Universitat Oberta de Catalunya

uoc.edu

Estudio comparativo de *pipelines* de análisis de *small non-coding RNAs (sncRNA)*

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

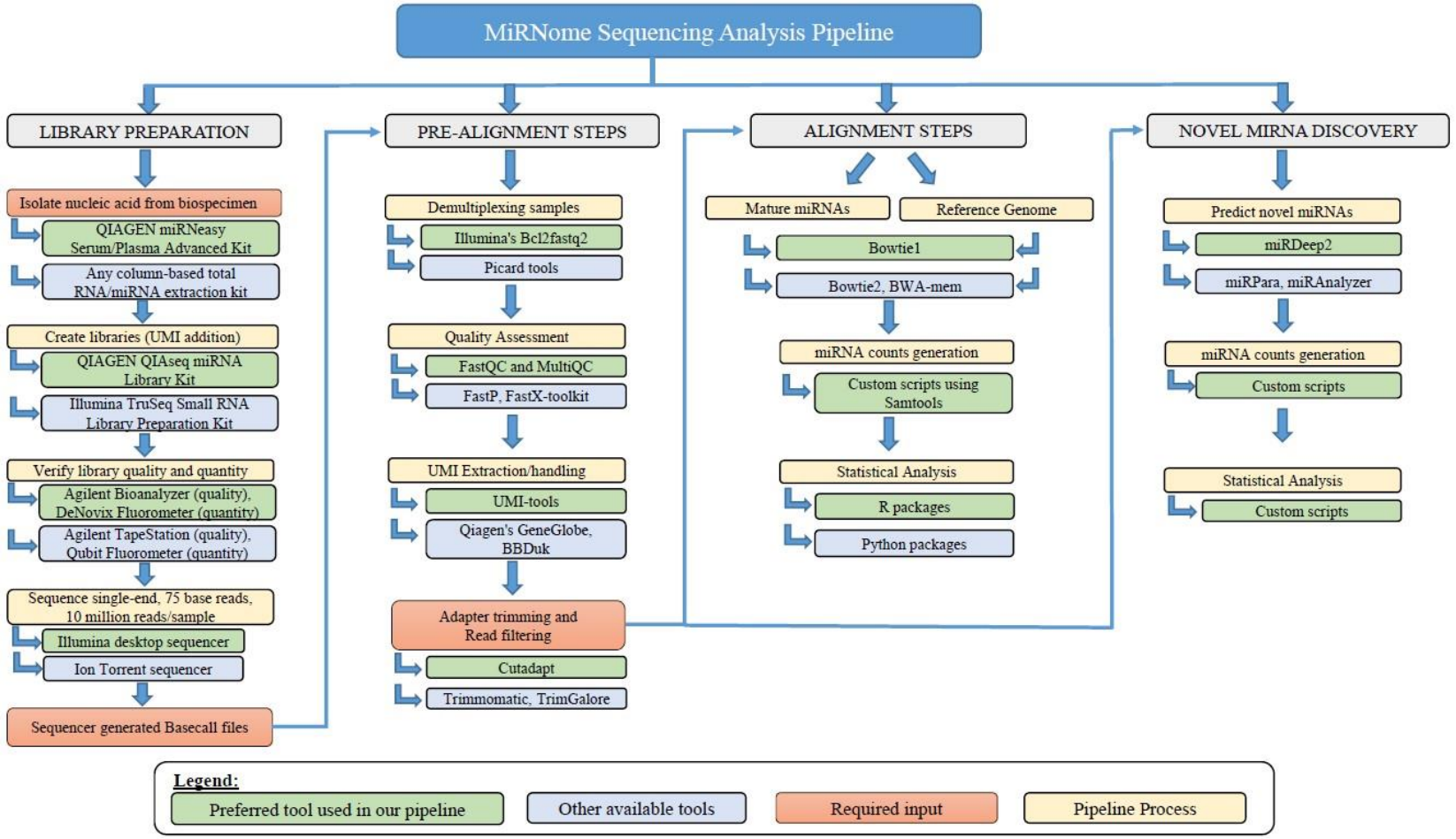
Tutora: Mireia Ferrer Almirall
Profesor responsable de la asignatura: Antoni Pérez Navarro

06/06/2022

1. Revisión de *pipelines* disponibles para el análisis de *small non-coding RNAs (sncRNA)*
    - Búsqueda bibliográfica: Pubmed, Web of Science, Github,…
    - Estudio comparativo de *pipelines*: estructura, tipos *sncRNA*, herramientas,…

2. Estudio comparativo del análisis de *sncRNA* realizado por diferentes *pipelines*
    - Selección de *pipelines*: estrategia alineamiento, tipos *sncRNA*, tipos de análisis,…
    - Análisis de *sncRNAs (miRNAs)* en un *dataset* con los *pipelines* seleccionados

Esquema general de un *pipeline* de análisis de *sncRNAs*



Adaptado de Potla *et al.* 2021

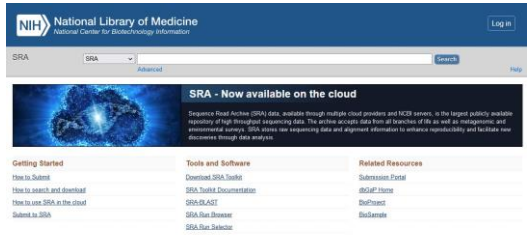
1. *Dataset* y análisis de calidad

A mouse tissue atlas of small noncoding RNA

Alina Isakova[a], Tobias Fehlmann[b], Andreas Keller[b,c], and Stephen R. Quake[a,d,e,1]

[a]Department of Bioengineering, Stanford University, Stanford, CA 94305; [b]Chair for Clinical Bioinformatics, Saarland University, 66123 Saarbrücken, Germany; [c]Department of Neurology, School of Medicine, Stanford University, Stanford, CA 94305; [d]Department of Applied Physics, Stanford University, Stanford, CA 94305; and [e]Chan Zuckerberg Biohub, San Francisco, CA 94158

Contributed by Stephen R. Quake, July 29, 2020 (sent for review February 10, 2020; reviewed by John C. Marioni and Igor Ulitsky)



Library strategy	ncRNA-Seq
Library source	transcriptomic
Library selection	size fractionation
Instrument model	Illumina NextSeq 500

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

Procesado de los ficheros *.fastq*:

FASTQC/MultiQC: análisis/control de calidad de muestras y visualización

Cutadapt 4.0: eliminación de adaptadores y secuencias de baja calidad

2. Selección *pipelines*

Published online 21 July 2021 *NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 3 1*
https://doi.org/10.1093/nargab/lqab068

miRge3.0: a comprehensive microRNA and tRF sequencing analysis pipeline

Arun H. Patil and Marc K. Halushka*

Department of Pathology, Division of Cardiovascular Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

SCIENTIFIC REPORTS
nature research

OPEN

COMPSRA: a COMprehensive Platform for Small RNA-Seq data Analysis

Jiang Li[1*], Alvin T. Kho[2], Robert P. Chase[1], Lorena Pantano[3], Leanna Farnam[1], Sami S. Amr[1*] & Kelan G. Tantisira[1,5*]

nf-core
https://nf-co.re

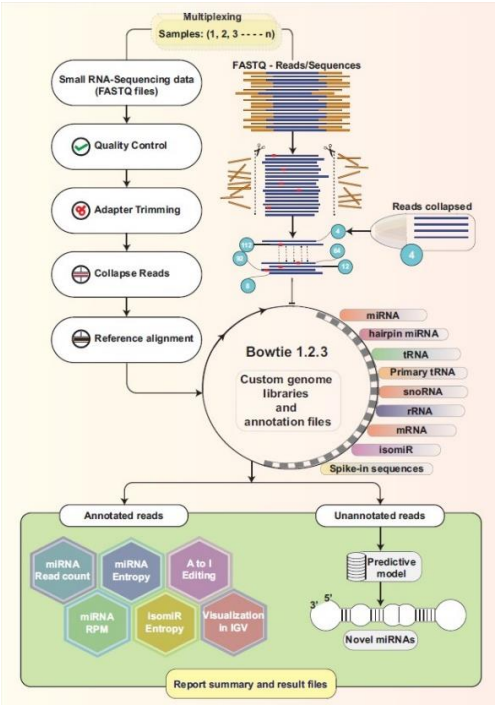
nf-core/smrnaseq
A small-RNA sequencing analysis pipeline

