Uso de datos genómicos para la identificación de miRNAs predictores de supervivencia en adenocarcinoma

Fernando Naya Català
Máster en bioinformàtica i bioestadística
Área del trabajo final

Jeroni Luna Cornadó
David Merino Arranz

02/01/2017


Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Uso de datos genómicos para la identificación de miRNAs predictores de supervivencia en adenocarcinoma*
Nombre del autor:	*Fernando Naya Català*
Nombre del consultor/a:	*Jeroni Luna Cornadó*
Nombre del PRA:	*David Merino Arranz*
Fecha de entrega (mm/aaaa):	01/2017
Titulación::	*Màster en bioinformática i bioestadística*
Área del Trabajo Final:	*TFM-Estadística i bioinformàtica 32 Aula 1*
Idioma del trabajo:	*Castellano*
Palabras clave	*Cáncer, microRNAs, supervivencia*

Resumen del Trabajo (máximo 250 palabras):

Objetivo

En este trabajo se muestra el estudio comparativo entre muestras tumorales adenocarcinosas de pacientes de alta y baja supervivencia con el objetivo de detectar moléculas de microRNA que estén diferencialmente expresadas y que sean predictoras de supervivencia.

Métodos

Usando información obtenida del repositorio The Genome Cancer Atlas, se extraen resultados de conteos de miRNAs y se procesan mediante software estadístico R para obtener expresión diferencial. Se realiza un análisis de supervivencia y se estudian los genes diana del panel de genes resultantes. Se realiza un análisis de enriquecimiento GO y KEGG y se buscan posibles interacciones microRNA-RNA mensajero.

Resultados

Se consigue una firma de expresión predictora de supervivencia consistente en 23 miRNA, repartidos entre todos los cánceres y 95 interacciones miRNA-mRNA para 15 miRNAs de los anteriores.

Conclusiones

Se muestra un método fiable y robusto de extracción de expresión diferencial de miRNAs a partir de datos genómicos. La determinación de los niveles de expresión de dichos miRNA podría constituir una herramienta para el diagnóstico y tratamiento personalizado de los pacientes.

Abstract (in English, 250 words or less):

Purpose:

This study aims to develop microRNA expression signature for adenocarcinome tumoral patients divided in two groups based on their survival values.

Methods:

Using TCGA (The Genome Cancer Atlas) genomic information, miRNA counts are extracted and processed with statistical software R in order to obtain differential expression information. Target study, Go and KEGG enrichment and interaction studies are done in order to extract knowledge of the comparison.

Results:

23 miRNA expression signature is found, as well as 95 possible interactions miRNA-mRNA are detected.

Conclusions: A reliable differential expression and survival analysis using genomic information is shown. The calculation of miRNA expression levels can consstitute a powerful tool for the diagnosis and personalized treatment of the cancer patients.

Índice

Lista de figuras

Ordenadas según el orden de aparición en la memoria:

1. CONTEXTUALIZACIÓN DEL TRABAJO

1.1 Contexto y justificación del trabajo

Los microRNAs (miRNAs) son moléculas de RNA no codificante pequeñas (20 nucleótidos) y altamente conservadas implicadas en la regulación de la expresión génica. “Son transcritas por RNA-polimerasas II y III, generando precursores que llevan a cabo eventos de escisión para crear miRNAs maduros, cuyo mecanismo de acción consiste en el reconocimiento de sus genes diana creando un silenciamiento de genes basado en la represión de la traducción o en la rotura del RNA mensajero (mRNA) maduro” (MacFarlane and Murphy, 2010).

Este estudio tiene como objetivo principal el desarrollo de un patrón de expresión diferencial de miRNAs en pacientes diagnosticados con distintos tipos de adenocarcinoma, cuyos datos clínicos se extraen del repositorio The Cancer Genome Atlas (TCGA) (Tomczak et al. 2015). Los niveles de expresión de miRNA entre muestras tumorales serán testados mediante un test t de significancia ($p<0.05$). Además, se usará el método estimador de Kaplan-Meier para discriminar aquellos miRNA que afecten a la supervivencia.

Una vez extraídos los miRNA que poseen un impacto potencial sobre la supervivencia, se procederá a su anotación mediante ontologías como Gene Ontology (GO; http://www.geneontology.org) y KEGG (http://www.genome.jp/kegg/), con el objetivo de explorar funciones biológicas y rutas metabólicas enriquecidas en las muestras para cada tipo de adenocarcinoma. Para ello, se analizarán los genes diana potenciales mediante bases de datos especializadas en validación de dianas y basadas en análisis funcionales.

Por último, se tendrá en cuenta la posibilidad de combinar datos de expresión de miRNAs y mRNAs, de nuevo extraídos del repositorio TCGA, con el fin de

remarcar posibles interacciones entre estos dos tipos de moléculas capaces de ocurrir en un contexto patológico determinado.

Los miRNAs constituyen una nueva fuente de estudio en la regulación de genes. Se ha demostrado que la desregulación de ciertos miRNAs contribuye al desarrollo y progresión de determinados tumores (Srinivasan et. Al, 2011).

Por otra parte, “los adenocarcinomas (COAD (Colon Adenocarcinoma), READ (Rectum Adenocarcinoma), LUAD (Lung Adenocarcinoma), PAAD (Pancreatic Adenocarcinoma), PRAD (Prostate Adenocarcinoma) i STAD (Stomach Adenocarcinoma)) son un subtipo de cáncer que se espera que aumente en 2017, acrecentados por los índices elevados de obesidad, que aumentan la probabilidad de estos casos en un 13%” (Cancer Facts & Figures, 2017). Además, “estos tipos de cáncer comparten características biológicas comunes, como mutaciones conductoras que puedan tener un papel clave en el desarrollo del propio càncer” (González-Pérez et al. 2013). En la figura 1 se muestra un diagrama de Venn estableciendo las relaciones en el conjunto de genes mutados compartidos por cada adenocarcinoma estudiado. Es por ello que en este estudio se analizan distintos tipos de adenocarcinoma con el objetivo de ver si pueden compartir también otro tipo de moléculas funcionales en un contexto tumoral.

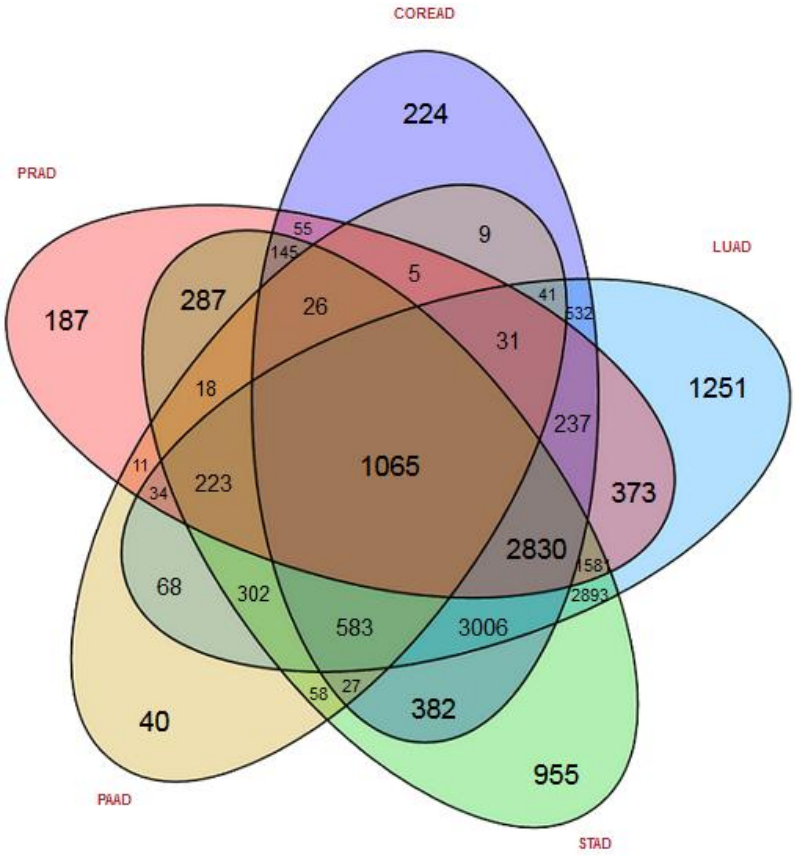

Figura 1: Diagrama de Venn mostrando la correlación existente entre los distintos tipos de adenocarcinoma en función de los genes anotados con mutaciones driver (Intogen).

En este contexto, la identificación de miRNAs que afecten a la supervivencia de los pacientes afectados de adenocarcinoma puede ayudar a la comprensión de este tipo de tumores, pudiendo llegar a ser usados como marcadores genéticos o como posibles dianas terapéuticas.

1.2 Objetivos del trabajo

Dada la importancia del cáncer en la sociedad, se plantea un estudio de expresión diferencial y de supervivencia con el objetivo de obtener un perfil de expresión asociada a la supervivencia de miRNAs específico de cada tipo de adenocarcinoma. Los objetivos se detallan a continuación:

Objetivos generales

1. Elaboración de un panel de miRNAs desregulados en procesos tumorales e implicados en la supervivencia de los pacientes usando datos del repositorio TCGA.
2. Estudio de genes diana relacionados con miRNAs identificados en el punto anterior mediante el uso de bases de datos especializadas.
3. Integración de la información interactómica miRNA-target en el contexto patológico de la enfermedad.
4. Diseño de una metodología que se pueda reproducir de manera sencilla con datos renovados (Reproducible Research).

Objetivos especificos

Objetivo 1

1. Selección de cánceres a estudio.
2. Obtención de la información clínica y de expresión necesaria para cada uno de los cánceres seleccionados.
3. Diseño del pipeline, comprendiendo la ejecución de herramientas de calidad, filtrado y normalización de la información seleccionada.
4. Diseño del pipeline, comprendiendo la ejecución de análisis estadísticos de expresión diferencial y supervivencia, en los distintos tipos de cáncer.
5. Análisis en conjunto que permita obtener resultados comunes en todos los tipos de cáncer, más robustos y extrapolables, con el objetivo de descubrir características comunes.

Objetivo 2

1. Selección de bases de datos especializadas en la anotación de genes diana de miRNAs.
2. Revisión de las bases de datos

3. Creación de una base de datos propia para el panel de genes obtenido en el objetivo 1 mediante la integración de resultados de las bases de datos seleccionadas y revisadas.
4. Completación, curado y optimización de la base de datos de genes diana creada.
5. Anotación de los genes diana mediante el uso de ontologías GO y KEGG.

Objetivo 3

1. Obtención de datos de genes diferencialmente expresados para cada tipo de cáncer mediante la misma metodología descrita en el objetivo 1.
2. Diseño del pipeline de integración de la información de expresión diferencial para ambos tipos de moléculas con el objetivo de determinar posibles interacciones.

Objetivo 4

1. Programación de herramientas automatizables y reproducibles mediante el uso de software accesible y herramientas bioinformáticas contrastadas.

1.3 Enfoque y método seguido

En el siguiente apartado se define el enfoque práctico del proyecto, dividido en cada una de las etapas que ha comprendido y se establece una primera aproximación a los recursos usados, estableciéndose una explicación más detallada de cada uno de ellos en el apartado Materiales y Métodos de esta memoria.

1.3.1 Revisión bibliográfica

La estrategia usada para el desarrollo de este estudio se empieza con una revisión bibliográfica. Se lleva a cabo una búsqueda bibliográfica de estudios y procedimientos similares, así como más información acerca de la enfermedad y

el papel regulador que los miRNAs puedan realizar sobre ésta (Mar-Aguilar et al. 2016, Xiao & Rajewsky, 2009, Zhang et al. 2015).

