

Transcriptómica de pequeños ARNs no codificantes

sRNA-seq

Rafael Fort

Sección Genómica Funcional - FCIEN

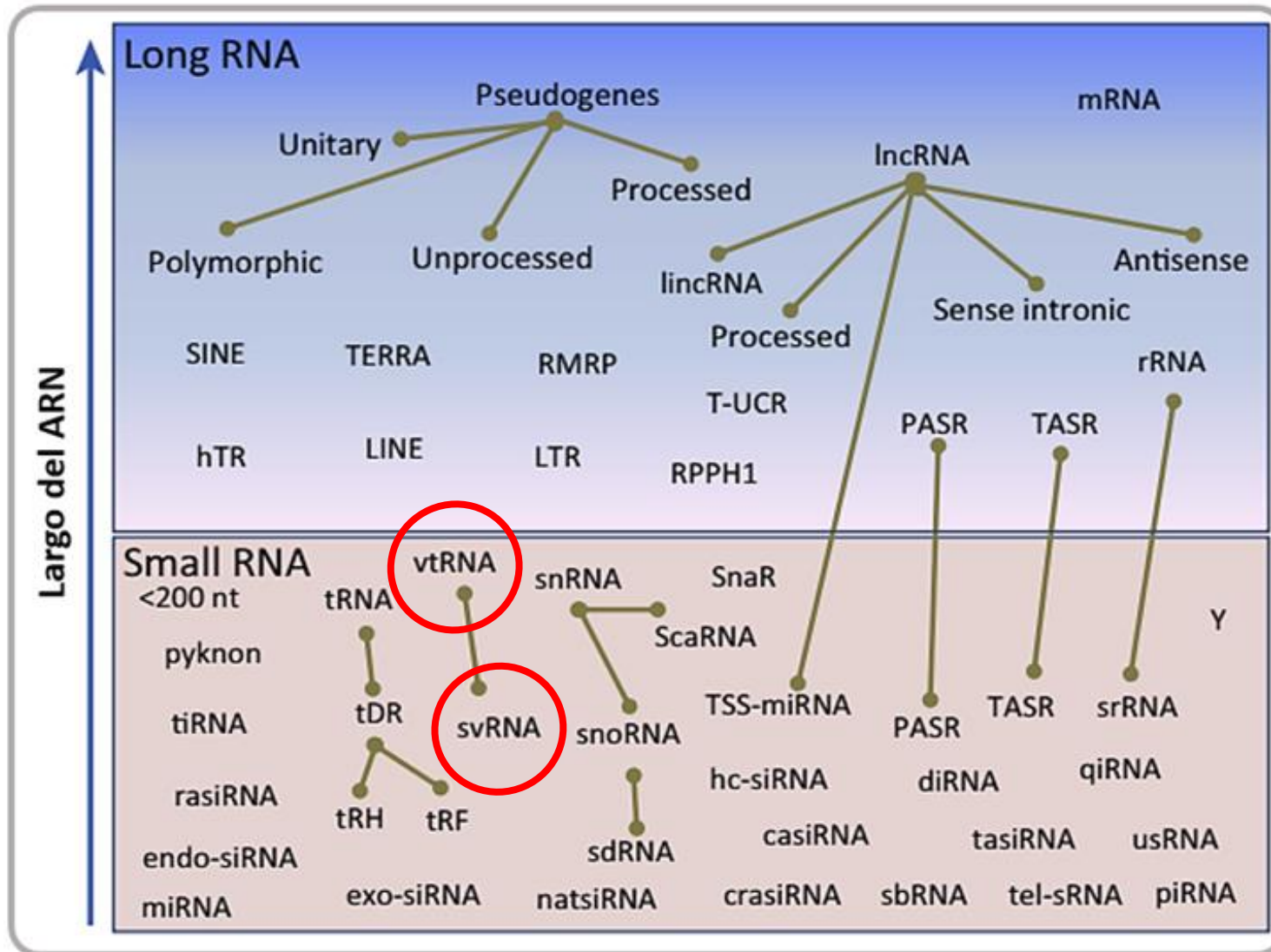
Depto. Genómica - IIBCE



Departamento de Genómica

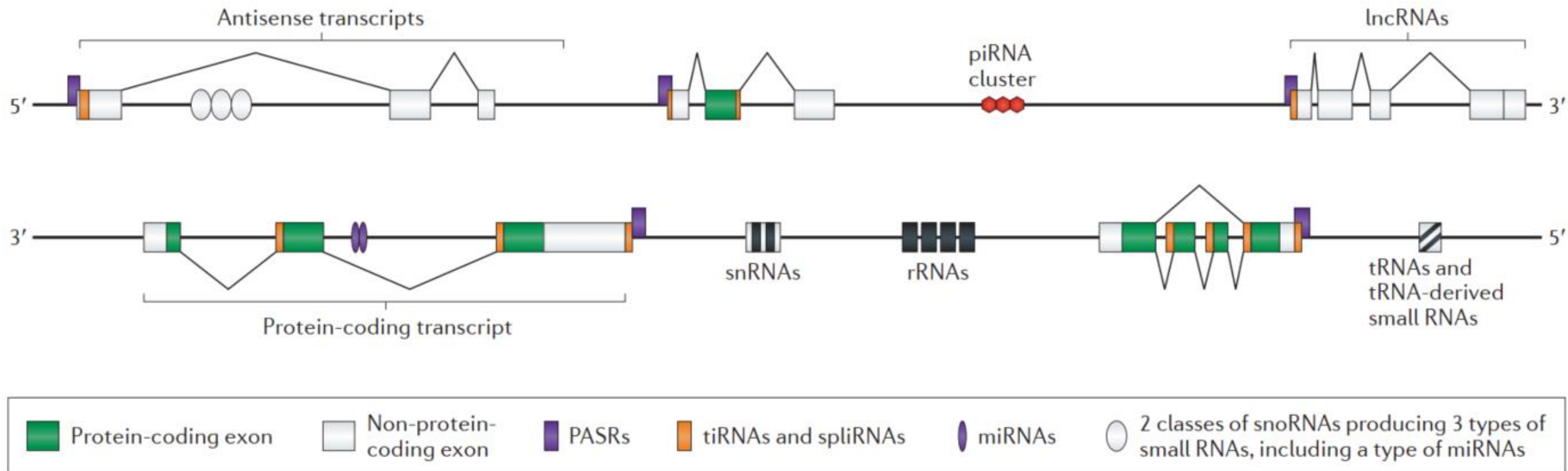
Instituto de Investigaciones Biológicas
Clemente Estable

Existe una gran diversidad de ARNs no codificantes

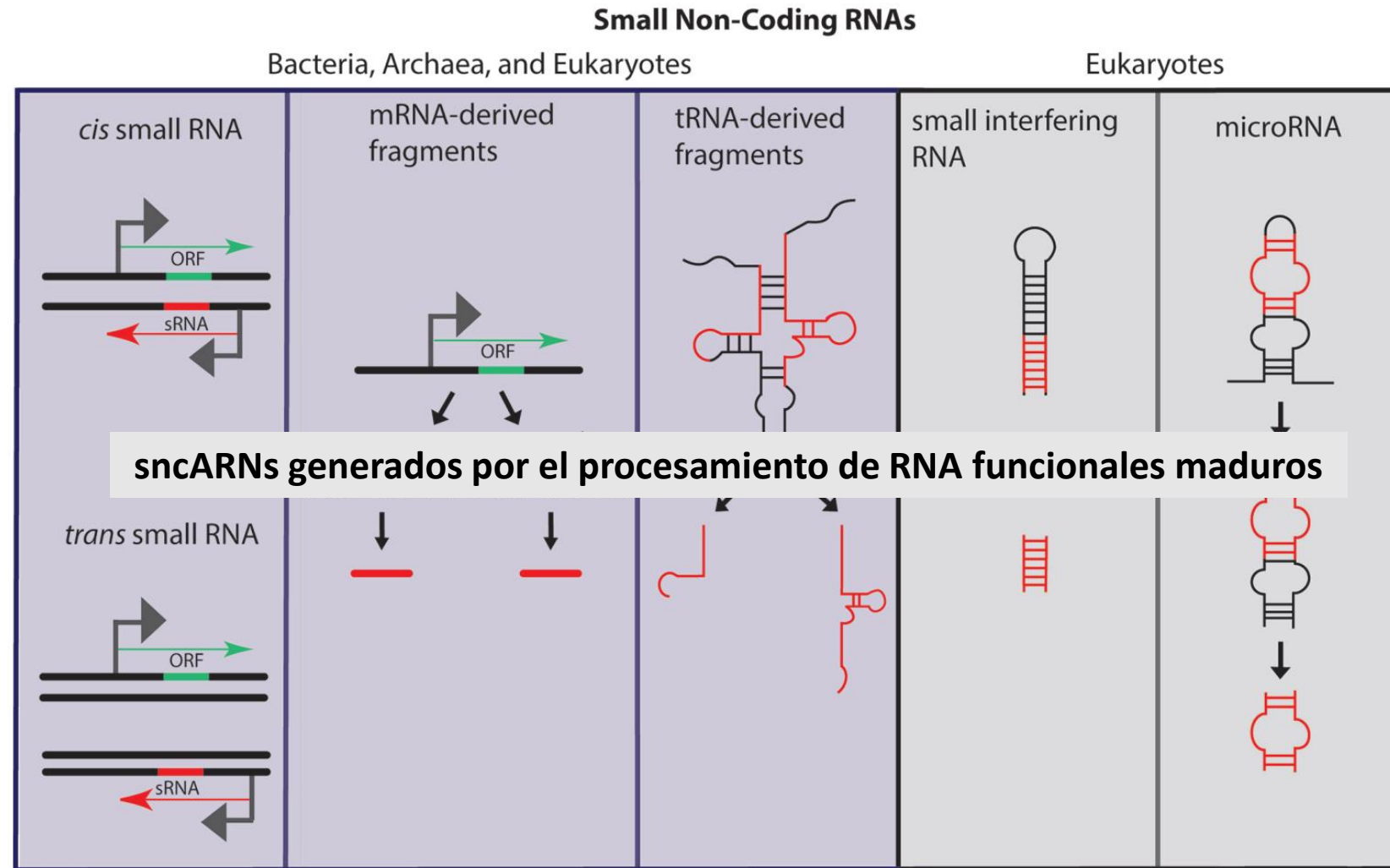
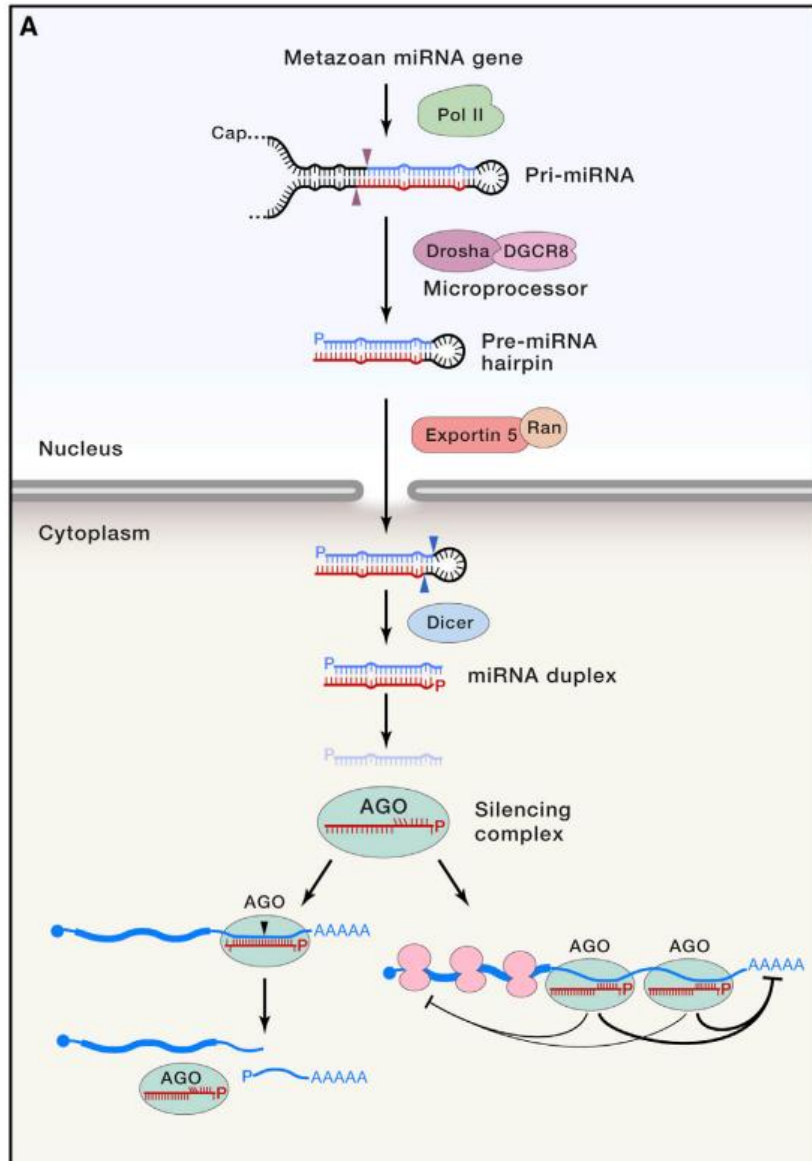


- Conocemos muchas **clases de ncARNs** presentes en el **transcriptoma celular**.
- Se clasifican según su **tamaño y función**, longitud.
 - **Small RNAs (sncARNs) <50 nt**
 - microARNs, piwiARNs, tRFs
 - **Mid-sized RNAs (mncRNAs) 50-400 nt**
 - **Long RNAs (lncARNs) >400nt**
 - lincARNs, rARN

Panorámica transcripcional de los ARNs no codificantes



microARNs y ARN pequeños derivados de ARN no codificantes más largos

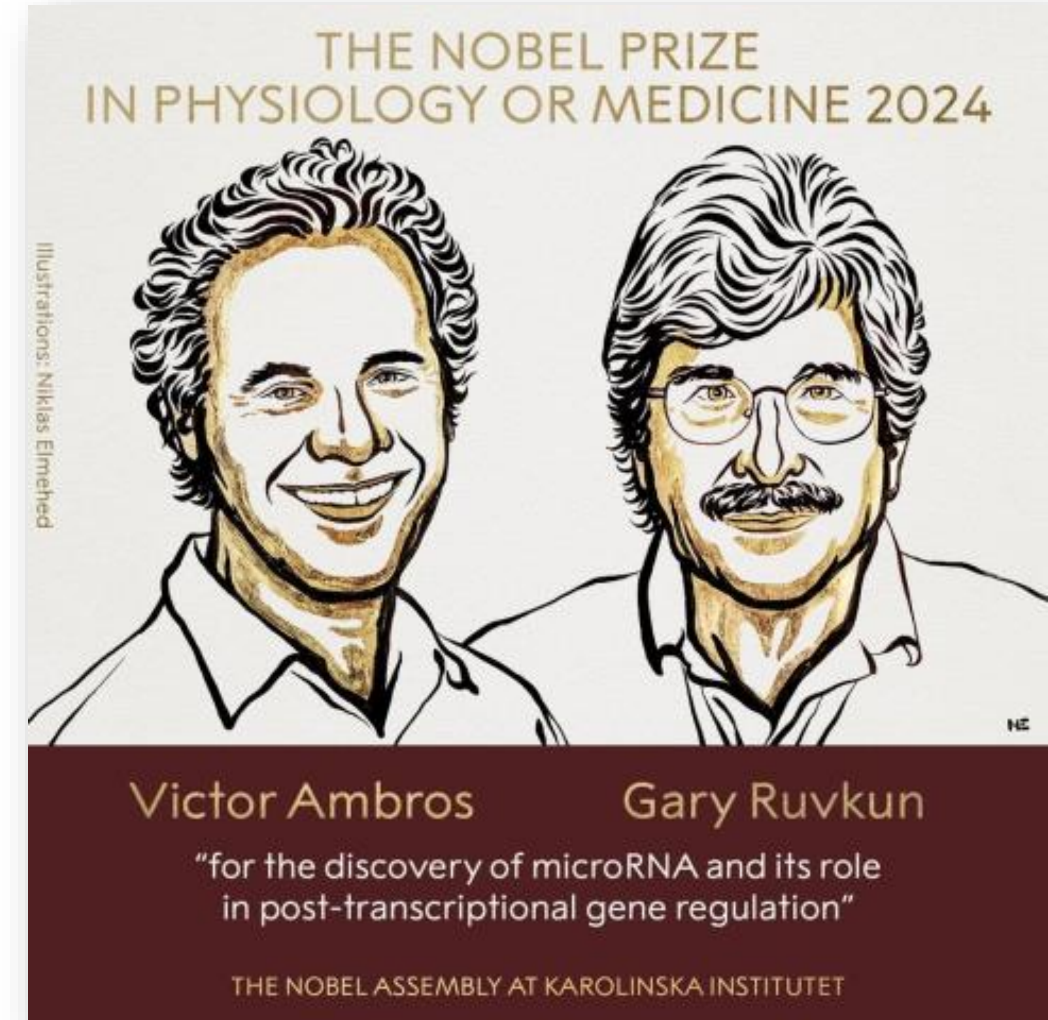
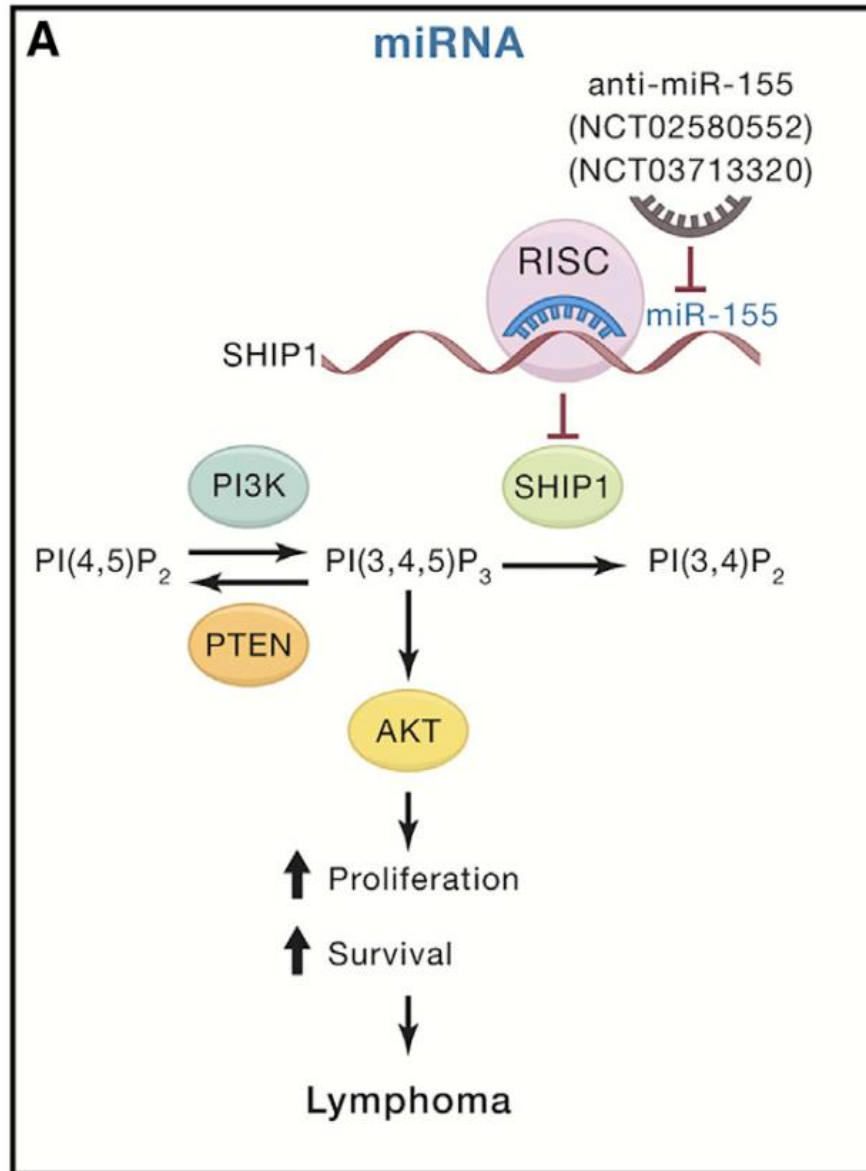
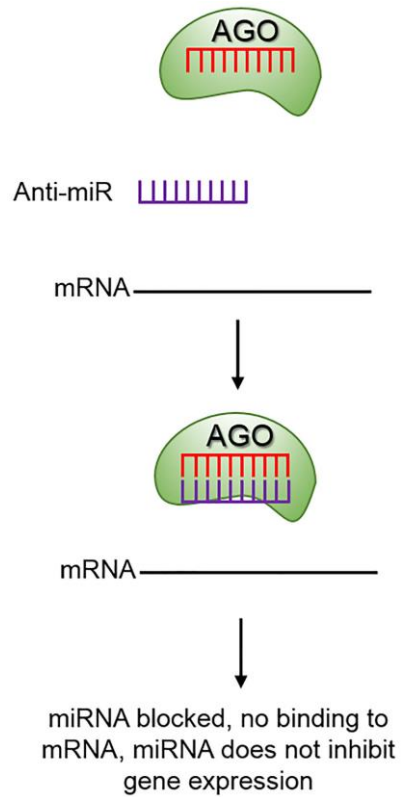