3. Análisis de expresión diferencial

Correlación de los perfiles de expresión de *miRNAs*

Análisis de *miRNAs* diferencialmente expresados

miRge3.0

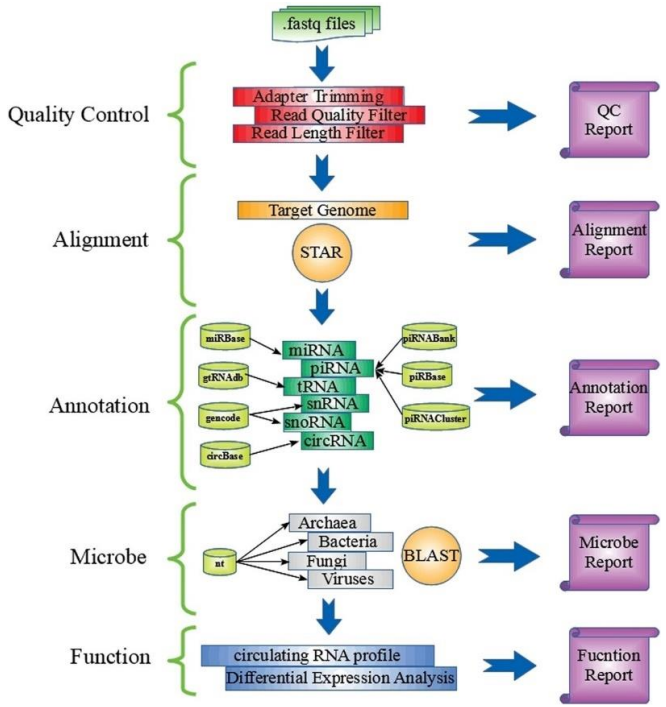


Tipo de *sncRNAs*: *miRNAs*

Alineamiento de lecturas frente a librerías específicas de anotación

Anotación mediante librerías específicas a partir de miRBase

COMPSRA

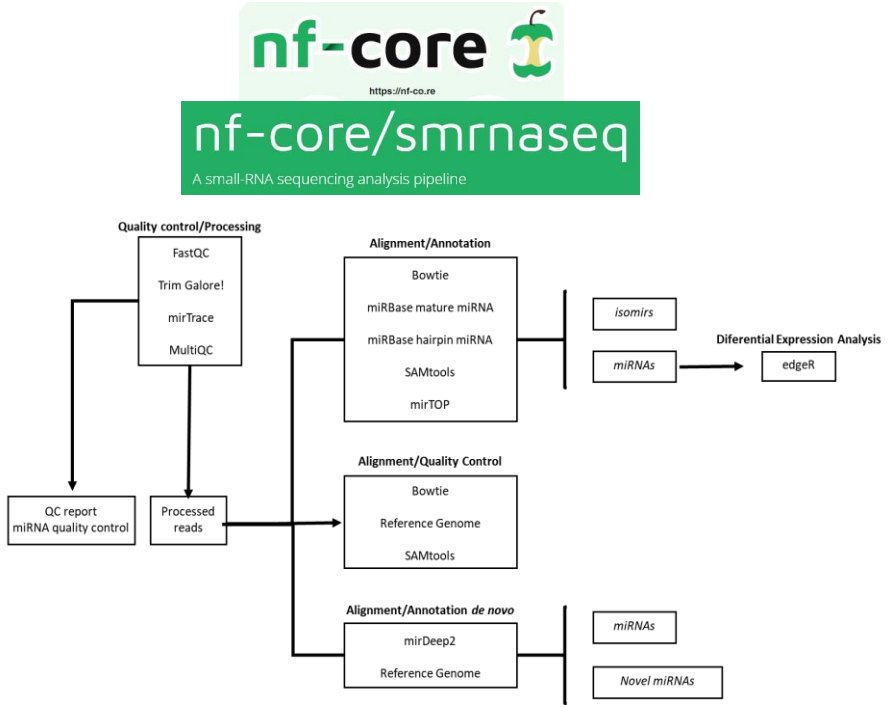


Tipo de *sncRNAs*: *miRNAs, piRNAs, snRNAs, snoRNAs, tRNAs, circRNAs*

Alineamiento de lecturas frente a genoma de referencia

Anotación mediante librerías específicas a partir de miRBase, piRNABank, piRBase, piRNACluster, gtRNAdb, GENCODE release 27, circBase.

nf-core/smrnaseq



Tipo de *sncRNAs*: *miRNAs*

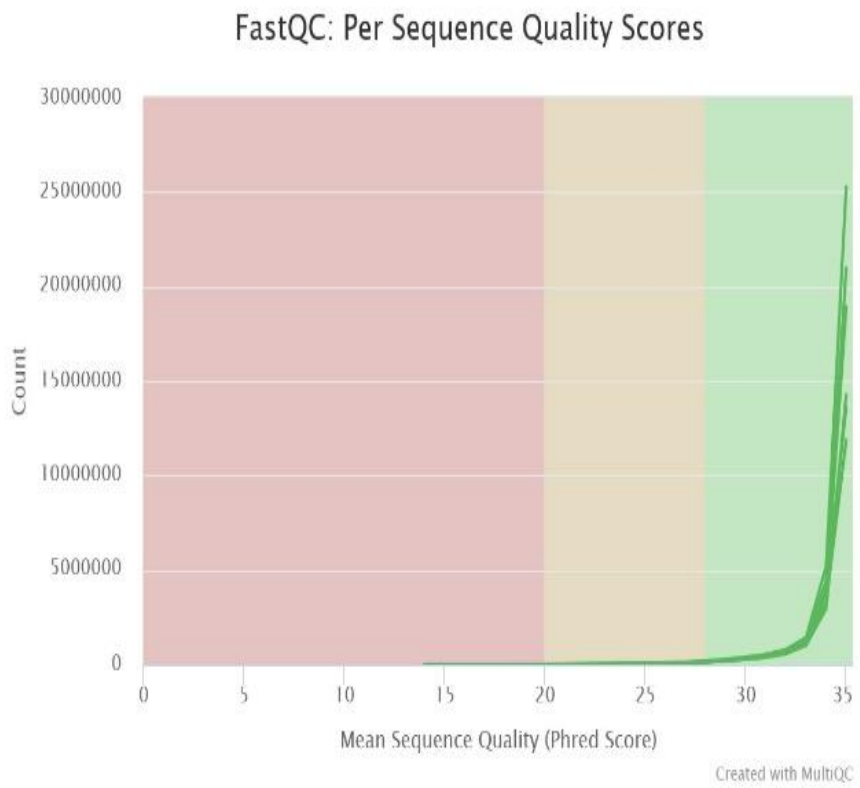
Alineamiento de lecturas frente a librerías de anotación de miRBase: identificación de *miRNAs* y análisis de *isomirs*

Anotación mediante librerías específicas a partir de miRBase

Alineamiento frente a genoma de referencia: control de calidad e identificación de novel *miRNAs*