1.3.2 Recursos informáticos

Para implementar la herramienta de análisis, se utiliza R como lenguaje de programación y R-Studio como entorno de trabajo, un entorno ampliamente utilizado en análisis de datos biológicos y con muchos recursos disponibles relativos a test estadísticos expresión diferencial y supervivencia.

En el análisis computacional es importante realizar una investigación reproducible y que las instrucciones del análisis de datos estén disponibles e inteligibles, de modo que cualquier investigador pueda recrear el mismo análisis con datos renovados. En este sentido, el entorno de programación R, dado su carácter de software libre y código abierto, ofrece a los investigadores la posibilidad de acceder a este estudio de una manera sencilla.

1.3.3 Construcción del panel de miRNAs diferencialmente expresados afectantes de supervivencia

1.3.3.1 Cohorte de pacientes y datos de miRNAs

El presente estudio se lleva a cabo mediante la formación de dos grupos de pacientes tumorales procedentes del repositorio TCGA diferenciados en sus datos de supervivencia. De un lado, el grupo de alta supervivencia comprende aquellos pacientes con unas características de supervivencia mayores, habiendo estado bajo seguimiento o habiendo llegado a su muerte, según sea el caso, en un periodo mayor de tiempo. Por otro lado, el grupo de baja supervivencia está compuesto de aquellos pacientes tumorales con una tasa de supervivencia o seguimiento inferior. Para un mayor detalle de los pacientes incluidos en el análisis y sus características de supervivencia ver el apartado Introducción de esta memoria.

Los datos para los pacientes anteriores se disponen en una matriz de expresión para cada tipo de cáncer, en el que cada columna representa a un paciente y cada fila un miRNA. El valor para cada celda por tanto viene determinado por el número de veces que se encuentra un miRNA en un paciente concreto. Esta matriz de expresión es el archivo input del análisis sobre el cual trabajaremos y contendrá un grupo de datos de cuentas para un total de 1881 miRNAs, todos ellos anotados en la miRbase v.21 (Kozomara & Griffiths, 2014).

1.3.3.2 Limpieza de datos

En las matrices de expresión obtenidas para cada tipo de adenocarcinoma, cada línea contiene el valor concreto de expresión (cuentas crudas) de un miRNA y en este paso se eliminan del muestreo aquellos pacientes que no tienen información en el archivo clínico, ya que a pesar de disponer de datos de expresión, si no se dispone de esta información clínica, no sería posible discernir si dichos datos pertenecen a un grupo u otro de supervivencia. También se eliminan todos aquellos miRNAs que no se encuentren presentes en más de 15 pacientes, siendo candidatos a ser artefactos de secuenciación y riesgo de falsos positivos.

1.3.3.3 Normalización

La normalización permite la comparación de niveles de expresión entre distintas muestras, e incluso dentro de las mismas muestras, la comparación entre réplicas distintas (Marioni et al., 2008). Se pretende asegurar que las diferencias en el número de lecturas obtenidas a la hora de comparar dos muestras reflejen la expresión diferencial existente, y no sesgos artificiales de secuenciación. En este caso la librería de R usada es edgeR (Robinson et al. 2010) y el método de normalización es TMM (trimmed mean of M-values normalization), un método simple y efectivo para datos de miRNA-seq que reduce la tasa de falsos positivos en el análisis de expresión diferencial (Robinson y Oshlack, 2010).

1.3.3.4 Análisis de expresión diferencial

A partir de los datos normalizados, la librería EdgeR es capaz de realizar el análisis de expresión diferencial usando datos de diversas plataformas, miRNA-seq incluída. La comparación se realiza entre los dos grupos formados, tomando como referencia el grupo de alta supervivencia.

1.3.3.5 Análisis de supervivencia

El análisis de supervivencia es un tipo de análisis estadístico basado en el seguimiento de los individuos de un estudio desde una experiencia inicial o exposición hasta la ocurrencia de un evento. El evento observado suele ser una variable dicotómica, es decir, puede tomar dos valores. En este caso la variable observada será el estado vital de los pacientes.

Para realizar el análisis de supervivencia, primero es necesario construir un modelo que permita discernir aquellos miRNA cuya alteración afecte significativamente a la supervivencia de los individuos en cada cáncer. Entre los diferentes tipos de modelos multivariantes, uno de los más extendidos en medicina es el modelo de riesgos proporcionales, también conocido como modelo de Cox (Cox, 1972).

La regresión de Cox busca cuales de las variables independientes introducidas en el modelo se relacionan con variaciones en la función de supervivencia, para lo cual calcula, para cada variable (miRNA diferencialmente expresado), la probabilidad de que no influya en el modelo construido (p-valor), y también un coeficiente que indique el peso de las variables en el modelo. Seleccionamos los valores de corte basándonos en un análisis ROC. A nivel de expresión diferencial de cada miRNA, se grafican la sensibilidad y la especificidad del modelo generándose una curva ROC. El punto de mayor sensibilidad y mayor especificidad será el seleccionado como corte, y será el que usaremos en un análisis de supervivencia Kaplan-Meier para separar aquellos pacientes con una abundancia alta del miRNA y aquellos con una abundancia pequeña.

1.3.3.6 Análisis de enriquecimiento GO y KEGG

Se analizan los genes potencialmente dianas que podrían ser regulados por los miRNAs obtenidos como afectantes de la supervivencia. Los genes diana son obtenidos a través de la plataforma mirWalk2.0 (Dweep & Gretz, 2015). Se realiza un análisis de enriquecimiento de términos GO mediante la herramienta PANTHER (Mi et al. 2013) y de rutas metabólicas mediante la herramienta Reactome (Fabregat et al. 2017).

1.3.4 Interacción miRNA-mRNA

Mediante la misma metodología del apartado 1.3.3.4, se obiene la expresión diferencial de genes usando una muestra representativa de los pacientes obtenidos para el análisis. Se describen las posibles interacciones entre miRNAs diferencialmente expresados y mRNAs diferencialmente expresados, atendiendo a si estos últimos son genes potencialmente dianas de los primeros.

1.4 Planificación del trabajo

1.4.1. Hitos, Objetivos y Tareas:

Objetivo 1: Panel de miRNAs desregulados que afecten supervivencia

Elaboración de un panel de genes miRNAs desregulados en procesos tumorales que afecten a la supervivencia total en pacientes de adenocarcinoma mediante el procesado de datos del repositorio TCGA.

Descripción	Nombre	Inicio	Fin	Prioridad	Jornadas
Objetivo 1	Panel de miRNAs	16/10/2017	25/11/2017	Alta	40
Tarea 1	Selección de cánceres a estudio	16/10/2017	17/10/2017	Alta	1
Tarea 2	Obtención de datos cínicos y de expresión	17/10/2017	20/10/2017	Alta	4
Tarea 3	Pipeline de preprocesado	21/10/2017	30/10/2017	Alta	10
Tarea 4	Pipeline de expresión diferencial y supervivencia	31/10/2017	25/10/2017	Alta	26
Tarea 5	Anàlisis en conjunto	20/10/2017	25/10/2017	Alta	6
Hito 1	Panel de miRNAs	16/10/2017	25/11/2017	Alta	40

Tabla 1: Descripción de tareas del objetivo 1

Tareas

1. Selección de cánceres a estudio. Elegir los tipos de cánceres que van a formar parte del estudio dentro del tipo adenocarcinoma.
2. Obtención de datos clínicos y de expresión: Obtención de los datos del repositorio TCGA referentes a la expresión diferencial de miRNAs y datos clínicos de los pacientes.
3. Pipeline de preprocesado: Diseño e implementación de un pipeline que comprenda los pasos de preprocesado, filtrado y normalización de las muestras obtenidas en las tareas anteriores.
4. Pipeline de preprocesado: Diseño e implementación de un pipeline que comprenda los pasos de expresión diferencial y análisis de supervivencia de las muestras anteriormente preprocesadas.

5. Análisis en conjunto de los diferentes tipos de cánceres estudiados con el objetivo de obtener resultados más robustos y fiables y de obtener patrones comunes en los distintos tipos de adenocarcinoma.

Objetivo 2: Genes diana.

Este objetivo consiste en la elaboración de una base de datos consistente en la anotación de los genes diana de lo miRNA obtenidos en el objetivo 1.

Descripción	Nombre	Inicio	Fin	Prioridad	Jornadas
Objetivo 2	Genes diana	25/11/2017	30/11/2017	Alta	5
Tarea 1	Selección de bases de datos especializada	25/11/2017	30/11/2017	Alta	5
Tarea 2	Revisión de bases de datos	25/11/2017	30/11/2017	Alta	5
Tarea 3	Creación de bases de datos propia y completación	25/11/2017	30/11/2017	Alta	5
Tarea 4	Anotación mediante ontologías	25/11/2017	30/11/2017	Alta	5
Hito 2	Panel de genes diana y anotación	25/11/2017	30/11/2017	Alta	5

Tabla 2: Descripción de tareas del objetivo 2

Tareas

1. Selección de bases de datos especializadas: Selección de las bases de datos que se usarán para la obtención del panel de genes diana atendiento a la contratabilidad de sus anotaciones.
2. Revisión de bases de datos: Evaluar la información de las bases de datos seleccionadas.
3. Creación de la base de datos de genes diana y completación: Construcción de las anotaciones de genes diana de nuestros miRNAs obtenidos mediante el escaneo de las bases de datos seleccionadas y revisadas y optimización de los resultados.
4. Anotación mediante ontologías: Anotación de los genes descritos como dianas mediante el uso de ontologías GO y KEGG.

Objetivo 3: Interactómica miRNA-target en Adenocarcinoma.

Este objetivo consiste en la obtención de posibles interacciones miRNA-mRNA mediante datos experimentales de expresión diferencial obtenidos en el objetivo 1.

Descripción	Nombre	Inicio	Fin	Prioridad	Jornadas
Objetivo 3	Interacción miRNA-mRNA en adenocarcinomas	16/10/2017	18/12/2017	Alta	62
Tarea 1	Obtención de datos de expresión diferencial de genes siguiendo el protocolo establecido en el objetivo 1	16/10/2017	25/11/2017	Alta	40
Tarea 2	Diseño de pipeline de interacción miRNA-mRNA	25/11/2017	18/12/2017	Alta	23
Hito 3	Panel de interacciones	25/11/2017	18/12/2017	Alta	23

Tabla 3: Descripción de tareas del objetivo 3

Tareas

1. Obtención de datos de expresión diferencial de genes siguiendo el protocolo establecido en el objetivo 1: Se pretende determinar genes diferencialmente expresados entre los pacientes mediante los mismos test usados para miRNAs.
2. Diseño de pipeline: Integración de la información de expresión miRNA-mRNA.

1.4.2. Calendario:

La calendarización de las tareas anteriormente descritas en cada uno de nuestros objetivos se puede visualizar en el diagrama de Gannt de la figura 2.