% de Genoma No-codificante:

Procariotas: 1-15%

Eucariotas: 50-80%

Herramientas para el bloqueo y modulación de ncARNs



Resource

2007

Cell

A Mammalian microRNA Expression Atlas Based on Small RNA Library Sequencing

Pablo Landgraf,¹ Mirabela Rusu,² Robert Sheridan,³ Alain Sewer,^{2,19,29} Nicola Iovino,^{1,27} Alexei Aravin,^{1,26} Sébastien Pfeffer,^{1,25} Amanda Rice,¹ Alice O. Kamphorst,¹ Markus Landthaler,¹ Carolina Lin,¹ Nicholas D. Socci,³ Leandro Hermida,² Valerio Fulci,⁴ Sabina Chiaretti,⁴ Robin Foà,⁴ Julia Schliwka,⁵ Uta Fuchs,⁶ Astrid Novosel,^{6,28} Roman-Ulrich Müller,^{1,7} Bernhard Schermer,⁷ Ute Bissels,⁸ Jason Inman,⁹ Quang Phan,¹⁰ Minchen Chien,¹¹ David B. Weir,¹¹ Ruchi Choksi,¹¹ Gabriella De Vita,¹² Daniela Frezzetti,¹² Hans-Ingo Trompeter,¹³ Veit Hornung,^{23,24} Grace Teng,¹⁴ Gunther Hartmann,¹⁸ Miklos Palkovits,¹⁵ Roberto Di Lauro,^{12,20} Peter Wernet,¹³ Giuseppe Macino,⁴ Charles E. Rogler,¹⁶ James W. Nagle,²² Jingyue Ju,^{11,21} F. Nina Papavasiliou,¹⁴ Thomas Benzing,⁷ Peter Lichter,⁵ Wayne Tam,¹⁷ Michael J. Brownstein,¹⁰ Andreas Bosio,⁸ Arndt Borkhardt,^{6,28} James J. Russo,¹¹ Chris Sander,³ Mihaela Zavolan,^{2,19,*} and Thomas Tuschl^{1,*}

Table 1. Continued

Human miRNA	Sequence	Clone #	Mouse miRNA	Sequence	Clone #	Rat miRNA	Sequence	Clone #
			miR-466c	GAUGUGUGUGU GCAUGUACAUA	11			
						miR-466c	UGUGAUGUGUG CAUGUACAUG	2
			miR-467e	AUAAGUGUGAG CAUGUAUAUGU	125			
			miR-466d	UGUGUGUGCGU ACAUGUACAUG	4			
			miR-466g	AUACAGACACA UGCACACACA	2			
			miR-466h	UGUGUGCAUGU GCUUGUGUGUA	4			
miR-885	UCCAUUACACU ACCCUGCCUCU	5						
miR-887	GUGAACGGGCG CCAUCCCGAGG	3						
miR-886	CGCGGGUGCUU ACUGACCCUU	16						

Desde el descubrimiento del primer miARN lin-4 en nematodos, se han detectado más de **38.000 miARN en 271 organismos, incluidos 1917 miARN humanos (miRBase v22)**

Última versión de **miRBase v22** están anotados **1917** miARN humanos

Perspective

The status of the human gene catalogue

<https://doi.org/10.1038/s41586-023-06490-x>

Received: 9 March 2023

Accepted: 27 July 2023

Published online: 4 October 2023

Paulo Amaral¹, Silvia Carbonell-Sala², Francisco M. De La Vega^{3,4}, Tiago Faial⁵, Adam Frankish⁶, Thomas Gingeras⁷, Roderic Guigo^{2,8}, Jennifer L. Harrow⁹, Artemis G. Hatzigeorgiou^{10,11}, Rory Johnson^{12,13,14,15}, Terence D. Murphy¹⁶, Mihaela Pertea^{17,18,19}, Kim D. Pruitt¹⁶, Shashikant Pujar¹⁶, Hazuki Takahashi²⁰, Igor Ulitsky^{21,22}, Ales Varabyou^{17,19}, Christine A. Wells²³, Mark Yandell²⁴, Piero Carninci^{20,25} & Steven L. Salzberg^{17,18,19,26}

- El número de genes que codifican proteínas se está estabilizando en alrededor de **19.500**.
- Aunque el número de isoformas de estos genes sigue siendo un tema de intenso estudio y discusión.
- Los ncRNAs se encuentran en una **etapa más temprana de comprensión** y su número sigue aumentando.

Table 1 | Annotation databases that catalogue long ncRNA genes

Resource	lncRNA genes
RefSeq ¹¹	17,948
GENCODE ¹⁰	19,933
NONCODE ⁷⁶	96,411
FANTOM CAT ⁷⁷	27,919
LNCipedia ⁷⁸	56,946
miTranscriptome ⁷⁹	58,648
CHESS ¹²	17,623
LncBook ⁸⁰	95,243

Figures as of late 2022. 'Long' refers to loci that are at least 200 nucleotides.

1990

1999

2009

2019

2029

Year

¿Por qué NGS? - small RNA-seq (*sRNA-seq*)

Ofrece una mayor sensibilidad y precisión que las tecnologías anteriores. Su sensibilidad y especificidad están por encima de las técnicas de *microarrays*.

- Secuenciación exacta, **resolución de nucleótidos**
- Identificación de **variaciones en longitud** o composición
- **No requiere información previa** del transcripto
- Se pueden detectar eventos de **edición de ARN**
- Identificación de **polimorfismos** (expresión específica de alelo)
- Proporciona una gran profundidad de cobertura (**rango dinámico**)

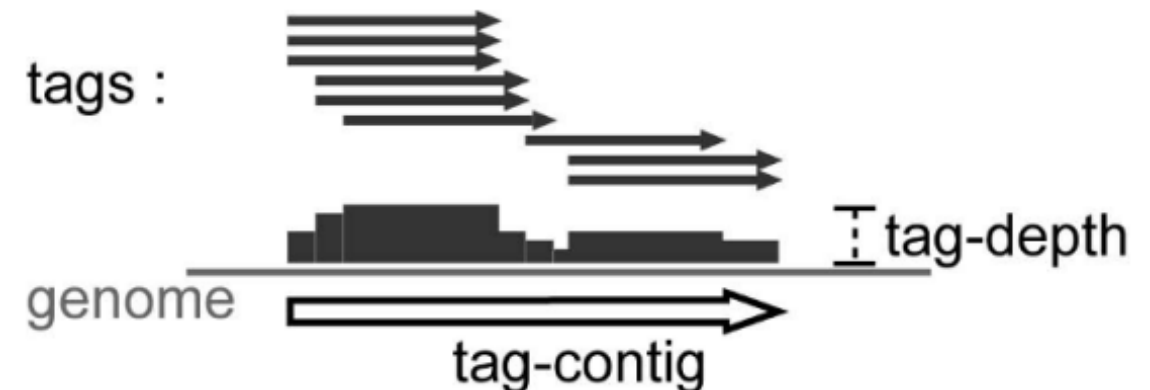
1. Aislamiento de ARN
2. Construcción de biblioteca de ADNc
3. Secuenciación

Published online 15 October 2008

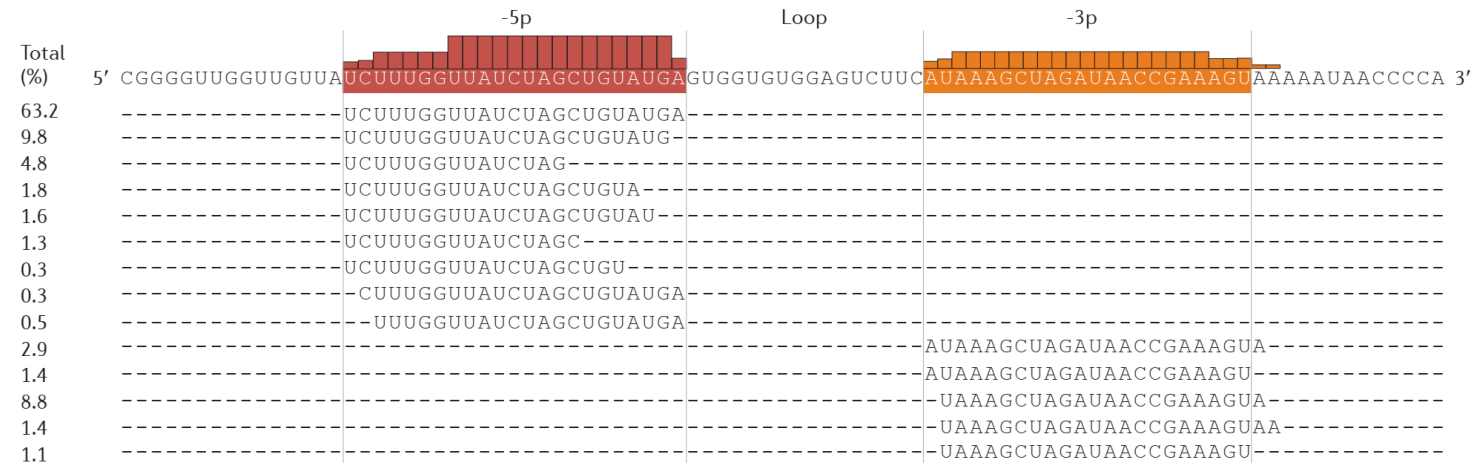
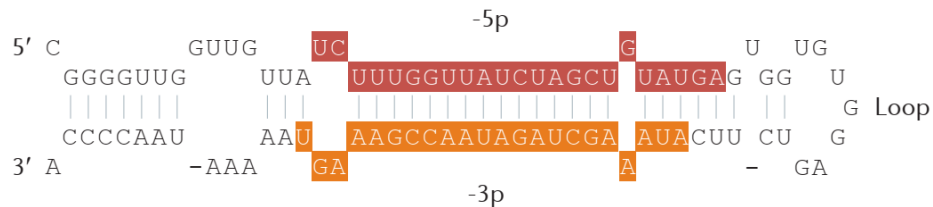
Nucleic Acids Research, 2008, Vol. 36, No. 21 e141
doi:10.1093/nar/gkn705

Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms

Peter A. C. 't Hoen^{1,*}, Yavuz Ariyurek¹, Helene H. Thygesen¹, Erno Vreugdenhil², Rolf H. A. M. Vossen¹, Renée X. de Menezes¹, Judith M. Boer¹, Gert-Jan B. van Ommen¹ and Johan T. den Dunnen¹



NON-CODING RNA

Colin C. Pritchard¹, Heather H. Cheng^{2,3} and Muneesh Tewari^{2,3,4,5}

- La **biogénesis de miARN maduros** se produce mediante un proceso de varios pasos (nucleasas DROSHA y DICER).
- Rango de tamaños (**16-25 nts**)
- El **número promedio de copias** de un miARN individual se ha estimado aproximadamente en ~500 por célula (mayor que mRNAs).
- **Rango dinámico** de al menos cuatro órdenes de magnitud.
- Algunos **miARN específicos de un tipo celular** están presentes hasta en **>10 000 copias por célula**.
- Se pueden producir **diferentes especies de miARN maduros** a partir de una única molécula de pre-miARN.

Consideraciones de la muestra para small RNA-seq (sRNA-seq)

- Los métodos de **procesamiento de muestras y extracción de ARN** pueden tener un **impacto sustancial** en los resultados de los miARN (degradación).
- Los productos disponibles comercialmente y ampliamente utilizados se basan generalmente en la extracción química utilizando **sales caotrópicas concentradas**, como el **tiocianato de guanidinio** (Trizol and QIAzol reagents).

Table 1 | **Sample considerations**

Sample type	Typical miRNA yield	Considerations
Cell lines	>1,000 ng	High-quality miRNA (usually)
Fluorescence-activated cell sorting	1–100 ng	Lower yield but less cell-type heterogeneity
Fresh tissue (for example, tumour tissue)	>1,000 ng	Cell-type heterogeneity
Formalin-fixed paraffin-embedded tissue	1–100 ng	More reliable than mRNA as an analyte in FFPE
Laser capture microdissected tissue	<1–10 ng	Less heterogeneity but lower yield, never completely pure
Plasma or serum	<1–10 ng	RNases, low yield, typically cannot measure quantity of extracted RNA
Urine	<1–10 ng	Can evaluate cell pellet versus supernatant

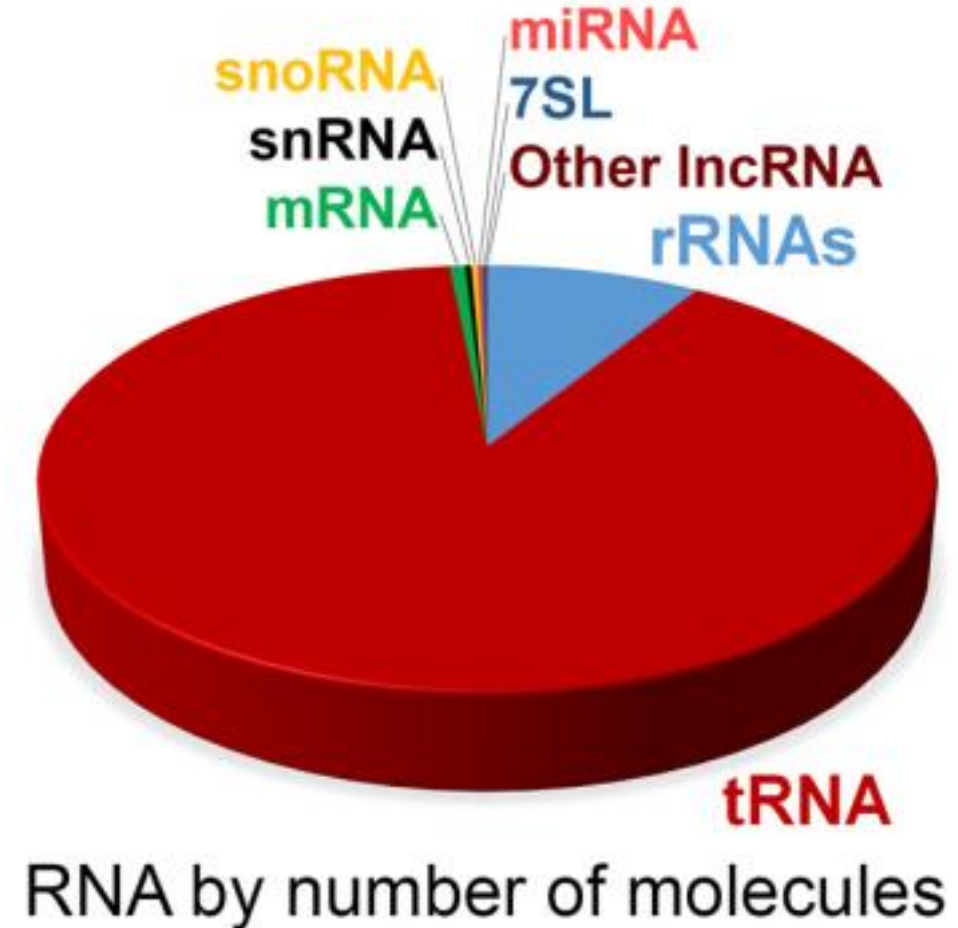
Los pasos de purificación siempre pueden introducir sesgos

Los miARN representan una pequeña fracción (~0,01%) de la masa total de ARN celular