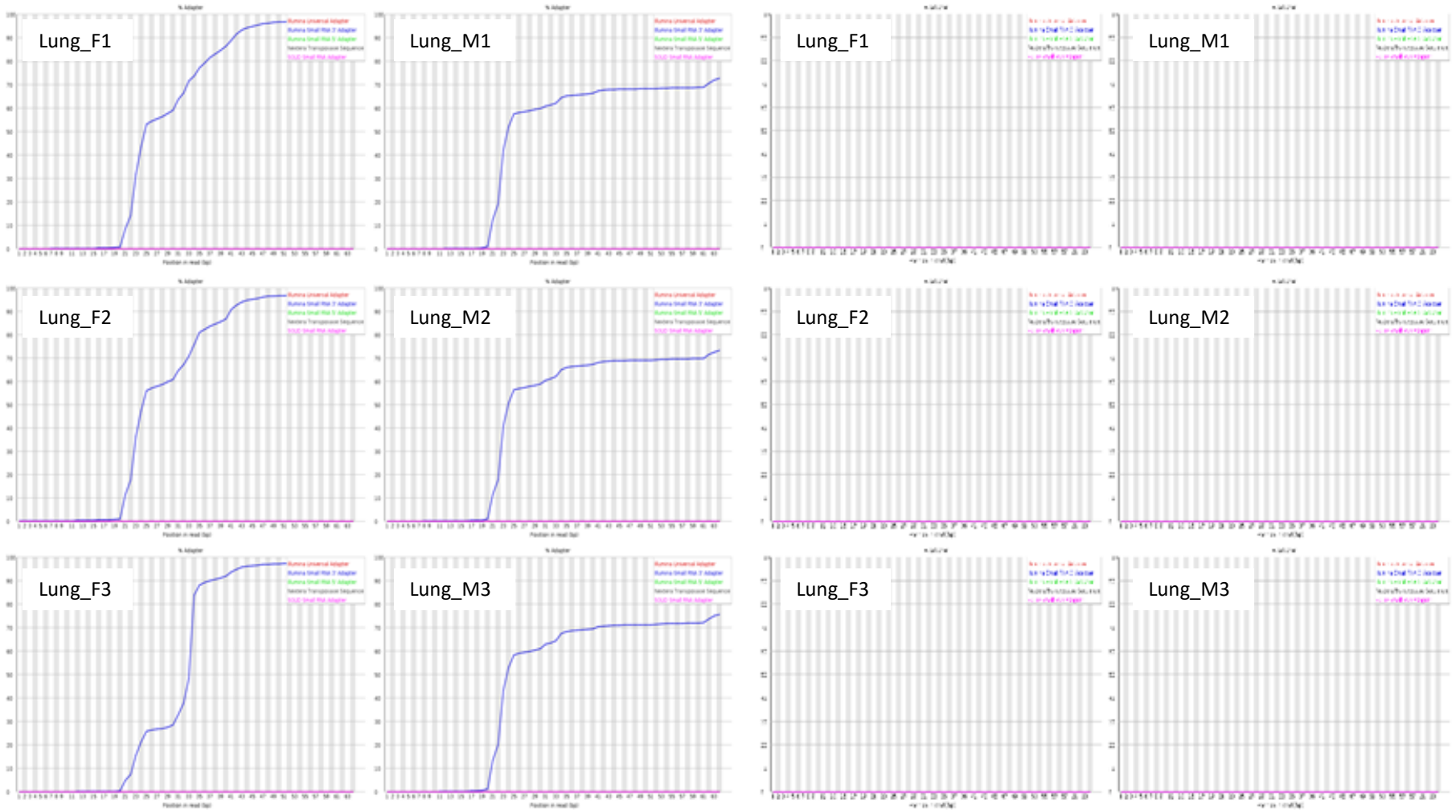
Análisis y control de calidad. FASTQC

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

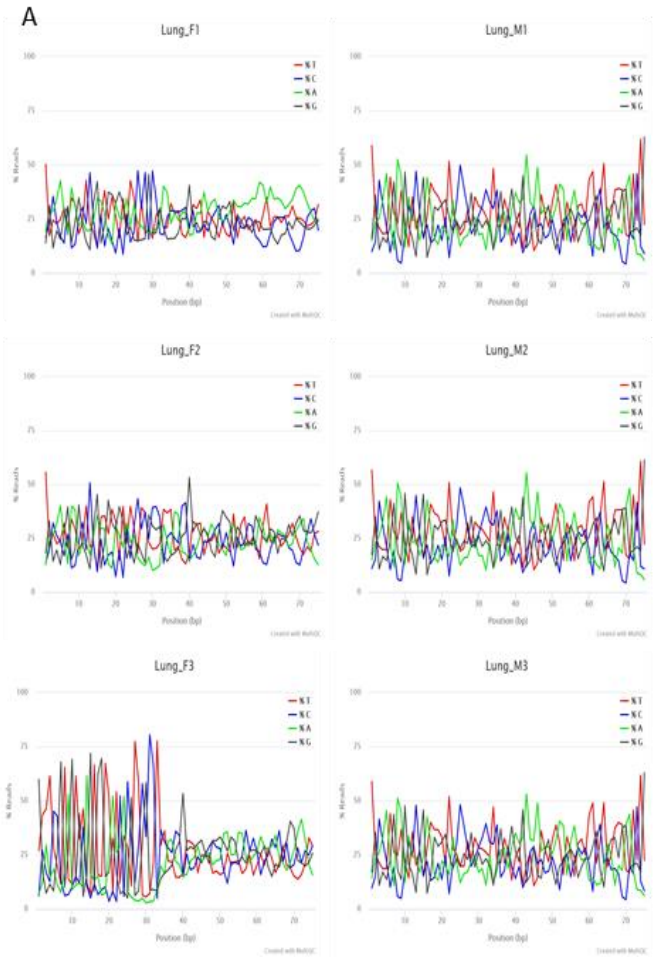


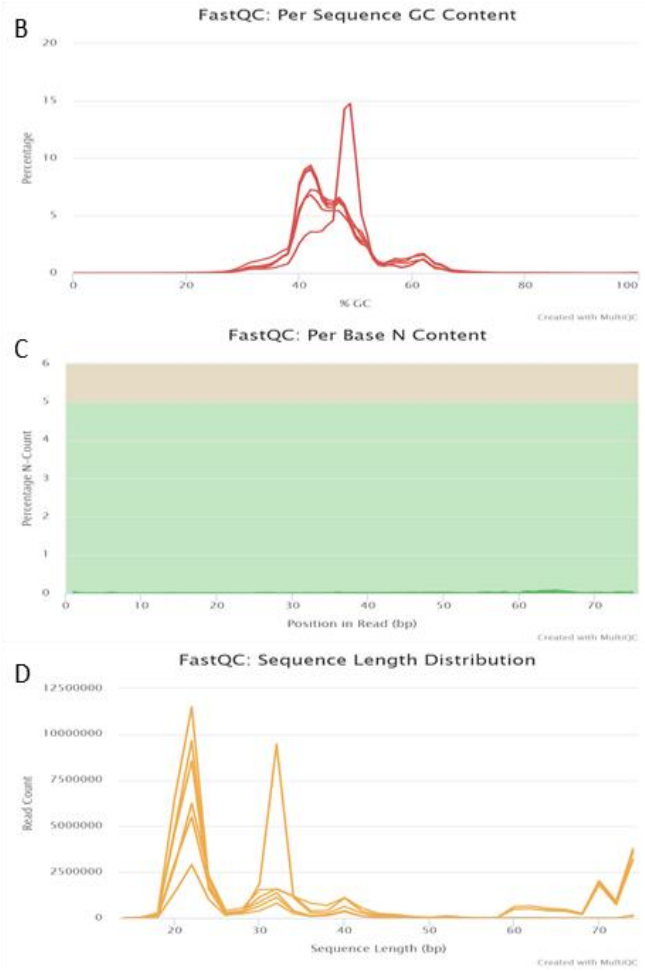
Los valores medios de calidad de las secuencias y de *Phred score* de todas las muestras son correctos

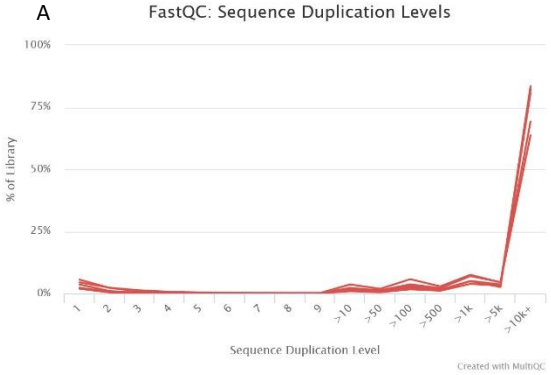
Identificación y eliminación de los adaptadores (Illumina small RNA 3' adapater). Cutadapt 4.0

