VN. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol. 2014;15(8):509-24.
15. Matera AG, Terns RM, Terns MP. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat Rev Mol Cell Biol. 2007;8(3):209-20.
16. Kiss T. Biogenesis of small nuclear RNPs. J Cell Sci. 2004;117(Pt 25):5949-51.
17. Ishizu H, Siomi H, Siomi MC. Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. Genes Dev. 2012;26(21):2361-73.
18. Kumar P, Anaya J, Mudunuri SB, Dutta A. Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. BMC Biol. 2014;12:78.
19. Hombach S, Kretz M. Non-coding RNAs: Classification, Biology and Functioning. Adv Exp Med Biol. 2016;937:3-17.
20. Seal RL, Chen LL, Griffiths-Jones S, Lowe TM, Mathews MB, O'Reilly D, et al. A guide to naming human non-coding RNA genes. EMBO J. 2020;39(6):e103777.
21. Uchida S, Dimmeler S. Long noncoding RNAs in cardiovascular diseases. Circ Res. 2015;116(4):737-50.
22. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature. 2008;453(7194):539-43.
23. Bofill-De Ros X, Yang A, Gu S. IsomiRs: Expanding the miRNA repression toolbox beyond the seed. Biochim Biophys Acta Gene Regul Mech. 2020;1863(4):194373.
24. Bussotti G, Notredame C, Enright AJ. Detecting and comparing non-coding RNAs in the high-throughput era. Int J Mol Sci. 2013;14(8):15423-58.

25. Wu X, Kim TK, Baxter D, Scherler K, Gordon A, Fong O, et al. sRNAnalyzer-a flexible and customizable small RNA sequencing data analysis pipeline. Nucleic Acids Res. 2017;45(21):12140-51.
26. Zhao S, Gordon W, Du S, Zhang C, He W, Xi L, et al. QuickMIRSeq: a pipeline for quick and accurate quantification of both known miRNAs and isomiRs by jointly processing multiple samples from microRNA sequencing. BMC Bioinformatics. 2017;18(1):180.
27. Kuksa PP, Amlie-Wolf A, Katanic Z, Valladares O, Wang LS, Leung YY. SPAR: small RNA-seq portal for analysis of sequencing experiments. Nucleic Acids Res. 2018;46(W1):W36-W42.
28. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38(3):276-8.
29. Potla PA, S. A. Kapoor, M. A bioinformatics approach to microRNA-sequencing analysis. Osteoarthritis and Cartilage Open. 2021;3(1).
30. Andrews S. FastQC: A Quality Control tool for High Throughput SequenceData Cambridge, UK.2010 [
31. Ewels P, Magnusson M, Lundin S, Kaller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32(19):3047-8.
32. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
33. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95.
34. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
35. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47(D1):D155-D62.
36. Fromm B, Hoye E, Domanska D, Zhong X, Aparicio-Puerta E, Ovchinnikov V, et al. MirGeneDB 2.1: toward a complete sampling of all major animal phyla. Nucleic Acids Res. 2022;50(D1):D204-D10.
37. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2).
38. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923-30.
39. Friedlander MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2012;40(1):37-52.
40. Aparicio-Puerta E, Gomez-Martin C, Giannoukakos S, Medina JM, Scheepbouwer C, Garcia-Moreno A, et al. sRNAbench and sRNAtoolbox 2022 update: accurate miRNA and sncRNA profiling for model and non-model organisms. Nucleic Acids Res. 2022.
41. Borchert GM, Holton NW, Williams JD, Hernan WL, Bishop IP, Dembosky JA, et al. Comprehensive analysis of microRNA genomic loci identifies pervasive repetitive-element origins. Mob Genet Elements. 2011;1(1):8-17.
42. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31(2):166-9.
43. 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;17:13.
44. Isakova A, Fehlmann T, Keller A, Quake SR. A mouse tissue atlas of small noncoding RNA. Proc Natl Acad Sci U S A. 2020;117(41):25634-45.
45. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17(2).
46. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.

47. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 2019;47(D1):D766-D73.
48. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24-6.
49. Kang W, Eldfjell Y, Fromm B, Estivill X, Biryukova I, Friedlander MR. miRTrace reveals the organismal origins of microRNA sequencing data. Genome Biol. 2018;19(1):213.