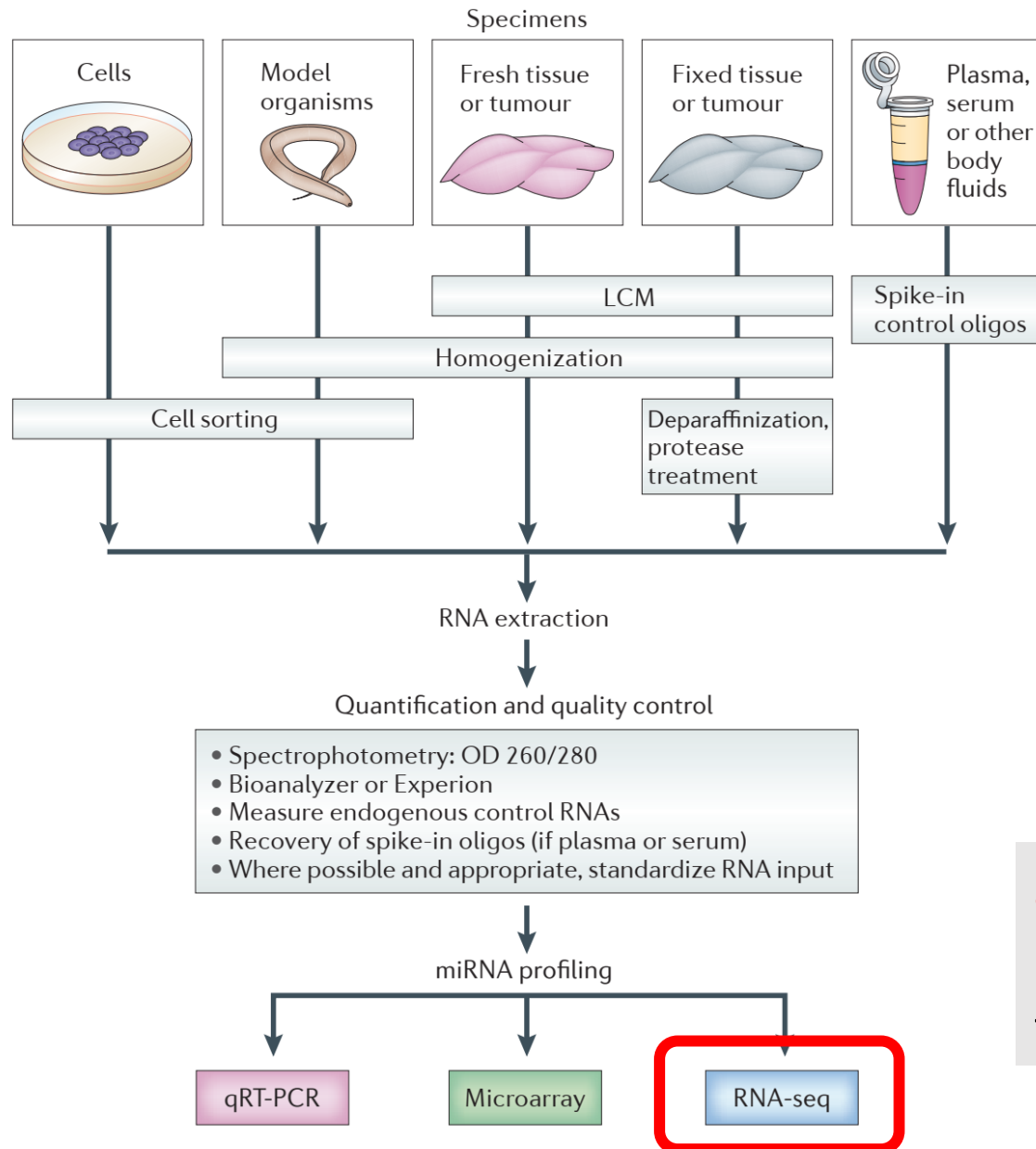
Estimación de los niveles de ARN en una célula de mamífero típica

El número total de moléculas de ARN se estima en aproximadamente 10^7 por célula

rRNA	8.9 %
tRNA	89 %
mRNA	0.9 %
hnRNA	0.01 %
circularRNA	0.02 %
snRNA	0.44 %
snoRNA	0.37 %
miRNA	0.30 %
7SL	0.18 %
Xist	0.001%
other lncRNA	0.04 %



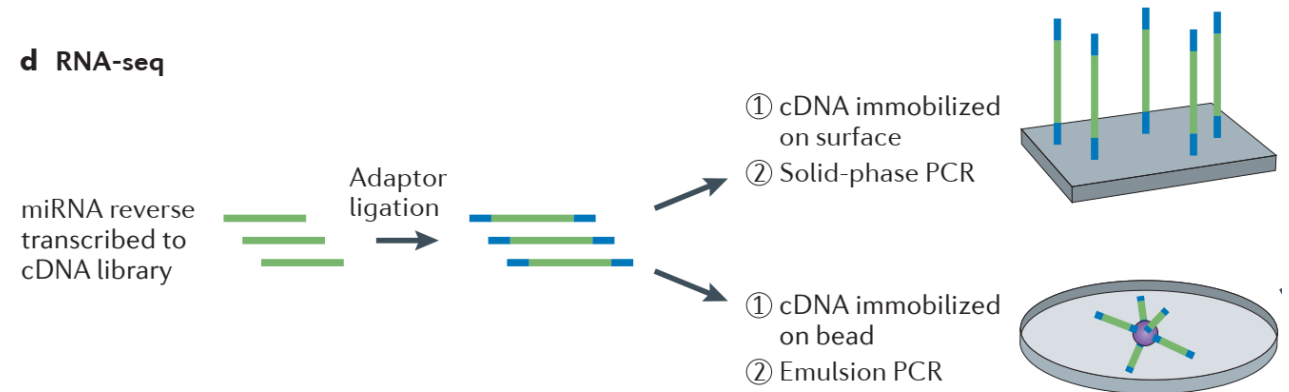
Que modelo, tejido/muestra, método de extracción y plataforma de análisis



sRNA-seq permite la identificación de secuencias de miARN y distinguir fácilmente entre miARNs que difieren en un solo nucleótido, así como **isomiR** de longitud variable

Construcción de una librería de cDNA

d RNA-seq



Sesgos

Métodos de preparación de bibliotecas de ADNc que favorecen la captura de algunos miARN sobre otros

Que modelo, tejido/muestra, *método de extracción*

Guo et al. BMC Genomics (2017) 18:50
DOI 10.1186/s12864-016-3470-z

BMC Genomics

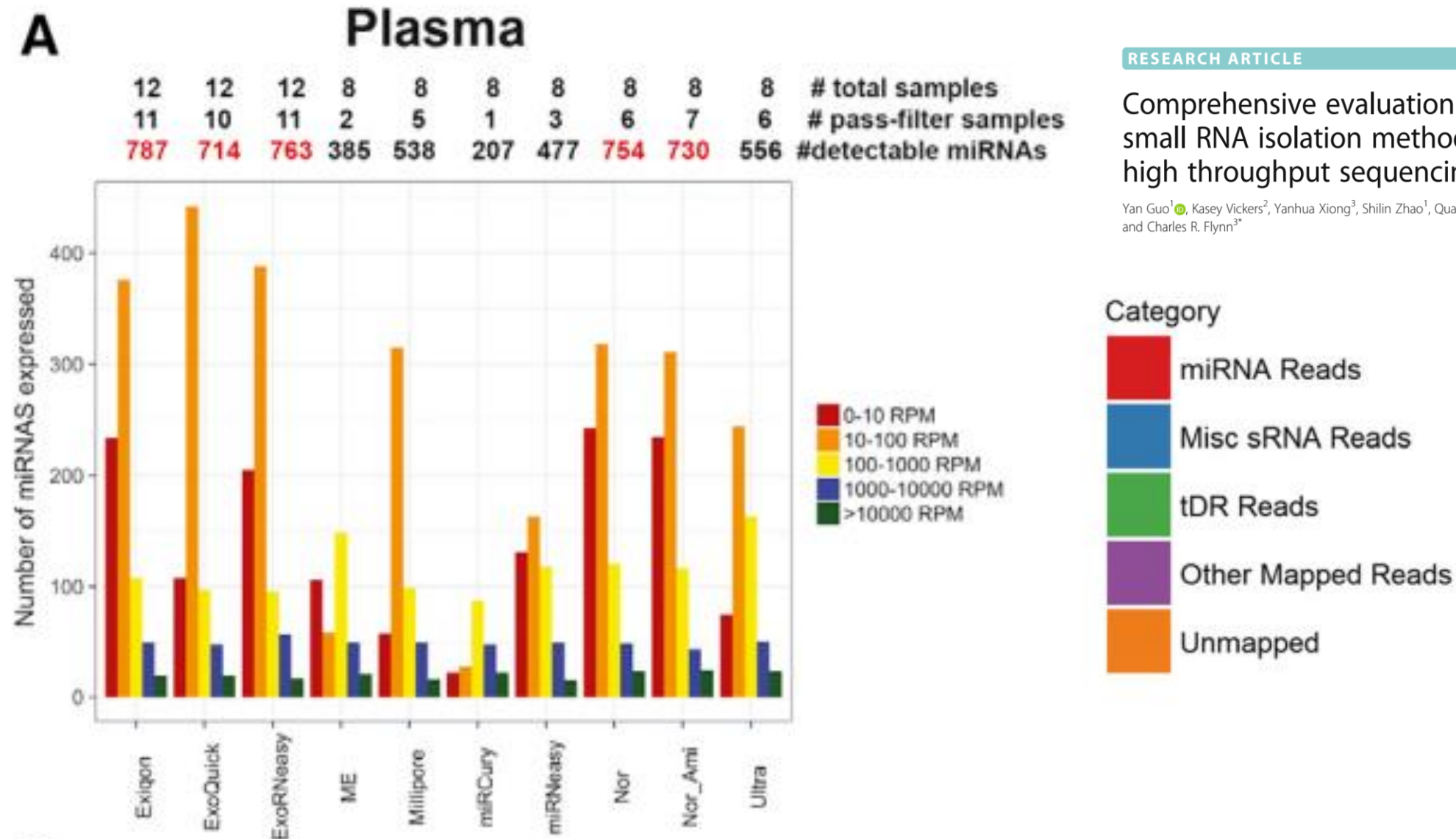
RESEARCH ARTICLE

Open Access



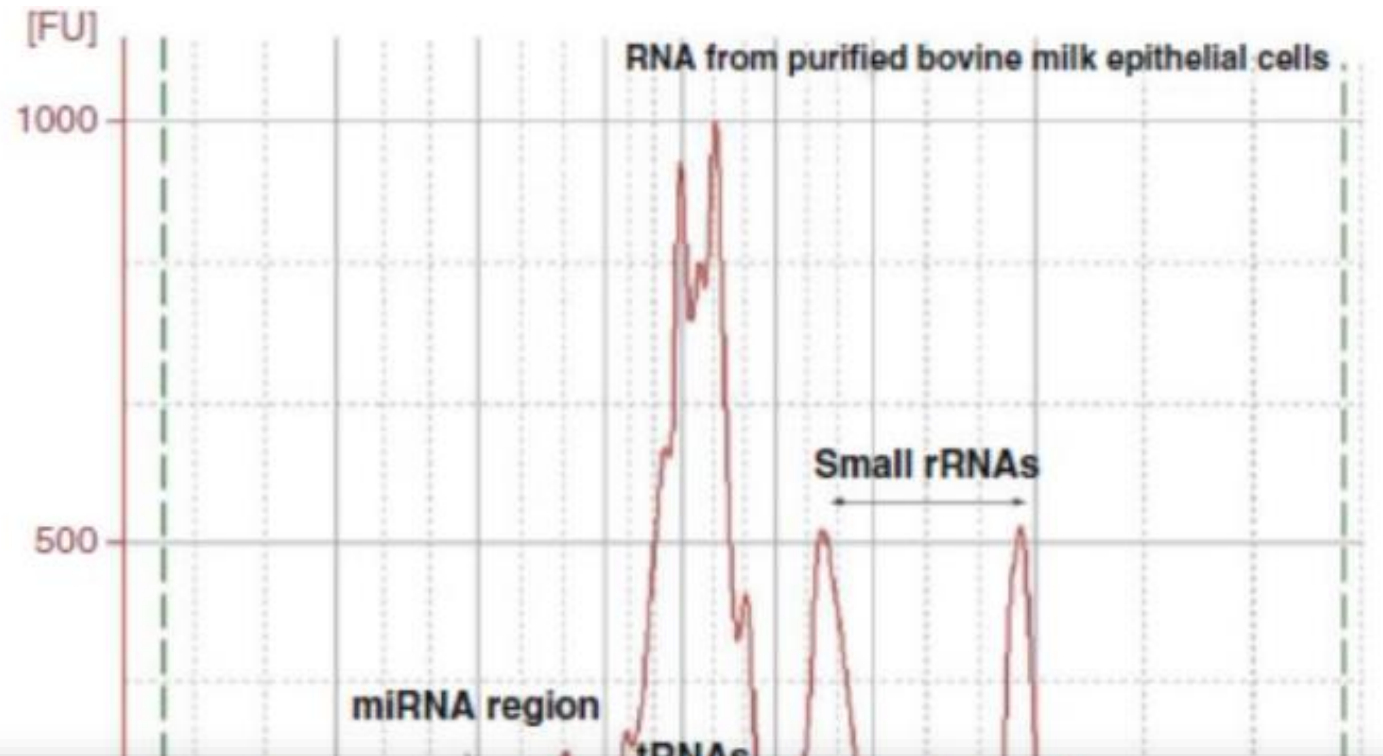
Comprehensive evaluation of extracellular small RNA isolation methods from serum in high throughput sequencing

Yan Guo¹, Kasey Vickers², Yanhua Xiong³, Shilin Zhao¹, Quanhui Sheng¹, Pan Zhang¹, Wanding Zhou⁴ and Charles R. Flynn^{3*}

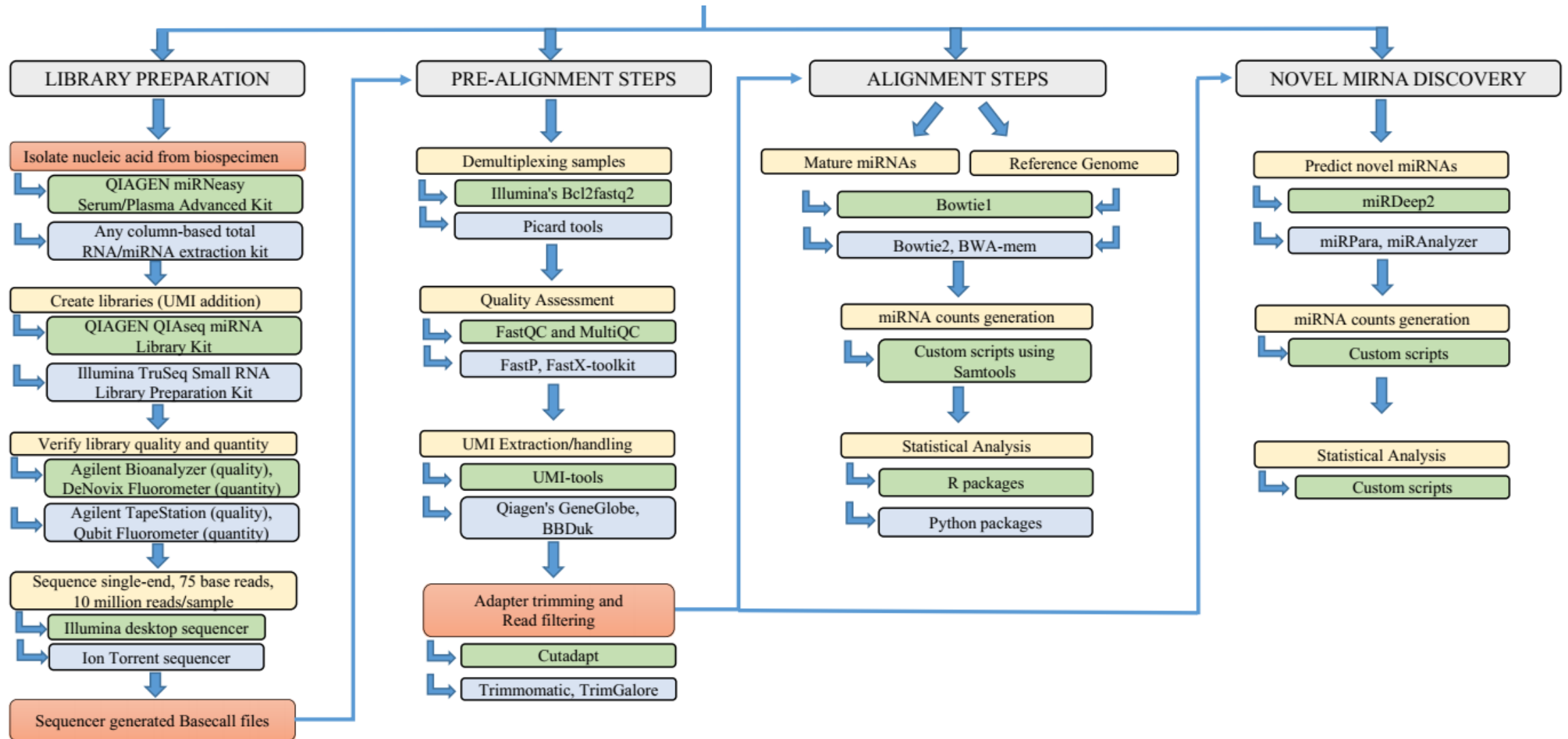


Evaluación de calidad, cantidad e integridad del ARN

- Es habitual evaluar el rendimiento y el grado de integridad general del ARN mediante espectrofotometría e instrumentos de electroforesis capilar automatizados, como el **Bioanalyzer 2100** (de Agilent).
- Proporción de ARN en la ventana de **15 a 40 nt** en relación con la abundancia total de ARN