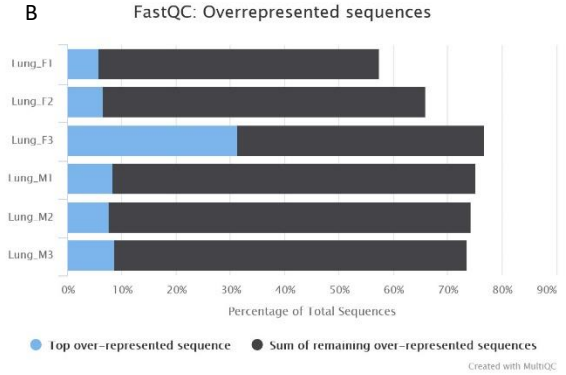


Análisis y control de calidad. FASTQC









Los valores atípicos son consecuencia del tipo de librería:

- Librerías de *sRNAseq*: secuencias cortas 20-76 (nucleótidos)

La muestra Lung_F3 sí es en realidad atípica:

- Alto contenido lecturas de 30 nucleótidos de longitud

Conclusión general del análisis de calidad:

<u>Las muestras son adecuadas para llevar a cabo un análisis de *sncRNAs*</u>

Resultados *pipelines*. Alineamiento de lecturas

miRge3.0

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

COMPSRA

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

nf-core/smrnaseq

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

Mas del 95% de las lecturas válidas y alineadas para los tres *pipelines*

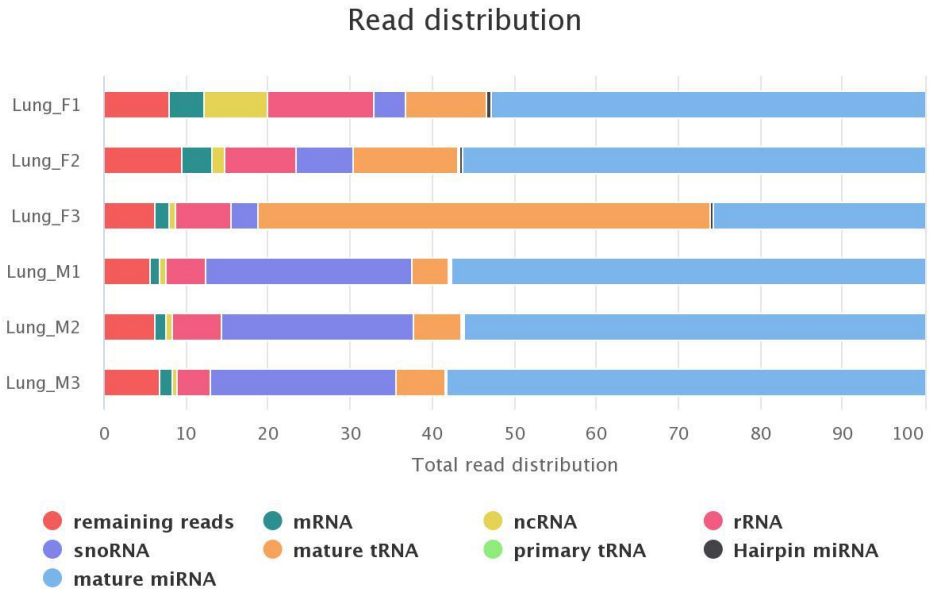
Lecturas alineadas frente a *miRNAs:*

- 52-57% para miRge 3.0
- 70% COMPSRA y nf-core/smrnaseq

La muestra Lung_F3 presenta valores atípicos en el resultado de los tres *pipelines:*

- *miRge 3.0 y* nf-core/smrnaseq : bajo número de lecturas alineadas frente a *miRNAs*
- COMPSRA: : elevado número de lecturas alineadas como *Multiple mapped reads*

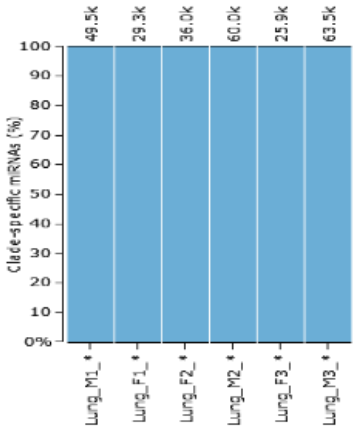
Resultados *pipeline* miRge3.0



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%

Lung_F3 : elevado número de lecturas alineadas frente a *tRNA*

Resultados *pipeline* nf-core/smrnaseq



Clade	Fraction (%)	miRNA families Detected	miRNA families Total
Primates		0	59
Rodents	100.0%	13	16
Birds/Reptiles		0	21
Fish		0	7
Echinoderms		0	9
Lophotrochozoa		0	26
Insects		0	60
Nematode		0	4
Sponges		0	8
Dicots	0.0%	1	175
Monocots		0	62
Gymnosperms		0	6
Lycopods		0	33
Bryophytes		0	67

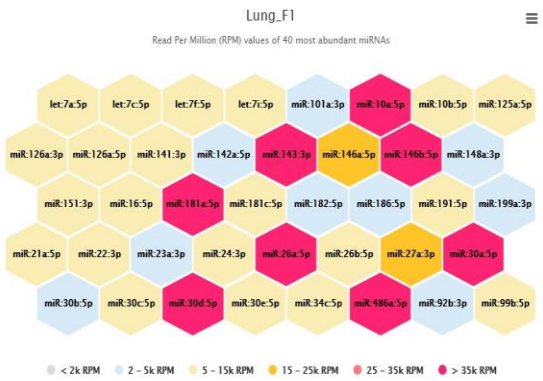
Análisis de calidad con miRTrace

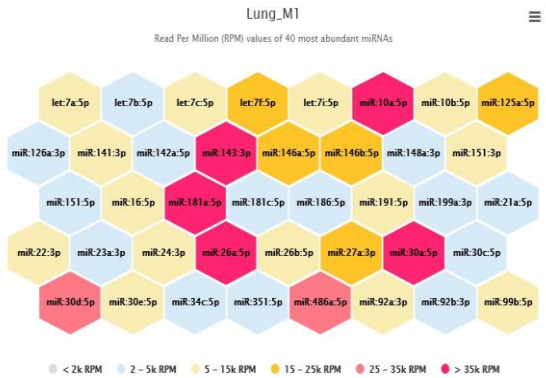
100% lecturas *miRNAs* orden Rodentia

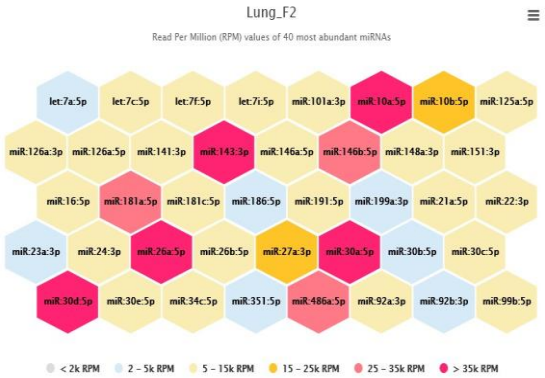
Resultados específicos *pipeline* miRge3.0

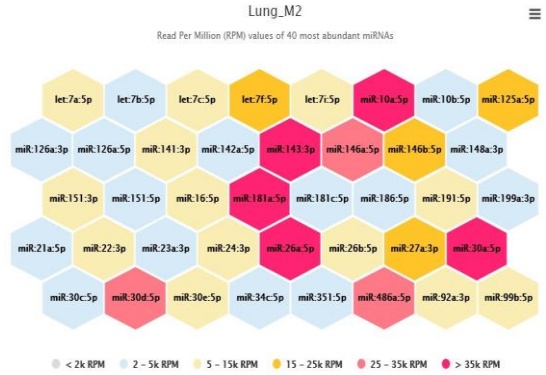
Top 40 *miRNAs* más abundantes en cada una de las muestras