Anexo

En la siguiente tabla (Tabla 13) se presentan algunos de los *pipelines* publicados en los últimos cinco años revisados para realizar este trabajo. Se incluye una descripción de sus principales características así como las referencias consultadas.

Herramienta	Descripción	Referencias
miRge3.0 (2021)	**Lenguaje /soporte:** python3, R, HTML, JAVA. Se puede ejecutar mediante *Command line interfaz* o *graphical user interfaz* (CLI o GUI). **Programas**: miREC, Cutadapt (v3.0), Pandas (v0.25.3), Bowtie (v1.3.0), ViennaRNA (v2.4.16), SAMtools (v1.7), biopython (v1.78), sklearn (v0.23.1), numPy (v1.18.4), SciPy (v1.4.1), reportlab (v3.5.42), DESeq2 (release 3.12). **Bases de datos:** bibliotecas de anotación de *sncRNA* específicas disponibles en: https://sourceforge.net/projects/mirge3/files/miRge3_Lib/ **Tipo de sncRNAs:** *miRNAs* y *tRFs*, incluye análisis de *isomiRs* y *novel miRNAs*. **Descripción**: - Preprocesamiento y control de calidad a partir de ficheros FASTQ (Cutadapt). - El alineamiento de las lecturas se realiza frente a bibliotecas de anotación específicas en de manera secuencial. (Bowtie). - Las lecturas no alineadas al final del proceso son utilizadas frente a un genoma de referencia para identificar novel *miRNA* mediante *support vector machine* (SVM) (ViennaRNAfold). - **Ficheros de resultados:** ficheros BAM de alineamientos. Fichero de contajes de *miRNAs*, fichero GFF3 de *miRNAs* e *isomirs*. Fichero de novel *miRNAs* tabla de recuentos de lecturas alineadas y sin alinear con el que realizar un análisis de expresión diferencial. Tabla de *isomiRs* en formato GFF3. Archivo BAM para visualizar en IGV. Tabla de *tRF*. Tabla de *novel miRNAs*.	https://doi.org/10.1093/nargab/lqab068 https://github.com/mhalushka/miRge3.0 https://sourceforge.net/projects/mirge3/

DIANA-mAP (2021)	**Lenguaje /soporte:** JAVA, python3, Perl, R. Se puede ejecutar mediante CLI aunque se recomienda el sistema Docker. **Programas**: FASTAQC, Cutadapt, DNApi, Bowtie, miRDeep2, DESeq2 **Bases de datos:** mirBase **Tipo de sncRNAs:** identificación y cuantificación de *miRNAs* **Descripción**: - Preprocesamiento y control de calidad a partir de ficheros FASTQ (Cutadapt y FASTQC). - El alineamiento de las *reads* se realiza frente a un genoma de referencia mediante el *script* de miRDeep2 (Bowtie) - La cuantificación de las *reads* alineadas también utiliza el *script* de mirRDeep2. Se alienan frente a *miRNAs* precursores y maduros procedentes de *mirBase*. - Permite el análisis de expresión diferencial mediante el paquete DESeq2. **Ficheros de resultados:** tabla de recuentos de los *miRNAs* identificados a partir de *reads* alineadas con la que realizar un análisis de expresión diferencial. Resumen estadístico. Gráfico de distribución de *reads*.	https://doi.org/10.3390/genes12010046 https://github.com/athalexiou/DIANA-mAP
COMPSRA (2020)	**Lenguaje /soporte:** JAVA. Se ejecuta mediante CLI. **Programas**: STAR v2.5.3a **Bases de datos:** bibliotecas específicas a partir de miRBase, piRNABank, piRBase, piRNACluster, gtRNAdb, GENCODE release 27, circBase. **Tipo de sncRNAs:** *miRNAs, piRNAs, snRNAs, snoRNAs, tRNAs, circRNAs*.	https://doi.org/10.1038/s41598-020-61495-0 https://github.com/cougarlj/COMPSRA

