Validación de una aplicación para la anotación de isomiRs

Rafael Alis Pozo
Master en Bioinformática y Bioestadística
Área 38

Lorena Pantano Rubino
María Jesús Marco Galindo
05/06/2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Validación de una aplicación para la anotación de isomiRs*
Nombre del autor:	*Rafael Alis Pozo*
Nombre del consultor/a:	*Lorena Pantano Rubino*
Nombre del PRA:	*María Jesús Marco Galindo*
Fecha de entrega (mm/aaaa):	05/06/2018
Titulación:	Master en Bioinformática y Bioestadística
Área del Trabajo Final:	*El nombre de la asignatura de TF*
Idioma del trabajo:	*TFM-Estadística y Bioinformática 38*
Palabras clave	*miRNA, IsomiR,*

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

Los micro RNAs (miRNAs) son una clase de *small* RNAs no codificantes de un tamaño alrededor de 22 nucleótidos con función represora de la expresión génica. Se ha demostrado que un único miRNA *locus* puede dar lugar a diferencias secuencias de miRNAs como resultado del proceso de maduración. Los *small* RNA que presentan estas variaciones en relación a la secuencia referencia se denominan isomiRs. La determinación de los niveles de expresión de miRNAs mediante secuenciación masiva (NGS) es una técnica muy extendida. Sin embargo, la anotación de las lecturas habitualmente se realiza sin tener en cuenta las isoformas a pesar de que se ha mostrado que la actividad biológica del miRNA se ve afectada por la presencia de modificaciones. En este trabajo hemos ayudado a implementar un formato de fichero para la anotación de *small* RNA-seq compatible con el standard GFF3. Además, hemos desarrollado herramientas bioinformáticas para comprobar el formato e importar

estos ficheros a un entorno Python. Por último, hemos comprobado la utilidad de este formato de fichero y de las herramientas desarrolladas analizando un set de datos experimentales.

Abstract (in English, 250 words or less):

Micro RNAs (miRNA) are a class of non-coding small RNAs with a length around 22 nucleotides and a repressor function of gene expression. It has been shown that a single miRNA *locus* might yield several distinct sequences as result of the maturation processes. The small RNA with these variations in relation to the reference miRNA are called isomiRs. The determination of the expression levels of miRNA by massive Next Generation Sequencing (NGS) is very popular. However, sequence annotation is usually performed ignoring the miRNA isomers, in spite of been shown that the miRNA biological activity is regulated by the modifications in the sequence. In this work we have helped to implement a file format to annotate small RNA-seq data, consistent with the GFF3 standard. Moreover, we have developed bioinformatic tools to check that file format and import the data to Python environment. Lastly, we have checked the usefulness of the file format and the developed tools analyzing a set of experimental data.

Índice

Lista de figuras

Lista de tablas

I. Introducción

1. Contexto y justificación del Trabajo

1.1. Micro RNAs

Los micro RNAs (miRNAs) son una clase de RNAs no codificantes de un tamaño alrededor de 22 nucleótidos. Tienen una función represora de la expresión génica a nivel post-transcripcional al inhibir la traducción mediante la degradación de RNAs mensajeros. A lo largo del genoma humano existen miles de genes que codifican un miRNA primario que, tras un proceso regulado por los complejos Drosha y Dicer acabará en un miRNA maduro con funciones represoras (Figura 1).

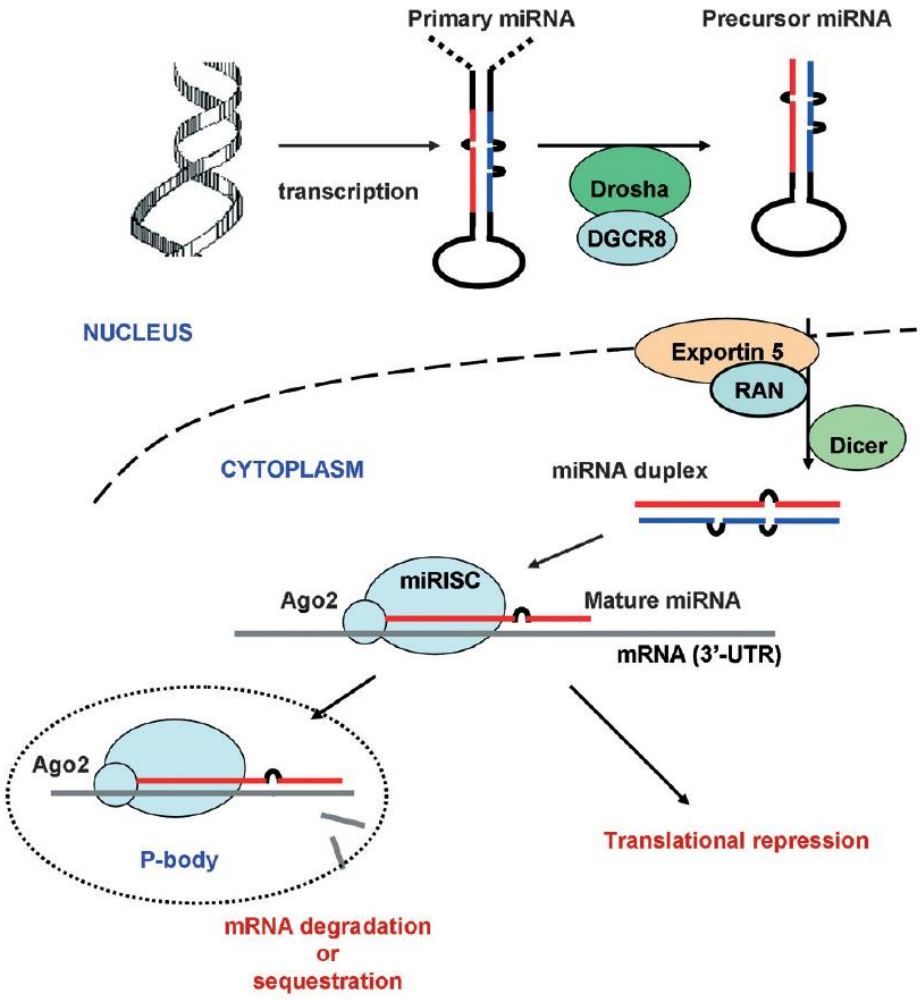

Figura 1. Biosíntesis de micro RNA (miRNA). Adaptado de Williams (2008).

Los miRNAs son transcritos en el núcleo como miRNAs primarios de unos 2000 nucleótidos de longitud. Son procesados por la enzima de tipo RNasa III Drosha en asociación con la proteína de unión a RNA DGCR8 para dar un miRNA precursor que será transportado a través de la membrana nuclear por Exportin 5 (Williams, 2008). En el citoplasma el miRNA precursor es cortado por Dicer, otra enzima RNasa III, resultando en un miRNA maduro de doble cadena. Únicamente la cadena guía del miRNA entra en el complejo de silenciamiento inducido por miRNA (miR-ISC), donde se une al RNA mensajero objetivo en la zona no traducida a 3' de forma complementaria parcialmente (Williams, 2008). Esto resulta en la inhibición de la traducción del RNA mensajero y por tanto la modulación de la producción de proteínas. Además, el complejo formado puede entrar en unos complejos proteicos especializados llamado P-bodies donde el RNA mensajero puede ser degradado (Williams, 2008).

Un gran número de estudios en diversas especies han demostrado un papel relevante de esta capa de regulación de la expresión génica en múltiples eventos biológicos como la aparición de cáncer (Liu *et al*, 2018a; Wong *et al*, 2018), respuesta inmune (Zhang *et al*, 2018) o el desarrollo (Gross *et al*, 2017; Liu *et al*, 2018b).

1.2. IsomiRs

Los miRNA han sido tradicionalmente anotados como una única secuencia validada experimentalmente y mediante métodos de predicción bioinformáticos. Sin embargo, se ha encontrado evidencia de que un único miRNA *locus* puede generar varias secuencias diferentes debido a la maduración y procesamiento de los miRNA (i.e. cortes imprecisos de Drosha y Dicer, adiciones a 3', edición de RNA y SNPs) (Guo *et al*, 2014; Morin *et al*, 2008). Las secuencias de miRNAs que resultan de estas variaciones de la secuencia referencia se denominan isomiRs. La **Figura 2** ejemplifica los diferentes tipos de isomiRs: adiciones o deleciones a 5' o 3', adiciones de bases dentro de la secuencia o polimorfismos de una base (SNP) en la secuencia del miRNA.

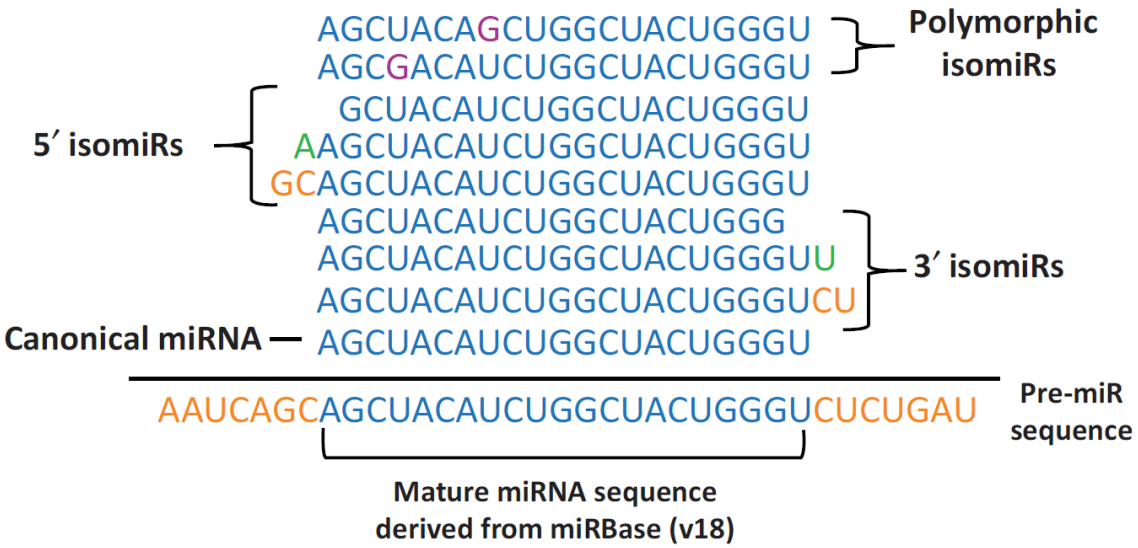

Figura 2. Representación esquemática de especies de isomiR. Adaptado de Neilsen et al (2012).

La **Figura 2** muestra las isoformas del miRNA humano (miR)-222 como ejemplo de la heterogeneidad de los isomiR. Las secuencias representadas se han obtenido de miRBase (Neilsen *et al*, 2012), la base de datos de referencia para la anotación de miRNAs. La secuencia canónica del miR-222 se representa en azul. Los isomiR 5' y 3' son variantes con secuencias diferentes en los extremos 5' y 3', respectivamente. Las modificaciones pueden coincidir con la secuencia génica (naranja) o no (verde). Los isomiR polimórficos (morado) contienen distintos nucleótidos dentro de las secuencias del miRNA.

La mayoría de los isomiR presentan modificaciones a 3'. Por tanto, ya que la acción de los miRNAs depende mayoritariamente de la región 5', no es de esperar que la mayoría de estos isomiR presenten diferentes especificidades en los mRNA a silenciar (Neilsen *et al*, 2012). Sin embargo, las modificaciones a 5' pueden tener grandes efectos en cuanto a la especificidad de su acción (Lee *et al*, 2010). El extremo 3' de los miRNAs juega un papel en la degradación de estas especies. Por tanto, los isomiR con modificaciones a 3' pueden presentar mayores o menores ritmos de degradación lo que puede tener impacto en la expresión de los genes que regulan (Neilsen *et al*, 2012).

Diversos estudios han mostrado diferentes papeles de las isoformas de un miRNA. Por ejemplo, la transfección de variantes 3' del miR-222 han mostrado tener un gran impacto en el fenotipo celular, promoviendo la apoptosis las

variantes más largas (Yu *et al*, 2017). La sobreexpresión de 5’ isomiR de miR-34 y miR-449 inducen distintos patrones de expresión génica y efecto biológicos que los miRNAs canónicos (Mercey *et al*, 2017).

1.3. Justificación del trabajo

La determinación de los niveles de expresión de miRNAs mediante secuenciación masiva (NGS) es una técnica muy extendida. Sin embargo, la anotación de las lecturas habitualmente se realiza sin tener en cuenta las isoformas. La cuantificación de los niveles de expresión de los isomiRs ayudaría a avanzar en el conocimiento del papel biológico de las isoformas de miRNAs. Para ello es necesario desarrollar nuevas herramientas bioinformáticas que permitan realizar estos análisis. Con este proyecto pretendemos implementar un formato de fichero GFF3 adaptado para las aplicaciones relacionadas con el análisis de isomiRs. Además, desarrollaremos herramientas bioinformáticas para la migración de formatos de datos de miRNAs y la comprobación de estos formatos en Python. Por último, comprobaremos la utilidad de este formato de fichero y de las herramientas bioinformáticas analizando datos experimentales. Los datos consistirán en una muestra de plasma de humano secuenciado por diferentes laboratorios anónimos. Esto nos ayudara a determinar qué tipos de isomiRs son debidos a diversidad técnica y cuales son reproducibles y, por lo tanto, reales. Mediante este proyecto ayudaremos a la comunidad científica a analizar datos de miRNA y a la democratización de análisis de isomiRs.

2. Objetivos del Trabajo

Establecimos los siguientes **objetivos generales**:

1. Desarrollar e implementar de un formato de fichero GFF3 para la anotación de la expresión de isomiRs.
2. Desarrollar herramientas bioinformáticas que permitan la importación de un fichero GFF3 de anotación de *small* RNA a un entorno Python para su análisis.
3. Utilizar las herramientas desarrolladas para analizar un set de datos experimentales.

Que desglosamos en **objetivos específicos**:

1.1. Registrar la opinión de desarrolladores que trabajan en isomiRs sobre el formato de fichero GFF3.

1.2. Modificar el formato ya propuesto de fichero GFF3 con las propiedades necesarias para satisfacer los comentarios de la comunidad de desarrolladores y usuarios.

2.1. Desarrollar código en Python para la importación de datos de expresión de *small* RNA a partir de un fichero en el nuevo formato GFF3.

2.2. Desarrollar código en Python para comprobar la coincidencia del formato de ficheros con el nuevo formato GFF3.

3.1. Obtener los niveles de expresión de miRNA e isomiRs en un set de datos experimentales de *small* RNA-seq.

3.2. Observar los niveles de expresión en función del tipo de isomiRs.

3.3. Analizar el efecto de los protocolos para la preparación de librerías en los niveles de expresión de isomiRs.

3.4. Observar la utilidad del tratamiento bioinformático y la utilidad del tipo de fichero desarrollado para el análisis de datos de *small* RNA-seq.

3. Enfoque y método seguido

Hemos optado por estrategias diferentes en cuanto a la forma de dar respuesta a cada uno de los objetivos generales. En el caso del primer objetivo, el desarrollo del formato de fichero se ha realizado en conjunto con una comunidad de desarrolladores que trabajan con datos de *small* RNA-seq e isomiRs (https://github.com/miRTop/incubator/blob/master/format/definition.md).

Hemos utilizado una herramienta como el repositorio GitHub para compartir y debatir el formato y llevar un registro de las versiones. A continuación, hemos desarrollado código en Python para implementar el tipo de fichero GFF3, incluyendo rutinas para la validación del formato, en un producto actualmente disponible para el tratamiento de datos de *small* RNA-seq, mirTop. Cabe destacar en este apartado que este formato de fichero lleva un tiempo en debate y desarrollo en la comunidad y que su definición ya estaba empezada al comienzo de este trabajo. En el marco de este proyecto hemos finalizado la definición del formato y su implementación.

Por último, para dar respuesta al tercer objetivo, hemos analizado datos mediante funciones y utilidades desarrolladas en Python. Hemos utilizado un set de datos de secuenciación masiva de *small* RNA proveniente de un proyecto compartido entre varios laboratorios con el objetivo de determinar el efecto de diversos protocolos para la preparación de librerías de DNA complementario (cDNA) previo a la secuenciación masiva. Gracias a este proyecto, dispondremos de los ficheros de lecturas correspondientes a muestras sintéticas y reales de plasma tratadas de diversas formas y secuenciadas por diversos laboratorios. De esta forma, hemos podido determinar la utilidad del nuevo formato de anotación de expresión. Además, hemos podio investigar el resultado del tratamiento bioinformático en los niveles de expresión de isomiRs discriminando por su tipo.