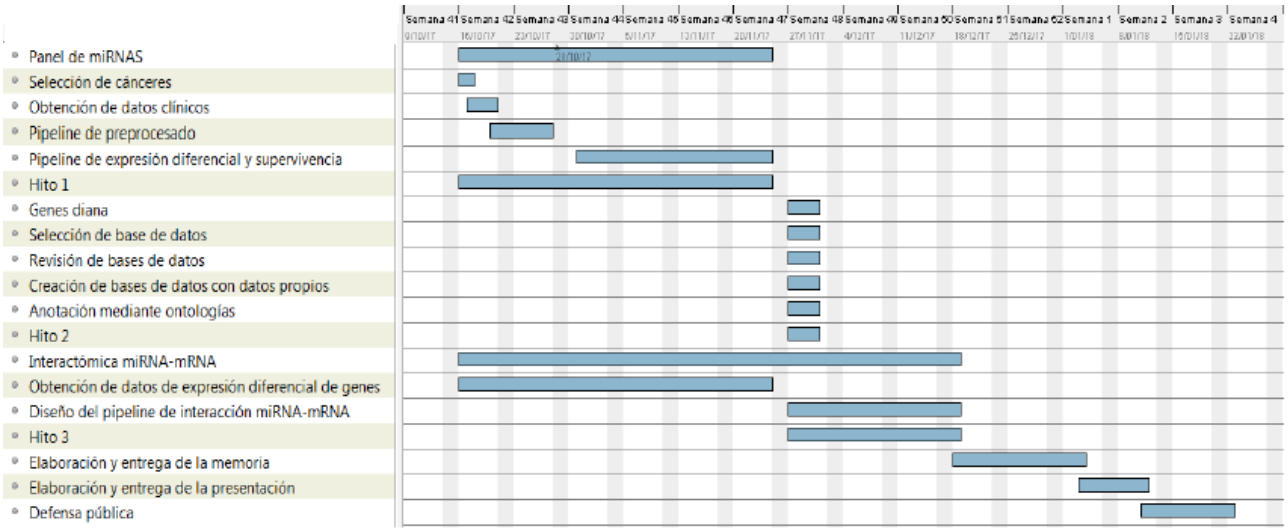

Figura 2: Figura 2: Diagrama de Gannt mostrando la planificación temporal del proyecto.

1.5 Breve sumario de los reusltados obtenidos

- Panel de miRNAs desregulados en ambos grupos del análisis, así como su significancia estadística y su valor de expresión para cada tipo de adenocarcinoma estudiado.
- Selección de miRNA des-regulados implicados en la supervivencia de los pacientes mediante un estimador Kaplan-Meier. Representación de las curvas más significativas para cada tipo de adenocarcinama estudiado.
- Listado de genes diana identificados bibliográficamente en bases de datos especializadas en genes validados de manera funcional.
- Análisis mediante ontologías GO y KEGG de los principales categorías funcionales y rutas metabólicas enriquecidas usando como input el panel de genes diana obtenidos en las muestras para cada tipo de adenocarcinoma estudiado.
- Meta-análisis comparativo de los tipos de adenocarcinoma estudiados, integrando la información de los resultados anteriores.

- Representación gráfica de las relaciones miRNA-mRNA, atendiendo al resultado de un segundo test de expresión diferencial relativo a genes con los mismos pacientes que el análisis de expresión diferencial de miRNAs.

1.6 Breve descripción de los otros capítulos de la memoria

- Introducción: Presentación de las moléculas de miRNA, estructural y funcionalmente. Cáncer en la sociedad. Plataforma The Genome Cancer Atlas (TCGA). Regulación génica mediada por miRNAs y cáncer. Esquema y resultados obtenidos en el muestreo de pacientes para cada tipo de cáncer usado. Entornos de trabajo y recursos bioinformáticos.
- Objetivos: Definición de los objetivos del trabajo, tanto generales como específicos.
- Materiales y métodos: Presentación de los recursos usados para llevar a cabo el estudio y de la estrategia usada para el mismo fin.
- Resultados y discusión: Presentación de los resultados más detacados y significativos obtenidos durante el trabajo, así como una argumentación sobre ellos.

2. INTRODUCCIÓN

2.1. Ciencias ómicas

Durante las últimas décadas, el alto volumen de datos derivados de la biología continúa creciendo de manera acelerada debido, en parte, al abaratamiento de los costes de secuenciación y al desarrollo de tecnologías de alto rendimiento. Estos hechos han promovido la aparición de las ciencias ómicas, tales como la genómica, transcriptómica, etc., que cubren las aproximaciones realizadas recientemente.

Este trabajo se centra en el uso de la bioinformática para el análisis de dicho volumen de datos transcriptómicos basados en moléculas de miRNA y mRNA en pacientes con los siguientes tipos de adenocarcinoma: COAD (Colon Adenocarcinoma), READ (Rectum Adenocarcinoma), LUAD (Lung Adenocarcinoma), PAAD (Pancreatic Adenocarcinoma), PRAD (Prostate Adenocarcinoma) i STAD (Stomach Adenocarcinoma), con el objetivo de obtener información sobre la expresión diferencial de las moléculas de miRNA, su papel en la supervivencia general de los pacientes, la interacción con sus genes diana y las funciones en las que estos genes están implicados.

2.2 Cáncer en la sociedad

La incidencia y mortalidad a causa del cáncer está creciendo globalmente como resultado del crecimiento y envejecimiento de la población. “En 2017, se estima que hayan sido diagnosticados 1688780 nuevos casos de cáncer y se prevé que mueran 600920 personas, solo en Norteamérica, a razón de 1650 personas al día. El cáncer es la segunda causa de muerte más común en Estados Unidos, solo superada por enfermedades cardíacas” (Cancer Facts & Figures, 2017).

En relación con los cánceres estudiados en este informe, en hombres, los cánceres de pulmón, colon, recto y próstata son los tipos que más se prevé que crezcan en 2017, ocupando el 41% del total de nuevos casos estimados y el 44% de muertes totales estimadas. En este último aspecto, el cáncer de páncreas también tiene asociado un alto aumento (7%) de muertes. En mujeres, los cánceres de pulmón, colon y recto representan el 20% de nuevos casos estimados y el 33% de muertes esperadas. De nuevo, se cree que las muertes por cáncer de páncreas pueden aumentar en un 7%. El cáncer de estómago no se prevé que haya aumentado en gran medida durante 2017.

		Hombres		Mujeres	
Nuevos Casos Estimados (2017)		Próstata	19%	Pulmón	12%
		Pulmón	14%	Colon y recto	8%
		Colon y recto	9%	Páncreas	3%
		Páncreas	<3%	Estómago	<3%
		Estómago	<3%		
Muertes estimadas (2017)		Pulmón	27%	Pulmón	25%
		Colon y recto	9%	Colon y recto	8%
		Próstata	8%	Páncreas	7%
		Páncreas	7%	Estómago	<3%
		Estómago	<3%		

Tabla 4: Porcentaje de nuevos casos estimados y de aumento de muertes estimadas en relación a los canceres estudiados (Cancer Facts & Figures, 2017).

2.3 Regulación génica mediada por miRNAs y cáncer

Los MicroRNAs (miRNAs) son regiones pequeñas (20-22 nucleótidos) no codificantes que juegan un papel fundamental en todas las rutas metabólicas de los organismos multicelulares, incluyendo a los mamíferos. “Bajo condiciones fisiológicas normales, la función de los miRNAs implica salvaguardar procesos biológicos clave, como la proliferación celular, la diferenciación o la apoptosis” (Reddy, 2015).

El cáncer es una enfermedad compleja y dinámica, que “envuelve una variedad de cambios en la estructura y la expresión de los genes” (Stahlhut & Slack, 2006). Tradicionalmente, el estudio del cáncer se ha centrado en las moléculas de RNA mensajero (mRNA) de genes codificantes de proteína, considerándose éstos como los principales efectores y reguladores de la tumorogénesis.

Sin embargo, la desregulación de los miRNAs ha sido reportada como el origen de la expresión diferencial de cientos de genes, hecho que pone de manifiesto la estrecha relación entre miRNAs y mRNAs. No obstante, todavía se conoce poco acerca de las dianas específicas de los miRNAs y sus funciones biológicas. Los métodos bioinformáticos, así como los datos genómicos creados y conservados por organismos e iniciativas públicas hacen que el estudio en esta área tenga una progresión mucho mayor.

Funcionalmente, los miRNAs son moléculas evolutivamente conservadas, de cadena simple, que se unen a los mRNA para prevenir la formación de proteínas de dos modos distintos, que comprenden la represión de la traducción o la rotura directa del mRNA. El nivel de complementariedad entre ambas moléculas determina el proceso seguido. Para producirse la unión, los precursores de miRNAs se unen al complejo RISC, una ribonucleoproteína que se une a la cadena simple de miRNA. Esta cadena simple actua como molde para el complejo RISC para reconocer los mRNA complementarios. Posteriormente, se produce el mecanismo de silenciamiento de genes, realizado por una endonucleasa Ago2 (en caso de una complementariedad

extensa, o alta) o por el silenciamiento de la traducción (en caso de una complementariedad limitada, o media). En la figura 3, se representa un esquema relativo a la biogénesis y la función de los miRNAs.

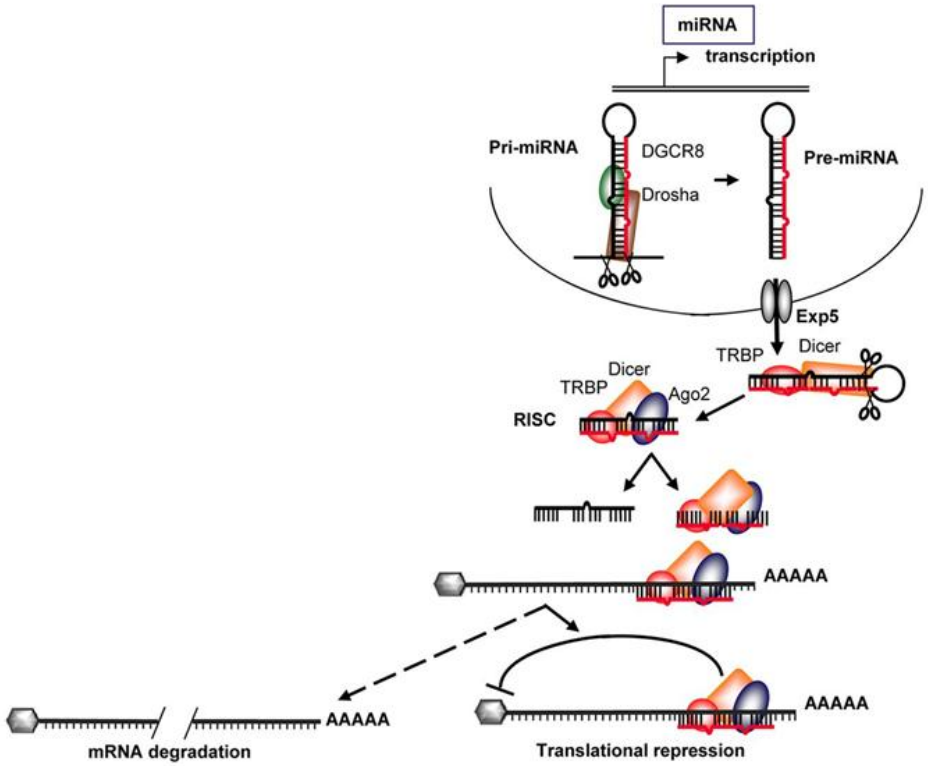

Figura 3: Biogénesis y mecanismo de acción de los miRNAs (Sylvanne M. Daniels, and Anne Gatignol Microbiol. Mol. Biol. Rev. 2012; 76:652-66s)

En 2002, el grupo de investigación del Dr. Croce reportó el primer caso de desregulación de miRNAs que afectaban a tumorogénesis (Calin et al., 2002). En este estudio, no se pudo encontrar ningún gen codificante de proteína relacionado con la Leucemia crónica linfocítica ligada a células B (CLL), pero sí se reconocieron dos moléculas de miRNAs, miR-15a y miR-16-1. Además, la expresión de estos miRNA se encontró reducida o incluso suprimida en más de un 65% de casos de CLL examinados. Esta pérdida en la expresión del miRNA así como la pérdida de la región cromosómica 13q14 relacionada con la CLL sugirió que estos miRNAs podían actuar como supresores de genes.