	Descripción: - Cinco módulos independientes: Quality Control (QC), Alignment, Annotation, Microbe y Function. - Preprocesamiento y control de calidad a partir de ficheros FASTQ. - El alineamiento de las *reads* se realiza frente a un genoma de referencia (STAR). - Si la *read* alinea en múltiples localizaciones del genoma sólo se cuenta una vez. Si la *read* alinea en una localización donde se produce *overlapping* de *sncRNAs* a cada secuencia se le asigna un recuento. - Las *reads* alineadas se utilizan en el módulo de anotación y posteriormente en el de función, que permite realizar análisis de expresión diferencial mediante *Wilcoxon Rank Sum Test*. El valor umbral de recuentos para la identificación de *sncRNAs* es de 5. - Las *reads* no alineadas se utilizan en el módulo Microbe (BLAST). **Ficheros de resultados:** cada módulo proporciona un fichero *report* con los resultados de su análisis.	
Manatee (2020)	**Lenguaje /soporte:** Perl, HTML, Shell, CSS. Se ejecuta mediante CLI. **Programas**: Bowtie v1, SAMtools, **Bases de datos:** es necesario proporcionar un fichero de anotaciones de *ncRNAs* en formato GTF obtenido a partir de bases de datos como miRBase v21 o Ensemble. **Tipo de sncRNAs:** *miRNAs, isomiRs, novel miRNAs* **Descripción**: - Los ficheros FASTA o FASTQ tienen que haber sido procesados previamente para eliminar los adaptadores. - El alineamiento de las *reads* se realiza frente a un genoma de referencia (Bowtie).	https://doi.org/10.1038/s41598-020-57495-9 https://github.com/jehandzlik/Manatee

	- Las *reads* que alinean en una sola localización del genoma se utilizan para formar los *cluster* UAR. Se requiere al menos 1 nucleótido de solapamiento entre la posición genómica de la *read* alineada y el transcrito anotado. - Las *reads* que alinean en múltiples posiciones son asignadas a aquellas que presentan un mayor *coverage*. El *coverage* se mide como el número de nucleótidos solapantes entre el transcrito anotado y la longitud de la *read*. El algoritmo prioriza la anotación con la mayor cobertura. - Las *reads* que no alinean frente al genoma o alinean en múltiples sitios por encima del valor umbral de 50 son alineadas frente al transcriptoma basado en el archivo de anotación. Se permite un número máximo de *mismatches* de 3. - Las *reads* identificadas y cuantificadas como *miRNAs* son utilizadas para la identificación de *isomiRs*: modificaciones post-transcripcionales, edición de extremos 5' y 3' y variaciones de nucleótidos. - Los *cluster* UAR identificados y que se encuentran en regiones sin anotaciones (5 *reads* alineadas con un espacio entre ellas menor de 50 nucleótidos) son utilizados para identificar potenciales novel *sncRNAs* loci. **Ficheros de resultados:** Tablas con recuentos de *sncRNAs* identificados, de *isomiRs* y de loci no anotados. Estas tablas pueden utilizarse en análisis posteriores como por ejemplo análisis de expresión diferencial con paquetes como DESeq2.	
Nf-core/smrnaseq (2021)	**Lenguaje /soporte:** Nextflow, Groovy, R, Python, HTML, Dockerfile. Se ejecuta mediante CLI. **Programas**: FastQC, Trim Galore! (Cutadapt), seqcluster, Bowtie1, SAMTools, edgeR, mirtop, mirDeep2, mirtrace, MultiQC **Bases de datos:** miRBase mature miRNA, miRBase hairpin. **Tipo de sncRNAs:** *miRNAs, isomiRs, novel miRNAs* **Descripción**:	https://nf-co.re/smrnaseq/1.1.0 https://github.com/nf-core/smrnaseq

	- Preprocesamiento y análisis de calidad a partir de ficheros FASTQ (FASTQC, Cutadapt, seqcluster). - El alineamiento de las *reads* se hace en primer lugar frente a miRBase. En primer lugar frente a secuencias de *miRNAs* maduros y posteriormente frente a secuencias de *hairpin miRNAs* (Bowtie). - Análisis estadístico de los resultados de los alineamientos (SAMtools). - Análisis de recuentos de *hairpin miRNAS* (edgeR). - Anotación de *miRNAs* e *isomiRs* identificados (mirtop) - Las *reads* son alineadas nuevamente, en este caso frente a un genoma de referencia (Bowtie, SAMtools). - Identificación de *miRNAS* y *novel miRNAS* mediante el script mirDeep2. Se utilizan los módulos de alineamiento y de cuantificación. **Ficheros de resultados:** Resumen estadístico de los alineamientos, análisis de expresión diferencial de *hairpin miRNAs*. Diferentes tablas con la identificación y los recuentos de *miRNAs* e *isomiRs* realizados por mirtop y mirDeep2.	
sRNAtoolbox /sRNAbench (2019)	**Lenguaje /soporte:** JAVA. Se ejecuta en servidor web. **Programas**: Bowtie, DESeq2, edgeR, NOISeq **Bases de datos:** miR-Base, mirGeneDB, archivos de anotación proporcionados **Tipo de sncRNAs:** *sncRNAs, miRNAs, isomiRs* **Descripción**: - Consiste en siete herramientas independientes interrelacionadas que permiten ejecutar un workflow con diferentes análisis: - sRNAbench: análisis del perfil de expresión de *sncRNAs*, novel *miRNAs*, análisis de *isomiRs*. Las *reads* pueden ser mapeadas frente a un genoma de referencia y los valores de expresión obtenidos utilizando los archivos de anotación, o las *reads* se mapean directamente frente a los archivos de anotación. Permite dos métodos para los casos	https://doi.org/10.1093/nar/gkz415 https://arn.ugr.es/srnatoolbox/