La naturaleza de este campo de investigación y del problema abordado junto con el marco temporal para la realización de este TFM ha hecho necesario implementar un enfoque recursivo en todas las fases del proyecto.

4. Planificación del Trabajo

Para dar respuesta a los objetivos en este trabajo establecimos diferentes tareas a cumplir en un marco temporal establecidos por un calendario. Además, establecimos una serie de hitos para monitorizar el avance del proyecto e hicimos un análisis de riesgos.

4.1. Tareas

Establecimos las siguientes tareas en función de los objetivos propuestos:

Tabla 1. Tareas establecidas, objetivo al que responden y numero de semanas a emplear en su cumplimiento.

Objetivo Específico	Tarea	Semanas
1.1	Definir los campos a registrar en el tipo de fichero GFF3.	1
1.1	Definir la información contenida en la cabecera en el tipo de fichero GFF3.	1
1.1	Definir los campos obligatorios y los atributos opcionales en el tipo de fichero.	1
1.2	Proponer el formato de fichero y reflejar los cambios requeridos por la comunidad.	2
2.1	Desarrollar el código en Python que transformar un fichero de anotación de expresión de isomiRs obtenido mediante *seqbuster* en el tipo de fichero definido por la comunidad.	4
2.2	Desarrollar el código en Python para implementar un módulo que compruebe el formato desarrollado.	4
3.1	Instalar y comprobar el pipeline *bcbio-nextgen*.	1
3.1	Analizar datos de secuenciación masiva de mediante la *pipeline bcbio-nextgen* para *small* RNA-seq.	1
3.1	Transformar los ficheros de salida del pipeline al formato GFF3	1
3.1	Compactar todos los ficheros de salida correspondientes a cada muestra en único fichero GFF3 que refleje los niveles de expresión en cada muestra.	1
3.2	Desarrollar una función que cargue en un entorno Python el fichero GFF3 en un tipo de datos que permita su análisis en ese entorno de programación	1
3.2	Analizar las distribuciones de los niveles de expresión de miRNA e isomiRs.	4
3.2	Analizar la necesidad de establecer criterios para filtrar los datos de expresión de isomiRs en función de las distribuciones anteriores.	4
3.2	Realizar un análisis de la expresión de los isomiRs en cada tipo de muestra en función de sus modificaciones.	4

3.3	Analizar los niveles de expresión de isomiRs en cada tipo de muestra en función del protocolo empleado para la preparación de cDNAs.	4
3.4	Analizar la validez del tratamiento bioinformático y la utilidad del tipo de fichero en función de los resultados obtenidos	1

4.2. Calendario

Hemos seguido el siguiente calendario para el desarrollo de las tareas correspondientes a cada objetivo específico:

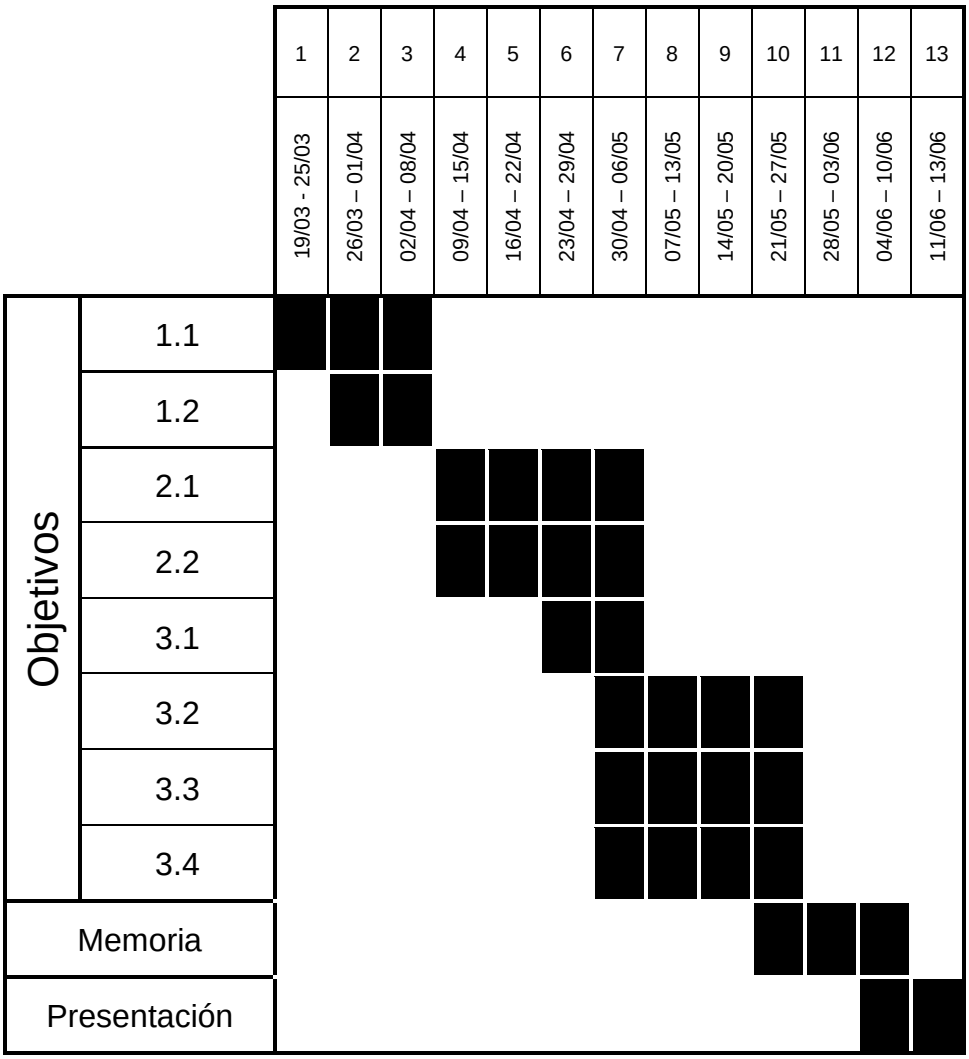

Figura 3. Calendario de cumplimiento de tareas en función del objetivo al que responden.

4.3. Hitos

Establecimos los siguientes hitos para marcar el desarrollo del proyecto:

Tabla 2. Hitos establecidos en el presente proyecto.

Hito	Descripción	Semana
1	Establecimiento del formato de fichero GFF3	3
2	Implementación de los módulos de transformación de formato y de comprobación de formato	6 - 7
3	Obtención de los datos de expresión de *small* RNA-seq en el set de datos experimental en el formato definido	7
4	Finalización del análisis de los datos de expresión	10

4.4. Análisis de riesgo

Identificamos diferentes amenazas para el desarrollo del programa de trabajo como la capacidad de cálculo y almacenaje de los recursos informáticos al alcance del alumno para instalar la *pipeline* a utilizar, *bcbio-nextgen*, y para realizar los análisis bioinformáticos en esta (*i.e. trimming* y limpiado de secuencias, alineación a genoma). El manejo y procesamiento de grandes volúmenes de datos requiere de una gran capacidad de procesamiento y de almacenamiento. En caso de presentarse este problema, el tratamiento de los datos de secuenciación se hubiera realizado mediante máquinas virtuales en entornos de programación en la nube (e.g. Amazon AWS o Google Cloud Platform). Sin embargo, no ha sido necesario recurrir a esta estrategia y se han podido cumplir las tareas con los recursos informáticos al alcance del alumno.

5. Breve sumario de productos obtenidos

Como resultado de este trabajo obtenido los siguientes productos:

1. Funciones para el importe y comprobación del formato de ficheros GFF3 para la anotación de datos de expresión de *small* RNA-seq en Python. Se adjuntan como fichero adjunto (Anexo 1.py).

2. La presente memoria y los documentos anexos que lo acompañan.

6. Breve descripción de los otros capítulos de la memoria

Los capítulos II, III y IV de la presente memoria detallarán los procedimientos desarrollados para dar respuesta a los objetivos 1, 2 y 3. Además, en estos capítulos se presentarán los resultados obtenidos como fruto de estos procedimientos. Por último, se discutirán estos resultados. En el capítulo V se presentarán las conclusiones globales del desarrollo del presente trabajo. El resto de capítulos comprenderán un glosario de términos, la bibliografía y un listado de documentos anexos, por este orden.

II. Procedimientos y resultados relativos al objetivo 1

1. Procedimientos relativos al objetivo 1

Para dar respuesta al objetivo 1 (*desarrollar e implementar de un formato de fichero GFF3 para la anotación de la expresión de isomiRs)*, recogimos el debate y propuestas que la comunidad de desarrolladores y científicos ha realizado en el hilo abierto a tal efecto en el repositorio GitHub del proyecto mirTop (https://github.com/miRTop/incubator/blob/master/format/definition.md). Los foros de discusión y debate en los que los usuarios expertos proponen modificaciones de software o de formatos es una herramienta fundamental en el desarrollo de aplicaciones en un campo de tan rápida evolución como el de la bioinformática. Hemos utilizado los *inputs* de la comunidad en este foro para establecer un formato para la definición de resultados de *pipelines* que analizan datos de secuenciación de *small* RNA. Este formato de fichero cumple con los requisitos del standard GFF3 (Generic Feature Format Version 3). Cabe destacar que el formato de fichero no está cerrado y puede modificarse en el futuro como resultado de las conclusiones de este trabajo, de las aportaciones de la comunidad y las necesidades futuras derivadas de la implantación del fichero en las diversas *pipelines* disponibles.

2. Resultados relativos al objetivo 1

En relación objetivo 1 (desarrollar e implementar de un formato de fichero GFF3 para la anotación de la expresión de isomiRs), hemos recogido la definición del formato de fichero realizado por la comunidad de desarrolladores y científicos que trabajan en secuenciación masiva de *small RNA*. El formato de fichero, compatible con el standard GFF3 de detalla a continuación:

Las líneas que forman parte del encabezamiento son precedidas por ## y puede contener las líneas descritas en la **Tabla 3**.

Tabla 3. Descripción del formato de la información contenida en el encabezado del tipo de fichero GFF3 para la anotación de *small* RNA.

Línea	Descripción	Cabecera de linea	Opción Múltiple Líneas	Opcional
1	Base de datos utilizada para la alineación, incluyendo versión y link.	`##source-ontology LINK TO DATABASE`	NO	SI
2	Comandos usados para la generación del fichero. Detalles de la eliminación de adaptadores, filtrado, alineamiento y otros detalles.	`## CMD:`	SI	SI
3	Versión del genoma o base de datos utilizada.	`## REFERENCE:`	NO	NO
4	Nombre de las muestras en la columna de Expression en los atributos. Separadas por espacios.	`## COLDATA:`	SI	NO
5	Versión del archico GFF de *small RNA* utilizado.	`## VERSION: 0.9`	NO	NO
6	Valores posibles de la etiqueta FILTER en la	`## FILTER:`	NO	SI

El cuerpo del fichero contiene necesariamente 9 columnas separadas por tabuladores. El contenido y formato de las columnas 1 a 8 se detallada en la **Tabla 4**.

Tabla 4. Descripción de las columnas 1 a 8 del cuerpo del fichero GFF3 para la anotación de *small* RNA.

Columna	Contenido	Descripción	Posibles valores
1	seqID	Nombre del precursor	
2	source	Base de datos utilizada para la anotación con la versión tras "_"	miRBase, mirDBgene,tRNA
3	type	Tipo de *small* RNA	ref_miRNA / isomiR
4	start	Posicion de inicio de la secuencia en el precursor	
5	end	Posición de final de la secuencia en el precursor	

6	score	Puntuación de mapeo de la secuencia u otra puntuación otorgada por la herramienta de alineamiento	
7	strand	En caso de alinear sobre precursores ha de ser “+”. En caso de alinear sobre el genoma puede ser “+” o “-“.	+/-
8	phase	Utilizado para sequencias de tipo CDS. Sin uso para *small* RNA.	.

La columna 9 del formato contiene una variedad de atributos de cada una de las secuencias. Cada atributo está precedido de una etiqueta y separada por “;”. La **Tabla 5** describe las posibles etiquetas.

Tabla 5. Etiquetas en la columna de atributos del fichero GFF3 para la anotación de *small* RNA.

Etiqueta	Descripción	Opcional
UID	Identidad única para la secuencia	NO
Read	Nombre de la secuencia	NO
Name	Nombre del *small* RNA maduro	NO
Parent	Nombre del *hairpin* precursor	NO
Variant	Cambios en la secuencia en relación con el miRNA de referencia. Varios valores posibles separados por “,”.	NO
Changes	Similar al anterior, indica los nucleótidos añadidos (mayúsculas) o eliminados (minúsculas). Varios valores posibles separados por “,”.	SI
Cigar	Cadena CIGAR	NO
Hits	Número de coincidencias en la base de datos.	NO
Alias	Nombres en bases de datos. Varios valores posibles separados por “,”.	SI
Genomic	Posiciones en el genoma. Formato: `chr:start-end,chr:start-end`	SI
Expression	Numero de lecturas de la secuencia. Varios valores separados por “,” en el mismo orden que `## COLDATA:` en el encabezado.	NO

Filter	PASS o REJECT en función de parámetros del alineador. Pueden tener subclases como `PASS:te` o `REJECT:lowcounts`	NO
Seed_fam	Formato 2-8 nucleótidos y el miRNA de referencia que comparte el *seed* de la secuencia.	SI

Los valores del atributo *Variant* describen las modificaciones de la secuencia en relación con el miRNA de referencia. Estas modificaciones pueden ser cambios de base (SNP), adiciones en el extremo 3' de la secuencia (iso_add) no consistentes con la secuencia del miRNA precursor y adiciones (consistentes con la secuencia del miRNA precursor) o deleciones en los extremos 5' y 3' (iso_5p, iso_3p).

La **Figura 4** representa las posibles modificaciones y el formato de anotación usado en esta etiqueta.

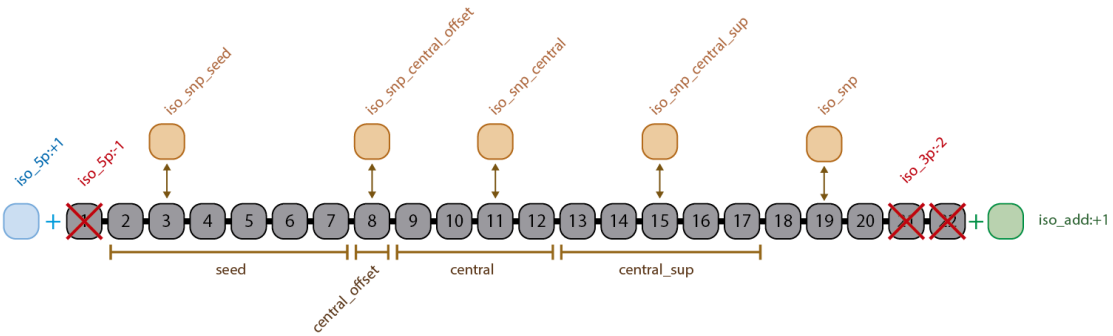

Figura 4. Representación de las posibles modificaciones con respecto al miRNA de referencia y la etiqueta que las identifica en la columna Variant del formato de fichero GFF3. En este ejemplo se representan modificaciones sobre un hipotético miRNA de 22 nucleótidos (representados como cajas de color gris). En azul se representan las adiciones (consistentes con la secuencia del miRNA precursor) y en rojo la deleciones a 5' o 3'. En verde se indican adiciones a 3' (no consistentes con la secuencia del miRNA precursor) y en naranja los posibles cambios de nucleótidos que se categorizan en función de la posición en la secuencia.

Las adiciones a 5' de la secuencia se etiquetan como iso_5p:+N mientras que las deleciones en este extremo se indican como iso_5p:-N, siendo N el número de bases añadidas. De igual forma se indican las adiciones y deleciones en el extremo 3' (iso_3p:+/-N). En estos casos, las adiciones son consistentes con la secuencia del miRNA precursor. Las adiciones a 3' de la secuencia que no son consistentes con la secuencia el miRNA precursor se anotan como iso_add:+N. Los cambios de base se indican mediante la etiqueta SNP que se acompañará de otros indicadores en el caso de encontrarse entre las posiciones 2 y 7 (iso_snp_seed), en la posición 8 (iso_snp_central_offset), entre las posiciones 9

y 12 (iso_snp_central) o 13 y 17 (iso_snp_central_sup). En caso de encontrarse en otras secciones de la secuencia se indicará como (iso_snp).