Small RNA-seq (sRNA-seq) pipeline



Small RNA-seq library - two-adaptor ligation-based methods

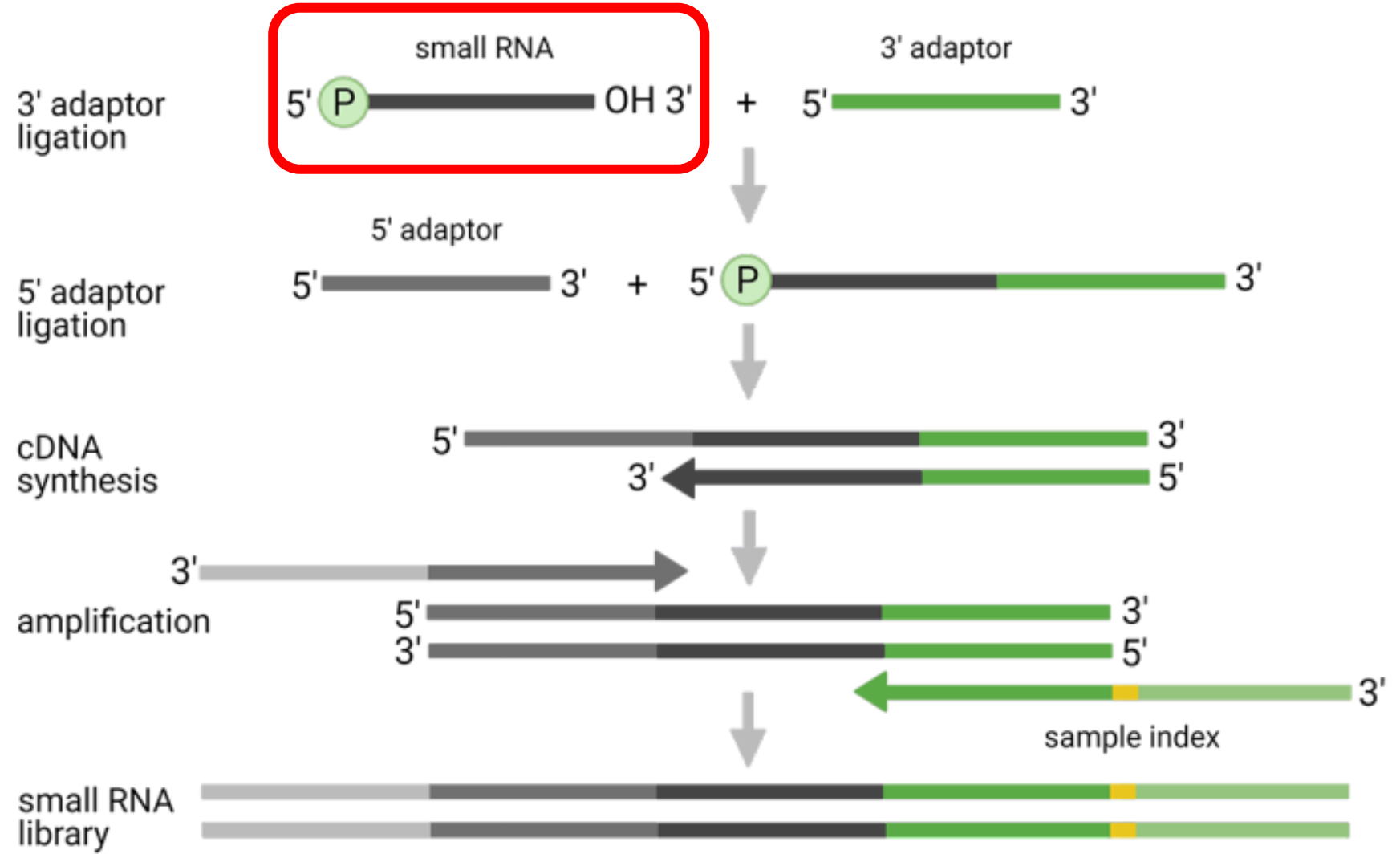
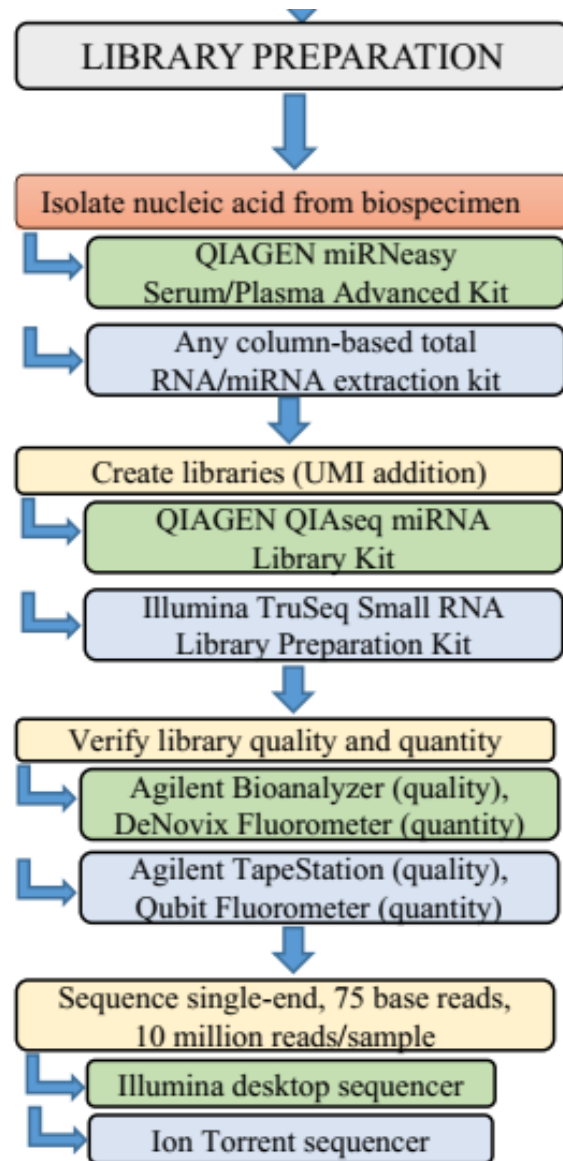


Figure 1. Original two-adaptor ligation protocol for small RNA-seq analysis.

Small RNA-seq library - *Sesgos*

A



Biases in small RNA deep sequencing data

Carsten A. Raabe^{1,*}, Thean-Hock Tang², Juergen Brosius¹ and
Timofey S. Rozhdestvensky^{1,*}

¹Institute of Experimental Pathology (ZMBE), University of Muenster, Von-Esmarch-Strasse 56, 48149 Muenster, Germany and ²Advanced Medical and Dental Institute (AMDI), Universiti Sains Malaysia, 13200 Penang,

B

Reverse Transcription
(first strand cDNA synthesis)

possible bias

1) RNA structure interference

2) RNA posttranscriptional modifications

3) RNA sequence (GC content)

PCR amplification

possible bias

1) cDNA size

2) cDNA GC-content

4) sequence and chemistry of 5'-adaptor

4) and modification
N or N₂'-OCH₃

5) sequence and chemistry of 3'-adaptor

C

Phosphatase;
TAP & 5'-adaptor
ligation

OH

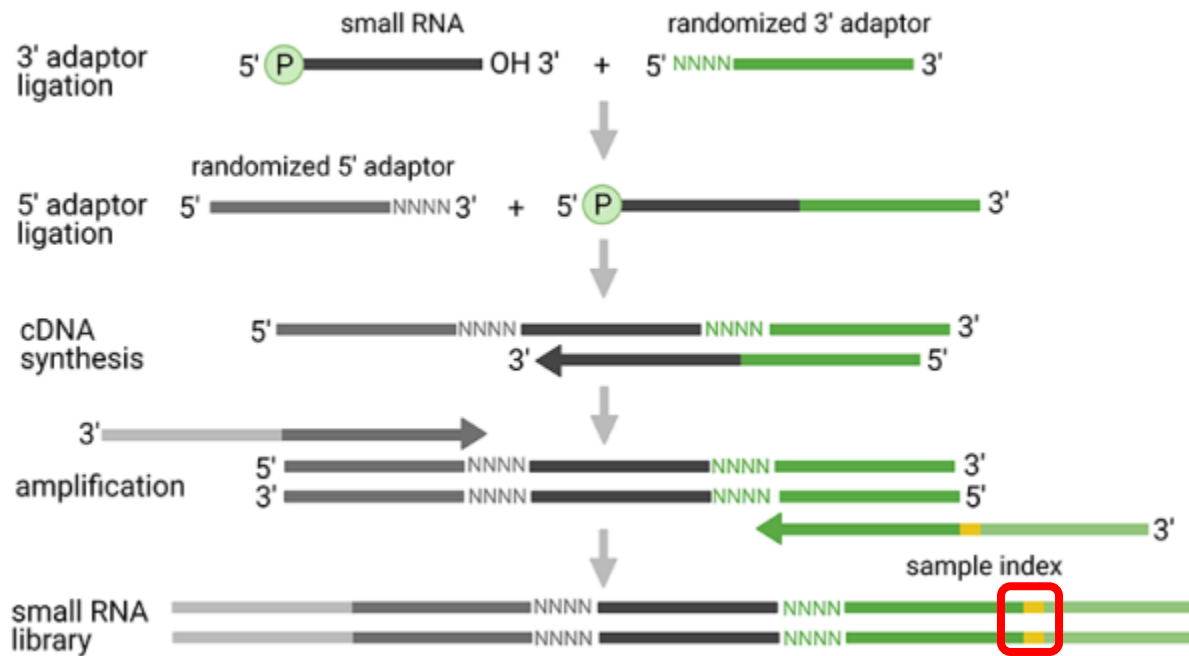
te los
rdadera
oría de
an o
algunos
amente

Small RNA-seq library - *Improved* two-adaptor ligation-based methods

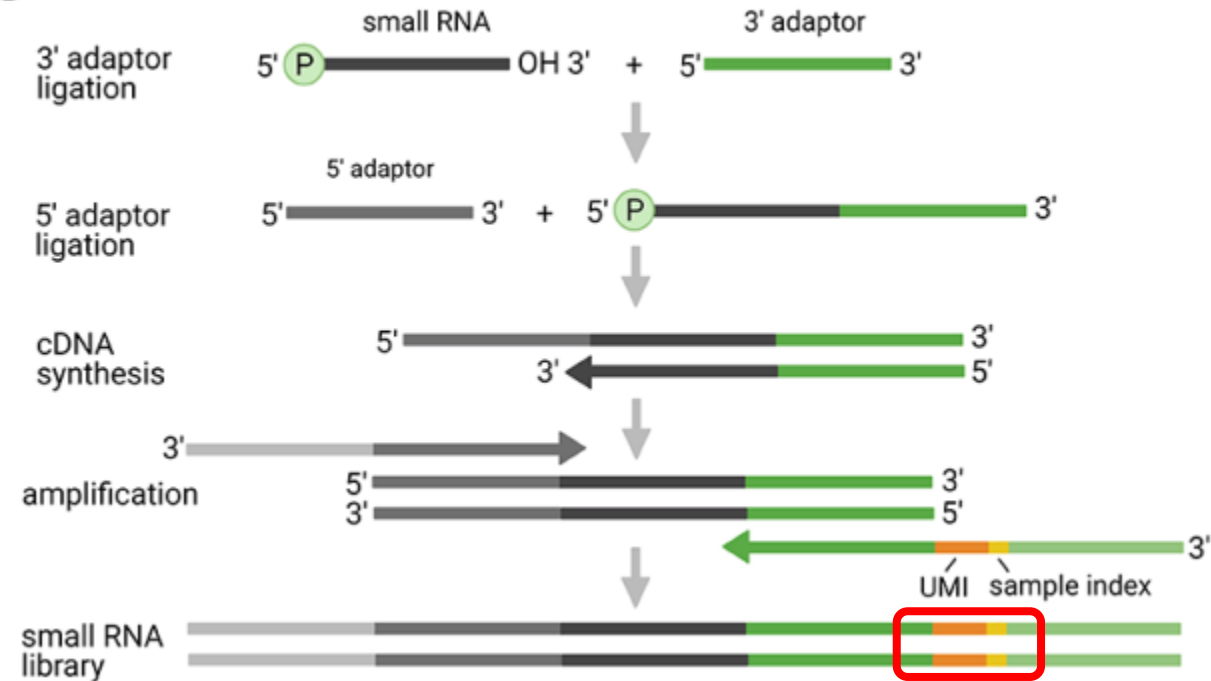
randomized adaptors

UMIs

A

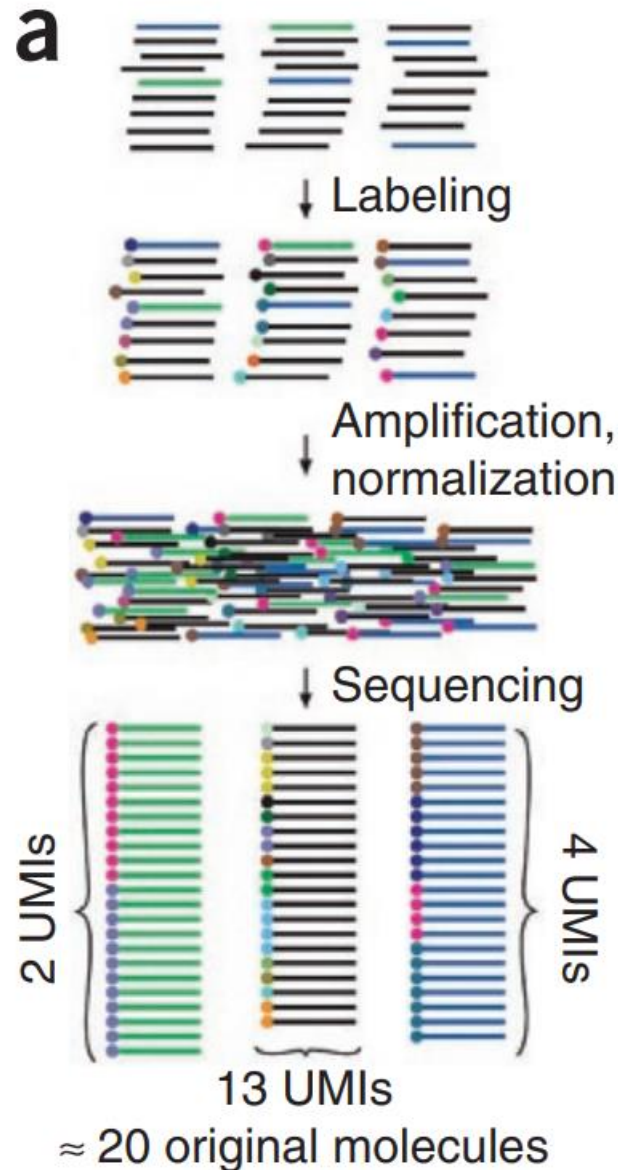


C

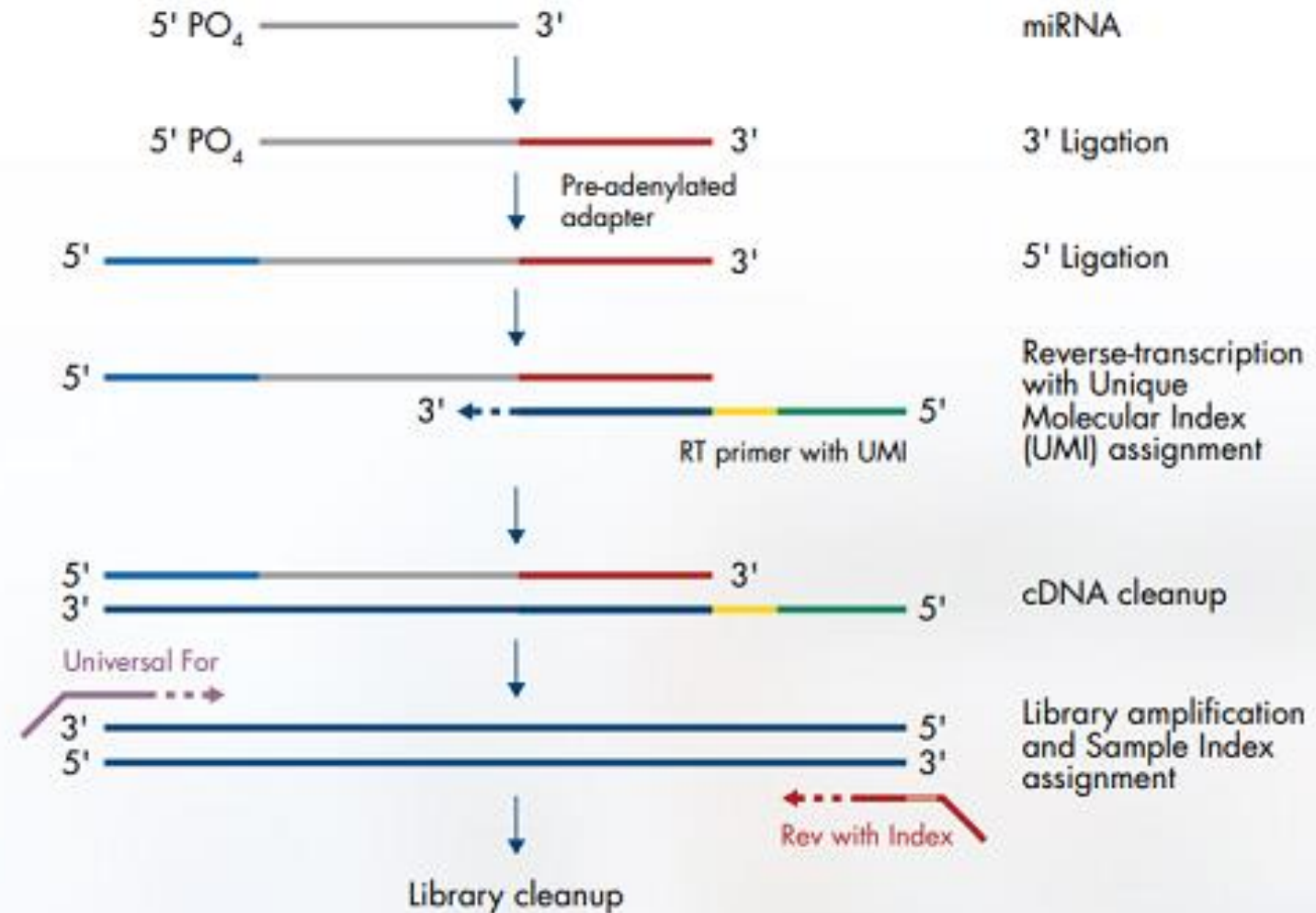


Barcode: Durante la preparación de la biblioteca antes de la secuenciación, se asigna un **código de barras específico** (8 nucleótidos [nt] de longitud) a cada muestra.