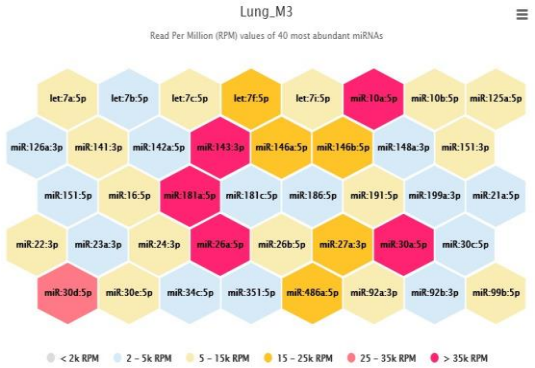
Los *miRNAs* más abundantes son en su mayoría los mismos para todas las muestras, destacando por ejemplo *miR10a-5p, miR143-3p, miR181a-5p, miR26a-5p o miR30a-5p*



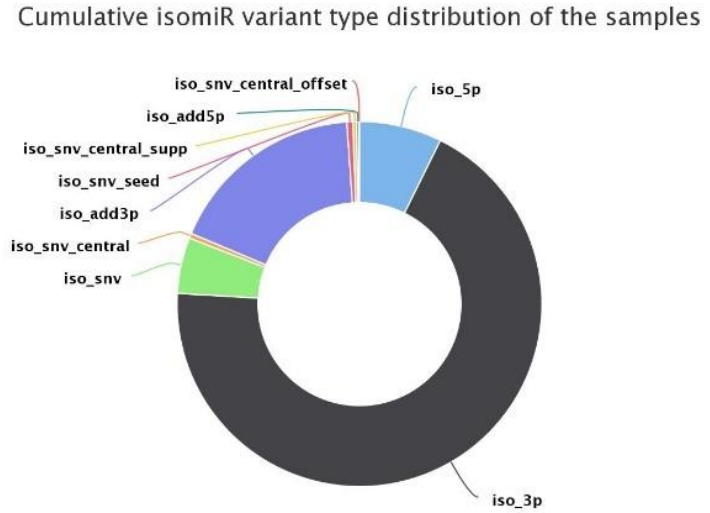






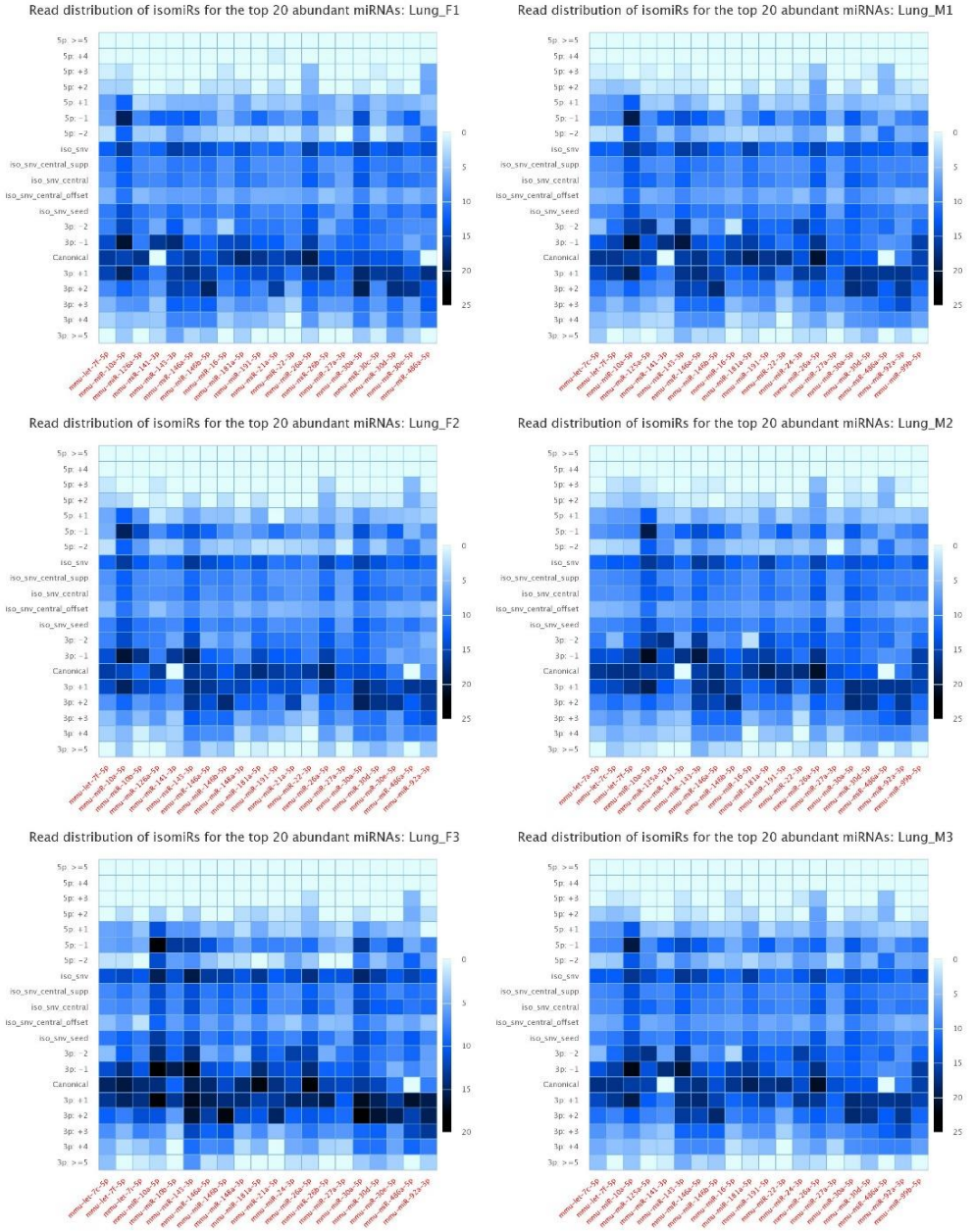


Resultados específicos *pipeline* miRge3.0



Proporción de *isomirs* con respecto a la región de la secuencia en la que se produce las variantes: region 3p

top 20 de *miRNAs* más abundantes: distribución de las lecturas en las regiones de la secuencia donde se producen las variantes



Resultados específicos *pipeline* miRge3.0

Análisis basado en *supporting vector machine* (SVM) *novel miRNAs*

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

miRNA localizado en el cromosoma 14. Identificado en todas las muestras excepto en Lung_F1

Análisis de expresión diferencial. Comparativa inicial entre *pipelines*

Análisis de expresión diferencial de *miRNAs*

Ficheros de contaje obtenidos a partir de cada *pipeline*

Fichero de contaje publicado por Isakova *et al.* 2020

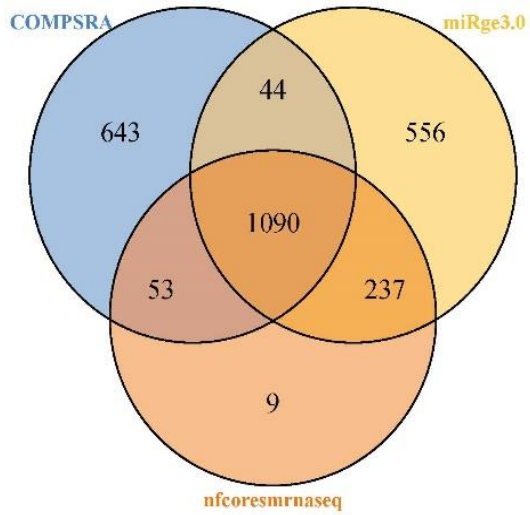
1090 *miRNAs* comunes a los 3 *pipelines*

miRNAs en un solo *pipeline*: COMPSRA y miRge3.0 más *miRNAs* exclusivos

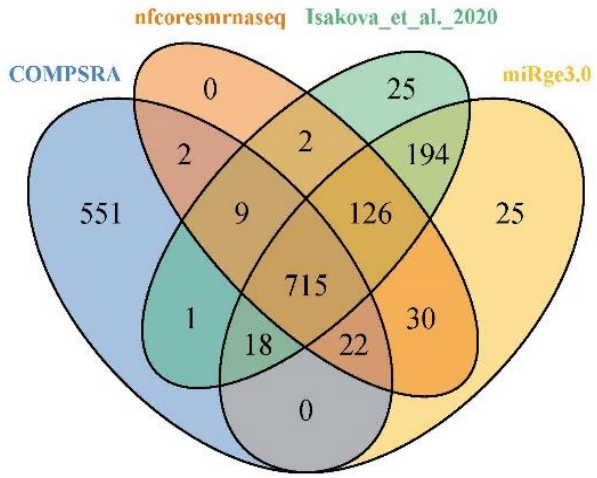
715 *miRNAs* comunes a los 3 *pipelines* con Isakova *et al.* 2020