La combinación de la información contenida en la etiqueta *Variant* en combinación con la información de la etiqueta *Changes* permite reconstruir la secuencia del isomiR. En caso de que la lectura corresponda a un miRNA de referencia ambas etiquetas contendrán “NA”. La descripción del formato definitivo está disponible en el repositorio de GitHub del proyecto mirTop (https://github.com/miRTop/incubator/blob/master/format/definition.md).

3. Discusión de los resultados relativos al objetivo 1

Mediante la consecución de este objetivo hemos establecido un formato de fichero para la anotación de datos de expresión de *small* RNA-seq consistente con el formato GFF3 y que es capaz de reflejar la heterogeneidad de los isomiRs. Este formato de fichero contiene toda la información necesaria para la identificación de la secuencia. Además, con el uso del atributo *Changes* en combinación con el atributo *Variant* se puede reconstruir la secuencia del isomiR anotado sin necesidad de contener la cadena de nucleótidos en ningún campo del fichero, con el subsecuente ahorro de espacio de almacenamiento.

Con este versátil formato pretendemos proponer a la comunidad un tipo de fichero que pueda ser adoptado por todas las pipelines para estandarizar los ficheros de resultados de estas. De esta forma, se facilitarían los subsecuentes análisis bioinformáticos ya que no seria necesario utilizar las librerías específicas de cada alineador o *pipeline* en lenguajes de alto nivel (R o Python) para poder obtener un objeto de datos manejable en estos entornos.

III. Procedimientos y resultados relativos al objetivo 2

1. Procedimientos relativos al objetivo 2

Para dar respuesta al objetivo 2 (*desarrollar herramientas bioinformáticas que permitan la importación de un fichero GFF3 de anotación de small RNA a un entorno Python para su análisis*), hemos desarrollado en Python (versión 2.7) código que permite importar datos desde un fichero con el formato implementado en el objetivo anterior. El código se encuentra como documento anexo a esta memoria (**Anexo 1**).

El código se compone de dos funciones, la función "*load_gff3"* carga el contenido del fichero GFF3 en un *dataframe* de Pandas que devuelve como resultado. La función "*load_check_gff3"* en primer lugar comprueba el formato de la cabecera del fichero GFF3, a continuación, carga el fichero en un *dataframe* mediante la función "*load_gff3"* y comprueba la integridad de los datos en el *dataframe* en función del formato definido. Si existen violaciones del formato, la función "*load_check_gff3"* devuelve el valor False e información del error mediante la consola. Si no existe error, la función devuelve el *dataframe* cargado desde el fichero.

La función "*load_gff3"* carga las columnas 1 a 8 en columnas del *dataframe*. La columna 9 de atributos se chequea para identificar las etiquetas presentes en el fichero y se cargan en en el número de columnas necesarias en el *dataframe*. Las columnas correspondientes a los atributos *Expression* y *Variant* se tratan de forma independiente.

La columna *Expression* contiene los recuentos de cada secuencia en cada uno de las muestras que contiene el fichero separadas por ",". El código separa los datos de expresión de cada muestra y los carga en columnas con el nombre de la muestra. Una muestra que no haya sido detectada en una muestra contiene

un 0 en el campo de expresión en el fichero GFF3. En la implementación que hemos realizado el valor en este caso es numpy.nan.

La columna *Variant* contiene etiquetas que informan las modificaciones que aparecen en la secuencia en relación con el miRNA de referencia. Esta información se ha traducido a una matriz adjunta al *dataframe* en la que las columnas corresponden con las variaciones posibles contenidas en la columna *Variant*. Un 1 en una fila representa la presencia de la modificación correspondiente a esta columna. La ausencia de modificación se indica insertando el valor numpy.nan.

2. Discusión de los procedimientos relativos al objetivo 2

Para dar respuesta a este objetivo hemos desarrollado dos funciones (*load_gff3, load_check_gff3*). La función *load_gff3* carga el fichero directamente en un *dataframe* Pandas sin comprobar el formato. La función *load_check_gff3* comprueba el formato del fichero en dos fases. En primer lugar, comprueba el formato de la cabecera. A continuación, llama a la función *load_gff3* y comprueba que el contenido del fichero es cumple con las especificaciones establecidas en el objetivo 1. De esta forma podemos comprobar y cargar el fichero GFF3 o por únicamente cargarlo si no es necesario comprobar el formato. Además, la función *load_check_gff3* devuelve el valor False si el formato no es correcto, por tanto, se puede usar para comprobar el formato de ficheros GFF3 independientemente del uso del *dataframe* resultante.

Cabe destacar que utilizamos un set de ficheros con diferentes errores de formato para comprobar la utilidad de las funciones en diversos escenarios. La versión final de las funciones es robusta a las diversas variantes del formato que pueden encontrarse dentro del standard GFF3, como por ejemplo espacios tras “;” en la columna de atributos.

El resultado de la carga de los ficheros es un *dataframe* de Pandas. Este tipo de objeto disfruta de la gama de funciones y métodos que presenta la librería Pandas. Esta combinación proporciona una gran versatilidad y potencia al manejo de los datos de secuenciación masiva de *small* RNA. Además, la

implementación de los tipos de isomiR en columnas independientes proporciona fácil y claramente la capacidad de análisis discriminando por tipo de isomiR.

En relación al objetivo 2, podemos concluir que las funciones desarrolladas en Python cumplen de forma robusta y satisfactoria su cometido y así como la implementación de los datos de expresión en un objeto de tipo *dataframe* de la librería Pandas.

IV. Procedimientos y resultados relativos al objetivo 3

1. Procedimientos relativos al objetivo 3

Para dar respuesta al objetivo 3 (*utilizar las herramientas desarrolladas para analizar un set de datos experimentales*), analizamos un set de datos experimentales provenientes de muestras de plasma sanguíneo de humanos utilizado en un estudio con el fin de explorar la reproducibilidad de los resultados de *small* RNA-seq (Giraldez *et al*, 2017). Los datos de secuenciación se trataron mediante el *pipeline bcbio-nextgen* para *small* RNA y fueron subsecuentemente transformados en un único fichero con el formato GFF3 definido en el objetivo 1, utilizando el programa mirTop. Este fichero fue cargado en un entorno Python mediante las funciones desarrolladas en el objetivo 2. A continuación, exploramos los datos y analizamos la reproducibilidad de los tipos de isomiR teniendo en cuenta el protocolo utilizado para la preparación de las librerías.

1.1. Muestras

En este estudio, diversos laboratorios utilizaron múltiples protocolos para preparar librerías de cDNA a partir de muestras sintéticas de *small* RNA y de plasma de diversos sujetos. Estos laboratorios utilizaron NGS (Next Generation Sequencing) para secuenciar estas muestras. Los resultados fueron comparados entre laboratorios y protocolos mostrando una alta reproducibilidad de las medidas (Giraldez *et al*, 2017). En nuestro diseño hemos utilizado los datos de secuenciación de las muestras de plasma provenientes de 5 laboratorios que utilizaron 2 protocolos comúnmente utilizados. La muestra de plasma se compuso de un *pool* del plasma de 11 sujetos aparentemente sanos (21-45 años). Detalles del protocolo utilizado para la preparación del *pool* de plasma puede encontrarse como información suplementaria del artículo indicado (Giraldez *et al*, 2017).

1.2. Preparación de librerías

Todos los laboratorios utilizaron dos protocolos en cuadruplicado para realizar librerías de cDNA a partir de las muestras de plasma. En ambos protocolos se partió de 2.1 µL de plasma par la construcción de las librerías siguiendo el protocolo establecido por ambos fabricantes (NEBNext, New England Biolabs, USA; TruSeq, Illumina, USA).

1.3. Secuenciación

Las librerías fueron secuenciadas usando secuenciadores Illumina HiSeq 2500, Illumina HiSeq 4000 o Illumina NextSeq 500 en lecturas de 50 pb en modo single-end. Se requirió a los laboratorios al menos 8 millones de lecturas por librería. Los resultados de la secuenciación se plasmo en ficheros FASTQ para su subsecuente análisis.

1.4. Preprocesamiento de las muestras, mapeo y cuantificación de la expresión

Los ficheros FASTQ fueron procesados mediante la *pipeline bcbio-nextgen* (https://github.com/bcbio/bcbio-nextgen). Tras instalar la *pipeline*, utilizamos los parámetros standard para *small* RNA-seq establecidos por la *pipeline*. Se utilizó *atropos* para eliminar las secuencias de adaptadores, el alineador *bowtie* para la alineación sobre el genoma hg19 y *seqcluster* para detectar transcritos cortos. Los miRNAs fueron detectados mediante *miraligner* usando la base de datos de humano *miRbase* versión 22. Los tRNAs fueron detectados y filtrados mediante la herramienta *tdrmapper*. La calidad de las lecturas se comprobó mediante FastQC.

El resultado final del proceso fueron ficheros de salida de *miraligner* con un formato y extensión propios (.*counts*).

1.5. Compactación de los datos mediante mirTop

Utilizamos el programa mirTop (http://mirtop.github.io) para obtener un fichero con el formato GFF3 definido en el objetivo 1. Además, colapsamos los 40 ficheros (5 laboratorios x 2 protocolos x 4 réplicas) en uno único fichero reflejando

el numero de lecturas (numero de veces que dicha secuencia es detectada en la muestra) de cada secuencia en cada fichero en la columna *Expression*. Este fichero se adjunta como documento anexo (**Anexo 2**).

1.6. Carga, normalización y análisis de réplicas.

En primer lugar, cargamos los datos en un *dataframe* a partir del fichero GFF3 utilizando las funciones desarrolladas en el objetivo 2. Los datos de recuentos por muestra se transformaron en veces por millón de lecturas totales (cpm, *counts per million*). La consistencia de las réplicas se comprobó mediante un análisis de correlación.

2. Resultados relativos al objetivo 3

Tras cargar el fichero GFF3 en un *dataframe* analizamos el total de lecturas por laboratorio y protocolo en función del tipo de de *small* RNA (**Figura 5**). Observamos que los laboratorios 2 y 5 obtuvieron un mayor número de lecturas que el resto de los laboratorios en ambos protocolos, mientras que el laboratorio 1 obtuvo un alto número de lecturas con el protocolo NEBNext mientas que fue bajo con el protocolo TruSeq (**Figura 5**).

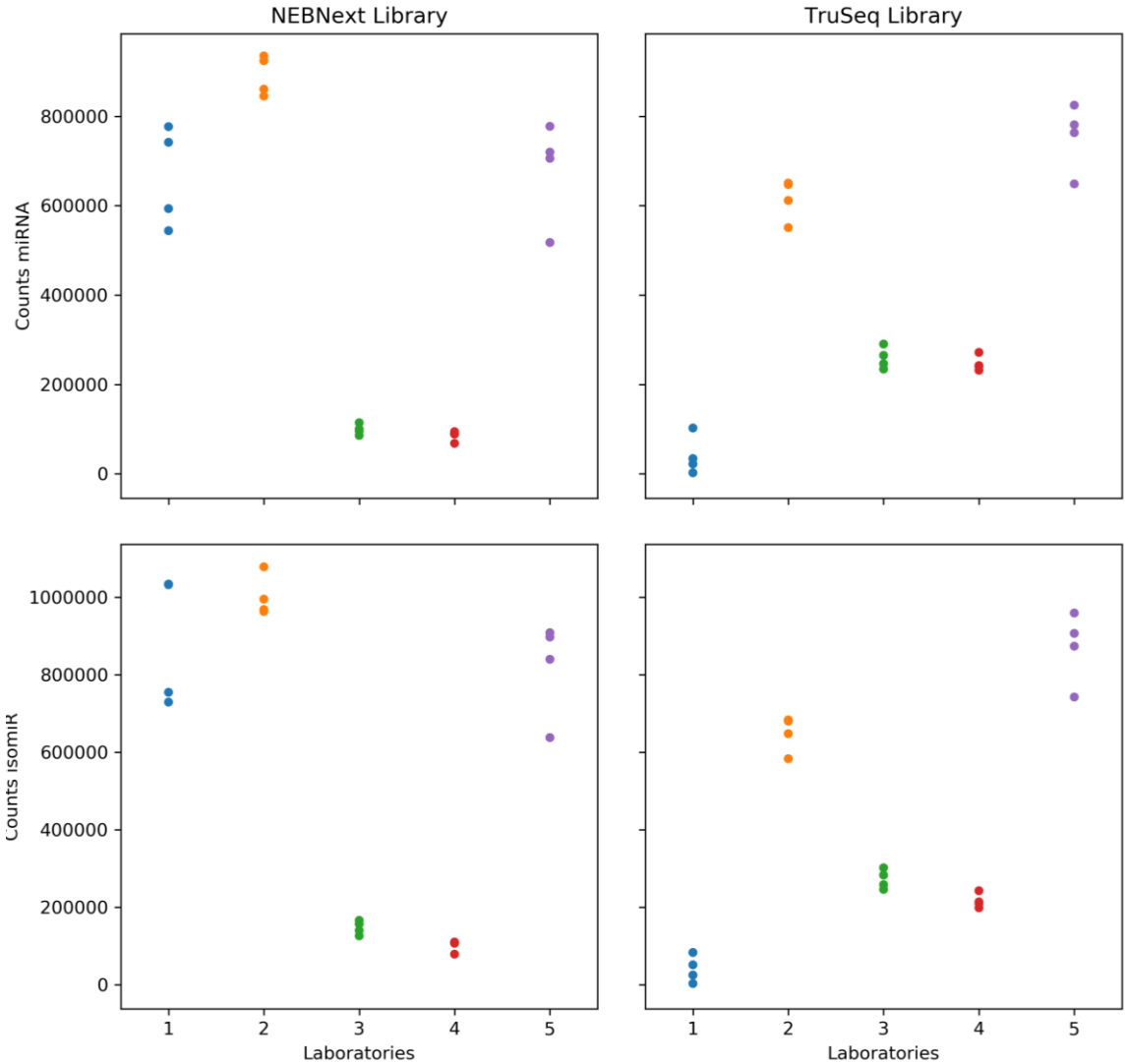

Figura 5. Total de lecturas por tipo de *small* RNA (miRNA arriba, isomiR abajo) en cada una de las réplicas (círculos) para cada laboratorio y protocolo.

A continuación, normalizamos los niveles de expresión, comprobamos la consistencia entre las réplicas realizadas por cada laboratorio en los dos protocolos mediante un análisis de correlación. En todos los casos observamos

una alta correlación entre todas réplicas (r ≥ 0.892). La **Tabla 6** muestra los coeficientes de correlación entre las muestras del laboratorio 1.

Tabla 6. Coeficientes de correlación de Pearson entre el logaritmo en base 10 de las lecturas en cada réplica realizada por el laboratorio 1 con el protocolo TruSeq.

	Réplica 1	Réplica 2	Réplica 3	Réplica 4
Réplica 1	1.000	.502	.461	.491
Réplica 2	.502	1.000	.901	.921
Réplica 3	.461	.901	1.000	.868
Réplica 4	.491	.921	.868	1.000

Los gráficos de correlación entre estas réplicas muestran la falta de consistencia de las replicas del laboratorio 1 en el protocolo TruSeq (**Figura 6**).