Desde este punto, las múltiples líneas de investigación han indicado que los miRNAs podían servir como herramientas para conocer perfiles tumorales. Cada tumor posee una firma de expresión de miRNA particular y esta firma puede proveer de una útil información sobre la malignidad del tumor o la supervivencia del paciente. Siguiendo con esta línea de investigación, en el presente estudio se describe la estrategia bioinformática seguida para el estudio de datos de expresión en los seis tipos de adenocarcinoma anteriormente nombrados.

2.4 Organización de la información genómica

La adaptación de las tecnologías de alto rendimiento ha facilitado el estudio de la expresión de múltiples miRNAs en una muestra, haciendo que perfilar un número sustancial de muestras tumorales sea relativamente sencillo. Una de estas técnicas es la miRNA-seq, o secuenciación de miRNAs. Esta técnica es una variación de la técnica de RNA-seq que se diferencia en el material input del análisis. Las muestras en miRNA-seq están enriquecidas con RNA pequeños (small RNA en inglés).

Por otro lado, uno de los objetivos de la bioinformática de hoy es la organización de la información de una manera que permita a los investigadores acceder a ella e introducir nuevos datos conforme se van produciendo. En este sentido, se han creado bases de datos o consorcios, como NCBI (NCBI Resource Coordinators, 2013) o TCGA (Tomsczak et al., 2015).

TCGA es un consorcio formado por National Cancer Institute (NCI) y National Human Genome Research Institute (NHGRI) que ha generado datos genómicos clave de una manera comprensible y multidimensional en 33 tipos de cáncer. Los datos de este repositorio están disponibles públicamente y sirven a la comunidad científica para mejorar la prevención, diagnóstico y tratamiento del cáncer.

Los datos TCGA tienen una doble funcionalidad en este estudio. Por una parte, TCGA ofrece los datos clínicos de los pacientes incluidos en el seguimiento para cada tipo de cáncer. Se obtienen datos tales como el tiempo de supervivencia, el estado vital y diferentes características fisiológicas de los pacientes. Por otra parte, TCGA ofrece los datos relativos a experimentos de miRNA-seq realizados sobre los pacientes del archivo clínico. Así, los archivos de cuentas con el número de veces que aparece un miRNA en un paciente están accesibles en el repositorio y se pueden descargar fácilmente.

2.5 Entornos de trabajo y recursos bioinformáticos

Para el análisis bioinformático o la creación de nuevas herramientas es necesario utilizar distintos lenguajes de programación. Entre ellos destacan python (Van-Rossum, 2003) o Perl (Wall et al. 2004). Sin embargo, uno de los lenguajes más utilizados para el análisis de datos es el lenguaje de programación R.

R es un software libre enfocado al análisis estadístico y gráfico, inspirado en el lenguaje S, uno de los lenguajes más utilizados en investigación por la comunidad estadística. Fue desarrollado inicialmente por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland en 1993 (Ihaka y Gentleman, 1996), y actualmente su desarrollo es responsabilidad del R Development Core Team.

Mediante el empleo de R, el análisis de datos puede realizarse de una manera reproducible. “En el análisis computacional es importante que los pasos y las instrucciones que se deben seguir para la realización de dicho análisis estén disponibles y sean inteligibles, de manera que cualquier investigador pueda repetir el mismo análisis con nuevos datos experimentales y verificar o actualizar el análisis con los mismos datos. Este concepto se denomina investigación reproducible (*Reproducible Research*)” (Donoho, 2010)

Los distintos paquetes de R y sus respectivos manuales se encuentran almacenados en varios repositorios. Los repositorios usados en este análisis han sido CRAN (Comprehensive R Archive Network) y Bioconductor, que contienen paquetes destinados al análisis de datos genómicos obtenidos mediante tecnologías de alto rendimiento.

El uso de R es mucho más cómodo usando la interfaz RStudio (Racine, 2012), ya que permite realizar todos los pasos del análisis de manera fácil y visual. Aporta información sobre los objetos que se encuentran en el entorno de trabajo, permite ver en la misma interfaz las gráficas que se vayan haciendo y también consultar los manuales de los distintos paquetes, en definitiva, da muchas facilidades a la hora de realizar una investigación.

Además de los lenguajes de programación R y Python anteriormente mencionados, con los que se lleva a cabo la mayor parte de este análisis, también se ha dispuesto de diferentes herramientas genómicas disponibles de manera libre, como son la plataforma MiRWalk2.0 (selección de genes diana de los miRNAs), PANTHER (análisis de enriquecimiento GO) y Reactome pathways (análisis de enriquecimiento KEGG).

3. MATERIALES Y METODOS

El esquema de trabajo seguido se representa en el diagrama de la Figura 4. Mediante los recursos informáticos anteriormente mencionados, como el lenguaje de programación R, se obtienen los datos relativos a la cuantificación de miRNA y los datos clínicos presentes en TCGA para cada uno de los adenocarcinomas estudiados. A su vez, estos datos se procesan en los dos grupos del análisis y una vez establecidos los pacientes que se usan, se obtienen datos de cuantificación de mRNA.

La metodología que se sigue tiene como objetivo estudiar la influencia de la expresión diferencial de las moléculas de miRNA en la supervivencia del paciente. También se estudia la expresión diferencial de los mRNA obtenidos.

Asimismo, se realiza un estudio de las posibles diana de cada uno de los miRNA diferencialmente expresados, un análisis de enriquecimiento GO y KEGG para cada tipo de adenocarcinoma y se comparan datos de expresión diferencial miRNA-mRNA diana con el fin de observar posibles interacciones entre estos dos tipos de moléculas.

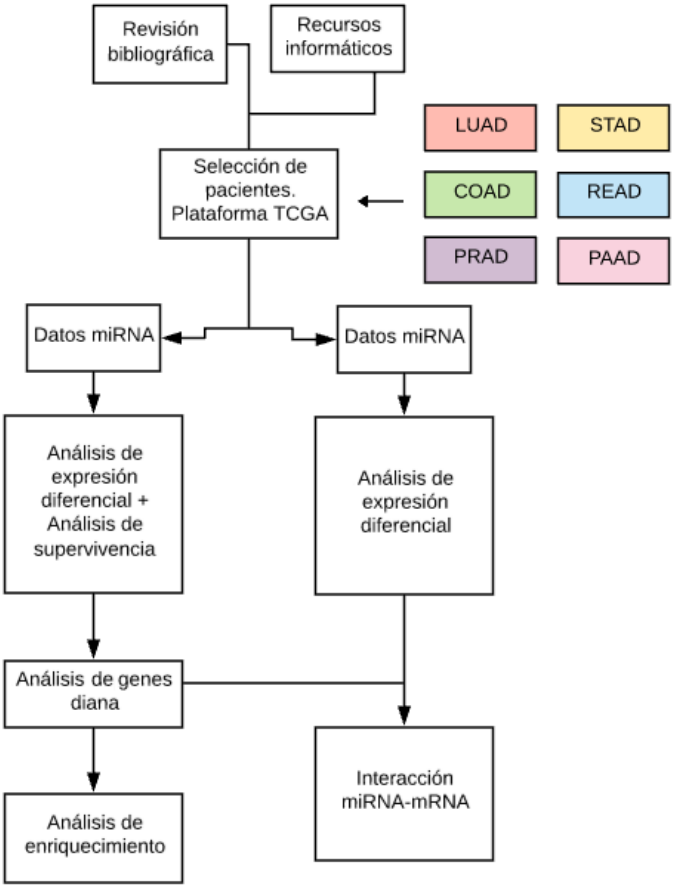

Figura 4: Esquema representativo de la metodología llevada a cabo para realizar el estudio.

3.1 Obtención de datos transcriptómicos

El repositorio TCGA ofrece la posibilidad de acceder a los distintos datos transcriptómicos de cuantificación de miRNAs en los distintos tipos de cánceres seleccionados y a los datos clínicos de los pacientes de dichos cánceres. De una manera más eficaz, la librería TCGAbiolinks (Colaprico et al., 2016) permite la obtención de los datos a través del lenguaje de programación R. Usando las funcionalidades de esta librería, para este estudio se seleccionan los datos de cuantificación de miRNAs para muestras del tipo tumor sólido primario (en inglés, primary solid tumor).

La librería otorga un archivo por paciente para todos los pacientes que componen el archivo clínico que se descarga. En el archivo, cada línea representa un miRNA y contiene dos columnas. La primera alberga las cuentas crudas de cada miRNA, y la segunda las cuentas normalizadas. Debido a la normalización que incluye el procedimiento, se eligen las cuentas crudas.

Mediante el lenguaje de programación Python y la línea de comandos de Linux, estos archivos son integrados en un único archivo.

En cada uno de los cánceres a estudio, las poblaciones de pacientes se dividen en dos grupos, según la supervivencia de los pacientes fuese alta o baja. Los datos clínicos usados para esta división son los correspondientes a los días desde que el paciente está bajo seguimiento (en caso de estar vivo; en inglés y en los datos clínicos, days to last follow up) y los días desde el inicio del seguimiento hasta la muerte (en caso no estar vivo; en inglés y en los datos clínicos, days to death). En caso de que ambos valores existan en los datos clínicos para un paciente, se elige el mayor valor entre ellos. Las poblaciones en ambos grupos fueron igualadas para evitar posibles diferencias en la cuantificación. Así, los grupos de pacientes que se compararon en este estudio son pacientes tumorales, siendo el dato de supervivencia su factor diferenciador. Para la creación de los grupos, se clasifican los pacientes desde el de mayor supervivencia al de menos supervivencia. Esta clasificación se divide en tres conjuntos y se escogen el primero (mayor supervivencia) y el último (menor supervivencia). De esta manera se eliminan pacientes intermedios que puedan modificar el análisis y se asegura una correcta diferenciación de los datos de supervivencia.

Se selecciona de la matriz de expresión los pacientes que vamos a usar correspondientes a cada grupo. La matriz resultante contiene datos de cuantificación de 1881 miRNAs, todos ellos anotados en miRbase v.21. Los datos clínicos se emparejan con los datos de esta matriz de expresión, eliminando aquellos pacientes sin valores de expresión conocidos o sin algún dato de supervivencia, clave en el análisis. En las tablas 5 y 6 se observan las características de supervivencia y el número de pacientes incluídos en el análisis para cada tipo de adenocarcinoma.

DATOS DE PACIENTES				
Tipo de adenocarcinoma	**Número de pacientes totales disponibles en TCGA****	**Número de pacientes incluidos en el análisis***	**Grupo de alta supervivencia**	**Grupo de baja supervivencia**
COAD	459	218	110	108
PAAD	185	113	56	57
PRAD	500	225	112	113
LUAD	522	210	108	102
READ	170	102	50	52
STAD	443	226	115	111
**** Número de pacientes de TCGA para los cuales se tienen datos clínicos**				
*** Número de pacientes con mayor y menor tasa de supervivencia excluyendo aquellos con valores no conocidos de days to last follow up y/o days to death**				

Tabla 5: Distribución y número de pacientes usado en el análisis

	DATOS DE LOS GRUPOS DE ALTA SUPERVIVENCIA			DATOS DE LOS GRUPOS DE BAJA SUPERVIVENCIA		
Tipo de adenocarcinoma	**Vivos**	**Muertos**	**Intervalo de supervivencia (días)**	**Vivos**	**Muertos**	**Intervalo de supervivencia (días)**
COAD	92	18	4502-952	67	41	0-442
PAAD	37	19	2741-614	19	38	0-314
PRAD	109	3	5024-1262	112	1	23-668
LUAD	71	37	7062-904	54	48	0-469
READ	41	9	3932-1003	44	8	0-425
STAD	89	26	3720-615	49	62	0-292

Tabla 6: Estado vital y datos cuantitativos de supervivencia en cada uno de los grupos comparados en el análisis

3.2 Filtrado de datos y normalización

Los archivos de datos transcriptómicos de cada uno de los cánceres seleccionados contienen miles de líneas, en los que cada una de ellas contiene el valor relativo al número de veces que un miRNA ha sido mapeado en un paciente determinado tras un experimento de secuenciación NGS mediante la plataforma miRNA-seq. Estos conteos pueden ser usados para realizar un análisis de expresión diferencial con el objetivo de determinar si algún miRNA está diferencialmente expresado en ambos grupos. En este estudio, se toma el grupo de alta supervivencia como referencia, relacionándose así todos los valores de expresión con este grupo.