COMPSRA se mantiene como el más diferente

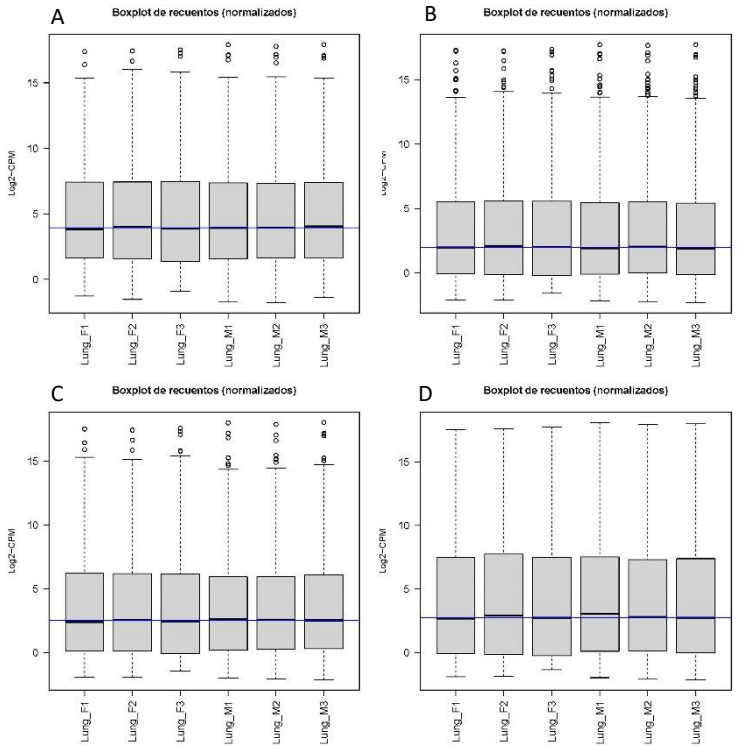
A

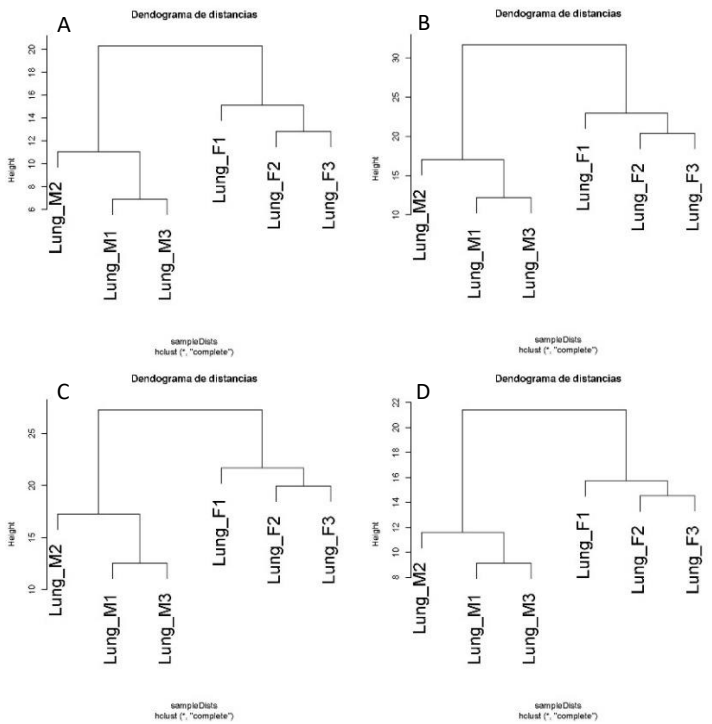


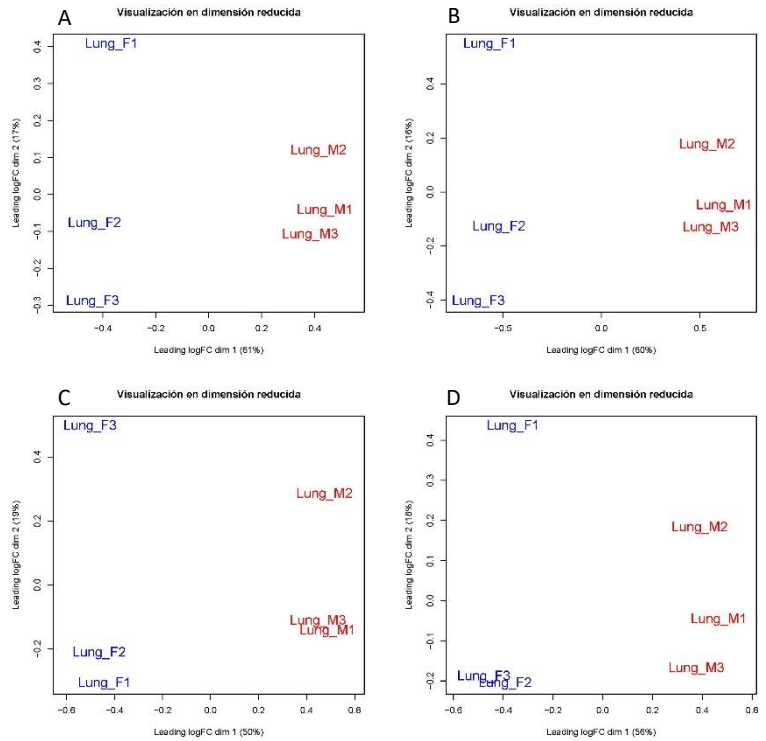
B



Análisis de expresión diferencial. Análisis exploratorio

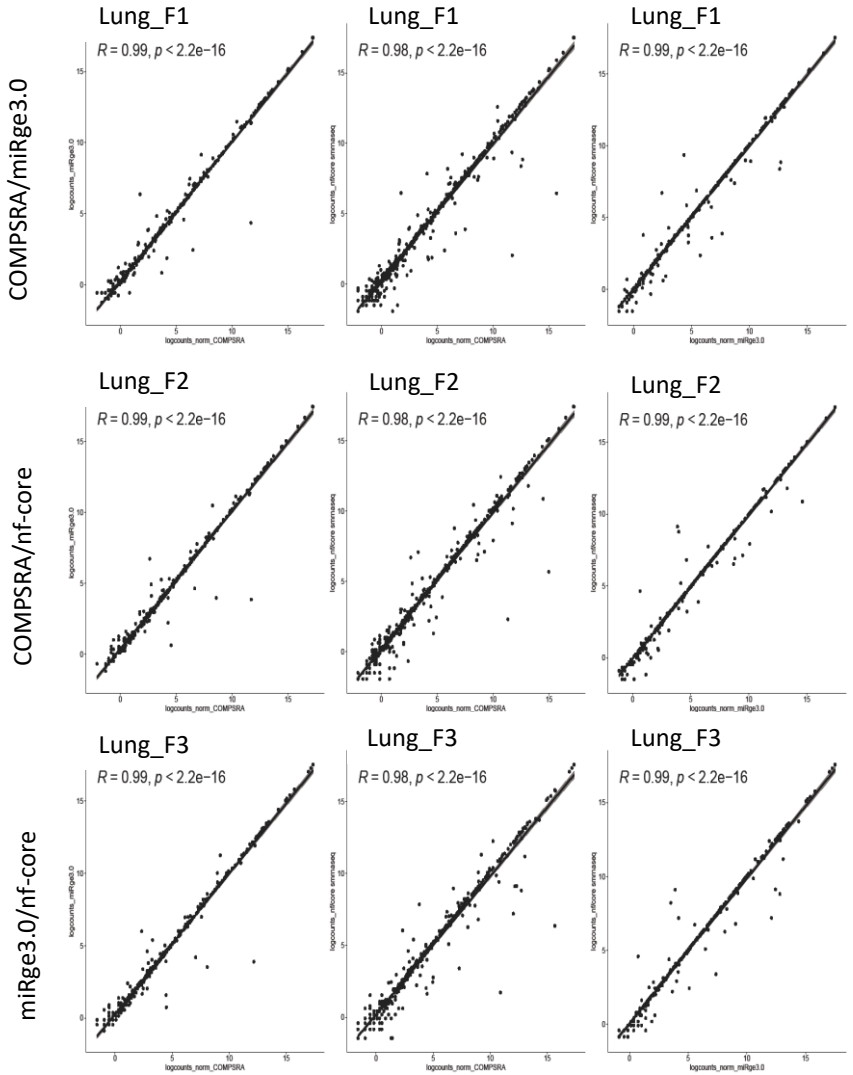






Valores normalizados correctos y agrupamiento adecuado de las muestras de acuerdo a los grupos macho y hembra en todos los casos

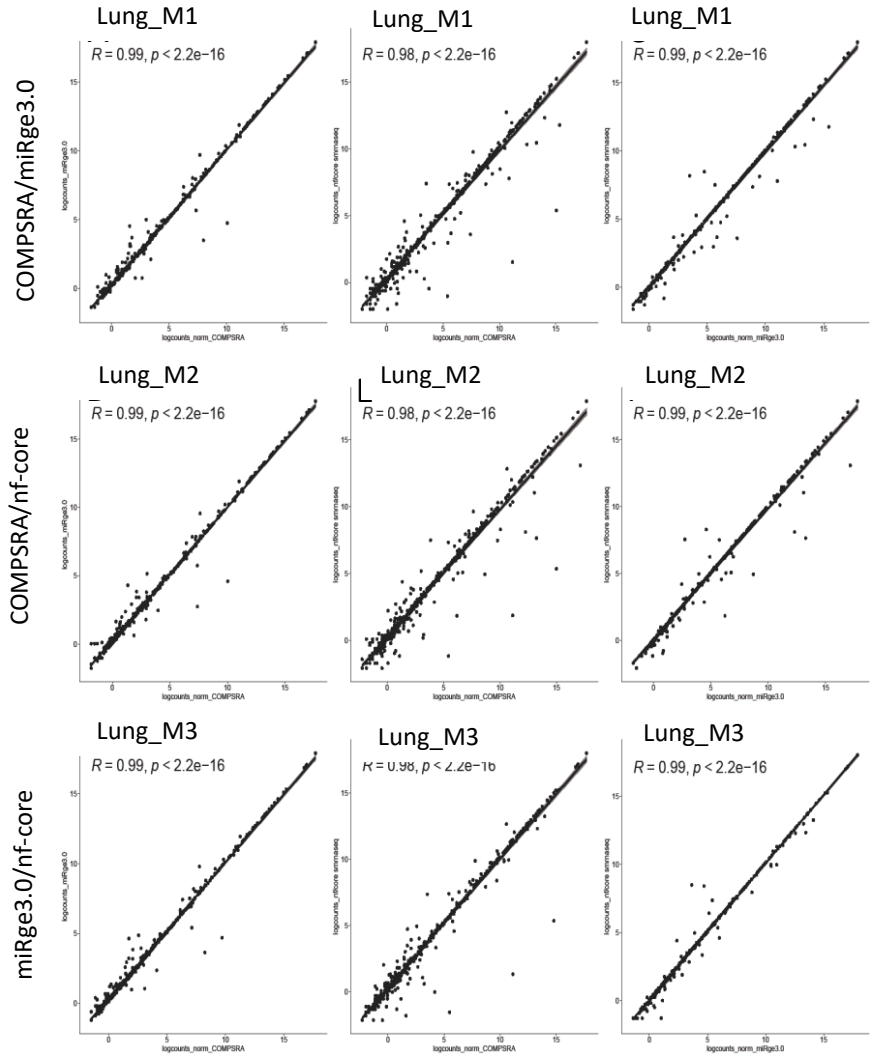
Análisis de expresión diferencial. Correlación de los perfiles de expresión



Perfil de expresión de *miRNAs* comunes entre *pipelines*

Valores de correlación entre 0,98 y 0,99 en todas las comparaciones