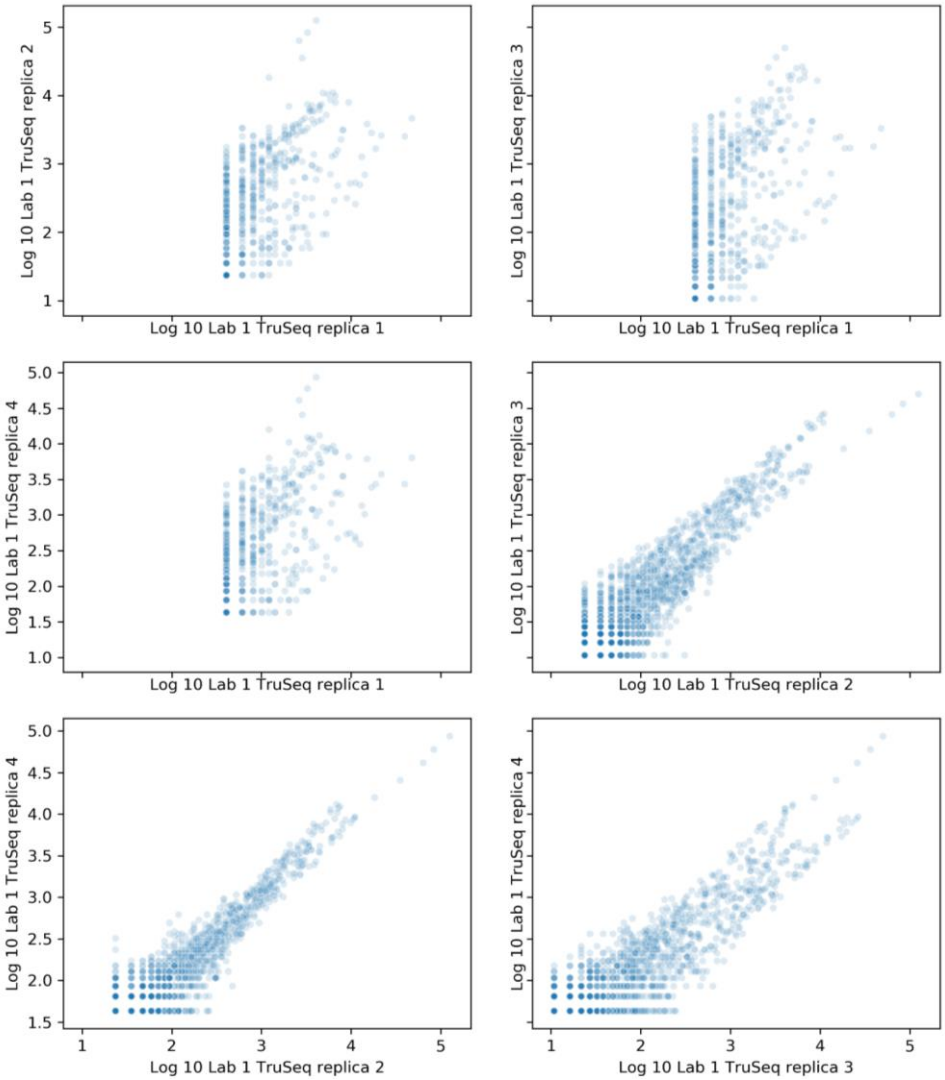

Figura 6. Gráficos de correlación de las réplicas normalizadas del laboratorio 1 utilizando el protocolo TruSeq.

Debida a esta falta de consistencia, las muestras del laboratorio 1 fueron excluidas del resto de análisis. Todos los gráficos de correlación se pueden encontrar como documento adjunto (**Anexo 3**)

A continuación, se promediaron los niveles de expresión de cada lectura para cada laboratorio y protocolo entre las cuatro réplicas siempre que haya sido detectada al menos en dos de las réplicas. La **Figura 7** muestra el número de lecturas contempladas así en cada réplica y el número final en la muestra promediada para cada laboratorio y protocolo.

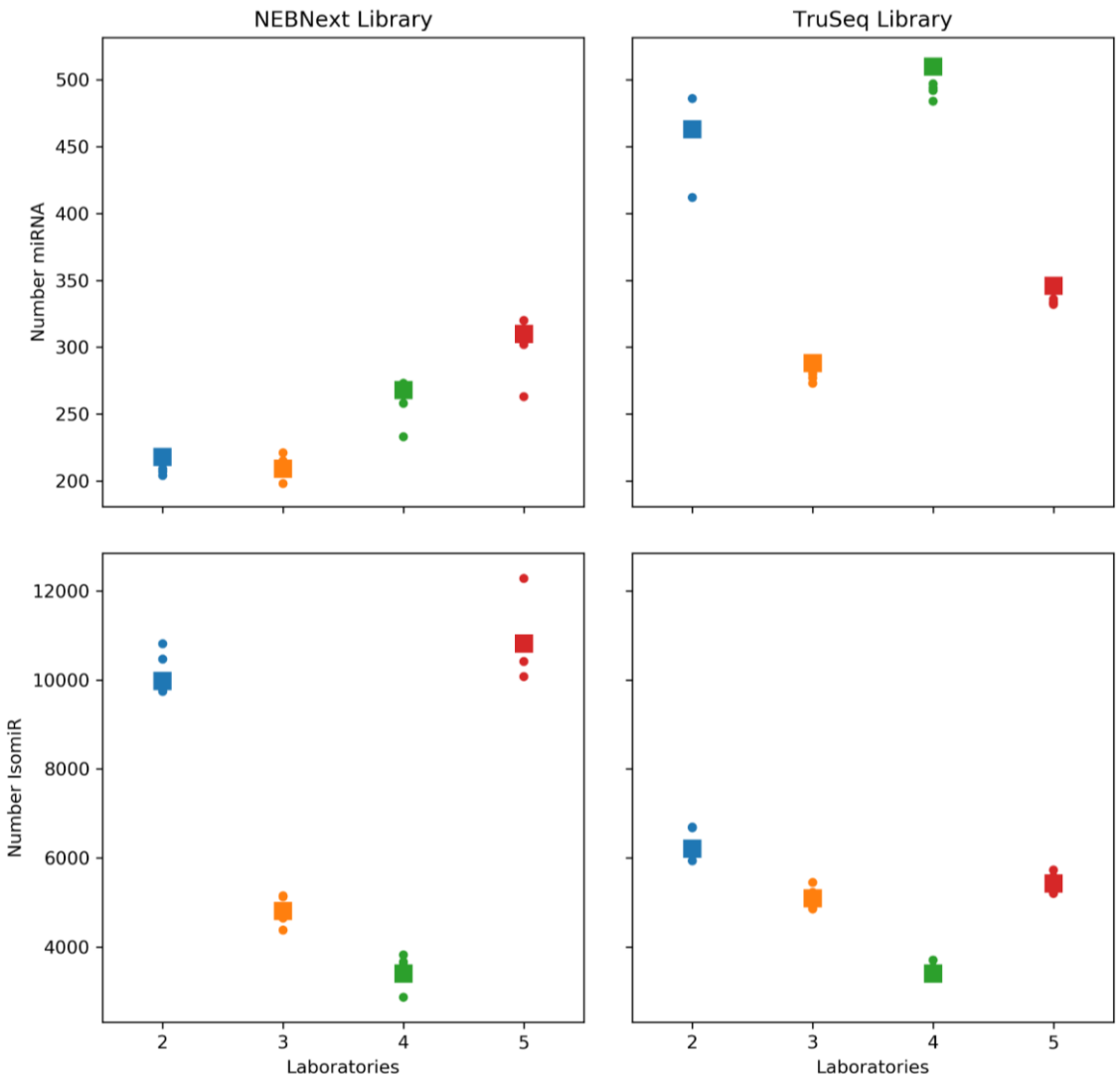

Figura 7. Numero de secuencias por tipo de *small* RNA (miRNA arriba, isomiR abajo) en cada una de las réplicas (círculos) y en la muestra promediada (cuadrado) en cada laboratorio y protocolo.

La **Figura 7** muestra que el número de lecturas que corresponden a los miRNA son mucho menores que los que corresponden a isomiRs, independientemente

del protocolo usado. Los laboratorios 2 y 4 detectaron más miRNA que el 3 y 5 mediante el protocolo TruSeq, mientras que los laboratorios 2 y 5 detectaron un gran número de isomiRs con el protocolo NEBNext en comparación con los otros dos laboratorios.

La **Figura 8** representa la suma de lecturas normalizadas en cada réplica y en la muestra promediada para cada laboratorio y protocolo.

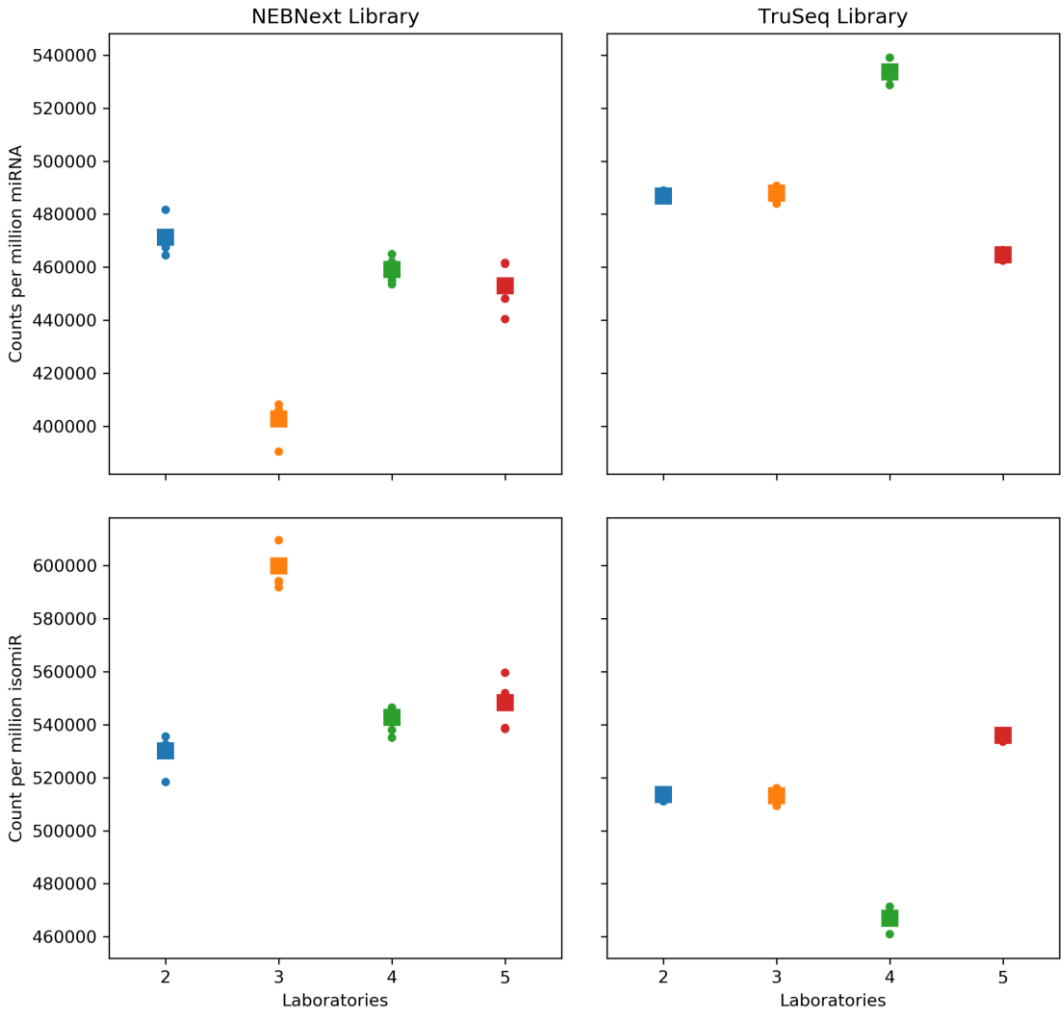

Figura 8. Suma de lecturas normalizadas por tipo de *small* RNA (miRNA arriba, isomiR abajo) en cada una de las réplicas (círculos) y en la muestra promediada (cuadrado) en cada laboratorio y protocolo.

El laboratorio 3 obtuvo una suma de lecturas normalizadas para miRNA más bajas y más altas para isomiRs en el protocolo NEBNext que el resto de laboratorios, mientras que el laboratorio 4 mostró un comportamiento inverso con el protocolo TruSeq (**Figura 8**).

Exploramos el número de secuencias y la suma de lecturas normalizadas obtenidas por cada laboratorio y protocolo en función de las modificaciones presentes en la secuencia de los isomiR. La **Figura 9** muestra estos datos agrupados por tipo general de isomiR. El documento **anexo 4** muestra estos datos desglosados para cada tipo individual de isomiR.

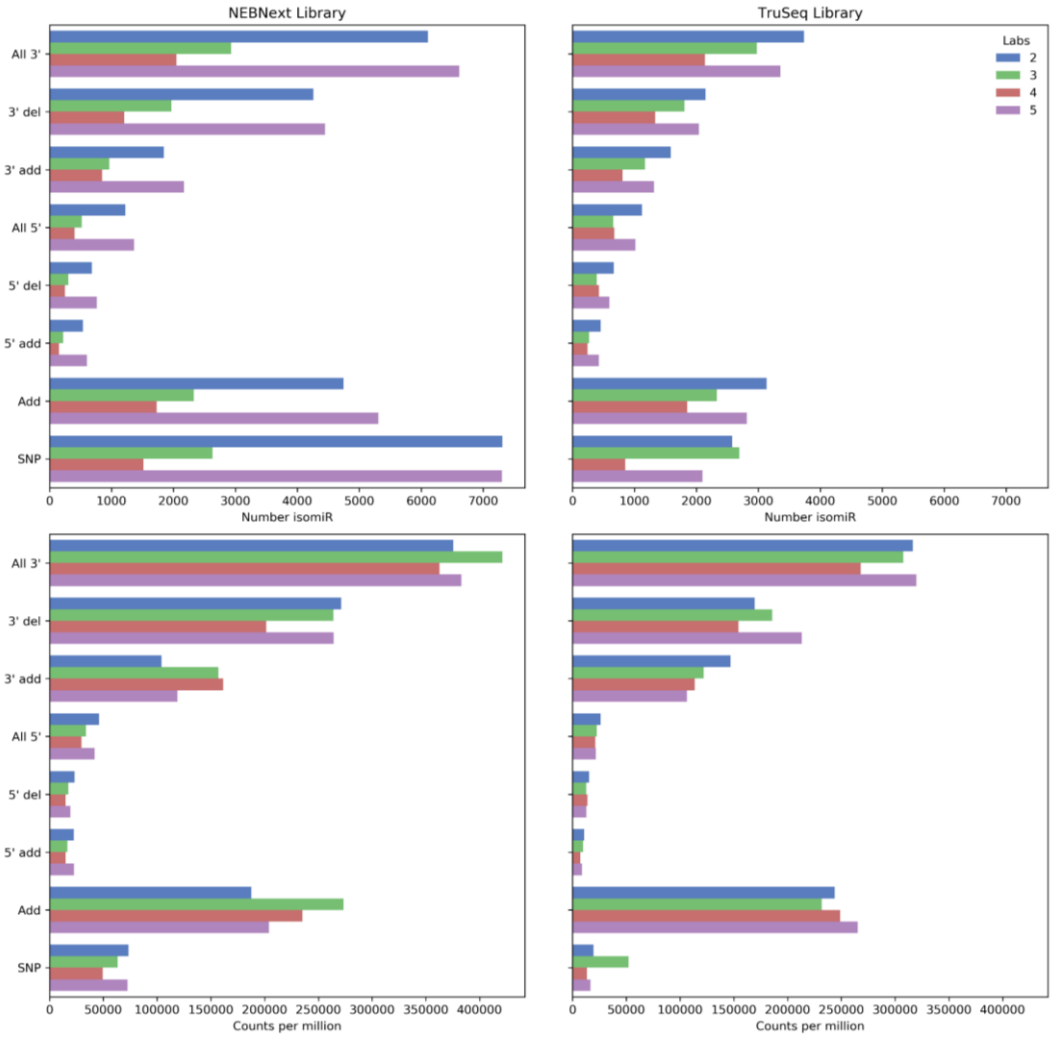

Figura 9. Numero de secuencias (arriba) y suma de lecturas (abajo) agrupadas por el tipo de modificación presente en los isomiR en cada uno de los laboratorios y protocolos. De arriba a abajo se muestran los valores en todos los tipos de modificaciones a 3' de la secuencia (All 3'), deleciones a 3' (3' del), adiciones a 3' (3' add), todas la modificaciones a 5' (All 5'), deleciones a 5' (5' del), adiciones a 5' (5' add), adiciones a 3' no consistentes con la secuencia del precursor (Add) y polimorfismos en la secuencia (SNP).

Observamos que el mayor número de isomiRs y de lecturas corresponden a las modificaciones en el extremo 3' del isomiR, ya sean consistentes con la secuencia del precursor (All 3') u otras adiciones (Add). Así mismo, el número de SNP es similar al de *Add* pero representan una menor cantidad de lecturas. Por otro lado, cabe destacar que los laboratorios 2 y 5 muestran un mayor número

de secuencias en todos los isomiRs que los laboratorios 3 y 4 en el protocolo TruSeq (**Figura 9**).

Cuando representamos el porcentaje sobre el total de secuencias y de lecturas de isomiRs por cada grupo general de modificación (**Figura 10**), las diferencias entre laboratorio son muy reducidas. Además, esta representación deja clara que las modificaciones tipo SNP son muy abundantes, pero representan una pequeña parte de las lecturas (**Figura 10**).