Para el filtrado de datos, se eliminan de las matrices de expresión todos aquellos miRNAs cuyo conteo en más de 15 pacientes fuera 0, para evitar

posibles sesgos causados por presencia/ausencia y no por una expresión diferencial real. La tabla 7 muestra el número de miRNAs a estudio antes y después de este filtrado. Como se observa, la reducción de las matrices es evidente, dejando entrever la posibilidad de que muchos de los miRNAs cuantificados sean artefactos o aparezcan puntualmente en un paciente. Al eliminarlos, reducimos el riesgo de falsos positivos que estos miRNAs aportan al análisis.

DATOS DE FILTRADO		
Tipo de adenocarcinoma	**miRNAs pre-filtrado**	**miRNAs post-filtrado**
COAD	1881	319
PAAD	1881	388
PRAD	1881	306
LUAD	1881	336
READ	1881	363
STAD	1881	311

Tabla 7: Número de miRNAs antes y después del filtrado de datos

La normalización permite la comparación de niveles de expresión entre distintas muestras, e incluso dentro de las mismas muestras en réplicas distintas (comparación de la expresión de distintos miRNAs en un individuo) (Marioni et al. 2008). Pretende asegurar que las diferencias en el número de lecturas obtenidas a la hora de comparar dos muestras, reflejen realmente la expresión diferencial de los miRNAs y no sesgos artificiales derivados de la secuenciación.

Existen varios métodos de normalización de datos genómicos, algunos de ellos como el ajuste según el tamaño de la biblioteca genómica (total de lecturas) son aproximaciones demasiado simples para varias aplicaciones biológicas, aunque sí que pueden servir para la normalización entre réplicas de muestras. El método de normalización elegido se trata de un método simple y efectivo para la estimación de los niveles de producción de ARN procedentes de datos de RNA-Seq, ya que baja la tasa de falsos positivos en el análisis de expresión diferencial (Robinson y Oshlack, 2010). Dicho método se trata de la normalización por TMM (en inglés, trimmed mean of M-values normalization),

que se aplica sobre los datos transcriptómicos preparados en matrices usando el paquete de R edgeR.

3.3 Expresión diferencial

El uso de tecnologías que miden la expresión génica es frecuente en el campo de la biología molecular para obtener una imagen de la actividad transcripcional en diferentes tejidos o poblaciones celulares. Estos perfiles después son comparados para identificar cambios en la expresión de genes asociados a un tratamiento o fenotipo de interés, lo que se conoce como análisis de la expresión diferencial.

Los modelos de conteos de lecturas se originaron de la idea de que cada lectura se muestrea independientemente de un grupo de lecturas y por tanto, el número de lecturas para un miRNA determinado siguen una distribución binomial, que puede ser aproximada a una distribución de Poisson (Huang et al., 2015). En base a esta distribución de Poisson para repetidos eventos de secuenciación, una de las aproximaciones que se propone es usar un modelo log-lineal para modelar la diferencia media existente entre los dos grupos y adoptar un test de similitud para calcular los p-valores, realizando a su vez correcciones por comparaciones múltiples (Marioni et al. 2008).

De esta manera, la estrategia usada parte de la creación de una matriz de datos normalizados únicamente con las cuentas de datos, seguida de la estimación del parámetro de dispersión binomial negativo para cada uno de los conteos. Con estos parámetros, se computa un estimado de Bayes empírico, con los niveles de expresión especificados por un modelo log-lineal. Por último, se realizan test exactos entre dos grupos de cuentas normalizadas y distribuidas de manera binomial negativa. Se testea para cada línea la diferencia en la media de los dos grupos y se le añade un valor de significancia (p-value) a cada variable, es decir, a cada miRNA, así como un factor de correcciones múltiples (FDR).

Se realiza el mismo sistema para muestras de miRNAs y para muestras de mRNAs y se seleccionan aquellas variables (mensajeros o miRNAs) que posean un valor estadístico p ($p < 0.05$) y un valor de FDR ($FDR < 0.01$) como diferencialmente expresadas. Además, la librería EdgeR otorga también valores de log2FC y CPM para cada variable.

3.4 Análisis de supervivencia

El análisis de supervivencia es un tipo de análisis estadístico basado en el seguimiento de los individuos de un estudio desde una experiencia inicial o exposición hasta la ocurrencia de un evento. El evento observado suele ser una variable dicotómica, es decir, puede tomar dos valores. En este caso la variable observada fue el fallecimiento de los pacientes.

Un aspecto importante del análisis de supervivencia es que tiene en cuenta los periodos de seguimiento de los individuos, por lo que no tendrá el mismo peso en el análisis un evento que ocurre a la semana de comenzar el estudio que uno que ocurre al final. También permite incorporar al modelo los datos de individuos que se incorporan tarde al estudio o que nunca sufren el evento de interés, ya sea debido a que se pierde su seguimiento en el estudio, porque el paciente sufre otro evento distinto al estudiado o porque no ocurra el evento durante el estudio (datos censurados) (Flynn, 2012). En este sentido, los datos de TCGA encajan en la definición del análisis, ya que dan pacientes vivos o muertos y con un valor de seguimiento cuantitativo.

El análisis de supervivencia se realiza únicamente con aquellos miRNAs diferencialmente expresados mediante un análisis univariante, para determinar si la alteración de alguno de ellos que eran significativos en el modelo anterior, es decir, si por sí mismo es capaz de predecir significativamente la supervivencia en un cáncer.

Para ello se empleó una de las técnicas más realizadas en los análisis de supervivencia, la función de Kaplan-Meier (Kaplan y Meier, 1958). Se trata de una técnica que considera en distintos puntos el número de pacientes que

permanecen en la población y el número de eventos acumulados que han ocurrido hasta ese punto, a partir de lo cual, va calculando la probabilidad acumulada de supervivencia en los distintos momentos del estudio. El método más habitual de observar los resultados es mediante la representación gráfica de la funcion de supervivencia Kaplan-Meier frente al tiempo, donde generalmente se suelen representar conjuntamente los distintos grupos a estudiar, para comparar las curvas y ver si existen diferencias. Finalmente para argumentar estadísticamente si existe una diferencia significativa entre dos o varias curvas de supervivencia se suele aplicar un test (en inglés, *log-rank test*), de donde se obtiene un p-valor a partir del cual se determina si existe dicha diferencia significativa entre las curvas. Este test fue realizado mediante la función *survdiff* del paquete *survival* (Therneau, 2014.) En las curvas representadas, se comparan niveles altos y bajos del miRNA, estableciendo como punto de corte aquel en el cual se reduzca la tasa de falsos positivos y falsos negativos. Este punto de corte se calcula mediante la librería de R pROC (Robin et al., 2015).

3.5 Análisis funcionales y de enriquecimiento

Una vez establecido el panel de miRNAs asociados con la supervivencia, se realiza una anotación de sus potenciales genes diana validados mediante la plataforma mirWalk2.0. Con la obtención de estos datos, se realiza un análisis de enriquecimiento con el objetivo de conocer posibles rutas metabólicas, procesos biológicos, funciones moleculares o componentes celulares en los que un contexto patólogico de adenocarcinoma pueda estar enriquecido. Las herramientas usadas para el análisis de enriquecimiento son PANTHER y Reactome.

El sistema de clasificación PANTHER combina la función génica, la ontología y herramientas estadísticas para permitir a los investigadores analizar datos de expresión diferencial de diferentes tipos de experimentos y sirve para llevar a cabo análisis de enriquecimiento de términos de Gene Ontology (GO).

Reactome es una base de datos curada de rutas metabólicas y reacciones biológicas en humanos. Además contiene diversas herramientas de análisis

que dan como resultado aquellas rutas diferenciadas en una muestra de genes. En este sentido, con esta herramienta se realiza el análisis de enriquecimiento de rutas metabólicas basado en la ontología KEGG.

4. RESULTADOS Y DISCUSIÓN

Los adenocarcinomas seleccionados se dividen en 2 grupos atendiendo a su valor de supervivencia. En las tablas 8, 9 y 10 se muestra una relación de los datos mencionados en apartados anteriores para cada una de las sub-poblaciones creadas en cada tipo de cáncer.

	Cohorte COAD (n = 215)		Cohorte PRAD (n = 225)	
Característica	**Alta supervivencia**	**Baja supervivencia**	**Alta supervivencia**	**Baja supervivencia**
Número de pacientes	107	108	112	113
Sexo				
Hombre	54	58	112	113
Mujer	53	50	0	0
Edad (años)				
Mediana	68	69	61	63
Rango	37-90	31-90	46-76	42-74
Estado patológico				
Estadío I	13	14	0	0
Estadío II	50	39	0	0
Estadío III	13	17	0	0
Estadío IV	7	21	0	0
Not Reported	17	17	112	113
Raza				
White	59	43	37	18
Black or African American	12	16	2	0
Asian	1	5	1	0
Not reported	35	44	72	95
Días de supervivencia (días)				
Mediana	1348	275,5	1784,5	395
Rango	952-4502	0-442	1262-5024	23-668
Media	1710,11	240,87	1942,16	356,1
Desviación estándar	871,73	150,36	668,43	193,71
Estado vital				
Vivo	89	67	109	112
Muerto	18	41	3	1
Ratio de supervivencia				
1 año	1	0,35	1	0,56
3 años	0,76	0	1	0
5 años	0,31	0	0,48	0

Tabla 8: Datos clínicos por grupo. COAD y PRAD

	Cohorte PAAD (n = 113)		Cohorte STAD (n =220)	
Característica	**Alta supervivencia**	**Baja supervivencia**	**Alta supervivencia**	**Baja supervivencia**
Número de pacientes	56	57	112	108
Sexo				
Hombre	34	34	71	69
Mujer	22	23	41	39
Edad (años)				
Mediana	64	65	65	68
Rango	39-78	35-85	30-90	39-90
Estado patológico				
Estadío I	10	8	18	0
Estadío II	43	49	14	8
Estadío III	1	0	44	46
Estadío IV	0	0	10	19
Not Reported	2	0	26	35
Raza				
White	49	53	69	65
Black or African American	2	0	1	5
Asian	4	4	24	19
Not reported	1	0	18	19
Días de supervivencia (días)				
Mediana	935,5	194	926	148
Rango	614-2741	0-314	615-3720	0-292
Media	1088,55	177,72	1114,53	137,03
Desviación estándar	519,58	97,96	607,78	99,84
Estado vital				
Vivo	37	19	86	46
Muerto	19	38	26	62
Ratio de supervivencia				
1 año	1	0	1	0
3 años	0,36	0	0,33	0
5 años	0,14	0	0,11	0