Unique Molecular Identifiers - (UMIs)



QIAseq™ miRNA Library Kit

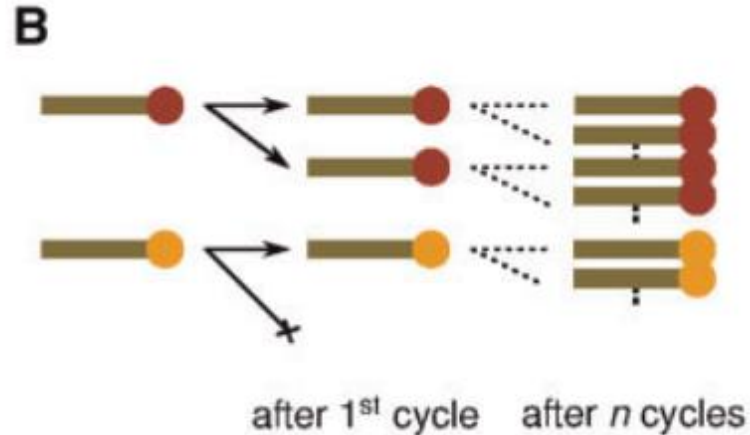
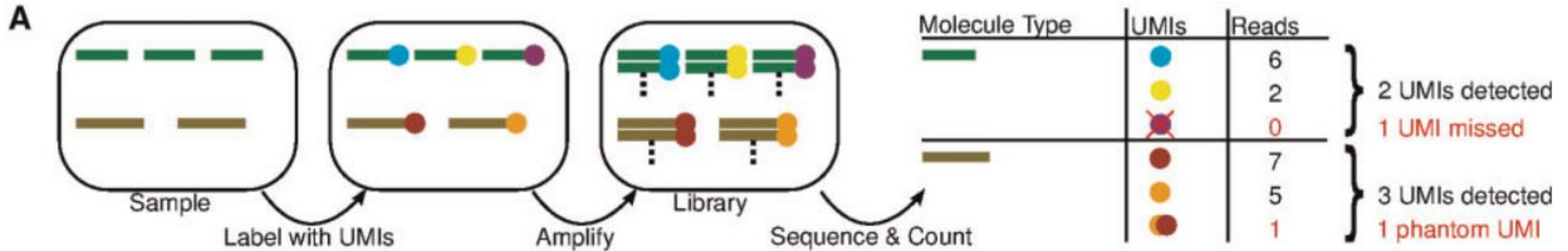


ón de la
miARNs.

molécula

de modo
muestra

Unique Molecular Identifiers - (UMIs)



Bioinformatics, 34(18), 2018, 3137–3144
doi: 10.1093/bioinformatics/bty283
Advance Access Publication Date: 16 April 2018
Original Paper

OXFORD

Gene expression

TRUmiCount: correctly counting absolute numbers of molecules using unique molecular identifiers

Florian G. Pflug^{1,*} and Arndt von Haeseler^{1,2}

Barcode (8 nt de longitud)
UMI (12 nt de longitud)

Kivioja T, et al. 2011; Pflug F.G. & von Haeseler A. 2018

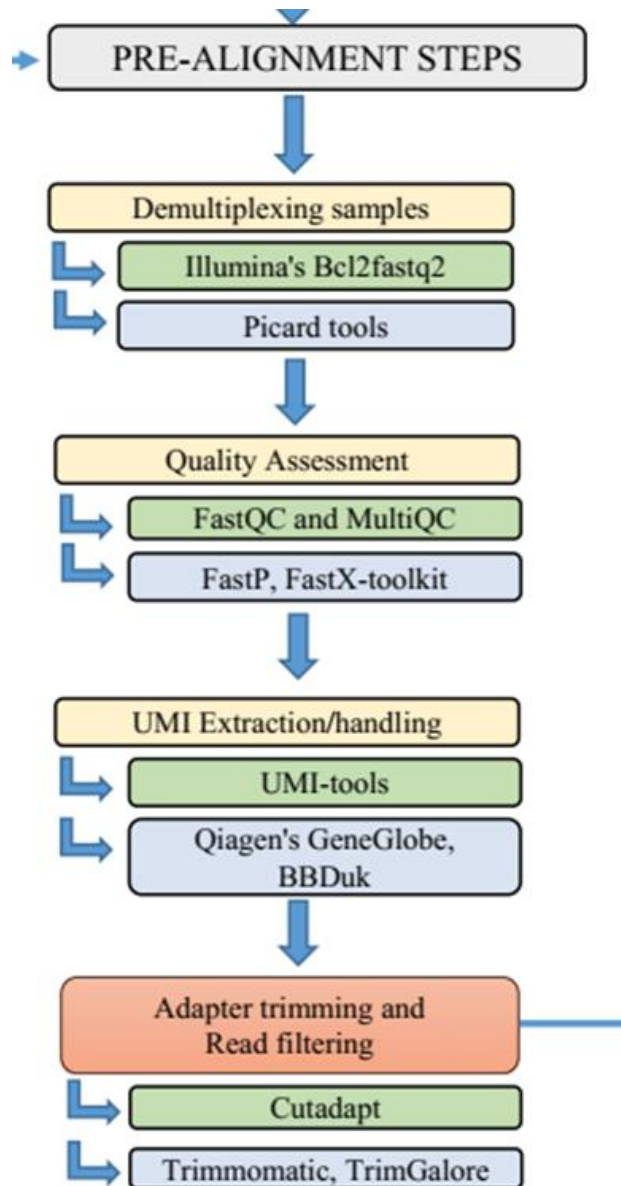
NGS-based RNA-seq bioinformatics analysis tools/pipelines

La mayoría de las herramientas/pipelines disponibles para la investigación del transcriptoma por NGS se concentran en la **detección/predicción/cuantificación** de la expresión de ARN pequeños, especialmente **miARN**.

TABLE 1 | Small non-coding RNA Tool comparison.

		miRDeep	miRDeep*	miRSpring	DARIO	CPSS	ncPRO-seq	CoRAL	RNA-CODE
Package	Online server				✓	✓	✓		
	Stand-alone	✓	✓	✓			✓	✓	✓
Applicable to	Raw data	✓	✓			✓	✓		✓
	Mapped data		✓	✓	✓		✓	✓	
Input format	FASTQ/FASTA	✓	✓			✓	✓		✓
	BAM/SAM		✓	✓	✓		✓	✓	
	BED				✓				
	GFF/GTF							✓	
Assembly	De novo								✓
	Reference genome sequence	✓	✓	✓	✓	✓	✓	✓	
Known miRNA detection		✓	✓	✓	✓	✓	✓	✓	✓
Known other ncRNA detection					✓	✓	✓	✓	✓
Novel ncRNA prediction		✓	✓		✓		✓	✓	
Expression analysis		✓	✓	✓	✓	✓	✓	✓	✓
miRNA target prediction			✓			✓			
miRNA target functional enrichment						✓			

NGS-based RNA-seq bioinformatics analysis – Quality check



Name	URL/link	Utility	Date of Access
Bcl2fastq2	https://support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html	Software: Conversion of.bcl files to.fastq files	12-Feb-2019
Bowtie1/2	http://bowtie-bio.sourceforge.net/index.shtml	Software: Read alignment	13-Mar-2019
miRDeep2	https://github.com/rajujewsky-lab/mirdeep2	Software: Novel miRNA prediction	20-Jun-2019
FastQC	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/	Software: Single sample Quality Control (QC) report creation	21-Mar-2019
MultiQC	https://multiqc.info	Software: Creation of multi-sample QC report	22-Mar-2019
UMI-tools	https://github.com/CATOxford/UMI-tools	Software: Handling of UMI-tagged reads	31-Mar-2019
Cutadapt	https://cutadapt.readthedocs.io/en/stable/guide.html	Software: Adapter trimming	25-Feb-2019
Samtools	http://www.htslib.org	Software: Processing aligned files (.SAM)	24-Jun-2019
miRBase	http://www.mirbase.org	Database: Mature and hairpin miRNA sequences	05-Jan-2019
vGRCh38	http://hgdownload.soe.ucsc.edu/goldenPath/hg38/chromosomes/	Database: Human reference genome sequence	06-Jan-2019

NGS-based RNA-seq bioinformatics analysis – Quality check

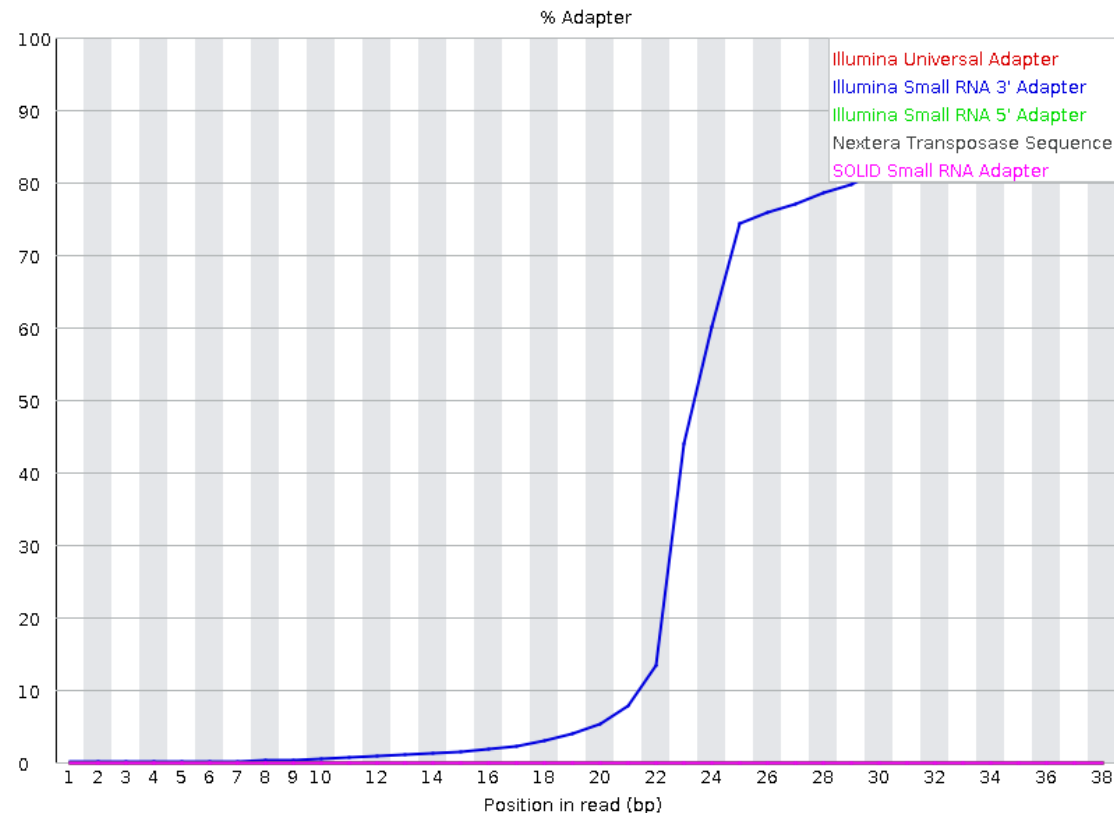
- Una valor **Q de 30 (99,9% de precisión)** se considera aceptable.
- Para datos de secuenciación de miARN, un **porcentaje mayor (>40%)** de lecturas que tienen 3' adaptadores.

FastQC Report

Summary

- ✓ [Basic Statistics](#)
- ✓ [Per base sequence quality](#)
- ✓ [Per tile sequence quality](#)
- ✓ [Per sequence quality scores](#)
- ✗ [Per base sequence content](#)
- ✗ [Per sequence GC content](#)
- ✓ [Per base N content](#)
- ✓ [Sequence Length Distribution](#)
- ✗ [Sequence Duplication Levels](#)
- ✗ [Overrepresented sequences](#)
- ✗ [Adapter Content](#)
- ! [Kmer Content](#)

✗ Adapter Content



cutadapt

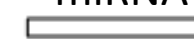
3' Adapter



or



miRNA



Read



Adapter



Removed sequence



FastQC Report

subsample

General Statistics

Copy table

Configure Columns

Sort by highlight

Plot

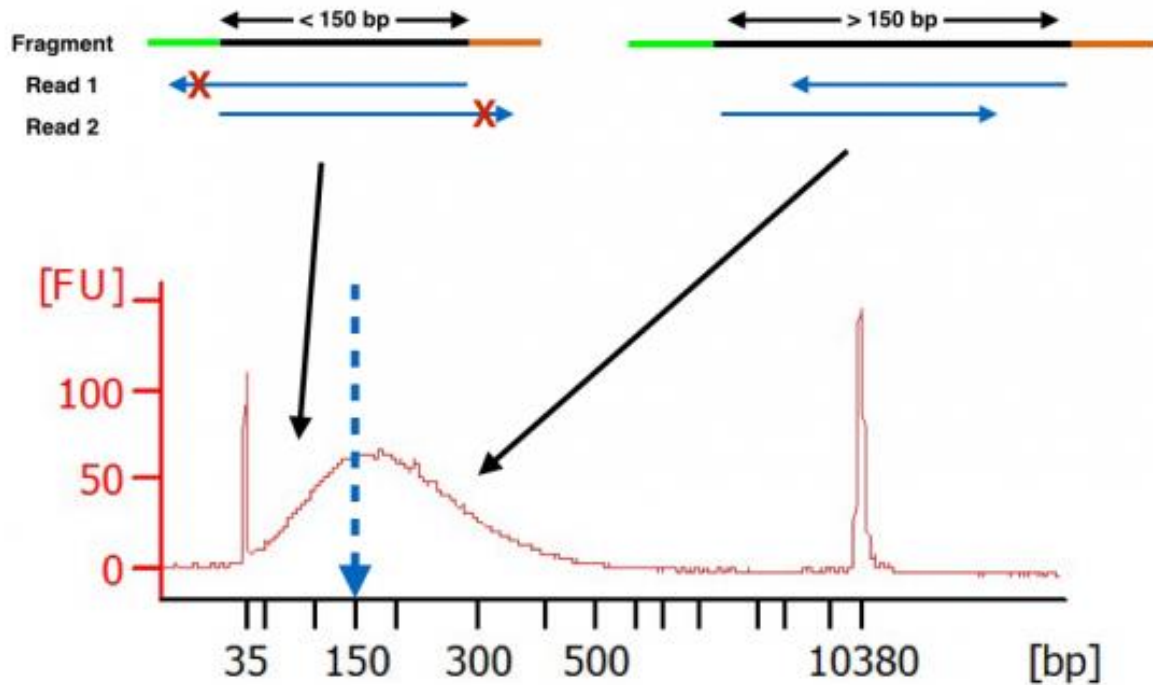
Showing 8/8 rows and 9/11 columns.

Sample Name	5'-3' bias	M Aligned	% Aligned	M Aligned	% Aligned	M Aligned	% Dups	% GC	M Seqs
Irrel_kd_1	1.18	35.6	86.4%	31.2	92.1%	33.2	55.9%	47%	36.1
Irrel_kd_2	1.14	30.4	86.0%	26.5	92.2%	28.4	53.6%	47%	30.8
Irrel_kd_3	1.19	23.6	85.7%	20.5	92.0%	22.0	50.1%	48%	23.9
Mov10_kd_2	1.13	51.9	86.0%	45.3	91.6%	48.3	60.5%	48%	52.7
Mov10_kd_3	1.13	30.7	86.0%	26.8	91.6%	28.5	54.6%	47%	31.1
Mov10_oe_1	1.09	38.1	80.2%	32.1	88.9%	35.5	56.5%	47%	40.0
Mov10_oe_2	1.18	35.4	81.0%	30.0	88.8%	33.0	55.9%	48%	37.1
Mov10_oe_3		20.3	81.5%	17.3	90.0%	19.1	50.1%	47%	21.2

er
er
ence
38

NGS-based RNA-seq bioinformatics analysis – *Single end vs Paired end*

Inserto de 20-50 nt – reads 150 pb



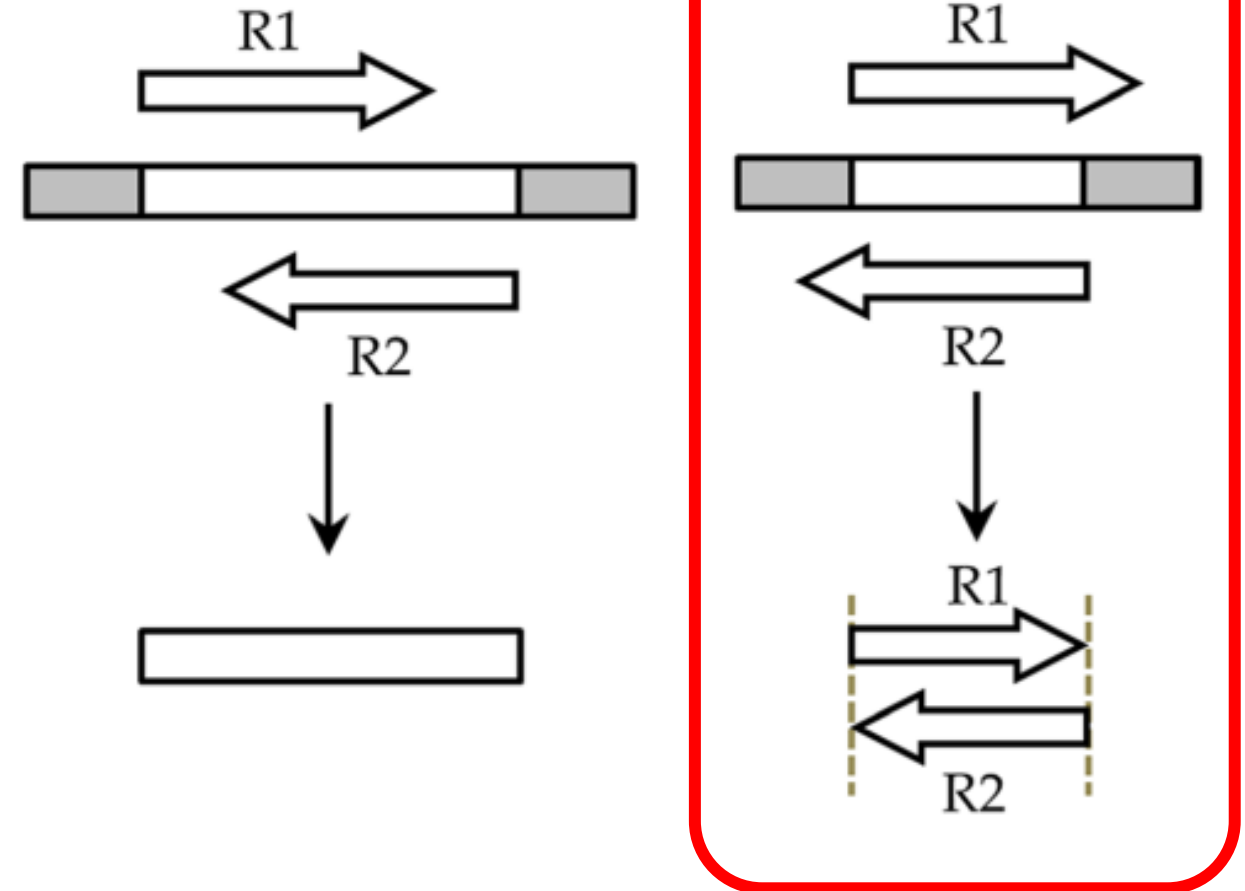
METHODOLOGY ARTICLE

Open Access



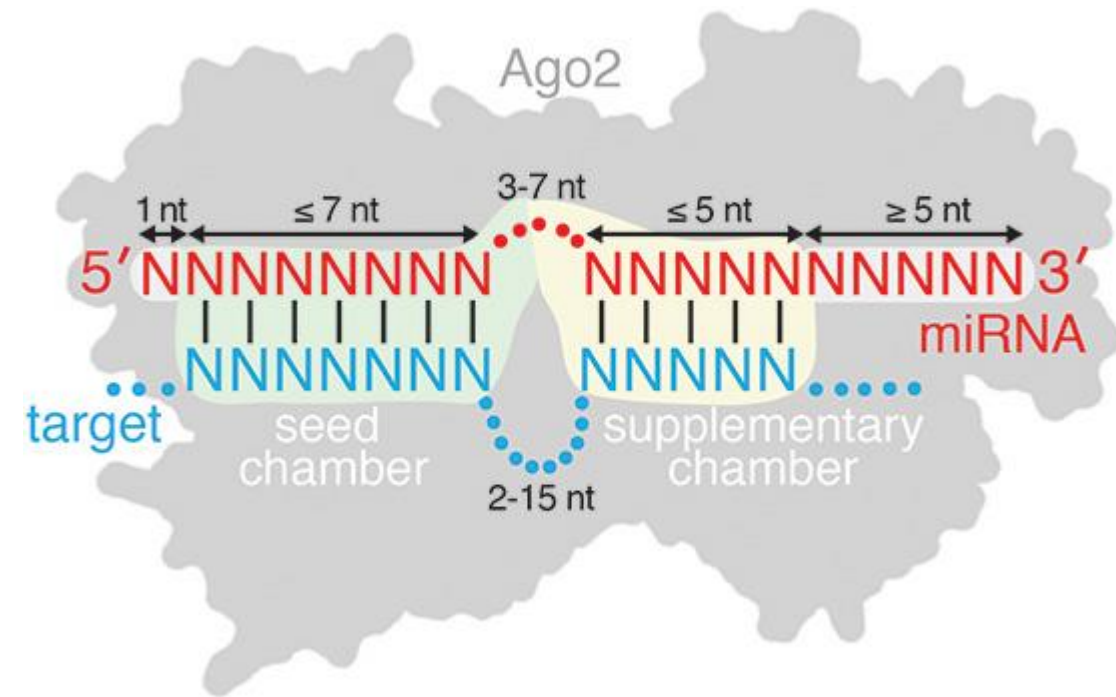
NGmerge: merging paired-end reads via novel empirically-derived models of sequencing errors