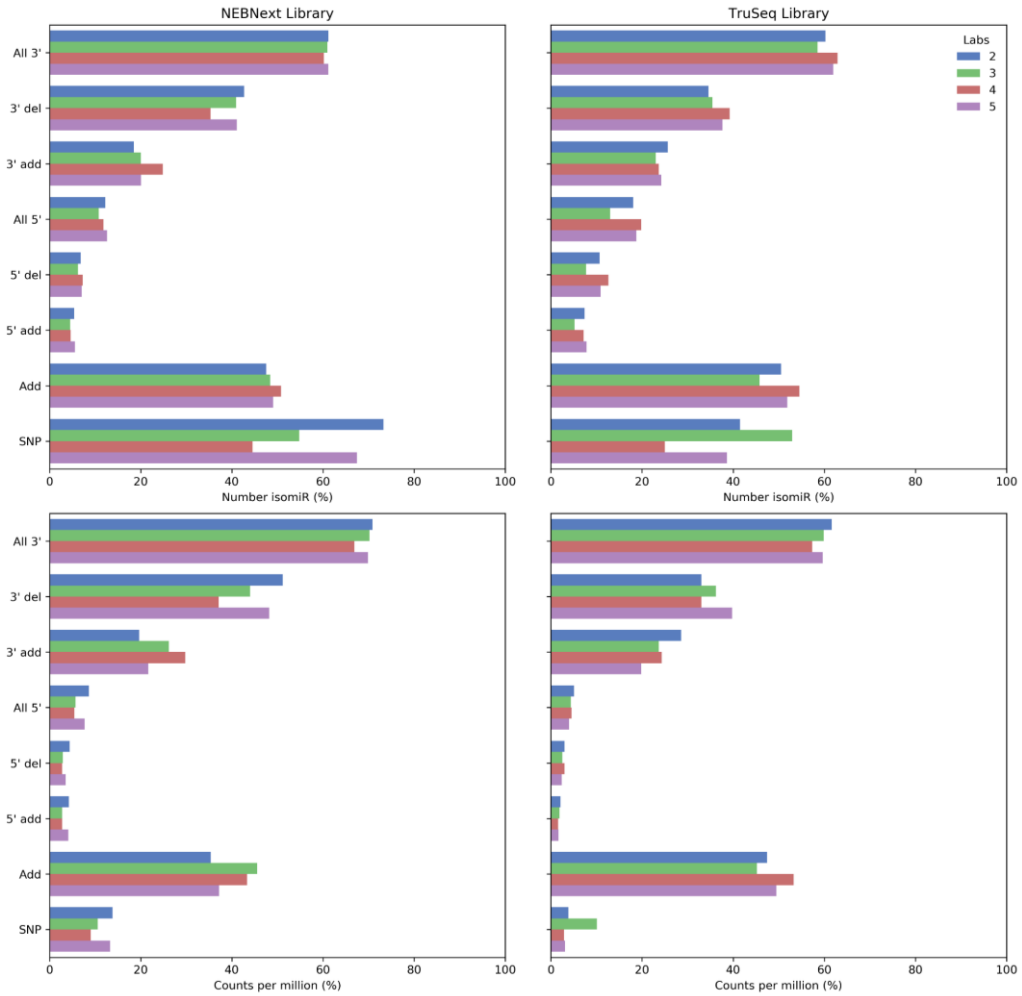

Figura 10. Porcentaje sobre el total del numero de secuencias (arriba) y suma de lecturas (abajo) agrupadas por el tipo de modificación presente en los isomiR en cada uno de los laboratorios y protocolos. De arriba a abajo se muestran los valores en todos los tipos de modificaciones a 3' de la secuencia (All 3'), deleciones a 3' (3' del), adiciones a 3' (3' add), todas la modificaciones a 5' (All 5'), deleciones a 5' (5' del), adiciones a 5' (5' add), adiciones a 3' no consistentes con la secuencia del precursor (Add) y polimorfismos en la secuencia (SNP).

En el documento **anexo 5** se pueden encontrar la representación del porcentaje sobre el total del número de secuencia y de la suma de lecturas desglosado para cada tipo individual de isomiR.

Exploramos la consistencia de las lecturas por secuencia entre las muestras tratadas por cada laboratorio mediante un análisis de correlación. La **Figura 11** muestra los gráficos de dispersión en estas muestras en función del protocolo y del tipo de *small* RNA.

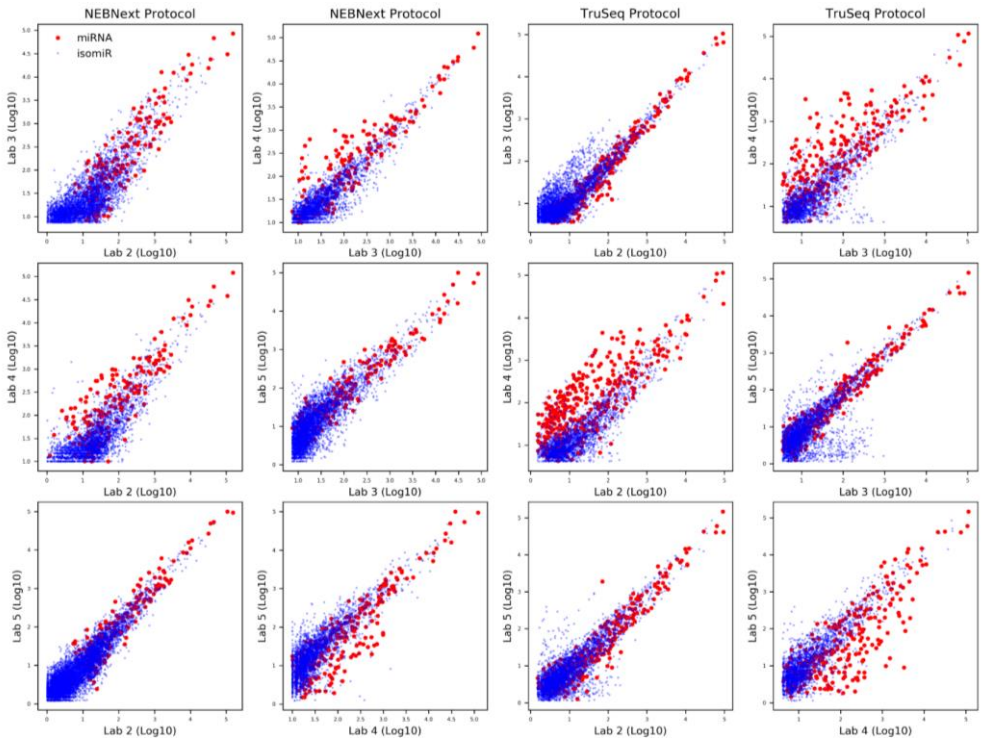

Figura 11. Gráficos de dispersión entre el logaritmo en base 10 de las lecturas obtenidas por cada laboratorio en función del protocolo de preparación de librerías y del tipo de *small* RNA (miRNA rojo, isomiR azul).

Observamos que la consistencia de los resultados entre laboratorios parece aceptable, como así parece mostrar los valores de los coeficientes de correlación (**Tabla 7**).

Tabla 7. Coeficientes de correlación de Pearson entre el logaritmo en base 10 de las lecturas provenientes de cada laboratorio en miRNA (casillas en rojo) e isomiR (casillas en azul) en cada uno de lo protocolos.

	Laboratorios	**2**	**3**	**4**	**5**
NEBNext	**2**		.900	.864	.947
	3	.802		.910	.958
	4	.789	.859		.902
	5	.908	.847	.843	
TruSeq	**2**		.973	.813	.953
	3	.867		.840	.969
	4	.832	.820		.827
	5	.858	.771	.859	

Sin embargo, al representar los gráficos de dispersión de las lecturas obtenidas por cada laboratorio entre ambos protocolos no observamos una buena consistencia (**Figura 12**).

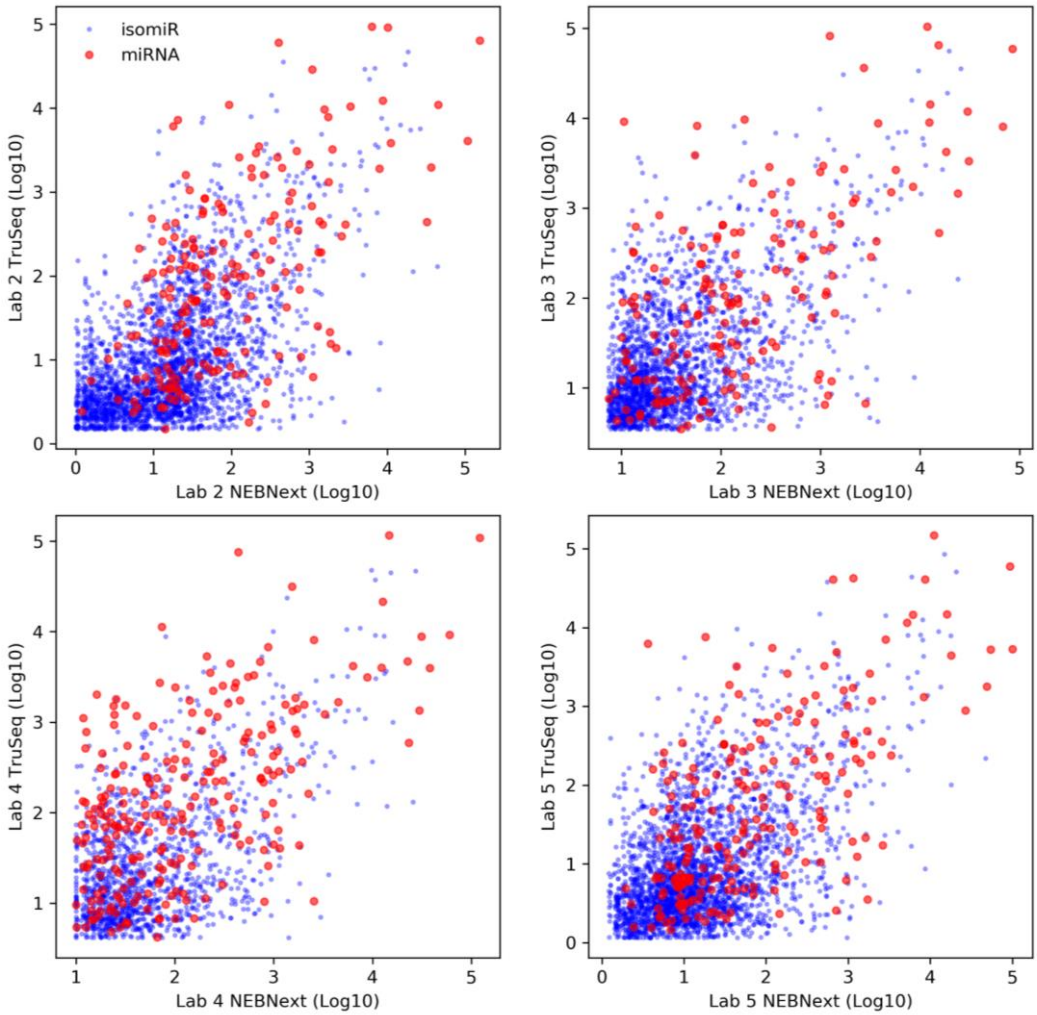

Figura 12. Gráficos de dispersión entre el logaritmo en base 10 de las lecturas obtenidas en ambos protocolos de preparación de librerías en función del laboratorio y del tipo de *small* RNA (miRNA rojo, isomiR azul).

Los coeficientes de correlación confirman esta observación (**Tabla 8**).

Tabla 8. Coeficientes de correlación de Pearson entre el logaritmo en base 10 de las lecturas en cada protocolo en función del laboratorio.

	Laboratorios			
	2	**3**	**4**	**5**
miRNA	.595	.617	.577	.638
isomiR	.544	.517	.533	.494

A continuación, exploramos las distribuciones del número de lecturas correspondientes a cada muestra (**Figura 13**).

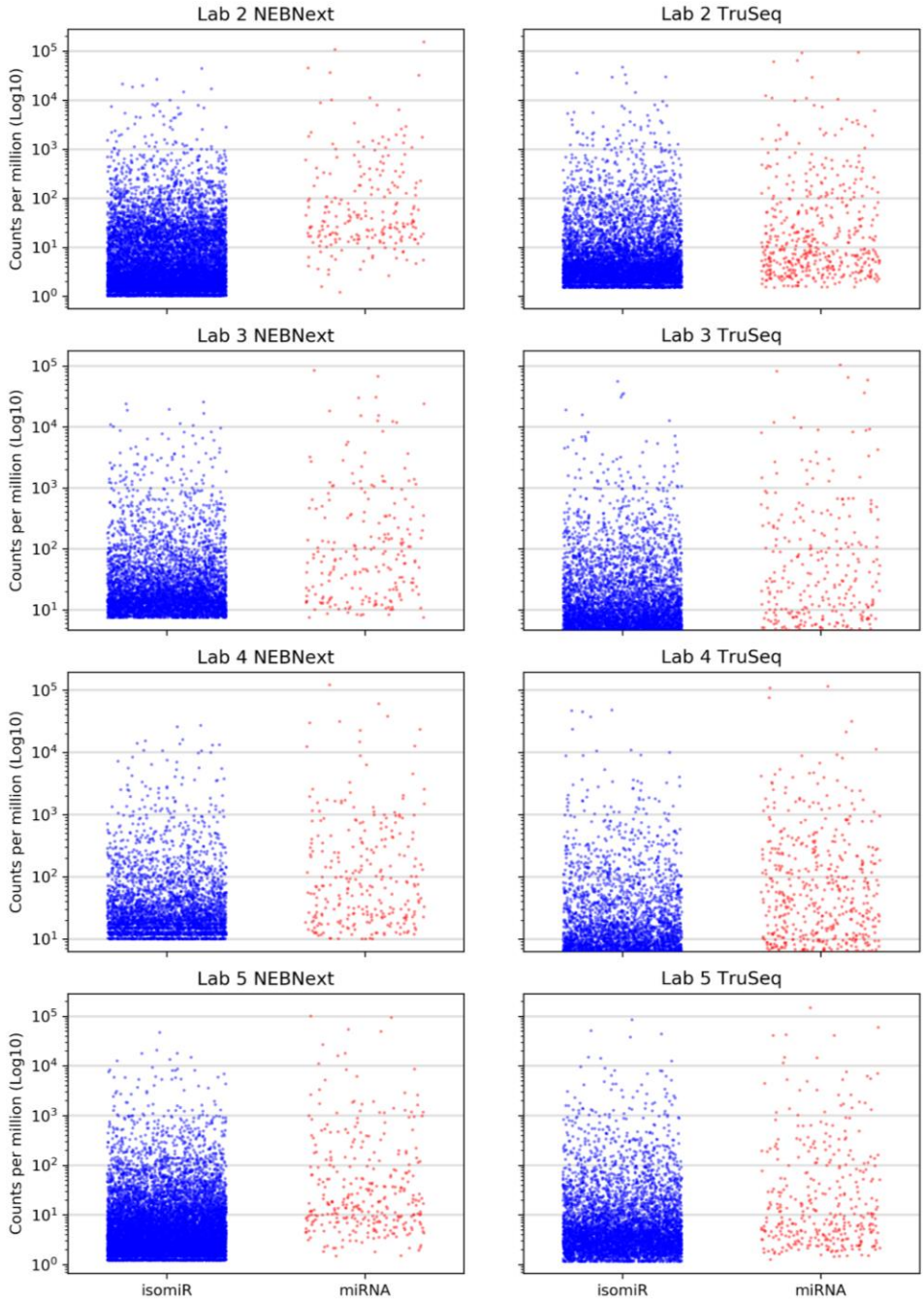

Figura 13. Distribución de los valores de expresión de isomiR y miRNA en cada uno de los laboratorios y protocolos. Nótese la escala logarítmica.

En la **Figura 13** se puede apreciar como una gran mayoría de las lecturas tienen unos niveles de expresión muy bajos.

Realizamos gráficos de tipo violín en las lecturas con menos de 200 veces detectada para observar la forma de la densidad de las distribuciones (**Figura 14**). Los gráficos de violín de las distribuciones completas y de secuencias con menos de 1000 lecturas se pueden observar en los documentos **anexos 6 y 7**.