Tabla 9: Datos clínicos por grupo. PAAD y STAD

	Cohorte LUAD (n = 210)		Cohorte READ (n = 102)	
Característica	**Alta supervivencia**	**Baja supervivencia**	**Alta supervivencia**	**Baja supervivencia**
Número de pacientes	108	102	50	52
Sexo				
Hombre	48	57	29	25
Mujer	60	45	21	27
Edad (años)				
Mediana	65	66	65	65
Rango	39-85	38-85	41-89	31-90
Estado patológico				
Estadío I	59	45	11	11
Estadío II	31	25	21	13
Estadío III	9	19	8	6
Estadío IV	4	7	4	8
Not Reported	5	6	6	14
Raza				
White	82	80	26	19
Black or African American	11	9	3	2
Asian	2	1	0	1
Not reported	13	12	21	30
Días de supervivencia (días)				
Mediana	1362	279,5	1298,5	230,5
Rango	904-7062	0-469	1003-3932	0-425
Media	1696,16	265,31	1495,5	211,65
Desviación estándar	923,76	147,99	671,63	157,85
Estado vital				
Vivo	71	54	41	43
Muerto	37	48	9	9
Ratio de supervivencia				
1 año	1	0,34	1	0,25
3 años	0,79	0	0,86	0
5 años	0,28	0	0,14	0

Tabla 10: Datos clínicos por grupo. LUAD y READ

Como se puede observar en las tablas anteriores, una vez procesados los datos del repositorio, las sub-poblaciones resultantes tienen una homogeneidad muy alta. Características como la edad media o el número de mujeres y de hombres en cada grupo son constantes a lo largo de todos los adenocarcinomas. Esto indica que los resultados obtenidos en el estudio no estarán influenciados por las características fisiológicas de los pacientes. Estas tablas informan por tanto de que la comparación arrojará resultados basados en diferencias moleculares y que los grupos formados no tienen diferencias significativas, por lo que los resultados serán más fiables y robustos.

Como comentarios a las tablas, se debe resaltar que algunos de los cánceres seleccionados, como por ejemplo PAAD o READ, no tienen un número de muertes muy alto, por lo que los análisis de supervivencia podrían estar influenciados por este hecho. Además, la diferenciación según el grado de supervivencia ha formado dos grupos bien diferenciados. La prueba se encuentra en los datos relativos a ratios de supervivencia. Los pacientes de alta supervivencia tienen un tiempo de vida de al menos 1 año en todos los cánceres, mientras que los pacientes de baja supervivencia tienen un ratio muy inferior, que apenas llega a 1 año en la mayoría de casos. Denotar también que en la mayoría de cánceres, las características de estadio y raza también están correctamente anotadas, por lo que sería sencillo relacionar los datos de supervivencia con alguna de estas variables, aunque en este trabajo, solo se asocia con la expresión diferencial de miRNAs.

Los resultados del test de expresión diferencial arrojaron resultados significativos para todos los tipos de adenocarcinoma. Con el objetivo de integrar la información, en la Figura 5 se muestran los volcano plots obtenidos en cada comparación. En rojo, se representan aquellos genes con un valor $p < 0.05$ y en verde aquellos que además de este valor tienen asignado un valor $FDR < 0.05$. Además se incorporan los nombres de los miRNAs con mayor significancia (menor FDR y mayor logFC) para el análisis. El resto de miRNAs extraídos en cada perfil de expresión, así como sus valores de logFC, pValue y FDR se pueden ver en la tabla S1, de la sección Material Suplementario de esta memoria.

En general, los datos de expresión diferencial arrojan datos poco significativos en cuanto a valores de logFC, debido a que todas las moléculas diferencialmente expresadas lo están en una proporción cercana a 1:1, aunque la significancia de los resultados hace que tengan que ser tenidos en cuenta, ya que una expresión diferencial significativa, aunque mínima, ya podría estar influyendo en los datos de supervivencia del paciente. Todos los miRNAs diferencialmente expresados significativos estadísticamente son tomados en consideración para el análisis de supervivencia.

Además, algunos de los adenocarcinomas comparten datos de miRNAs diferencialmente expresados, como es el caso del hsa-miR-508, significativo (tanto estadísticamente como por logFC) y presente en los adenocarcinomas COAD, READ y PRAD.

El siguiente paso fue llevar a cabo el análisis de supervivencia. A partir de los miRNAs diferencialmente expresados obtenidos anteriormente, se empleó un método ROC para establecer un punto de corte que clasificara cada población de cada uno de los cánceres en dos grupos: uno con un nivel alto de miRNA (high level) y otro con un nivel bajo de miRNA (low level). Si el valor del punto de corte es significante, es decir, si no representa un desequilibrio en el número de pacientes de un grupo y otro, se lleva a cabo el análisis de supervivencia Kaplan-Meier. Tras el análisis de superviencia, se obtiene una firma de 23 miRNAs diferencialmente expresados que afectan a la supervivencia total de los pacientes (Log-rank method, $p<0.005$).

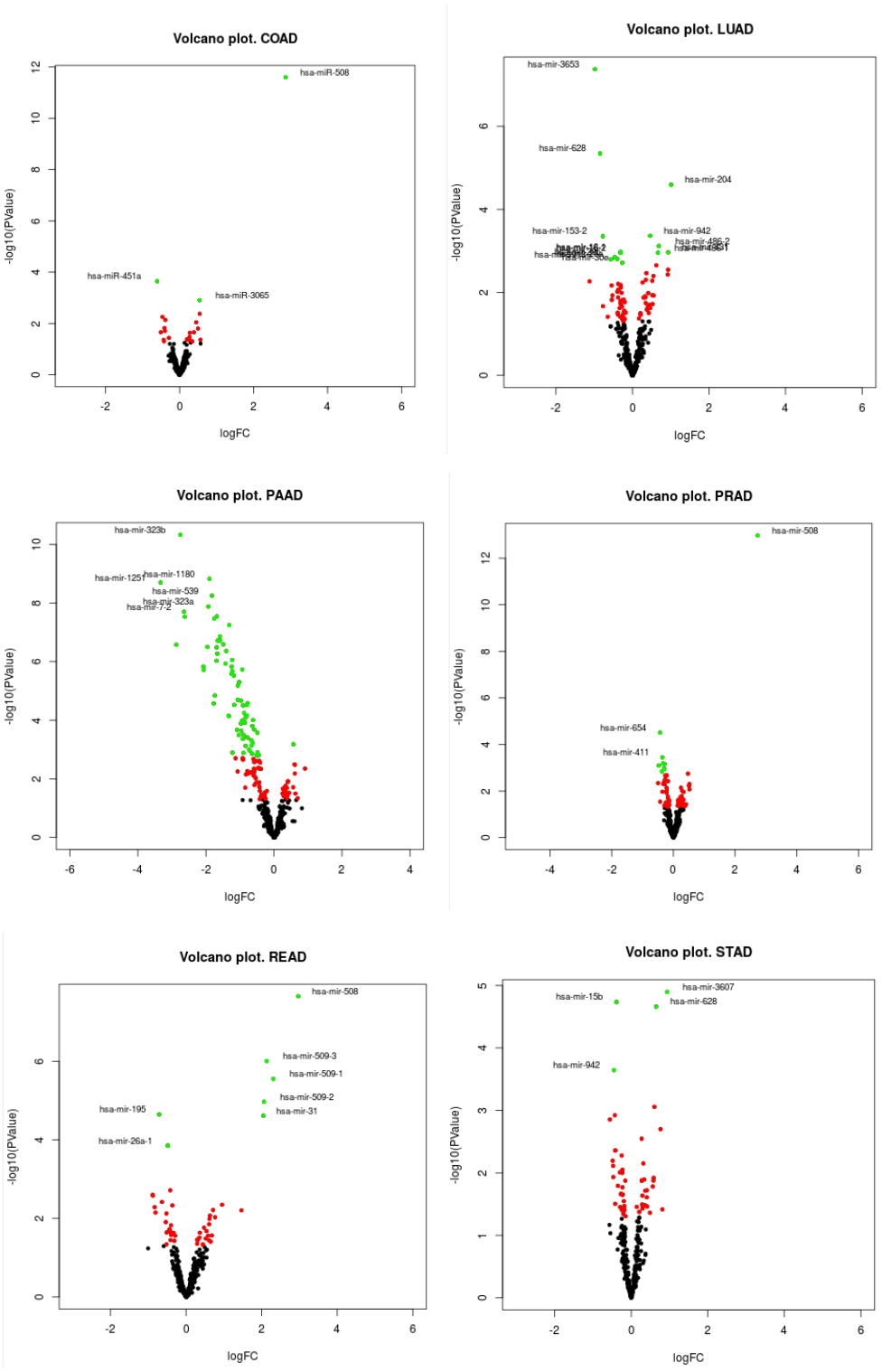

Figura 5: Volcano plots resultados del test de expresión diferencial por tipo de adenocarcinoma

En la Figura 6 se muestran los resultados para los miRNAs más significativos de cada tipo de cáncer. El resto de gráficas para cada uno de los miRNAs significantes se pueden observar en la Figura S2, en el apartado Material suplementario de esta memoria. De acuerdo con el resultado de la Figura 6, la alta expresión de hsa-miR-3653, hsa-miR-642a y hsa-miR-509-3 en sus respectivos cánceres podría ser un factor protectivo en la supervivencia de los pacientes, ya que el grupo de baja expresión muestra unos datos menores de tiempo y ratio de supervivencia. Por el contrario, en el caso del hsa-miR-3065 y hsa-miR-3607, es el grupo de baja expresión el que tiene unos índices de supervivencia más elevados, cosa que podría indicar que estos miRNAs están realizando su mecanismo de silenciamiento de genes sobre genes clave para el crecimiento celular o claves para detener el proceso tumoral. Como se observa, el estimador Kaplan-Meier para el cáncer PRAD no otorga unos resultados del todo fiables, ya que no aparece ningún evento de supervivencia en grupo high level. En la tabla 11 se relacionan los ratios de supervivencia a 5 años en cada uno de los miRNA representados.

miRNA	Cancer	high level	low level
hsa-miR-3065	COAD	55%	70%
hsa-miR-642a	PAAD	45%	20%
hsa-miR-3653	LUAD	60%	30%
hsa-miR-509-3	READ	70%	40%
hsa-miR-3607	STAD	25%	45%

Tabla 11: Ratios de supervivencia a 5 años para los miRNAs más significativos en cada tipo de cáncer