	de mapeo múltiple: ajustar el recuento de lecturas por el número de veces que la lectura mapea en el alineamiento. O asignar cada *read* sólo una vez a la secuencia con mayor valor de expresión. - sRNAde: análisis de expresión diferencial mediante DESeq2, edgeR o NOISeq. - sRNAblast: análisis de las *reads* que no mapean frente a bases de datos. Posible contaminación vírica o bacteriana. - miRNAconsTarget: predicción de secuencias diana a partir de bases de datos como Miranda, PITA, TargetSpy. - sRNAjBrowser: visualización de los valores de expresión en el genoma. - sRNAjBrowserDE: visualización del análisis de expresión diferencial - sRNAfuncTermss: análisis de enriquecimiento funcional. - sRNAfuncTargets: análisis de enriquecimiento funcional.	
miRDeep 2.0 (2019)	**Lenguaje /soporte:** Perl, HTML, Shell. Se ejecuta mediante CLI. **Programas**: Bowtie **Bases de datos:** fichero de anotación de pre-*miRNAs*, *miRNAs* como por ejemplo los de mirBase y fichero con secuencias estrella y el código de 3 letras de la especie de interés. **Tipo de sncRNAs:** *miRNAs y novel miRNAs* **Descripción**: - Los ficheros de entrada son en formato FASTA o FASTAQ con *reads* preprocesadas. - El programa tiene dos módulos o scripts: mapper.pl y quantification.pl - El alineamiento de las *reads* en mapper.pl se realiza frente a un genoma de referencia. Se permite 1 *mismatch* a hasta 5 posiciones diferentes en el genoma. Diferentes parámetros pueden ser configurados en la línea de comando. - El script quantification.pl alinea las *reads* frente a los ficheros de anotación proporcionados para cuantificar la expresión de *miRNAs*. Diferentes parámetros pueden ser configurados en la línea de comando.	https://github.com/rajewsky-lab/mirdeep2

miRMaster 2.0 (2021)	**Lenguaje /soporte:** Python 3.7.6, Django 2.2.10, Postgres 11.1, Redis 5.0., Celery 4.4.7, Bootstrap 4.5.3, Angular JS 1.5.11, jQuery 3.4.1. Se ejecuta en un servidor web. **Programas**: Bowtie 1.2.3, STAR 2.7.5a, mirtop 0.4.23 **Bases de datos:** miRBase (version 22.1), Ensembl ncRNA (version 100), RNACentral (for piRNAs) (version 15), GtRNAdb (version 18.1), circBase (accessed 25.10.20), NONCODE (version 5) and NCBI RefSeq for the reference genomes as well as viruses and bacteria. **Tipo de sncRNAs:** *miRNAs, isomiRs, novel miRNAs* y otros *sncRNA.* **Descripción**: - Preprocesamiento y análisis de calidad a partir de ficheros FASTQ (FASTQC, Cutadapt). Puede procesar *unique molecular identifier* (UMI). Se pueden incluir archivos de anotación de las muestras para análisis posteriores como de expresión diferencial (Bowtie). - El alineamiento de las *reads* se realiza en primer lugar frente a un genoma de referencia. Se permiten hasta cinco localizaciones genómicas sin *mismatches* y una longitud de *read* de 18 nucleótidos. - Un segundo alineamiento de las *reads* se realiza frente bases de datos de distintos *sncRNA* sin permitir *mismatch: microRNA, tRNA, piRNA, rRNA, scaRNA, lncRNA, snoRNA, snRNA, miscRNA and circRNA* (Bowtie). - Para cuantificar *miRNAs* las *reads* se mapean frente a pre-*miRNAs* permitiendo 1 *mismatch*. Se filtran permitiendo dos nucleótidos de diferencia en el extremo 5' y cinco en el 3'. - Para cuantificar *isomiRs* las *reads* se mapean permitiendo 2 *mismatchs* y filtrando de nuevo. - Filtrado: solo aquellos *sncRNA* con más de tres *reads* y que se expresan en al menos el 50% de muestras de una condición son retenidos. - Es posible realizar análisis de expresión diferencial: *t-test, Wilcoxon Mann-Whitney test* o ANOVA.	https://doi.org/10.1093/nar/gkab268 https://ccb-compute.cs.uni-saarland.de/mirmaster2