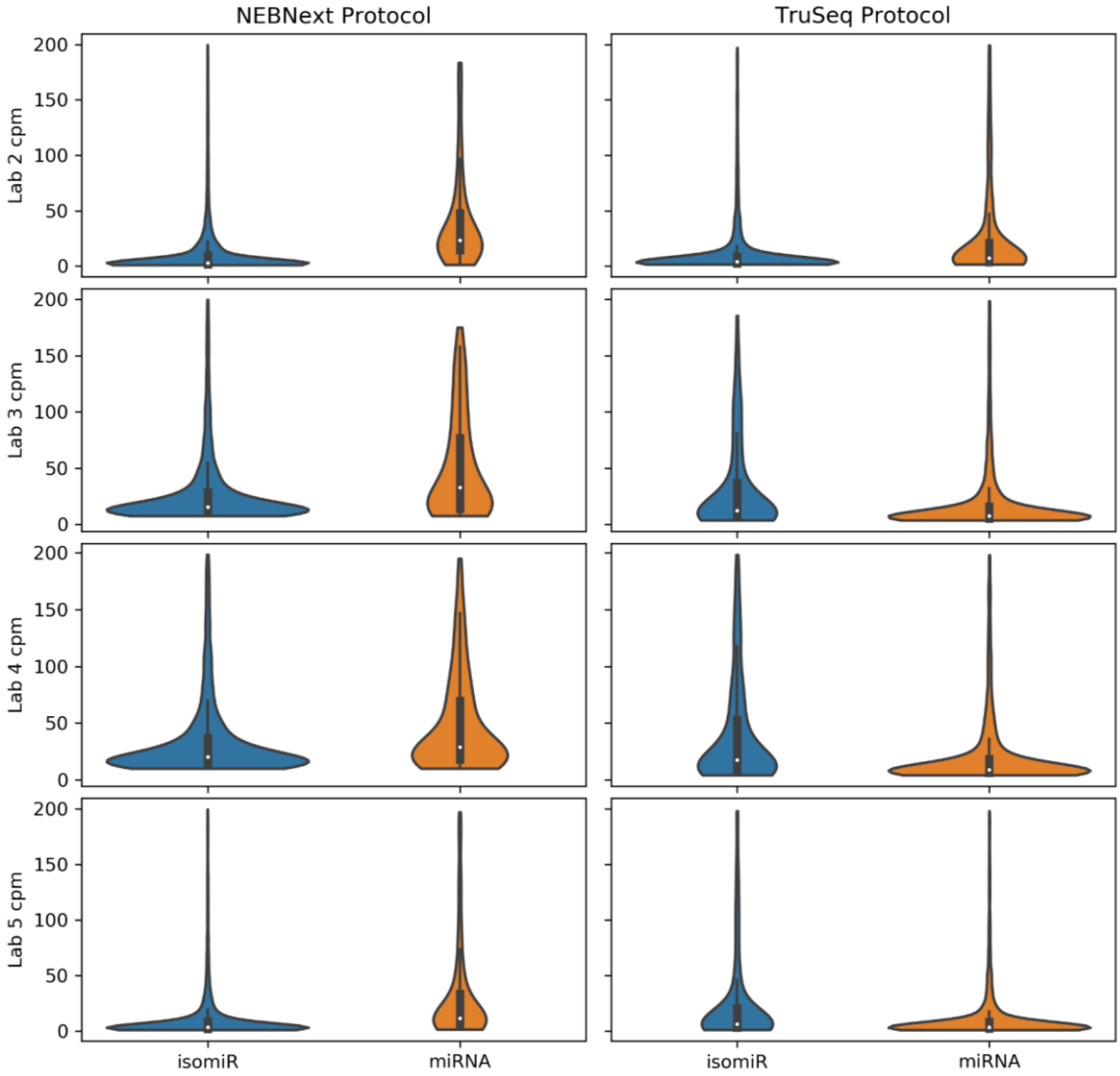

Figura 14. Gráficos de violín en cada uno de las muestras en función del tipo de *small* RNA en secuencias con menos de 200 lecturas.

Las distribuciones en todas las muestras muestran que la mayoría de las secuencias, y en especial las que corresponden a isomiR, tienen un número de lecturas muy bajo. Una parte de los isomiR detectados pueden deberse a errores de secuenciación y/o en la preparación de las librerías. Para comprobar esta hipótesis aplicamos una serie de filtros definidos por el número de laboratorios que han detectado una lectura y el número de lecturas mínimo para ser considerado. La **Figura 15** muestra el número de isomiRs y la suma de lecturas

contempladas al aplicar diferentes filtros. En esta figura podemos observar que al aplicar un criterio cada vez más restrictivo se reduce de manera importante el número de isomiRs contemplados, aunque la suma de lecturas no disminuye de forma tan acusada.

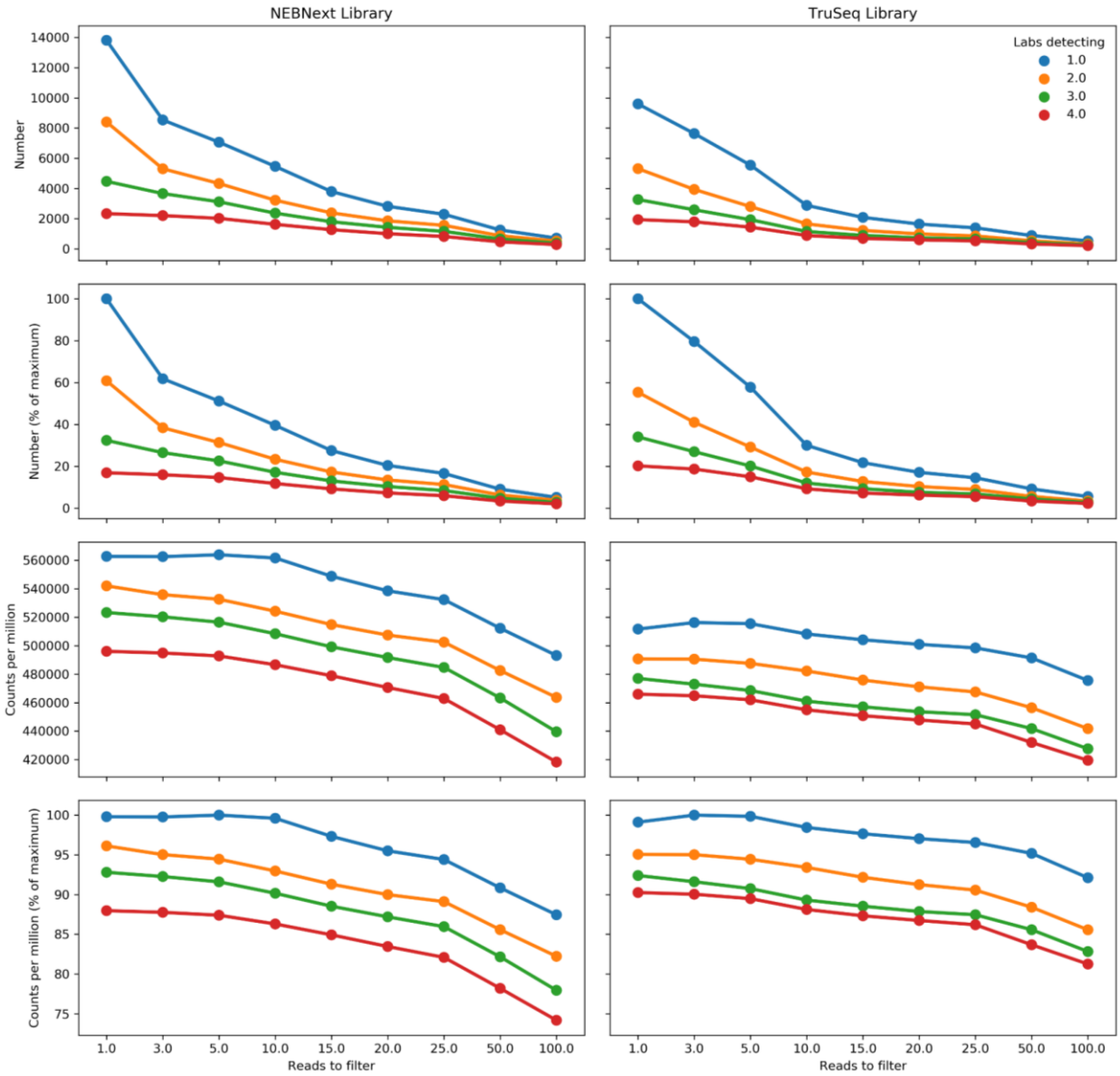

Figura 15. Numero de isomiRs (2 filas superiores) y suma de lecturas (2 filas inferiores) en función del número de laboratorios que detectan una secuencia y el número de lecturas mínimas para ser tomada en cuenta, en cada protocolo (NEBNext columna izquierda, TruSeq columna derecha).

Las figuras representando el número de isomiRs y suma de lecturas en función de los filtros para cada tipo de modificación se encuentra como documento adjunto (**Anexo 8**).

La **Figura 16** muestra las distribuciones de las lecturas de los diferentes isomiRs en función de los criterios establecidos de número mínimo de laboratorios detectando la secuencia y número mínimo de lecturas.

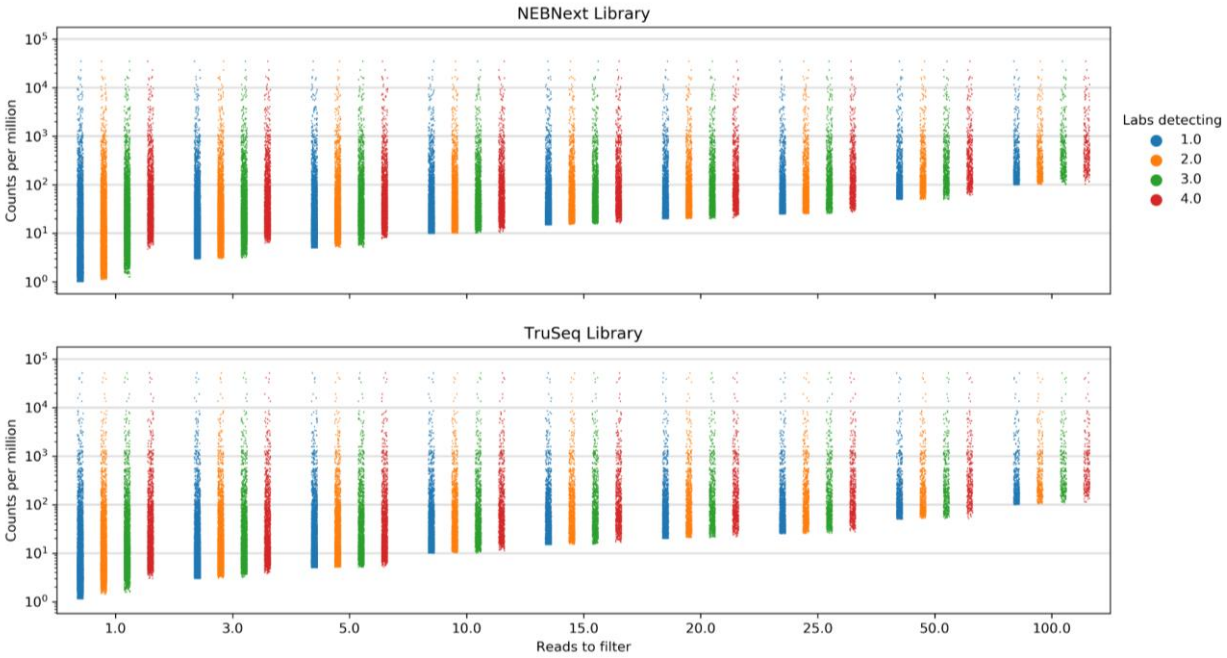

Figura 16. Distribuciones de lecturas de isomiRs en función del número de laboratorios detectando la secuencia y el mínimo de lecturas para ser tenido en cuenta, en cada uno de los protocolos.

Las distribuciones de lecturas en función de los tipos de isomiRs para cada filtrado se pueden consultar en el documento **anexo 9**.

Para analizar la reproducibilidad de los diferentes tipos de modificaciones en los isomiRs, establecimos 3 niveles de fiabilidad en función del número mínimo de laboratorios que detectan la secuencia: 2, 3 o 4 laboratorios. A continuación, exploramos el porcentaje del número de secuencias y el porcentaje de lecturas que cumplen cada criterio de fiabilidad sobre el total de isomiRs, en cada protocolo y en cada tipo de modificación.

La **Figura 17** muestra los valores agrupando los isomiR por tipos. Se puede observar como los isomiR de tipo SNP reducen el número y la suma de lecturas de forma más acusada al subir el nivel de confianza que el resto de tipos de isomiR, en ambos protocolos.

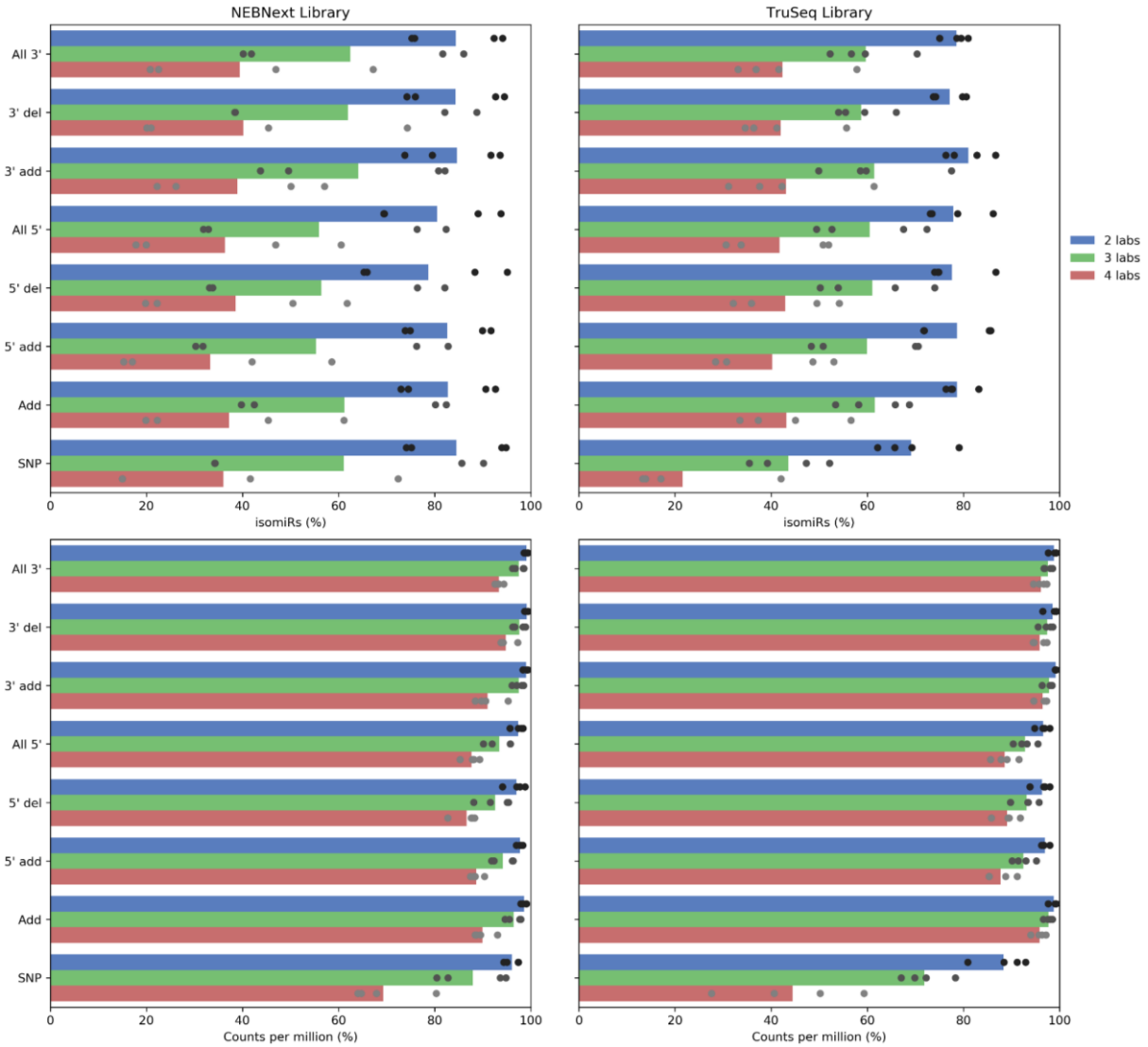

Figura 17. Porcentaje de secuencias y lecturas de isomiRs por tipo, cumpliendo 3 criterios de fiabilidad en función del número mínimo de laboratorios detectando la secuencia (azul 2 laboratorios, verde 3 laboratorios y rojo 4 laboratorios). De arriba a abajo se muestran los valores en todos los tipos de modificaciones a 3' de la secuencia (All 3'), deleciones a 3' (3' del), adiciones a 3' (3' add), todas la modificaciones a 5' (All 5'), deleciones a 5' (5' del), adiciones a 5' (5' add), adiciones a 3' no consistentes con la secuencia del precursor (Add) y polimorfismos en la secuencia (SNP). Se representan los valores correspondientes a cada laboratorio como puntos individuales y el valor promedio como la longitud de las barras.