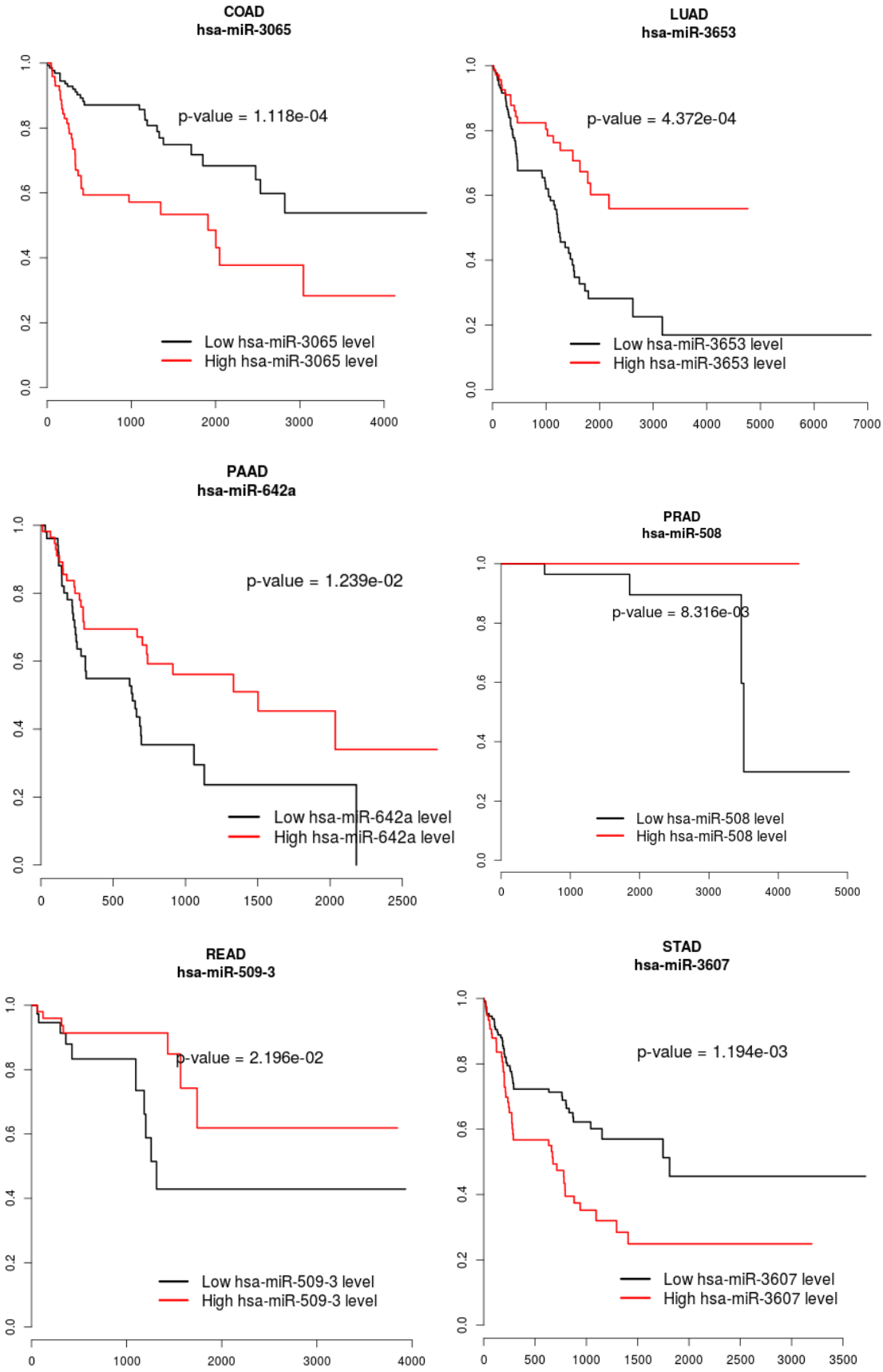

Figura 6: Estimador Kaplan-Meier basado en un punto de corte ROC divisor de los pacientes en dos grupos según el nivel de expresión del miRNA

Con el objetivo de explorar la función y el mecanismo biológico de la firma de expresión de miRNAs en cada tipo de adenocarcinoma, se analizan los genes diana potenciales que podrían estar regulados por los miRNAs. Los datos de genes diana fueron introducidos en la herramienta PANTHER de enriquecimiento de genes. Los resultados de esta herramienta no se muestran en la memoria, debido a la alta cantidad de información procesada y obtenida.

En la Tabla 12 se representan aquellas rutas metabólicas significativas en la herramienta Reactome que se encuentren diferencialmente expresadas en al menos 3 de nuestros cánceres y sus respectivos Pvalues y conteo de genes. En la figura S3 del material suplementario se encuentra la relación de genes dianas del panel de miRNA encontrados en cada una de las rutas.

Ruta Reactome	cancer	Conteos Genes	Pvalue
Oxidative Stress Induced Senescence	COAD	10	<0,01
	LUAD	33	<0,01
	PRAD	8	<0,01
	PAAD	24	<0,01
	STAD	23	<0,01
MET activates RAP1 and RAC1	COAD	3	<0,01
	PRAD	3	<0,01
	PAAD	5	<0,01
	STAD	4	<0,01
Regulation of TP53 Degradation	COAD	4	<0,01
	LUAD	16	<0,01
	PRAD	5	<0,01
Regulation of TP53 Expression and Degradation	COAD	4	<0,01
	LUAD	17	<0,01
	PRAD	5	<0,01
Major pathway of rRNA processing in the nucleolus and cytosol	COAD	15	<0,01
	LUAD	67	<0,01
	PRAD	11	<0,01
Cellular Senescence	LUAD	52	<0,01
	PAAD	35	<0,01
	STAD	31	<0,01
Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation	LUAD	25	<0,01
	PAAD	16	<0,01
	STAD	14	<0,01
Oncogene Induced Senescence	COAD	5	<0,01
	LUAD	16	<0,01
	PAAD	11	<0,01
	STAD	9	<0,01

Tabla 12: Rutas metabólicas implicadas en al menos 3 de los cánceres a estudio

Los resultados anteriores arrojan 8 rutas metabólicas enriquecidas en las muestras del estudio y en diversos cánceres a la vez. El nexo común entre las rutas es el proceso de senescencia, bien inducida por oncogenes o por un estrés oxidativo. Diversos oncogenes y genes supresores de tumores han sido probados como reguladores de esta senescencia, que limita la capacidad replicativa de las células previniendo la proliferación de células que se encuentran en diferentes estados de malignicidad (Dimri, 2005).

Por último, para estudiar la función de los miRNAs y su interacción con los genes, se sometió a los mismos pacientes a un test de expresión diferencial de transcritos. Una vez obtenidos los resultados, se combinó esta información con la información de los genes diana que se había obtenido en pasos anteriores,

en busca de posibles interacciones que pudieran ocurrir en el contexto patológico de cada adenocarcinoma estudiado. Los resultados se muestran en la tabla 13. En toal, detectamos 95 interacciones de 15 miRNAs identificados como afectantes de la supervivencia. Los miRNAs hsa-miR-508 y miR-3065 son los más representados y todos los tipos de cáncer incluyen alguna interacción excepto el READ.

Cancer	miRNA DE	log2FC	Gen diana DE	log2FC
COAD	Hsa-miR-3065	0,54	CXCL5	1,73
COAD	Hsa-miR-508	2,87	HIST1H2AG	-1,67
COAD	Hsa-miR-508	2,87	ZNF460	-1,76
COAD	Hsa-miR-508	2,87	HIST1H2AH	-3,52
LUAD	Hsa-miR-153-2	-0,77	THSD7A	1,34
LUAD	hsa-miR-16-1	-0,31	FABP7	-3,54
LUAD	hsa-miR-29a	-0,40	LAMC2	1,28
LUAD	hsa-miR-29a	-0,40	KREMEN2	1,94
LUAD	hsa-miR-29c	-0,46	IGFBP1	5,76
PRAD	hsa-miR-508	2,73	HMGA2	2,48
PRAD	hsa-miR-508	2,73	PPP1R15B	0,44
PRAD	hsa-miR-508	2,73	MYPN	-6,33
PRAD	hsa-miR-508	2,73	KLHDC8A	1,02
PRAD	hsa-miR-508	2,73	NLRP9	2,67
STAD	hsa-miR-15b	-0,38	FASN	-0,80
STAD	hsa-miR-15b	-0,38	TRAM1	0,47
STAD	hsa-miR-15b	-0,38	REXO1	-0,49
STAD	hsa-miR-15b	-0,38	CLSPN	-0,79
STAD	hsa-miR-15b	-0,38	BAMBI	-1,85
STAD	hsa-miR-15b	-0,38	SLC1A5	-0,89
STAD	hsa-miR-15b	-0,38	PIK3R1	0,59
STAD	hsa-miR-15b	-0,38	TMEM100	1,31
STAD	hsa-miR-15b	-0,38	XKR7	-3,83

STAD	hsa-miR-628	0,65	FOXE1	-3,36
STAD	hsa-miR-942	-0,45	RAB22A	0,79
STAD	hsa-miR-942	-0,45	TMPRSS11B	-5,04
STAD	hsa-miR-942	-0,45	TADA2A	0,63
PAAD	hsa-miR-1251	-3,33	MAVS	-0,29
PAAD	hsa-miR-1251	-3,33	FAM46A	-0,59
PAAD	hsa-miR-1251	-3,33	OCRL	-0,39
PAAD	hsa-miR-1251	-3,33	ASTN2	-1,68
PAAD	hsa-miR-1251	-3,33	ORAI2	-0,39
PAAD	hsa-miR-1251	-3,33	GABRB1	-1,24
PAAD	hsa-miR-1251	-3,33	CABP7	-1,97
PAAD	hsa-miR-1251	-3,33	CCDC65	-0,72
PAAD	hsa-miR-1251	-3,33	ADAMTS18	-1,17
PAAD	hsa-miR-1251	-3,33	MARCH4	-1,65
PAAD	hsa-miR-1251	-3,33	TMEM151B	-1,94
PAAD	hsa-miR-3065	-0,68	KCNN1	-3,07
PAAD	hsa-miR-3065	-0,68	TMEM74B	-1,57
PAAD	hsa-miR-3065	-0,68	ENPP2	-1,47
PAAD	hsa-miR-3065	-0,68	UNC119B	-0,58
PAAD	hsa-miR-3065	-0,68	RIMS2	-2,34
PAAD	hsa-miR-3065	-0,68	MYEF2	-0,69
PAAD	hsa-miR-3065	-0,68	SYT4	-2,62
PAAD	hsa-miR-3065	-0,68	RAB39B	-1,66
PAAD	hsa-miR-3065	-0,68	IRGQ	-0,32
PAAD	hsa-miR-3065	-0,68	CXCL10	1,07
PAAD	hsa-miR-3065	-0,68	DMRT2	-1,04
PAAD	hsa-miR-3065	-0,68	SOX1	-5,23
PAAD	hsa-miR-3065	-0,68	FFAR4	-1,12
PAAD	hsa-miR-3065	-0,68	PNMA2	-1,02
PAAD	hsa-miR-323b	-2,75	ZFP42	-2,37

PAAD	hsa-miR-487b	-1,68	SOBP	-0,71
PAAD	hsa-miR-487b	-1,68	MAGI2	-0,67
PAAD	hsa-miR-487b	-1,68	CDH7	-1,83
PAAD	hsa-miR-487b	-1,68	SERTAD3	0,43
PAAD	hsa-miR-487b	-1,68	ZDHHC22	-2,51
PAAD	hsa-miR-592	-1,69	NRL	-0,55
PAAD	hsa-miR-592	-1,69	PGPEP1	-0,38
PAAD	hsa-miR-592	-1,69	HS6ST3	-2,37
PAAD	hsa-miR-592	-1,69	PPP1R10	-0,27
PAAD	hsa-miR-642a	-1,76	MPP2	-1,73
PAAD	hsa-miR-642a	-1,76	SMYD1	-2,26
PAAD	hsa-miR-642a	-1,76	USP22	-0,34
PAAD	hsa-miR-642a	-1,76	CACNA1B	-3,18
PAAD	hsa-miR-642a	-1,76	SPTLC3	0,50
PAAD	hsa-miR-642a	-1,76	WDR37	-0,30
PAAD	hsa-miR-642a	-1,76	ATP9A	-0,57
PAAD	hsa-miR-642a	-1,76	ASIC4	-2,14
PAAD	hsa-miR-642a	-1,76	LHX5	-3,35
PAAD	hsa-miR-642a	-1,76	CRY2	-0,51
PAAD	hsa-miR-642a	-1,76	TMX4	-0,63
PAAD	hsa-miR-642a	-1,76	GNAZ	-1,60
PAAD	hsa-miR-642a	-1,76	CTSV	0,98
PAAD	hsa-miR-642a	-1,76	FRMD6	0,63
PAAD	hsa-miR-642a	-1,76	CDH8	-1,49
PAAD	hsa-miR-642a	-1,76	PANK1	-0,42
PAAD	hsa-miR-642a	-1,76	JPH3	-1,97
PAAD	hsa-miR-642a	-1,76	EMC10	-0,64
PAAD	hsa-miR-642a	-1,76	WNT4	-2,36
PAAD	hsa-miR-642a	-1,76	CLDN1	0,78
PAAD	hsa-miR-642a	-1,76	GPR37L1	-0,84