John M. Gaspar



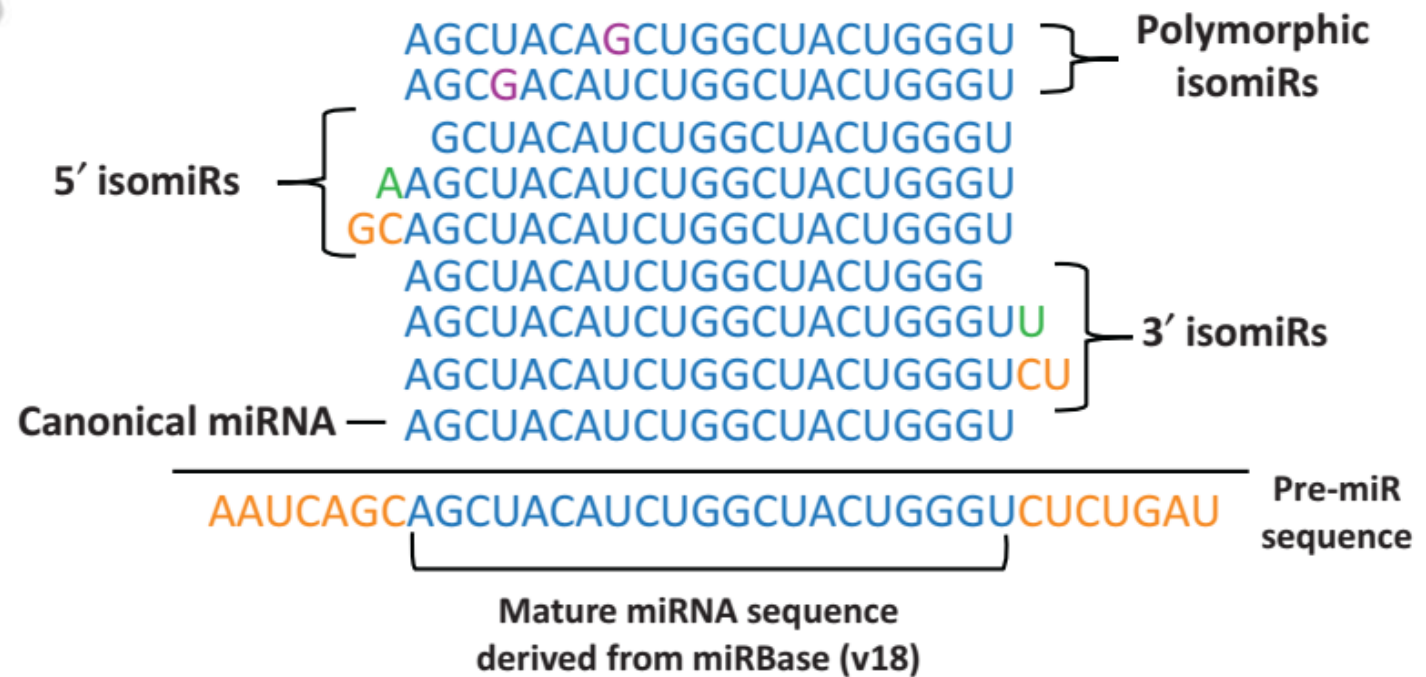
Permite la corrección de errores y la determinación precisa de la cobertura de lectura

QC de las lecturas secuenciadas – consideraciones del *Trimming*



Solo remover adaptadores

Conservar longitudes de lectura para miARNs esperadas para eliminar lecturas que sean demasiado cortas (<18 pb) o demasiado largas (>30 pb)

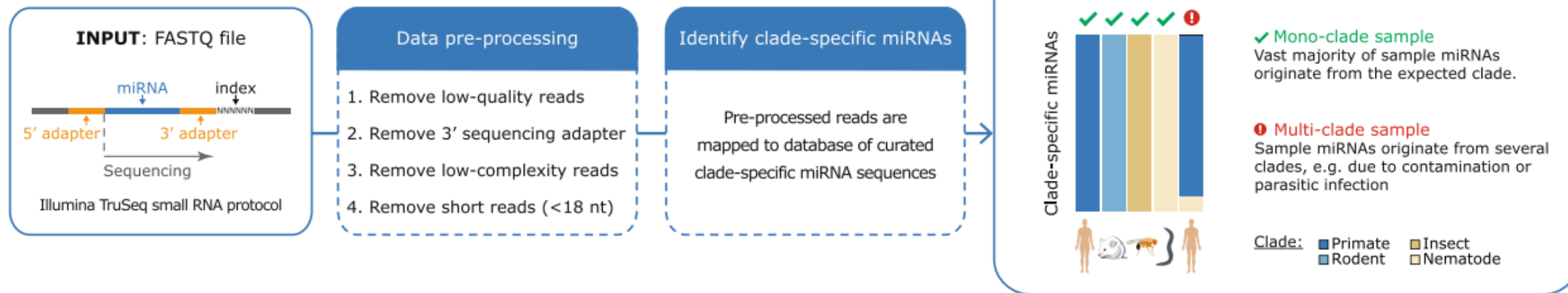




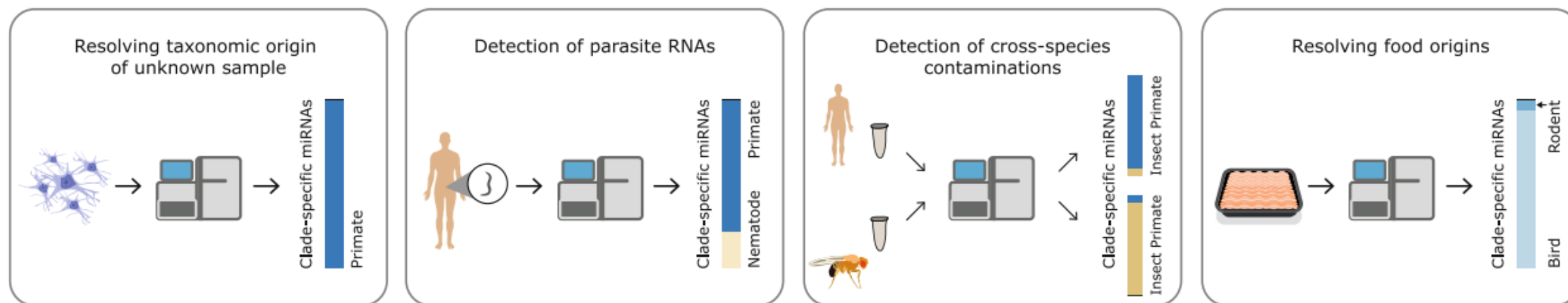
miRTrace reveals the organismal origins of microRNA sequencing data

Wenjing Kang¹, Yrin Eldjell¹, Bastian Fromm¹, Xavier Estivill², Inna Biryukova¹ and Marc R. Friedländer^{1*}

A Flowchart of miRTrace tracing miRNA sequencing data to their taxonomic origins

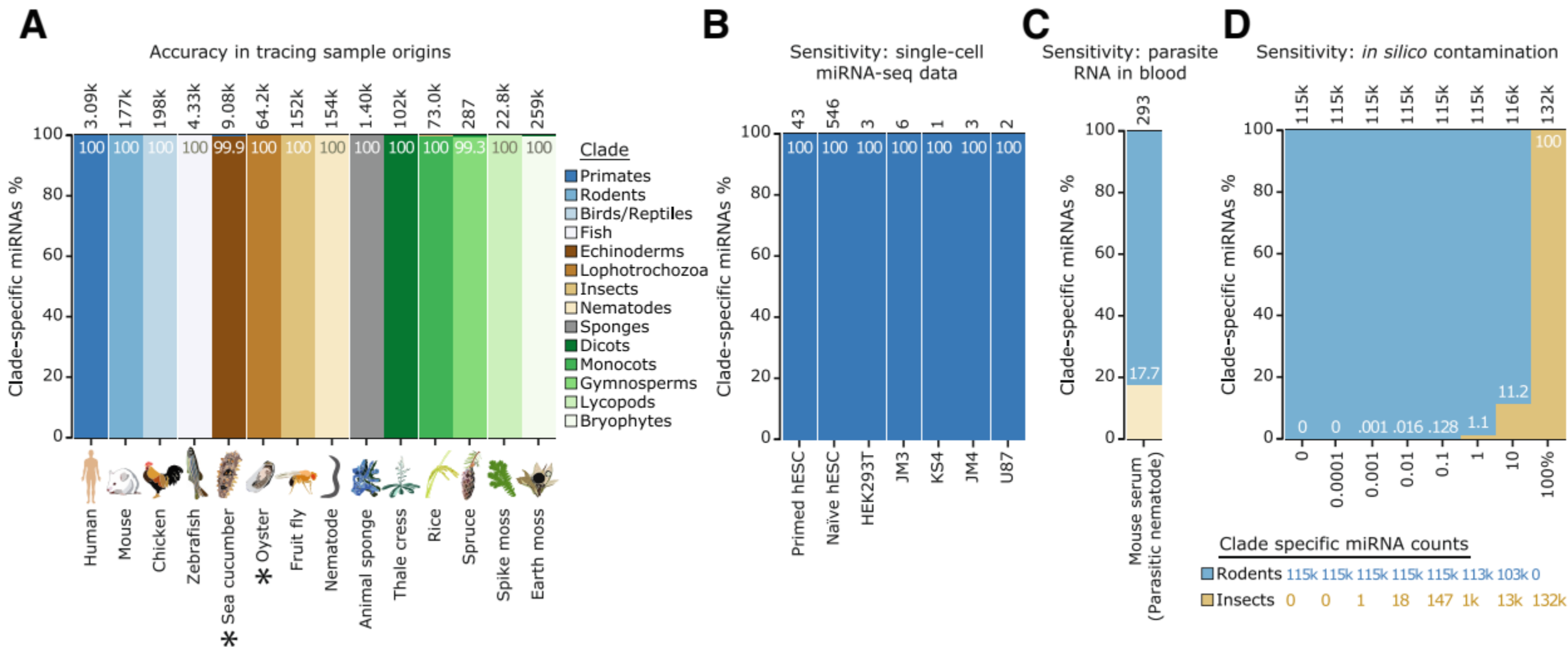


B miRTrace possible applications

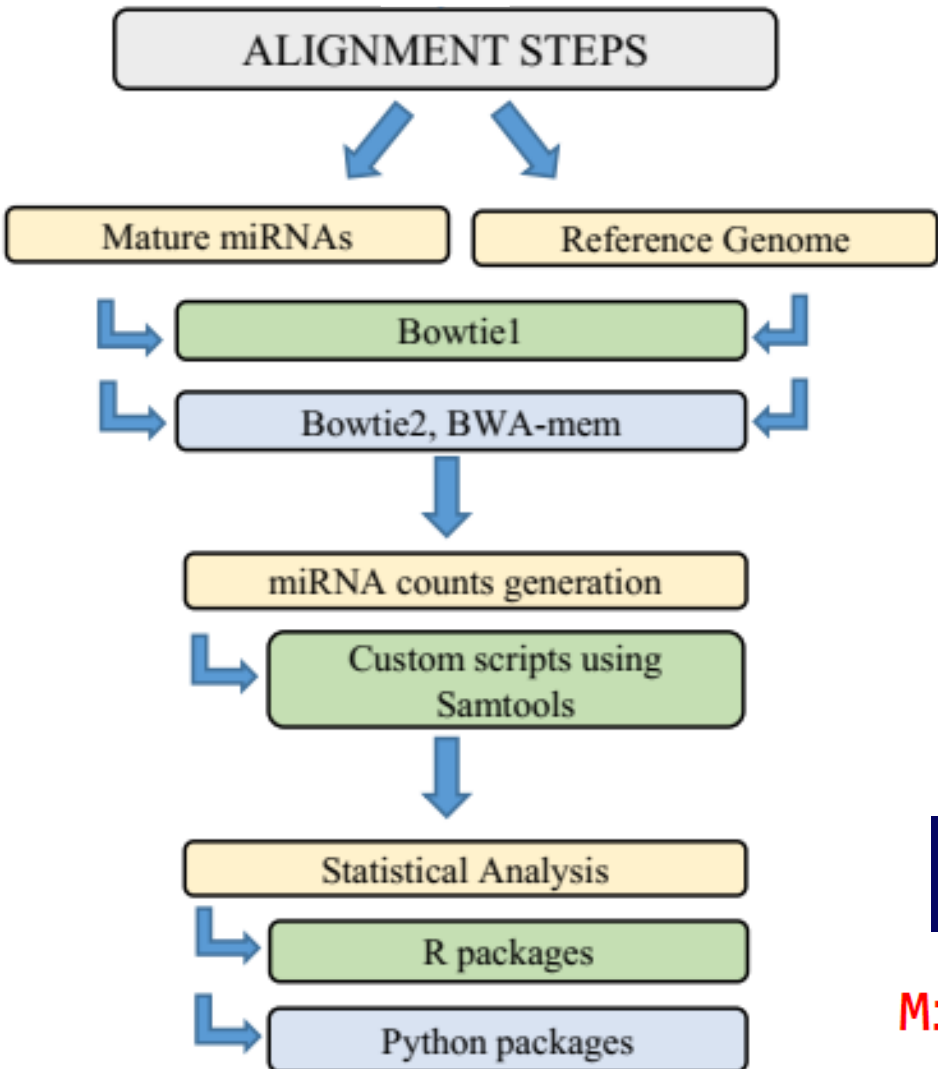




miRTrace reveals the organismal origins of microRNA sequencing data

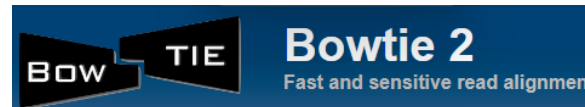
Wenjing Kang¹, Yrin Eldjell¹, Bastian Fromm¹, Xavier Estivill², Inna Biryukova¹ and Marc R. Friedländer^{1*}

Alineamiento de lecturas basada en referencias



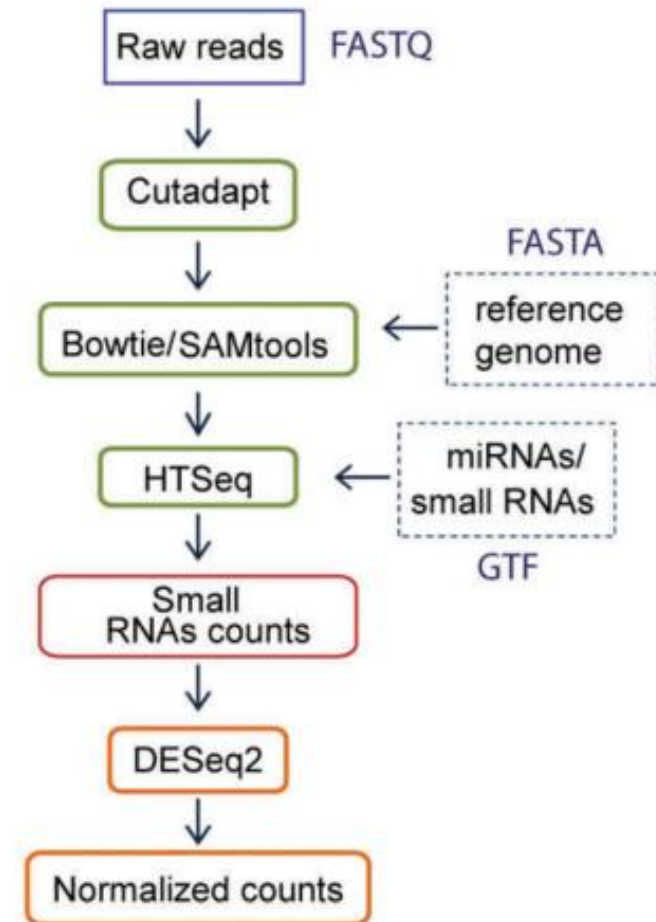
Fast gapped-read alignment with Bowtie 2

Ben Langmead^{1,2} & Steven L Salzberg¹⁻³



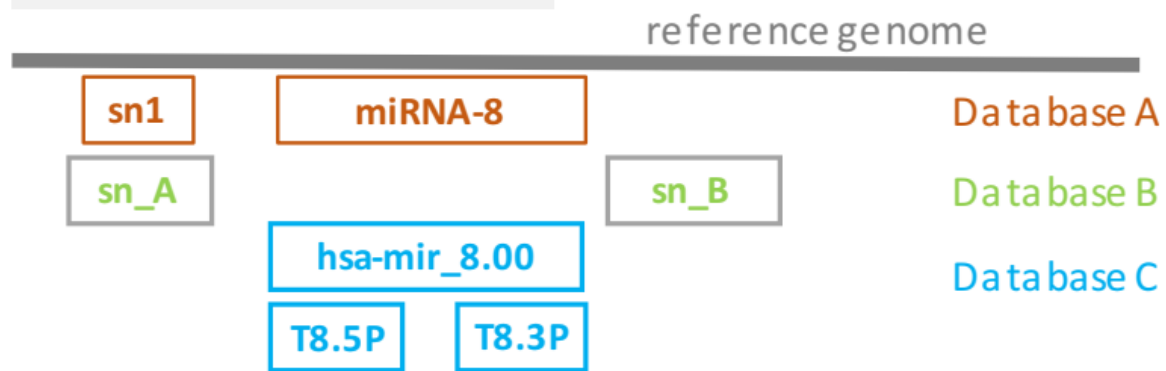
MirGeneDB 2.1

GtRNAdb

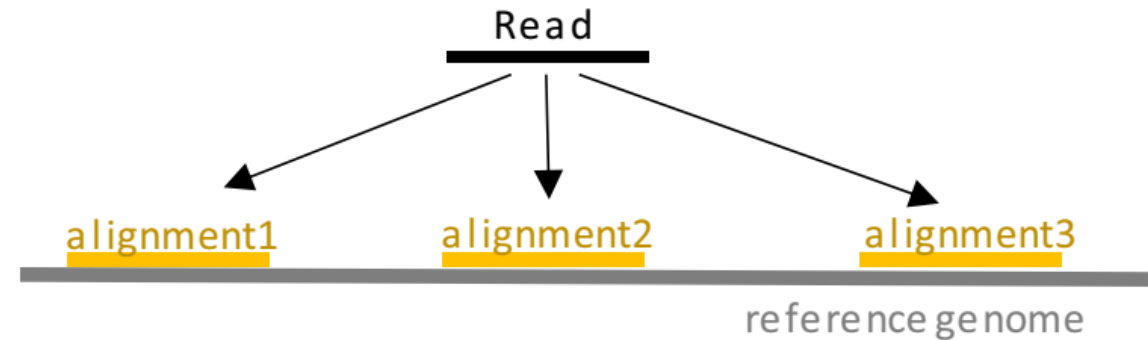


Desafíos del análisis de datos de *sRNAseq*

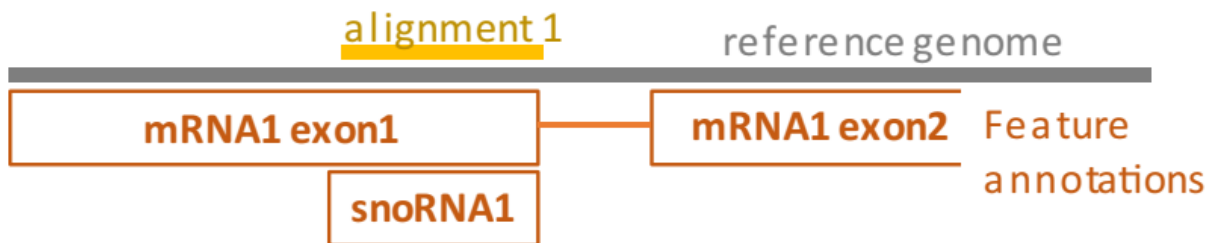
a ANNOTATION SOURCES Which ones to use (input)?