La **Figura 18** muestra que los isomiR con adiciones de 3 o más bases a 3' de la secuencia, consistentes con la secuencia del miRNA precursor presentan una baja reproducibilidad en ambos protocolos. Con ninguno de los criterios se han aceptado deleciones de 4 bases a 3', por lo que estas modificaciones son de muy baja reproducibilidad.

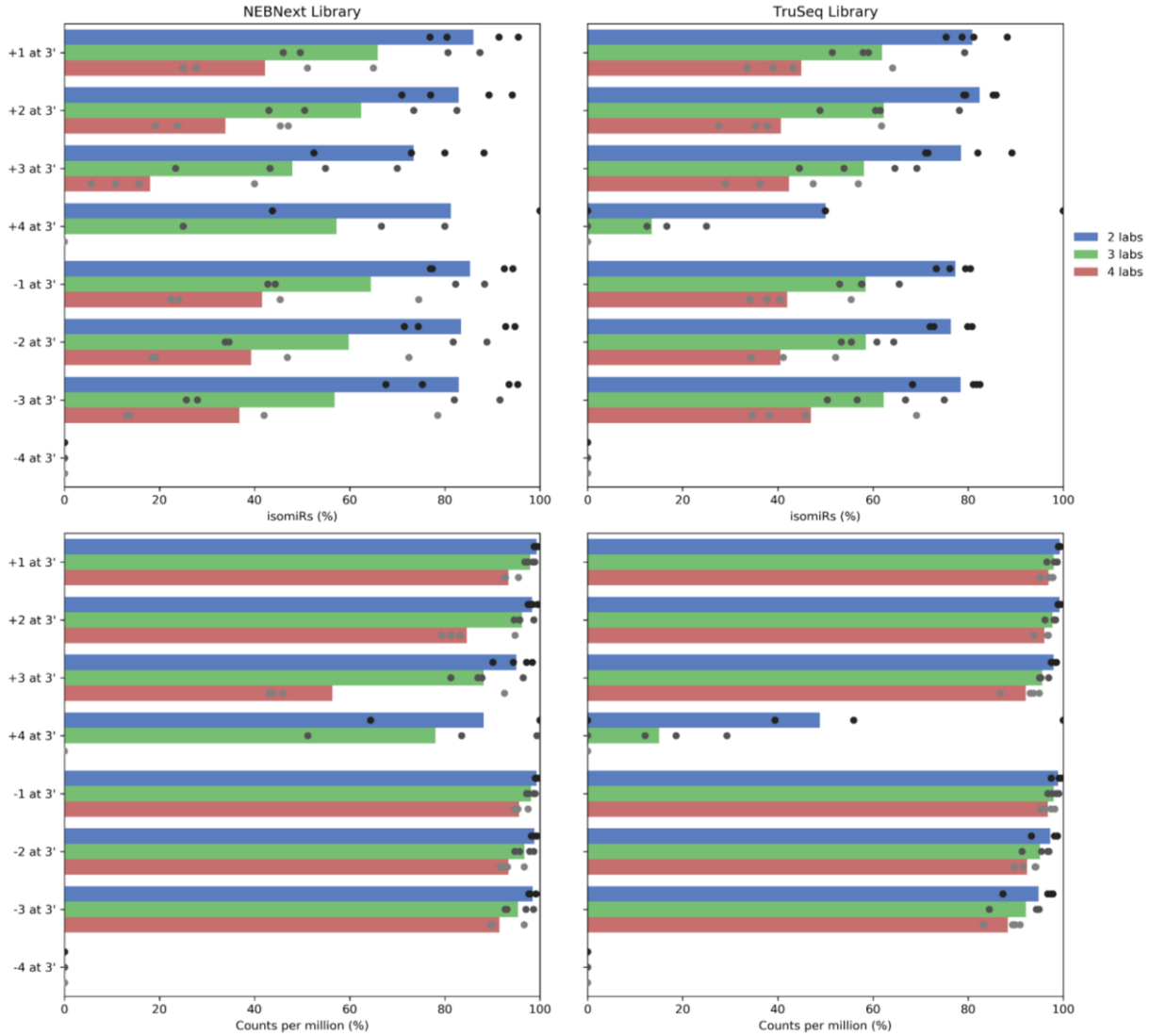

Figura 18. Porcentaje de secuencias y lecturas de isomiRs con modificaciones a 3' consistentes con la secuencia del miRNA precursor, cumpliendo 3 criterios de fiabilidad en función del número mínimo de laboratorios detectando la secuencia (azul 2 laboratorios, verde 3 laboratorios y rojo 4 laboratorios). De arriba a abajo se muestran las adiciones de 1 a 4 bases y las deleciones de 1 a 4 bases en el extremo 3' de la secuencia. Se representan los valores correspondientes a cada laboratorio como puntos individuales y el valor promedio como la longitud de las barras.

La **Figura 19** muestra que los isomiR con adiciones de 3 bases a 5’ de la secuencia presentan una baja reproducibilidad en ambos protocolos. Con ninguno de los criterios se han aceptado deleciones de 4 bases a 5’, por lo que estas modificaciones son de muy baja reproducibilidad.

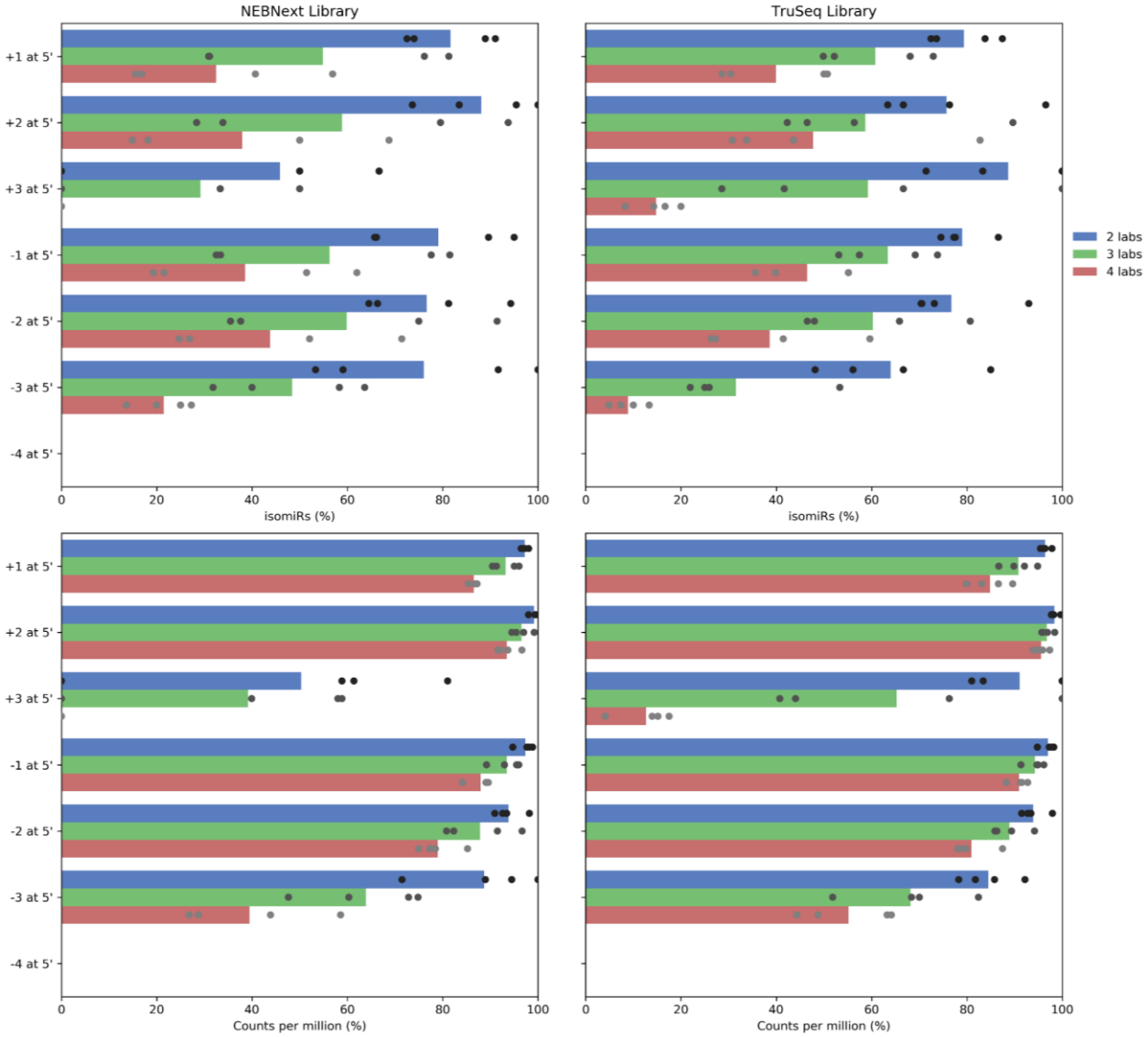

Figura 19. Porcentaje de secuencias y lecturas de isomiRs con modificaciones a 5 consistentes con la secuencia del miRNA precursor,’, cumpliendo 3 criterios de fiabilidad en función del número mínimo de laboratorios detectando la secuencia (azul 2 laboratorios, verde 3 laboratorios y rojo 4 laboratorios). De arriba a abajo se muestran las adiciones de 1 a 3 bases y las deleciones de 1 a 4 bases en el extremo 5’ de la secuencia. Se representan los valores correspondientes a cada laboratorio como puntos individuales y el valor promedio como la longitud de las barras.

La **Figura 20** muestra que las adiciones de 3 bases dentro de la secuencia muestran baja reproducibilidad en el protocolo NEBNext.

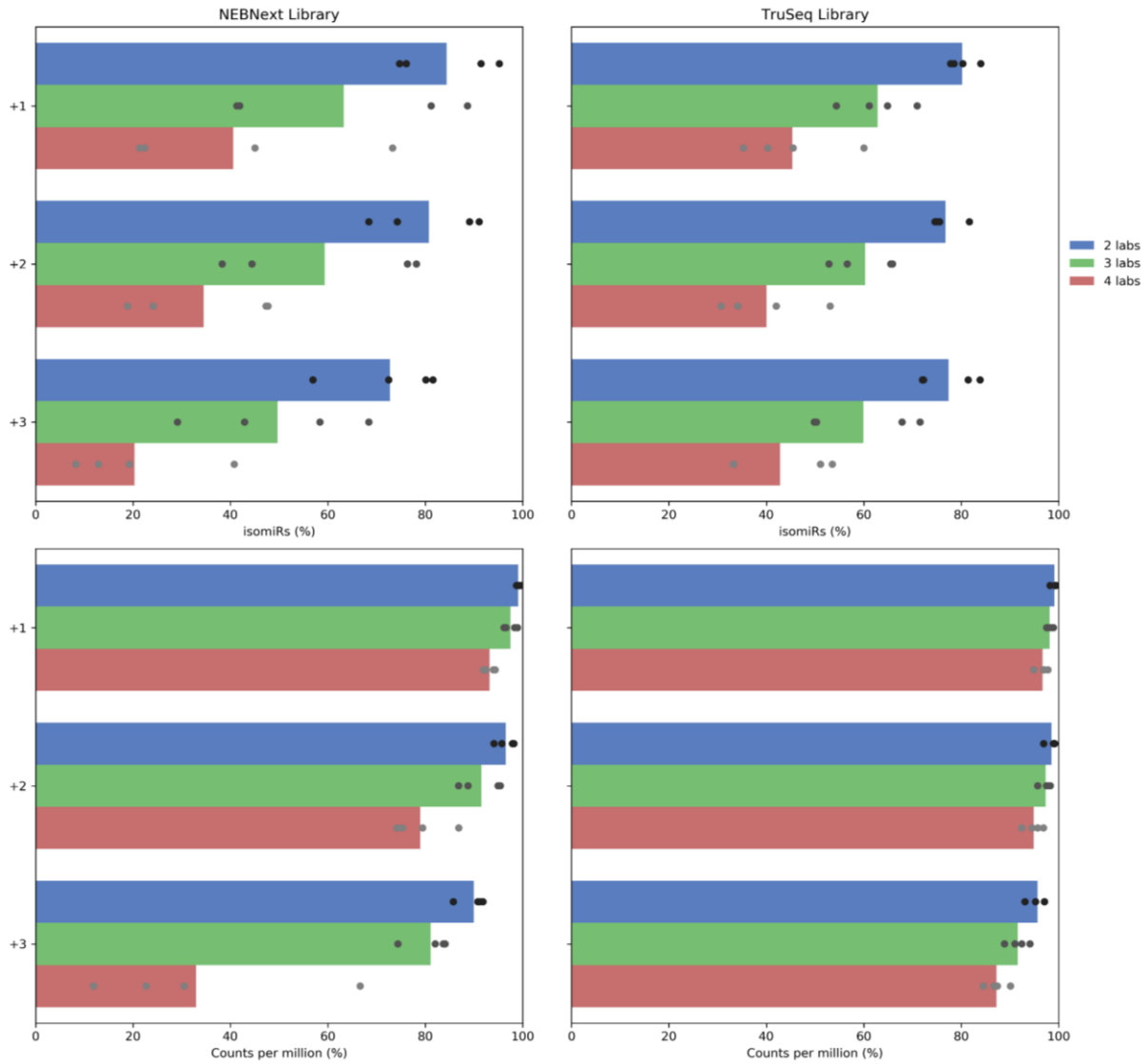

Figura 20. Porcentaje de secuencias y lecturas de isomiRs con adiciones a 3' de la secuencia no consistentes con la secuencia del miRNA precursor, cumpliendo 3 criterios de fiabilidad en función del número mínimo de laboratorios detectando la secuencia (azul 2 laboratorios, verde 3 laboratorios y rojo 4 laboratorios). De arriba a abajo se muestran las adiciones de 1 a 3 bases. Se representan los valores correspondientes a cada laboratorio como puntos individuales y el valor promedio como la longitud de las barras.

La **Figura 21** muestra que todos los polimorfismos en la secuencia son poco reproducibles en el protocolo TruSeq ya que la aplicación del criterio de 2 laboratorios reduce considerablemente el número de secuencias aceptadas, así como la suma de lecturas.