La cuantificación llevada a cabo por los tres *pipelines* a partir del alineamiento de lecturas es muy similar.



Análisis de expresión diferencial. *miRNAs* diferencialmente expresados

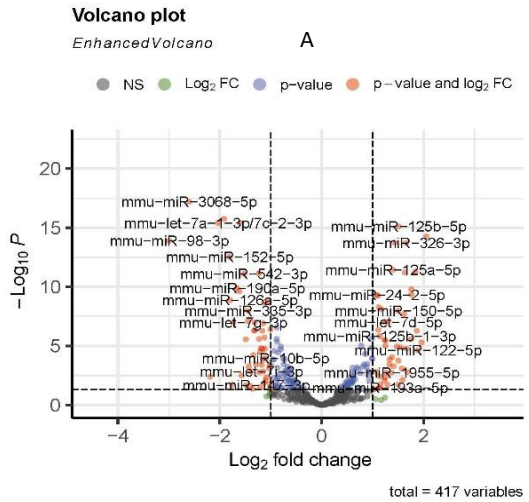
miRNAs diferencialmente expresados ***DESEQ2***: *log2FC* absoluto mayor de 1 y *padj* menor de 0.05

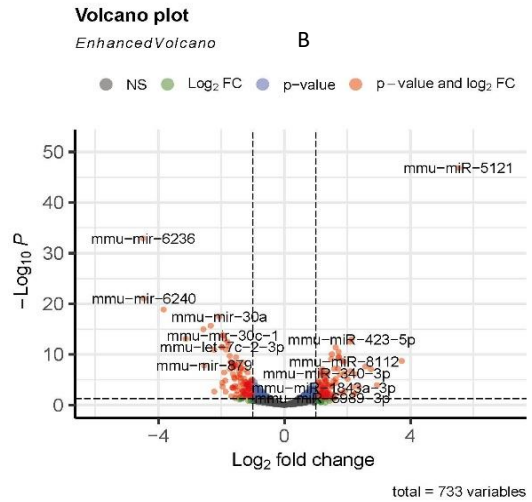
Pipeline	Total	*Upregulated*	*Downregulated*	No significativos
miRge3.0	417	46	47	324
COMPSRA	733	77	93	563
nf-core/smrnaseq	558	56	65	437
Isakova *et al.* 2020	355	26	33	296

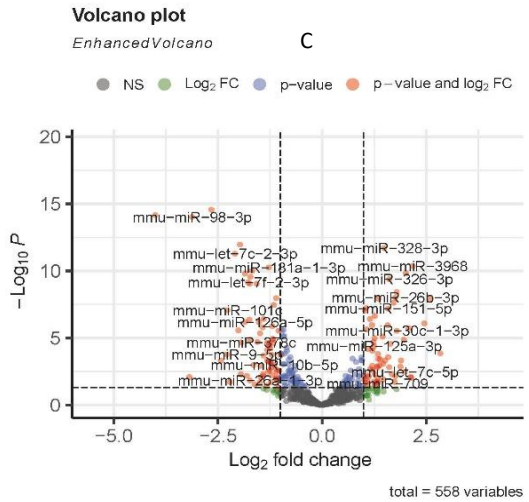
COMPSRA identifica el mayor número de *miRNAs* diferencialmente expresados