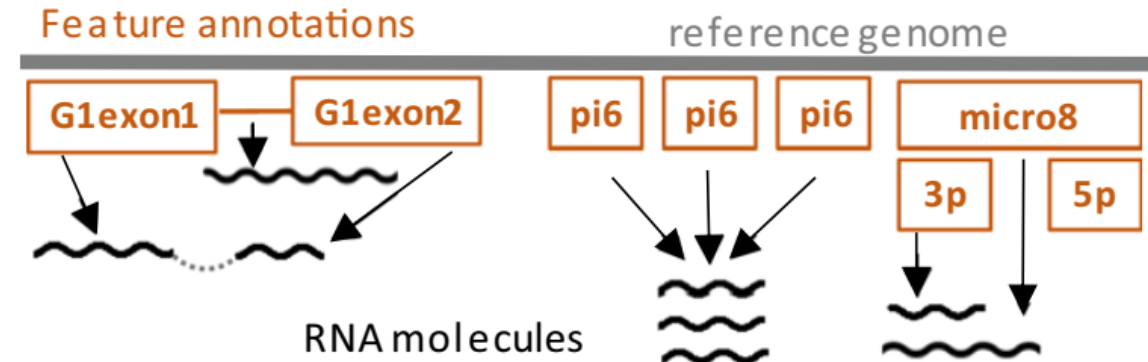
b MULTI-MAPPERS ASSIGNATION How to assign?



c MULTI-OVERLAPS ASSIGNATION How to assign?



d DIVERSE STRUCTURE What to quantify (output)?



miRBase y MirGeneDB

La última versión de **miRBase (v22)** contiene secuencias de microARN de **271 organismos**: 38.589 precursores en horquilla y 48.860 microARN maduros.

miRBase

MirGeneDB 2.1

[Home](#) [Browse](#) [Search](#) [Download](#) [Information](#)

Welcome to MirGeneDB 2.1 - the curated microRNA Gene Database

MirGeneDB is a database of manually curated microRNA genes that have been validated and annotated as initially described in [Fromm et al. 2015](#) and [Fromm et al. 2020](#). MirGeneDB 2.1 includes more than 16,000 microRNA gene entries representing more than 1,500 miRNA families from 75 metazoan species and published in the [2022 NAR database issue](#). All microRNAs can be browsed, searched and downloaded.

- [miRNA.dead](#) - List of entries that have been removed from the database

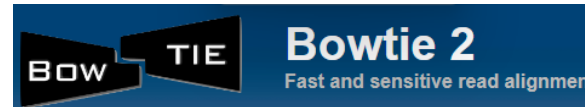
Genome coordinates

aae.gff3	ame.gff3	ath.gff3	bmo.gff3	bta.gff3	cbr.gff3	cel.gff3	cfa.gff3	cre.gff3	dme.gff3	dps.gff3	dre.gff3
ebv.gff3	fru.gff3	gga.gff3	hcmv.gff3	hsa.gff3	kshv.gff3	mdo.gff3	mghv.gff3	mml.gff3	mmu.gff3	osa.gff3	ptc.gff3
ptr.gff3	rno.gff3	sme.gff3	tni.gff3	vvi.gff3	xtr.gff3	zma.gff3					

Alineamiento de lecturas basada en referencias

Fast gapped-read alignment with Bowtie 2

Ben Langmead^{1,2} & Steven L Salzberg¹⁻³



MirGeneDB 2.1

GtRNAdb

ALIGNMENT STEPS

Mature miRNAs

Reference Genome

Bowtie1

Bowtie2, BWA-mem

miRNA counts generation

Custom scripts using Samtools

Statistical Analysis

R packages

Python packages

Raw reads FASTQ

Cutadapt

Bowtie/SAMtools

FASTA
reference
genome

HTSeq

miRNAs/
small RNAs
GTF

Small
RNAs counts

DESeq2

Normalized counts

Evaluation of microRNA alignment techniques

MARK ZIEMANN, ANTONY KASPI, and ASSAM EL-OSTA
Epigenetics in Human Health and Disease Laboratory, Baker IDI Heart and Diabetes Institute, The Alfred Medical Research and Education Precinct, Melbourne, Victoria 3004, Australia
Epigenomics Profiling Facility, Baker IDI Heart and Diabetes Institute, The Alfred Medical Research and Education Precinct, Melbourne, Victoria 3004, Australia

TABLE 1. SmRNA/microRNA-seq analysis pipelines in common use

Tool	Alignment engine	Reference sequence	Limited species	Local computer	Open source	Citation
miExpress	Smith-Waterman	miRbase	All miRbase	Yes	Yes	Wang et al. 2009
DSAP	Smith-Waterman	miRbase	All miRbase	Web-server only	NA	Huang et al. 2010
MIReNA	MEGABLAST	Whole genome	Any	Yes	Yes	Mathelier and Carbone 2010
miRDeep	MEGABLAST	Whole genome	Any	Yes	Yes	Friedländer et al. 2008
miRDeep2	Bowtie1	Whole genome	Any	Yes	Yes	Friedländer et al. 2012
miRanalyzer	Bowtie1	miRbase and whole genome	34 species	Web-server only	No	Hackenberg et al. 2011
Shortran	Bowtie1	Whole genome	Any	Yes	Yes	Gupta et al. 2012
mirTools2	SOAP2	Whole genome	32 species	Yes, and web-server		Wu et al. 2013b
MiRNAkey	BWA	miRbase	All miRbase	Yes	Yes	Ronen et al. 2010
UEA sRNA workbench	PatMaN	Whole genome	Any	Yes	Yes	Stocks et al. 2012
ShortStack	Any	Whole genome	Any	Yes	Yes	Axtell 2013

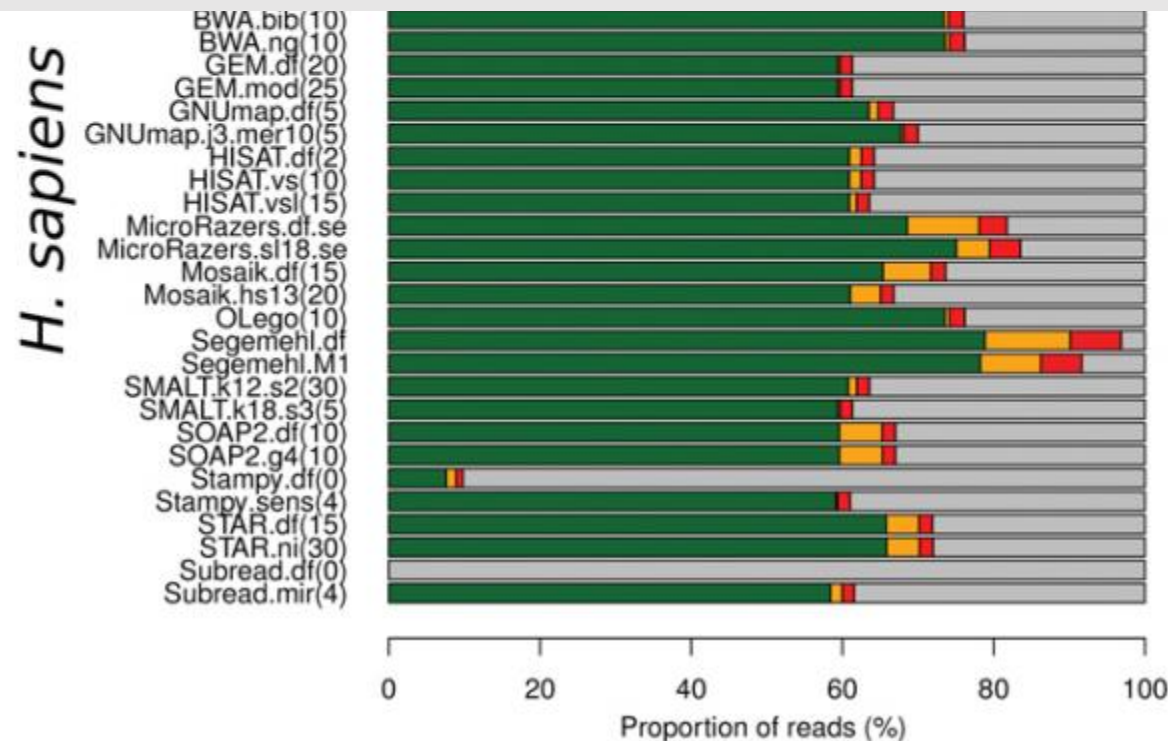
Alineamiento de lecturas cortas

TABLE 3. Alignment software evaluated in the present study

Aligner version	Parameter	Parameter	Reference
BBMap v34.x	df	Default	B Bushnell (unpubl.) ^a This study, rice and human Langmead et al. 2009
	k8	$k = 8$	
Bowtie1 v0.12.8	df	-k1	
	m2	-k1 -m2	
	m1	-k1 -m1	
	best	-k1 -m1 -best	Langmead and Salzberg 2012
	best strata	-best -strata -k 1 -m 1	
	try hard	-tryhard -best -k 1 -m 1	
Bowtie2 v2.1.0	df	Default	
	bib	-local -q -D 20 -R 3 -N 0 -L 8 -i S,1,0.50	
	vs	-end-to-end -very-sensitive	Li and Durbin 2009
	vsl	-very-sensitive-local	
BWA v0.7.10-r806	df	Default	
	bib	-n 1 -o 0 -e 0 -k 1	
	ng	-o 0	
GEM v1.819	df	Default	Marco-Sola et al. 2012
	mod	$m = 0.2$ $D = 1$ $e = 0.2$ -min-matched-bases = 0.5	
GNUMAP v3.0.2	df	Default	Clement et al. 2010 This study, rice This study, human
	j2.mer6	$j = 2$ $m = 6$	
	j3.mer10	$j = 3$ $m = 10$	
HISAT v0.1.5	df	Default	Kim et al. 2015
	vs	-end-to-end -very-sensitive	
	vsl	-very-sensitive-local	

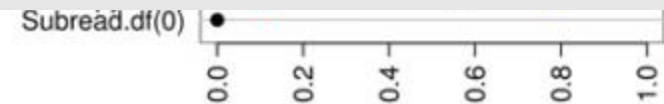
Alineamiento de lecturas cortas - consideraciones

El análisis de datos de smRNA-seq reales y simulados demuestra que la selección del **mapeador afecta los resultados** y la interpretación de la expresión diferencial.

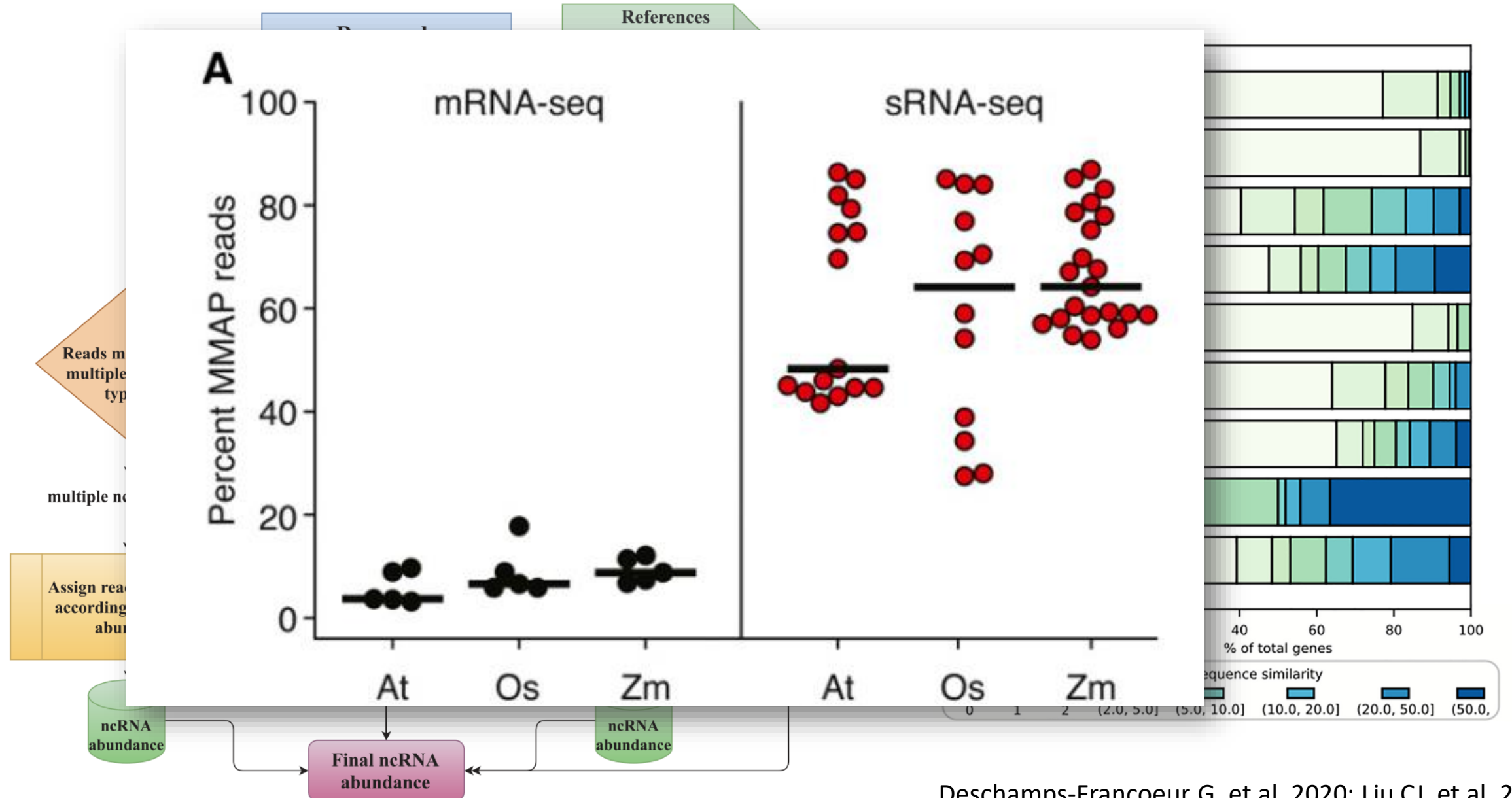


Aligner accuracy:

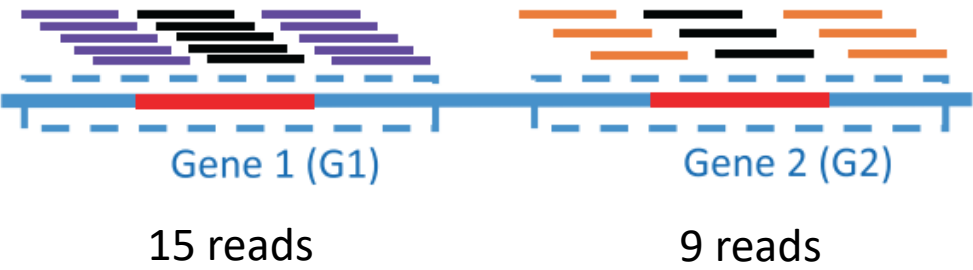
- Non-templated terminal extension
- Subject to varying read length
- Reads containing mismatches
- Reads containing indels
- Multiply mapped reads