	- La predicción de *miRNAs* se realiza tras el alineamiento frente al genoma a partir de los candidatos a pre-*miRNAs*. Se buscan en ventanas de 70 nucleótidos y que no solapan con ningún *miRNA* anotado. **Ficheros de resultados:** A partir del módulo mapper.pl: fichero de *reads* procesadas y fichero de *reads* mapeadas. A partir del script quantification.pl: Un archivo con identificadores de miRNA y su recuento de *reads*, un archivo de firma llamado miRBase.mrd, un archivo llamado expression.html que da una visión general de todos los *miRNAs*, los datos de entrada y un directorio llamado pdfs que contiene para cada *miRNA* un archivo pdf que muestra su firma y estructura.	
Jasmine (2020)	**Lenguaje /soporte:** JAVA. Se ejecuta mediante CLI **Programas**: Trimmomatic, Cutadapt, Fastax_collapser, Bowtie **Bases de datos:** miRBase, MiRGeneDB **Tipo de sncRNAs:** *miRNAs, isomiRs* **Descripción**: - Preprocesamiento y análisis de calidad a partir de ficheros FASTQ (Trimmomatic, Cutadapt, Fastax_collapser) - El alineamiento de las *reads* se realiza frente a bases de datos (Bowtie). Al ser una herramienta de análisis de *isomiRs* se recomienda 1 *mismatch*. - Análisis de *isomiRs* basados en su clasificación de acuerdo con 4 niveles consecutivos en función de parámetros como similitud, longitud, *mismatches* **Ficheros de resultados:** fichero de *miRNAs* con recuentos, fichero de *isomiRs* con recuentos. Tabla de polimorfismos e isoformas. Los archivos de recuentos se generan de acuerdo al sistema de clasificación basada en niveles. Para cada nivel se genera un fichero.	https://doi.org/10.1093/bioinformatics/btz806
miARma-Seq (2019)	**Lenguaje /soporte:** Perl, Python, HTML, R. Se ejecuta mediante CLI.	https://doi.org/10.1016/j.ymeth.2018.09.002

	Programas: FASTQC, Cutadapt, Reaper, Minion, Bowtie1, Bowtie2, mirDeep2, FeatureCounts, edgeR, NOISeq, RNAfold v2.2.4, Samtools v1.3 **Bases de datos:** miRGate, Ciri v2.0.1 **Tipo de sncRNAs**: *miRNAs, novel miRNAs, circRNAs* **Descripción**: - Preprocesamiento y análisis de calidad a partir de ficheros FASTQ (FASTQC, Cutadapt, Reaper, Minion) - El alineamiento de las reads se realiza frente al genoma de referencia (Bowtie) - La cuantificación de los *miRNAs* se hace a partir de las *reads* alienadas. La cuantificación se realiza mediante FeatureCounts. - El análisis de expresión diferencial es posible realizarlo mediante edgeR o NOISeq. - Mediante miRGate es posible realizar un análisis de interacción de los *miRNAs* identificados con posibles *mRNAs*. **Ficheros de resultados:**	https://github.com/eandresleon/miARma-seq
sRNApipe (2018)	**Lenguaje /soporte:** Perl. Se ejecuta mediante Galaxy server **Programas**: BWA, SAMtools **Bases de datos:** ficheros con secuencias de referencia (genoma, transcriptoma, *TEs*, *rRNAS*, *miRNAs*,…) **Tipo de sncRNAs:** *miRNAs, siRNAS, piRNAs, rRNAs, snRNAs* **Descripción**: - Los archivos FASTQ deben contener *reads* procesadas sin adaptadores. - Es necesario proporcionar archivos con secuencias de referencia (genoma, transcriptoma*, TEs, rRNAS, miRNAs*,…)	https://doi.org/10.1186/s13100-018-0130-7 https://github.com/GReD-Clermont/sRNAPipe