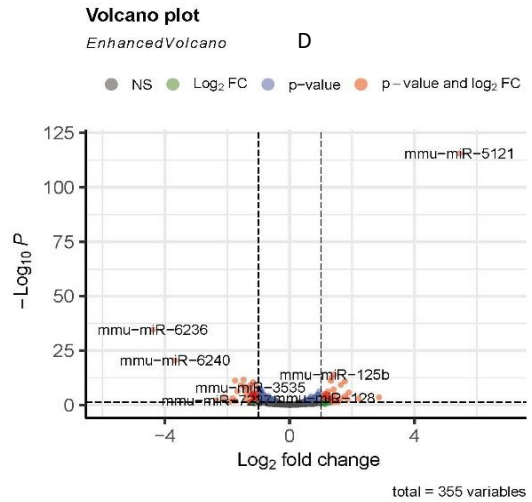
Significativos (log2FC absoluto mayor de 1 y padj menor de 0.05):

- COMPSRA identifica el mayor número
- el número de *upregulated* y *downregulated* es bastante similar para un mismo *pipeline*





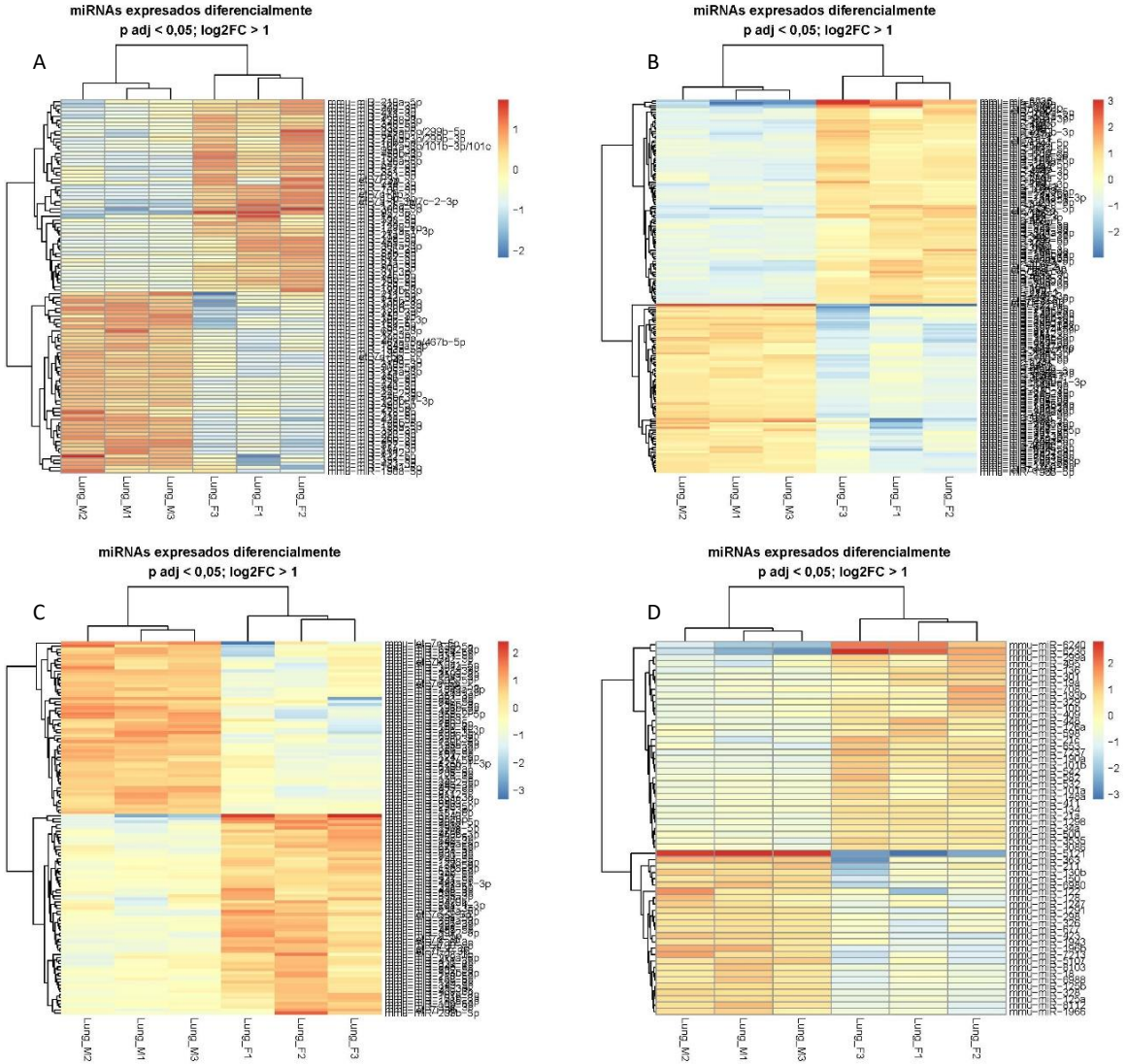




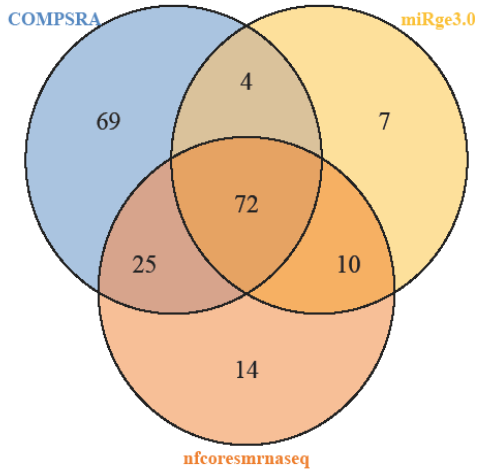
Análisis de expresión diferencial. Comparativa *miRNAs* diferencialmente expresados y significativos

Heatmap*: miRNAs* diferencialmente expresados y significativos (*log2FC* absoluto mayor de 1 y *padj* menor de 0.05)

Se pueden identificar perfiles de expresión claramente específicos de sexo

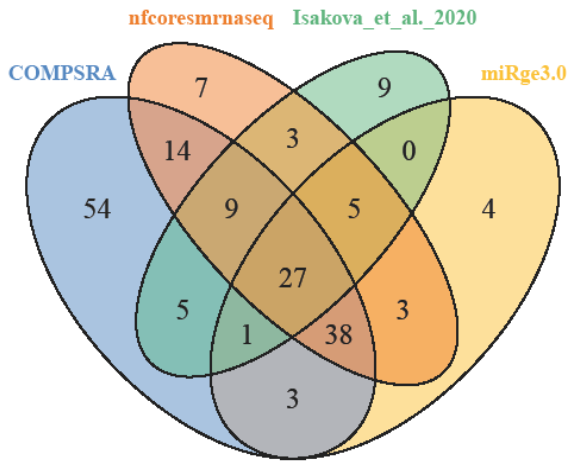


Análisis de expresión diferencial. Comparativa *miRNAs* diferencialmente expresados y significativos



miRNAs

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



miRNAs

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

- La variedad en los *pipelines* disponibles hace que tanto el tipo de *sncRNAs* que se quieren analizar como el tipo de análisis que se quiera realizar sean aspectos importantes a tener en cuenta a la hora de seleccionar un *pipeline* de trabajo.
- Una posibilidad es analizar los datos con más de un *pipeline* y consensuar resultados, lo que puede hacer que estos sean más fiables o completos.
- Los tres *pipelines* utilizados han tenido un rendimiento en el alineamiento de lecturas similar, y superior al 95%, independientemente de si la estrategia de alineamiento está basada en un genoma de referencia o en librerías de anotación específicas.
- Sería necesario evaluar otros *pipelines* o utilizar otros *dataset* más complejos para confirmar si existen o no diferencias significativas entre ambas estrategias de alineamiento.
- 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.
- 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 análisis de expresión diferencial.

https://github.com/asroco/TFM

Agradecimientos

Servidor para la ejecución de los *pipelines*

Tutora: Mireia Ferrer Almirall

Muchas gracias por su atención