Evaluación de la distribución de mapeos de ncRNA



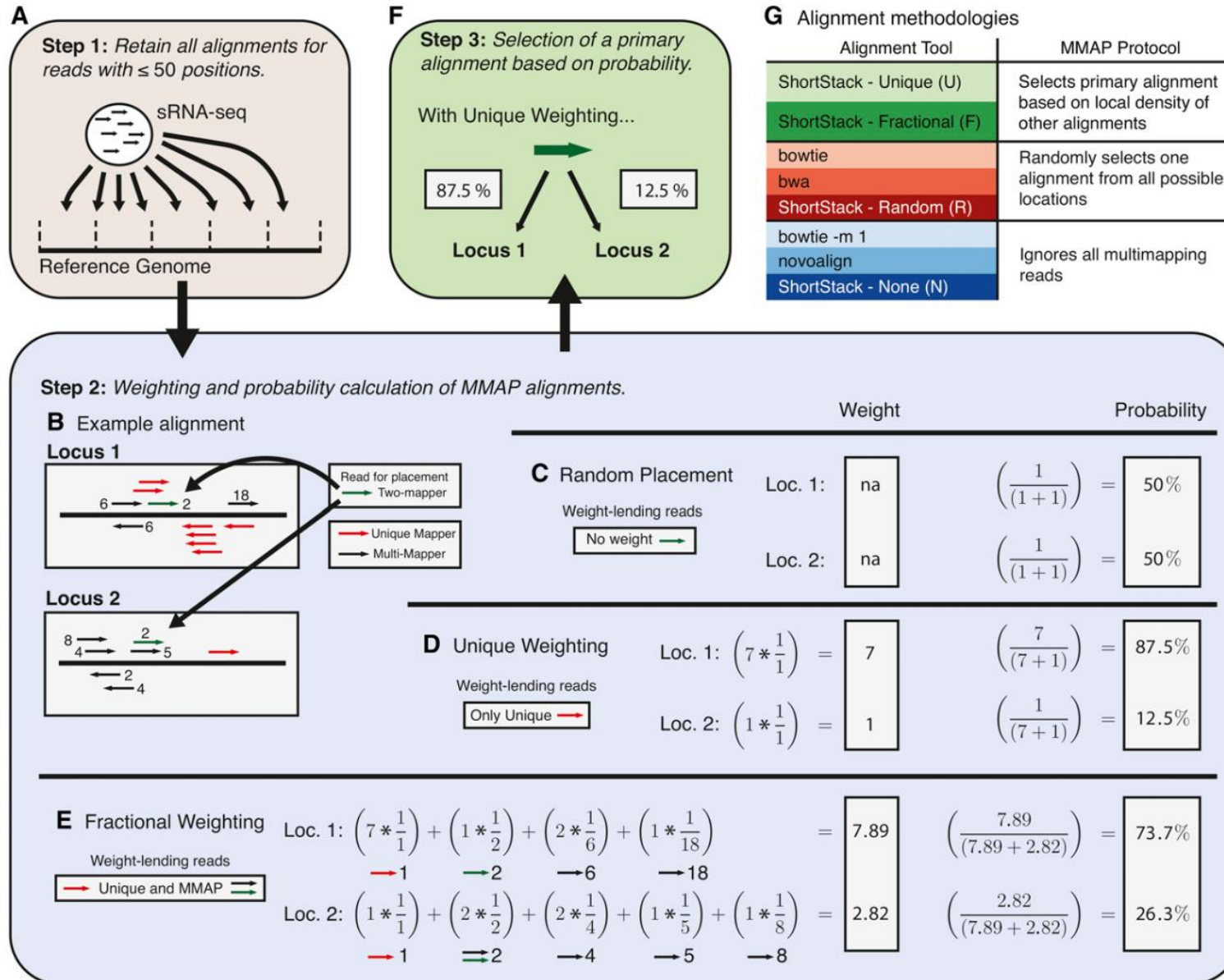
Evaluación de la distribución del mapeo de lecturas de *Multi-Mapping*



Las lecturas de **MMAP** se producen cuando hay múltiples posibles alineamientos con la mejor puntuación en el genoma de referencia.

Approach to handle multireads	Read distribution representation	Counts
Ignore		G1: 10 reads G2: 6 reads
Count once per alignment		G1: 18 reads G2: 14 reads
Split them equally		G1: 14 reads G2: 10 reads
Rescue based on uniquely mapped reads		G1: 15 reads G2: 9 reads
Expectation-maximization		G1: 15 reads G2: 9 reads
Read coverage based methods		G1: 15 reads G2: 9 reads

Evaluación de la distribución del mapeo de lecturas *Multi-Mapping*



BIOINFORMATICS

ShortStack: Comprehensive annotation and quantification of small RNA genes

MICHAEL J. AXTELL¹

Department of Biology, and Huck Institutes of the Life Sciences, Penn State University, University Park, Pennsylvania 16802, USA

Alineamiento y asignación de lecturas – *Cross-mapping ncRNAs*

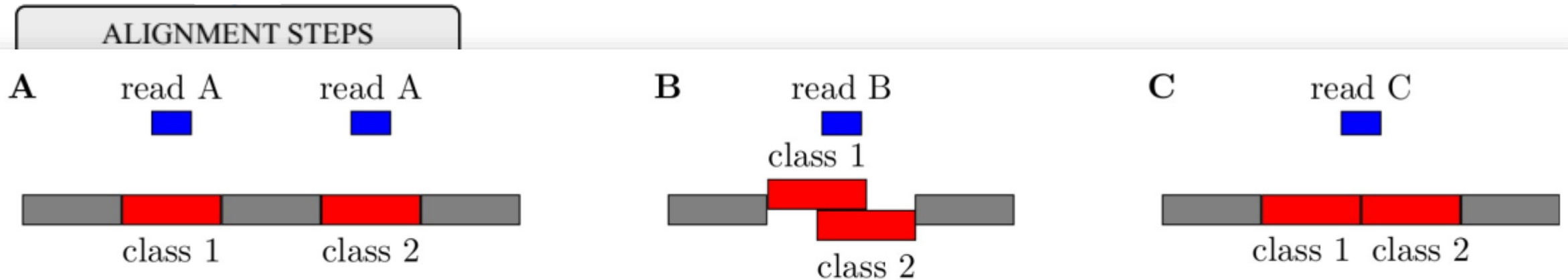


Fig 1. Ambiguous cases of sRNA class quantification. In each figure, the read is in blue, intergenic regions are grey, and known annotated sRNAs are red. (A) The read A maps at two different locations, and each mapping overlaps with a different sRNA annotation class. (B) Two sRNA annotation classes overlap in the annotation file. (C) The read overlaps two different sRNAs annotation classes that do not overlap each other.

Custom scripts using
Samtools

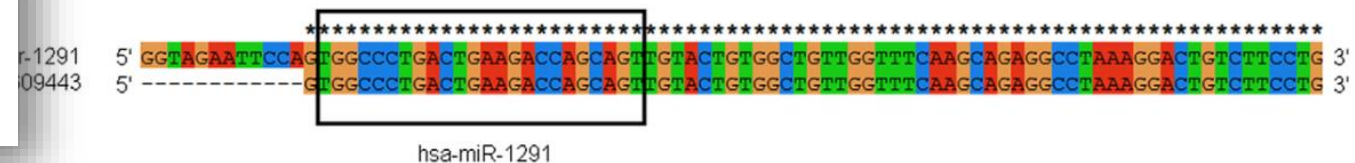
OPEN ACCESS Freely available online

PLoS one

Cross-Mapping Events in miRNAs Reveal Potential miRNA-Mimics and Evolutionary Implications

Li Guo¹, Tingming Liang², Wanjun Gu³, Yuming Xu¹, Yunfei Bai¹, Zuhong Lu^{1,3*}

Python packages



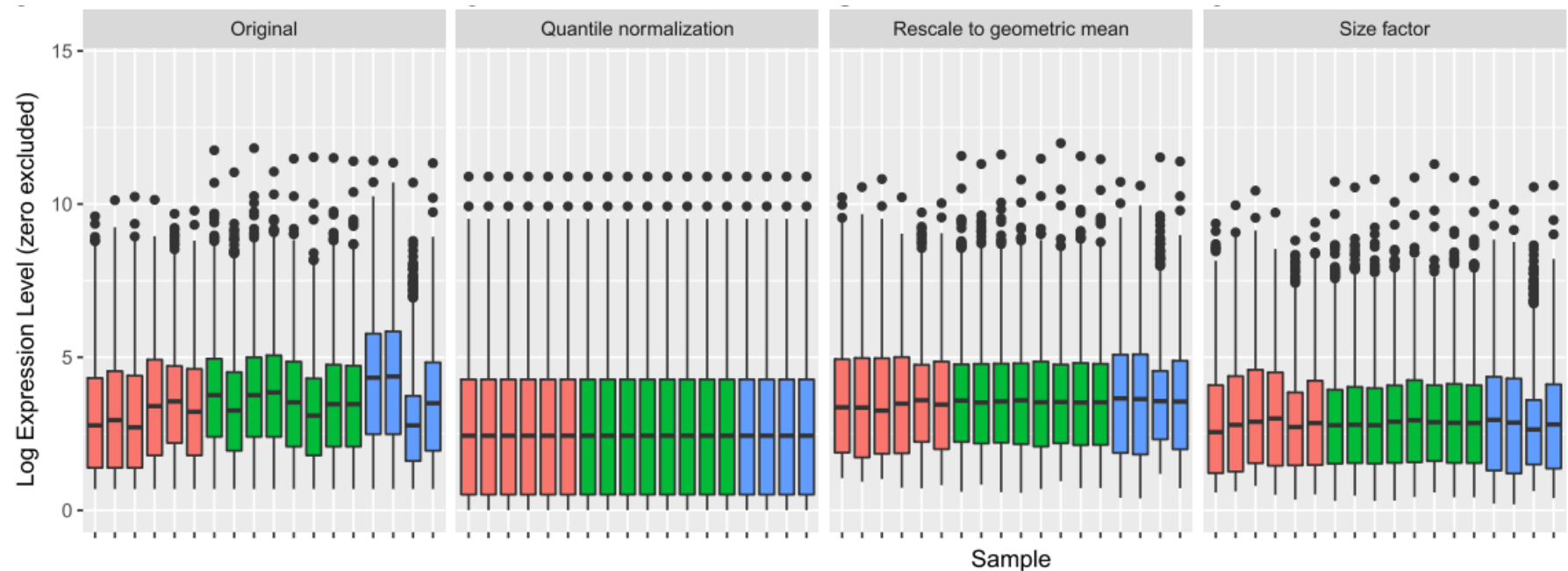
Normalización de los datos

- La normalización (el **proceso de reducir el error técnico o la variación entre muestras**) es fundamental para un análisis de expresión preciso.
- El objetivo de la normalización es ajustar los datos para **eliminar la variación entre muestras no relacionada con la condición biológica** que se estudia, facilitando así la identificación de las diferencias biológicas relevantes.
- La **expresión media global** de los datos de miARN como normalizador.
- Se basa en la suposición de que, aunque miARN específicos pueden variar entre muestras biológicas debido a causas biológicas, se espera que el patrón general de expresión de los miARN sea invariable.
- Otro enfoque es el uso de **miARN de control sintéticos añadidos (spike-in)** que se introducen en la muestra de ARN en un rango de cantidades conocidas y que normalmente no están presentes en la muestra (por ejemplo, miARN de plantas).
- **RPM o CPM**: un enfoque común es **representar la abundancia de un miARN determinado** en una muestra como el porcentaje del total de lecturas (o, en algunos casos, del total de lecturas mapeables).

Normalización de los datos

- La normalización (el proceso de reducir el error técnico o la variación entre muestras) es fundamental para un análisis de expresión preciso. ***Distinguir la verdadera diferencia biológica y el ruido aleatorio***
- **Profundidad de cobertura** — definida como el número promedio de lecturas por molécula
- *“Currently, a method is often chosen by the analyst based on personal preference and convenience rather than objective criteria governed by the data.”*

Su desempeño
depende en gran
medida de los datos
bajo estudio



Normalización de los datos

- **Scaling-based methods:**
 - Calcula un **factor de escalado basado en los datos de cada muestra** y divide sus recuentos por este factor.
- **Regression-based methods:**
 - Se basa en una **regresión lineal**, que normalmente adopta una **covariable que representa la variación de profundidad** e incluye esta covariable en un marco de regresión para el análisis de la expresión diferencial.
- Existen algunos paquetes de software diseñados para normalizar los datos de RNA-Seq antes del estudio de expresión diferencial (DE).
 - *Trimmed Mean Method* in **edgeR** (Robinson et al. 2010)
 - **DESeq/DESeq2** que utiliza *negative binomial model* (Robinson and Oshlack 2010)
- Trimmed Mean Method (TMM), quantile normalization (QN), scaling normalization (SC).

Normalización de los datos - DANA

- 'Data-driven Normalization Assessment' (DANA) eliminates as much as possible the variations of proportionality that are not true ("effective") biological signals
- TCGA-UCEC endometrial cancer RNA-seq data

C
100%
90%
80%
70%
60%
cc⁺; Biological signal preservation

Published online 21 February 2022

Nucleic Acids Research, 2022, Vol. 50, No. 10 e56
<https://doi.org/10.1093/nar/gkac064>

Depth normalization of small RNA sequencing: using data and biology to select a suitable method

Yannick Düren^{1,2}, Johannes Lederer¹ and Li-Xuan Qin^{2,*}

¹Department of Mathematical Statistics, Ruhr-University Bochum, Universitätsstraße 150, 44801 Bochum, Germany and ²Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

SPECIAL SERIES: INFORMATICS TOOLS FOR CANCER RESEARCH AND CARE

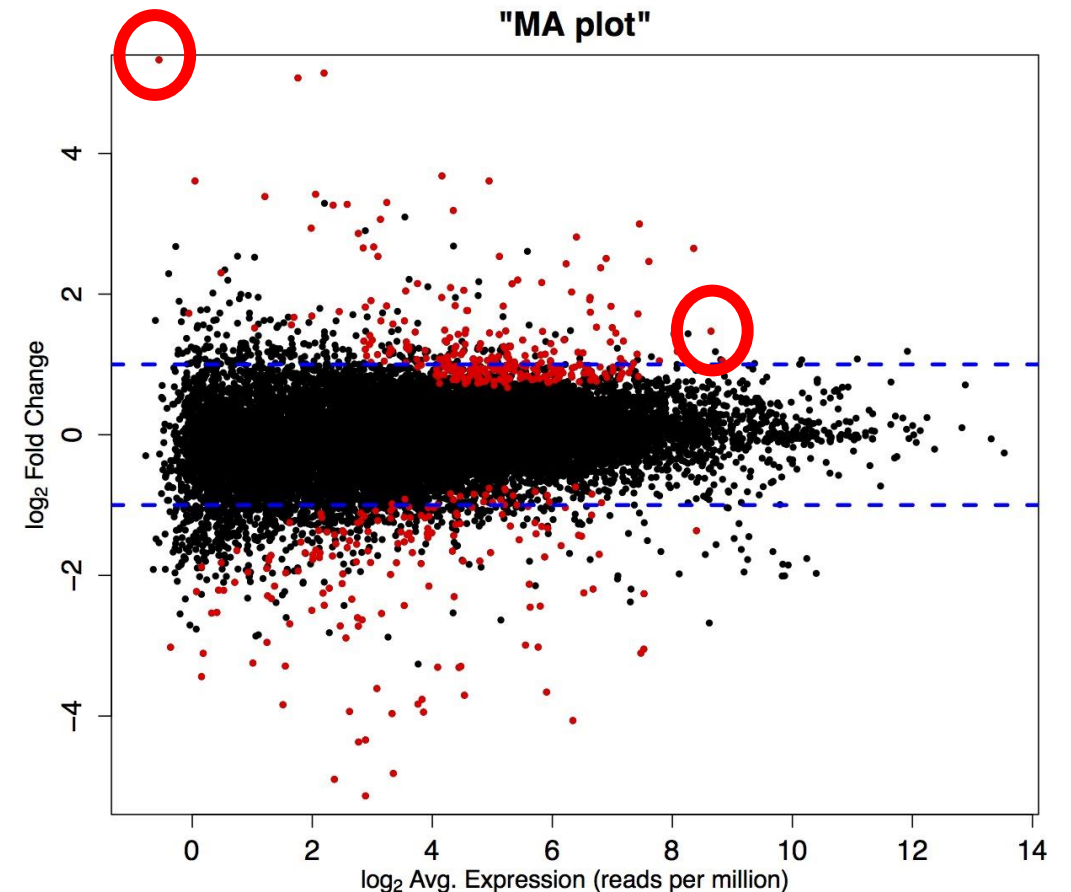
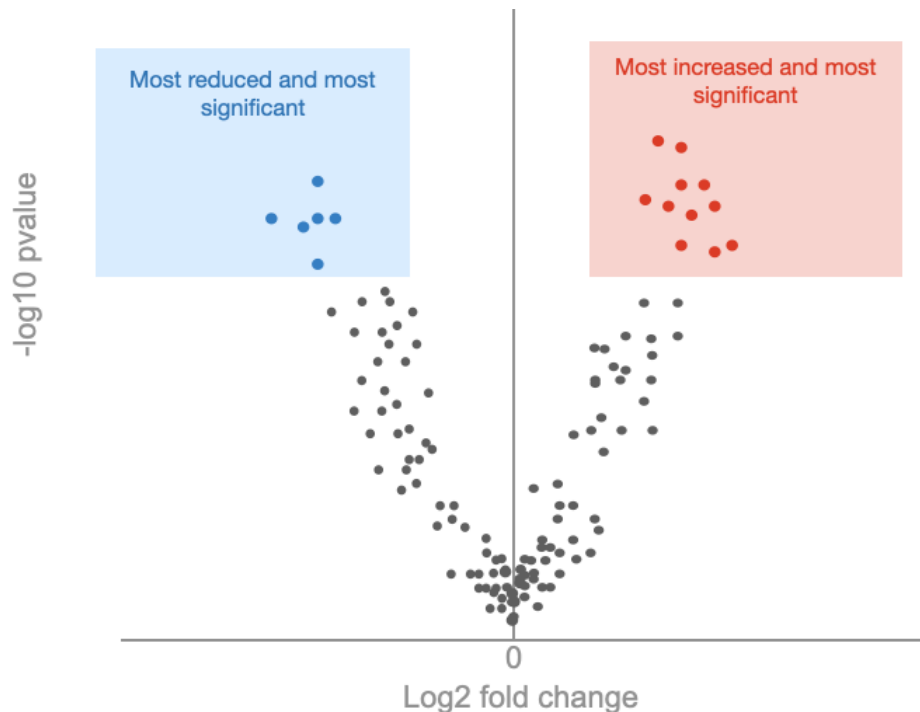
Statistical Assessment of Depth Normalization for Small RNA Sequencing

Li-Xuan Qin, PhD¹; Jian Zou, MS¹; Jiejun Shi, PhD¹; Ann Lee, MD²; Aleksandra Mihailovic³; Thalia A. Farazi, MD, PhD³; Thomas Tuschl, PhD³; and Samuel Singer, MD²

Among them, **Trimmed mean of M-values (TMM)** was a better scaling method, whereas the median and the upper quartiles were consistently the worst performers.

Expresión diferencial - DEGs

- El cálculo de la expresión diferencial de miARN entre grupos (comparaciones entre dos o más grupos)
- La magnitud del *Fold Change* que constituye una expresión diferencial significativa dependerá del contexto experimental.
- La significancia estadística y el *false discovery rate*



Non-coding RNA