	- El alineamiento de las reads (18-29 nt) se realiza en primer lugar frente a un genoma de referencia (BWA) permitiendo el máximo número de *mismatches* posible. Solo las *reads* mapeadas se utilizan en análisis posteriores. - Posteriormente estas *reads* se mapean frente al restos de secuencias de referencia. Las *reads* que no mapean en un alineamiento se utilizan en el siguiente alineamiento. - Se crean cuatro grupos: *reads* que mapean frente al genoma pero no frente a las secuencias de referencia, *reads* que mapean frente a *miRNAs, reads* que mapean frente *siRNAs* (21 nt) y *reads* que mapean frente a *piRNAs* (23-29 nt). Para cada grupo se obtiene un fichero con los recuentos. - Los cuatro grupos también se mapean frente al archivo de *TEs* - Permite el análisis *ping pong signature*. **Ficheros de resultados:** se obtienen en formato html. Resultados de *reads* mapeadas (*all genome-mappers, unique mappers, length distribution*). Distribución por categorías analizadas tanto en gráfico como por cromosoma**.**	
Oasis2 (2018)	**Lenguaje /soporte:** Java, J2EE, mysql, Python, R, PHP and JavaScript. Se ejecuta en un servidor web. Dispone de API para envío de datos al servidor. **Programas**: FASTQC, Cutadapt, STAR, mirDeep2, featureCounts, Kraken, Deseq2, miRanda, mirTarBase, miRecords, randomForest. **Bases de datos:** biblioteca específica Oasis-DB (*miRBase, piRNAbank, Ensembl, predicted novel miRNAs, and sRNA families*) **Tipo de sncRNAs:** *miRNAs, novel miRNAs, piRNAs, snoRNAs, snRNAs, rRNAs*. **Descripción**: - Disponible para al análisis en 14 especies diferentes - Preprocesamiento y análisis de calidad a partir de ficheros FASTQ (FASTQC, Cutadapt). Longitud de *reads* entre 15 y 32 nucleótidos. - Se realizan cuatro alineamientos diferentes y sucesivos (STAR):	https://doi.org/10.1186/s12859-018-2047-z http://oasis.dzne.de/index.php

	- Un primer alineamiento frente a la base de datos de *miRNAs* de la especie en Oasis-DB. Las *reads* de 15-19 nucleótidos sin *mismatches* y las *reads* de 20-32 nucleótidos permitiendo 1 *mismatch*. Las *reads* que no alinean frente a esta base de datos se alinean frente *snRNAs, snoRNAs, rRNA*s, y *piRNAs* de Oasis-DB con los mismos parámetros para identificar otros tipos de *sncRNAs*. - En un segundo alineamiento las *reads* no alineadas en el primer paso se alinean frente al genoma de referencia para identificar novel *miRNAs*. Se utiliza el script de mirDeep2 con los parámetros por defecto (1 *mismatch* y no más de 5 potenciales regiones de localización). Los nuevos *miRNAs* identificados son añadidos a Oasis-DB y en caso de múltiples localizaciones se crea una familia de *miRNAs*. - El tercer alineamiento se realiza con las restantes *reads* que no han alineado frente al genoma de referencia. Son alineadas frente a genomas de virus y bacterias de *RefSeq*. Se utiliza el programa Kraken basado en la búsqueda a través de *k-mers*. - El último alineamiento se realiza con las *reads* que no han alineado en ningún paso anterior frente a las bases de datos de *miRNAs* en Oasis-DB del resto de especies soportadas. - A partir de las tablas de recuentos generadas se puede realizar análisis de expresión diferencial mediante el paquete DESeq2, análisis funcionales (GeneMania, G:Profiler, STRING, DAVID) y de anotación y predicción de dianas para los *miRNAs* identificados (miRanda, MirTarBase, miRecords).	

Tabla 13. *Pipelines* revisados durante la realización de este trabajo.