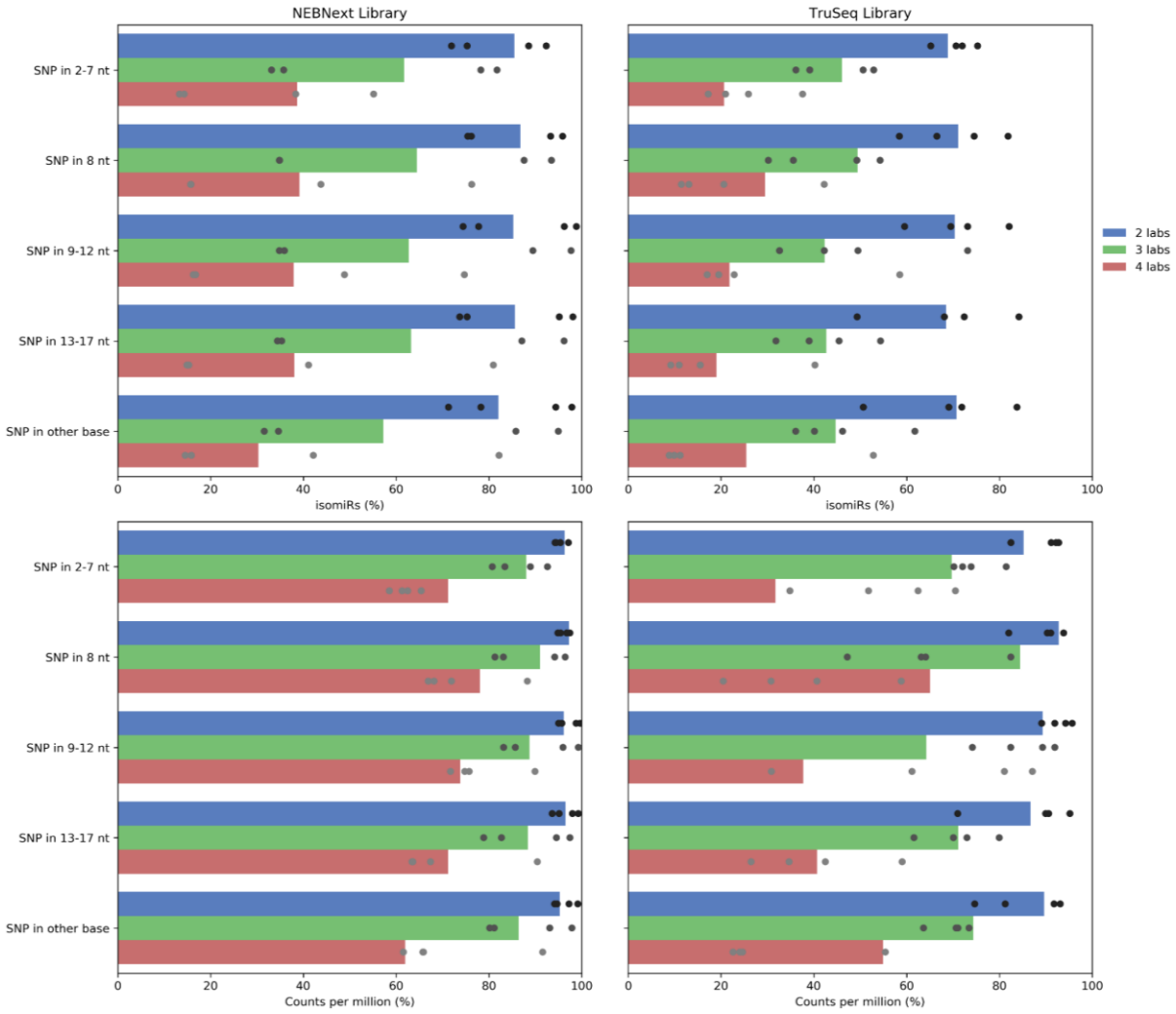

Figura 21. Porcentaje de secuencias y lecturas de isomiRs con polimorfismo en la secuencia, cumpliendo 3 criterios de fiabilidad en función del número mínimo de laboratorios detectando la secuencia (azul 2 laboratorios, verde 3 laboratorios y rojo 4 laboratorios). De arriba a abajo se muestran polimorfismos en diferentes zonas del *small* RNA (nucleótidos de 2 a 7, 8º, de 9 a 12, de 13 a 17 o en otro nucleótido). Se representan los valores correspondientes a cada laboratorio como puntos individuales y el valor promedio como la longitud de las barras.

3. Discusión de los resultados relativos al objetivo 3

Para dar respuesta al objetivo 3 (*utilizar las herramientas desarrolladas para analizar un set de datos experimentales*), analizamos un set de datos experimentales provenientes de muestras de plasma sanguíneo de humanos utilizado en un estudio con el fin de explorar la reproducibilidad de los resultados de *small* RNA-seq (Giraldez *et al*, 2017).

En este análisis pudimos utilizar las herramientas desarrolladas en los anteriores objetivos para registrar los niveles de expresión de miRNA e isomiRs en la muestra de plasma analizada por 5 laboratorios que usaron dos protocolos diferentes (NEBNext y TruSeq) para la obtención de las librerías de cDNA, por cuadruplicado, que fueron secuenciadas una vez cada una. Utilizamos la pipeline bcbio-nextgen en su configuración para small RNA-seq y el resultado final del proceso fueron 40 ficheros de salida de *miraligner* con un formato y extensión propios (*.counts*). La utilidad mirTop fue utilizada para transformar estos 40 ficheros en un único fichero con el formato de anotación propuesto en este trabajo que refleja los resultados de las 40 secuenciaciones. Las herramientas desarrolladas en el objetivo 2 fueron utilizadas para cargar este fichero en un *dataframe* de Pandas en un entorno Python, una vez comprobada la corrección del formato.

Una vez tuvimos los datos en un *dataframe* Pandas utilizamos la potencia del lenguaje Python y las librerías disponibles para el análisis de datos para explorar los datos de secuenciación de las muestras de plasma. Las lecturas en cada réplica se normalizaron, expresándolas como lecturas por millón de lecturas totales. A continuación, el número de lecturas por cada secuencia para cada laboratorio y protocolo se expresó como el promedio de todos los valores en cada réplica siempre y cuando la secuencia hubiera sido detectada por al menos 2 laboratorios.

Al explorar la correlación entre las réplicas en cada laboratorio y protocolo, observamos que los datos provenientes de laboratorio 1 posiblemente contenían algún tipo de error en la obtención de las librerías TruSeq y/o en la secuenciación. Optamos por eliminar los datos provenientes de este laboratorio

al no ser posible discernir la fuente de los errores. Las muestras de los restantes laboratorios mostraron buenas correlaciones entre las lecturas lo que indica que la variabilidad inducida por repetir el proceso de preparar las librerías y es baja.

A continuación, exploramos el número y suma de lecturas para cada tipo de *small* RNA. Observamos que el número de isomiR es mucho mayor que el número de miRNA de referencia (**Figura 7**), aunque ambos tipos de *small* RNA representan un parecido numero de lecturas (**Figura 8**). Esto indica que la mayoría de los isomiR presenta un número bajo de lecturas. Las formas de las distribuciones de las lecturas que se muestran en la **Figura 13** refuerzan esta observación.

Posteriormente exploramos el número de secuencias y las lecturas correspondientes a cada tipo general de isomiR en cada protocolo y laboratorio. Observamos que un gran número de isomiR presentan modificaciones a 3' consistentes con la secuencia del miRNA precursor (iso_3p), así como de adiciones en este extremo no consistentes con el precursor (iso_add) (**Figuras 8 y 9**). Además, observamos que un gran número de secuencias presentan modificaciones SNP pero que estas representan un bajo número de lecturas, especialmente en el protocolo TruSeq (**Figuras 8 y 9**). Esto denota que las secuencias con SNPs presentan un bajo número de lecturas, lo que podría indicar una parte de estos isomiR son producto de errores de secuenciación.

A continuación, exploramos la consistencia de los resultados al comparar las lecturas para cada secuencia obtenidas por cada laboratorio discriminando por protocolo (**Figura 11**). Observamos que en tanto para miRNAs como para isomiRs existe una aceptable concordancia entre las muestras, tanto en el protocolo NEBNext como en el protocolo TruSeq (**Tabla 7**). Sin embargo, al comparar los datos obtenidos por cada laboratorio en función del tipo de *small* RNA (**Figura 12**) observamos que la correlación no es aceptable (**Tabla 8**). Estos resultados indican que el protocolo de preparación de librerías de cDNA (NEBNext *vs* TruSeq) induce una mayor variabilidad en los resultados que la secuenciación. En línea con esta observación, la forma de la distribución de lecturas en miRNA e isomiRs es ligeramente diferente entre protocolos (**Figura 14**).

En este punto del análisis nos propusimos evaluar en que medida los errores acumulados en los procesos de preparación de librerías y de secuenciación podría ser responsable de la presencia de el alto número de modificaciones con tan bajas lecturas. Para este fin, establecimos diferentes niveles de confianza de las secuencias detectadas en función del número de laboratorios que han detectado la secuencia y del número de lecturas mínimo para que la observación sea considerada. Observamos que la aplicación de un grado mayor de confianza reduce considerablemente el número de isomiRs aceptados, aunque las lecturas no se reducen en la misma proporción (**Figura 15**). Este resultado está en concordancia con las observaciones anteriores sobre la característica de las distribuciones de isomiRs. En la misma línea, las distribuciones de las lecturas de los isomiR presentan un comportamiento similar (**Figura 16**).

Al filtrar las secuencias con un bajo número de lecturas es posible que eliminemos errores tanto de retrotranscripción como de secuenciación. Sin embargo, podemos estar eliminando secuencias genuinas muy poco expresadas. Al aplicar el criterio de número de laboratorios que detectan la secuencia, independientemente de las lecturas de esta, conservamos las secuencias que han sido detectadas por un número creciente de laboratorios lo que otorga una confianza creciente a la secuencia. No obstante, al combinar este criterio con la aceptación de secuencias bajo numero de lecturas aumenta la posibilidad de que la secuencia corresponda a un error de retrotranscripción o de secuenciación. El diseño experimental y los resultados de este trabajo no nos permiten establecer una adecuada combinación de ambos criterios para establecer el nivel de confianza de la secuencia. Por tanto, para los siguientes análisis, optamos por aplicar únicamente el criterio de número de laboratorios detectando la secuencia, independientemente de las lecturas de esta.

Para finalizar, exploramos cual es el efecto de aplicar el criterio de confianza en las secuencias en el numero de secuencias y en el total de lecturas aceptadas para cada tipo en los isomiR. En primer lugar, exploramos el efecto del criterio de confianza sobre el porcentaje secuencias y suma de lecturas aceptadas en función del tipo general de modificación (**Figura 17**). Observamos que a mayor grado de confianza en la secuencia el numero de secuencias en todos los tipos

se reduce considerablemente pero el total de lecturas no se reduce en la misma proporción, excepto en el caso de los isomiR con SNP que experimentan una reducción más acusada que el resto (**Figura 17**). Esta afirmación es cierta en ambos protocolos, aunque la reducción es aún más acusada en el protocolo TruSeq. Esta observación está en línea con resultados anteriores y abunda en la idea de que una proporción de los isomiR SNP son poco reproducibles y posiblemente fruto de errores de secuenciación o de preparación de las librerías. Esto podría ser importante ya que un considerable número de isomiRs presentan estas modificaciones (**Figura 10**).

Tanto las modificaciones a 3' como a 5' consistentes con la secuencia del precursor mostraron una baja reproducibilidad en adiciones o deleciones de 3 o más bases (**Figura 18** y **Figura 19**), aunque estas modificaciones están presentes en una muy baja proporción de los isomiR (**anexo 3**). Las adiciones de 3 bases a 3' de la secuencia no consistentes con la secuencia del precursor mostraron baja reproducibilidad al utilizar el protocolo NEBNext, no así con el protocolo TruSeq (**Figura 20**). Esto puede indicar un sesgo inducido por el proceso de retrotranscripción utilizado en cada protocolo. Todos los isomiR con polimorfismos en sus secuencias mostraron una menor reproducibilidad cuando se usó el protocolo TruSeq (**Figura 21**).

En relación al objetivo 3, podemos concluir que la combinación del formato de fichero GFF3 desarrollado para la anotación de resultados de *small* RNA-seq y las funciones desarrolladas en Python anteriormente son útiles para el análisis de datos experimentales de *small* RNA-seq. Además, podemos afirmar que permiten explorar los datos discriminando fácilmente por el tipo de isomiR. Por ejemplo, en un experimento real con réplicas por condición sería muy fácil usar solo los isomiRs con un tipo concreto de modificación que aparecen en un número determinado de las condiciones. Esto ayudaría a eliminar errores experimentales.

V. Conclusiones

Durante el presente trabajo hemos ayudado a implementar un formato de fichero para la anotación de datos de *small* RNA-seq, hemos desarrollado funciones para su importar y comprobar estos ficheros a un entorno Python y hemos utilizado estas herramientas y para explorar un set de datos de experimentales aprovechando la versatilidad del formato. Este trabajo ha sido estructurado a través de los objetivos, procedimientos y resultados presentados en anteriores capítulos.

Durante el transcurso de este trabajo hemos tenido la primera toma de contacto con un proyecto puramente bioinformático. Este proyecto nos ha permitido alcanzar un alto grado familiarización con las herramientas bioinformáticas que hemos necesitado para su desarrollo. Estas herramientas son: el entorno Linux, programación y comandos en BASH, instalación y configuración de al pipeline *bcbio-nextgen*, programa mirTop y el entorno de programación Python.

Hemos cumplido los objetivos que nos habíamos propuesto para este proyecto. A pesar de haber encontrado alguna dificultad con la instalación de la *pipeline* y el tratamiento bioinformático de los ficheros FASTQ, hemos podido solventar estas dificultades y obtener los ficheros de anotación de la expresión de *small* RNA para los subsecuentes pasos.

Este proyecto está enmarcado en el contexto de un trabajo final de master y con una limitación temporal. Sin embargo, a partir de este punto hay múltiples opciones para futuros proyectos. Por ejemplo, un posible proyecto para el futuro podría ser el desarrollo de una librería en Python para ampliar el repertorio de funciones y herramientas para el tratamiento de estos ficheros y de los datos que contienen. Otro posible proyecto posible podría ser implementar estas funciones en un entorno web para hacer más fácil el análisis de datos a usuarios no expertos en entornos como Python o R.

VI. Glosario

cDNA: DNA complementario

cpm: counts per million

GFF3: generic feature format version 3

NGS: next generation sequencing

SNP: polimorfismos de una base

VII. Bibliografía

Giraldez, M. D., Spengler, R. M., Etheridge, A., et al. (2017). Accuracy, Reproducibility And Bias Of Next Generation Sequencing For Quantitative Small RNA Profiling: A Multiple Protocol Study Across Multiple Laboratories. *bioRxiv*.

Gross, N., & Kropp, J. (2017). MicroRNA Signaling in Embryo Development. *6*(3).

Guo, L., & Chen, F. (2014). A challenge for miRNA: multiple isomiRs in miRNAomics. *Gene, 544*(1), 1-7.

Lee, L. W., Zhang, S., Etheridge, A., et al. (2010). Complexity of the microRNA repertoire revealed by next-generation sequencing. *Rna, 16*(11), 2170-2180.

Liu, H., Lei, C., He, Q., et al. (2018a). Nuclear functions of mammalian MicroRNAs in gene regulation, immunity and cancer. *17*(1), 64.

Liu, H., Yu, H., Tang, G., et al. (2018b). Small but powerful: function of microRNAs in plant development. *37*(3), 515-528.

Mercey, O., Popa, A., Cavard, A., et al. (2017). Characterizing isomiR variants within the microRNA-34/449 family. *FEBS Lett, 591*(5), 693-705.

Morin, R. D., O'Connor, M. D., Griffith, M., et al. (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res, 18*(4), 610-621.

Neilsen, C. T., Goodall, G. J., & Bracken, C. P. (2012). IsomiRs--the overlooked repertoire in the dynamic microRNAome. *Trends Genet, 28*(11), 544-549.

Williams, A. E. (2008). Functional aspects of animal microRNAs. *Cell Mol Life Sci, 65*(4), 545-562.

Wong, C. M., Tsang, F. H., & Ng, I. O. (2018). Non-coding RNAs in hepatocellular carcinoma: molecular functions and pathological implications. *Nat Rev Gastroenterol Hepatol, 15*(3), 137-151.

Yu, F., Pillman, K. A., Neilsen, C. T., et al. (2017). Naturally existing isoforms of miR-222 have distinct functions. *Nucleic Acids Res, 45*(19), 11371-11385.

Zhang, X., Ma, X., Jing, S., et al. (2018). Non-coding RNAs and retroviruses. *15*(1), 20.

VIII. Anexos

Anexo 1.py	Funciones en load_gff3 y load_check_gff3 en Python.
Anexo 2.zip	Fichero en formato GFF3 con los datos de expresión de las 40 muestras utilizadas en este estudio.
Anexo 3.pdf	Gráficos de correlación de las réplicas de cada laboratorio y protocolo.
Anexo 4.pdf	Numero de secuencias y suma de lecturas agrupadas por el tipo de modificación presente en los isomiR en cada uno de los laboratorios y protocolos.
Anexo 5.pdf	Proporción de secuencias y de lecturas agrupadas por el tipo de modificación presente en los isomiR en cada uno de los laboratorios y protocolos.
Anexo 6.pdf	Gráficos de violín de las distribuciones de los niveles de expresión en función cada laboratorio y protocolo.
Anexo 7.pdf	Gráficos de violín de las distribuciones de los niveles de expresión en función cada laboratorio y protocolo en secuencias con menos de 1000 lecturas.
Anexo 8.pdf	Numero de isomiR y suma de lecturas en función del filtro para cada tipo de modificación.
Anexo 9.pdf	Distribuciones de lecturas de isomiR en función del filtro para cada tipo de modificación.