PAAD	hsa-miR-642a	-1,76	KSR2	-1,38
PAAD	hsa-miR-642a	-1,76	ST8SIA3	-1,29
PAAD	hsa-miR-642a	-1,76	GLIPR1L2	-1,08
PAAD	hsa-miR-642a	-1,76	SP9	-2,13
PAAD	hsa-miR-7-2	-2,64	PAK3	-1,75
PAAD	hsa-miR-7-2	-2,64	GALNT8	-3,16
PAAD	hsa-miR-7-2	-2,64	FGF2	0,85
PAAD	hsa-miR-7-2	-2,64	KCNB1	-2,27
PAAD	hsa-miR-7-2	-2,64	PCLO	-1,25
PAAD	hsa-miR-7-2	-2,64	KIAA1147	-0,35
PAAD	hsa-miR-7-2	-2,64	ZBTB8B	-1,41

Tabla 13: Interacciones entre miRNAs diferencialente expresados y genes diferencialmente expresados en la comparación

Una vez definida toda la información salida de nuestro análisis, el siguiente paso del análisis consiste en una integración de toda ella para un miRNA en concreto. Esta integración podría haberse realizado para cualquier miRNA, ya que en esencia, el análisis bioinformático es un paso preliminar a la validación in vitro de estas moléculas.

El miRNA analizado en este caso es hsa-miR-508, que forma parte del panel identificado con la supervivencia en los cánceres COAD y PRAD. La curva de supervivencia construidas para este miRNA en el cáncer COAD (Figura 7) nos otorga diferencias significativas entre un nivel alto y un nivel bajo de expresión de la molécula.

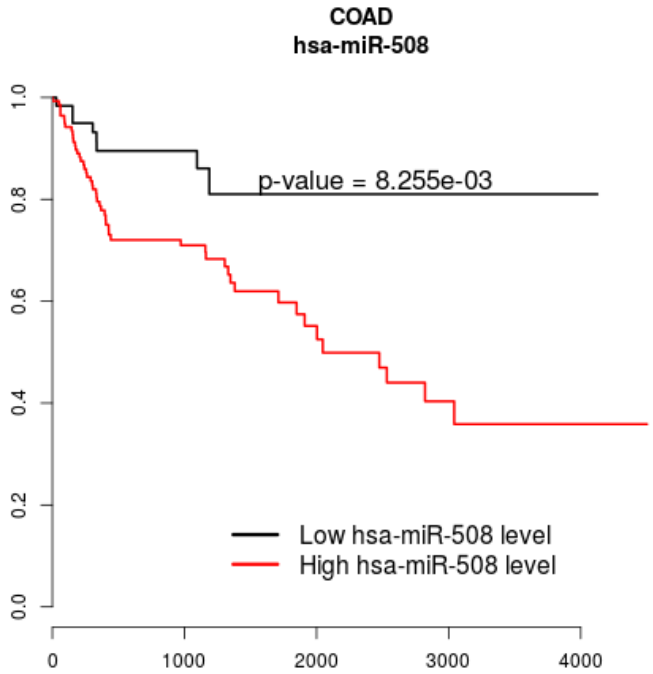

Figura 7: Curva de supervivencia del miRNA hsa-miR-508. Cáncer COAD.

Como se aprecia en la figura anterior, la alta expresión del miRNA provoca una disminución del tiempo y el ratio de supervivencia de los pacientes, por lo que cabe la posibilidad de que este miRNA esté realizando un silenciamiento de genes relacionados con la supervivencia. Concretamente, y observando los parámetros de interacción, en el contexto patológico del adenocarcinoma de colon coexiste la sobreexpresión del hsa-miR-508 (log2FC = 2,87) en pacientes de alta supervivencia y la infraexpresión del gen HIST1H2AH en este mismo grupo de pacientes (log2FC = -3.52). Además de este hecho, si observamos las curvas de supervivencia de este gen ya anotadas para el COAD a través de la herramienta The human Protein Atlas (http://www.proteinatlas.org/, Figura 8) observamos que los pacientes con un bajo contenido en este gen, como es nuestro caso, presentan unos índices de supervivencia menores. Por tanto, el silenciamiento de este gen puede estar provocando un déficit de la proteína que codifica relacionada directamente con la supervivencia. Por último, el gen HIST2H1AH es uno de los genes que se encuentran relativos a la ruta metábolica Oxidative Stress Induced Senescense, ruta que se encuentra en 5 de nuestros adenocarcinomas a estudio.

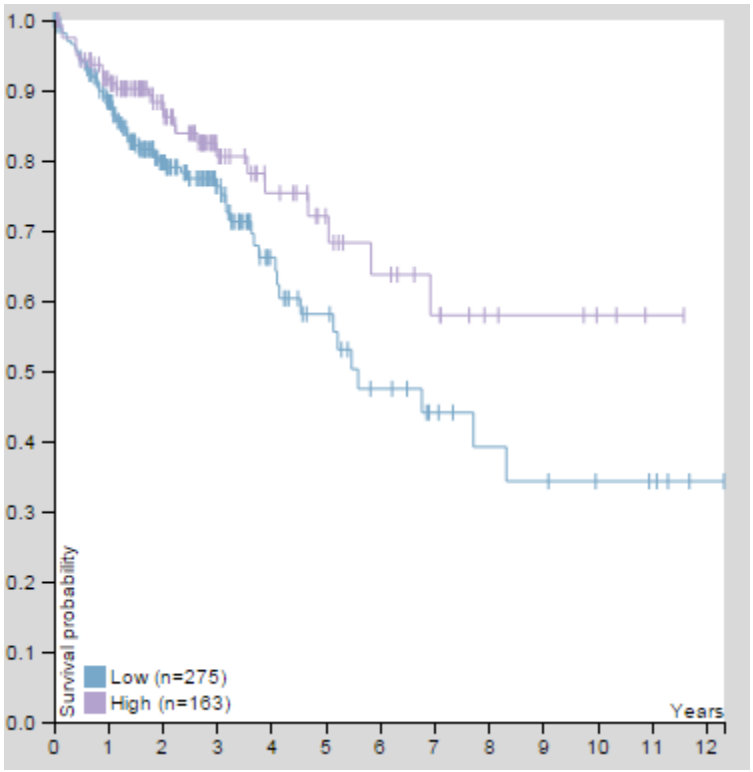

Figura 8: Tabla de supervivencia (The Human Protein Atlas) para HIST1H2AH

Tras estas observaciones, la teoría de que el hsa-miR-508 está altamente relacionado con la supervivencia se hace más robusta, pero serán necesarios más experimentos de validación experimentales para poder contrastar esta hipótesis.

5. CONCLUSIONES

- En la comparación estudiada, existe un panel de miRNAs que afectan a la supervivencia de los pacientes en un contexto patólogico de enfermedad.
- Algunos de los miRNA de este panel se encuentran en dos o más tipos de adenocarcinoma.
- Se pueden detectar de manera cualitativa interacciones entre miRNA sobrexpresados y mRNA infraexpresados en pacientes de alta supervivencia, estableciéndose así una hipótesis de que estos miRNA pueden estar ejerciendo su función y de que los genes pueden estar implicados en la respuesta tumoral.
- La senescencia celular tiene un efecto clave en la diferenciación de las poblaciones y podría también tenerlo en la respuesta tumoral.

En un futuro, se podría considerar como línea de trabajo la posible validación de las interacciones miRNA-mRNA, mediante correlaciones entre los perfiles individuales de expresión de ambas moléculas y se podría estudiar su actividad in vivo, mediante el uso y seguimiento de moléculas cancerosas conteniendo un miRNA estudiado.

En definitiva, estas aproximaciones al conocimiento de la relación entre moléculas de miRNA y distintas variables clínicas, como por ejemplo la supervivencia podrían suponer, como futuras líneas de trabajo, la determinación de marcadores de pronóstico. Por lo tanto, la determinación de los niveles de expresión de dichos miRNA podría constituir una herramienta para el diagnóstico y tratamiento personalizado de los pacientes.

6. BIBLIOGRAFÍA

American Cancer Society. Cancer Facts & Figures 2017. Atlanta: American Cancer Society; 2017.

Bullard, J.H., Purdomt, E., Hansen, K.D., Dudoit, S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments BMC Bioinformatics2010 11:94

Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, Rassenti L, Kipps T, Negrini M, Bullrich F, Croce CM. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. (2002) Proc Natl Acad Sci U S A.; 99(24):15524-9.

Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., … Noushmehr, H. (2016). TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Research, 44(8), e71.

Dimri, G. P. (2005). What has senescence got to do with cancer? Cancer Cell,7(6), 505–512.

Donoho, D. L. (2010). An invitation to reproducible computational research. Biostatistics, 11(3), 385-388.

Dweep, H., Gretz, N. miRWalk2.0: a comprehensive atlas of microRNA-target interactions (2015) Nature Methods, 12(8): 697-697.

Fabregat A., Jupe S., Matthews L., Sidiropoulos K., Gillespie M., Garapati P., Haw R., Jassal B., Korninger F., May B., Milacic M., Roca CD., Rothfels K., Sevilla C., Shamovsky V., Shorser S., Varusai T., Viteri G., Weiser J., Wu G., Stein L.,7, Hermjakob H., D'Eustachio P. The Reactome Pathway Knowledgebase. (2017) Nucleic Acids Res. gkx1132

Flynn, R. (2012). Survival analysis. Journal of clinical nursing, 21(19pt20), 2789-2797.

Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J, Tamborero D, Schroeder MP, Jene-Sanz A, Santos A & Lopez-Bigas N (2013) IntOGen-mutations identifies cancer Nature Methods, 10, 1081-1802.

Huang, H.-C., Niu, Y., & Qin, L.-X. (2015). Differential Expression Analysis for RNA-Seq: An Overview of Statistical Methods and Computational Software. Cancer Informatics, 14(Suppl 1), 57–67.

Ihaka, R., Gentleman, R. (1996). R: a language for data analysis and graphics. Journal of computational and graphical statics, 5 (3): 299-314.

Kaplan, E.L., Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American statistical association, 53(282), 457-481.

Kozomara A., Griffiths-Jones S.; miRBase: annotating high confidence microRNAs using deep sequencing data (2014) Nucleic Acids Research, Volume 42, Issue D1, D68–D73

MacFarlane, L.-A., & Murphy, P. R. (2010). MicroRNA: Biogenesis, Function and Role in Cancer. Current Genomics, 11(7), 537–561.

MAR-AGUILAR, F., RODRÍGUEZ-PADILLA, C., & RESÉNDEZ-PÉREZ, D. (2016). Web-based tools for microRNAs involved in human cancer. Oncology Letters, 11(6), 3563–3570.

Marioni, J. C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome research, 18(9), 1509-1517.

Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. (2013) Nat Protoc. 8(8):1551-66.

Racine, J. S. (2012). RStudio: A Platform-Independent IDE for R and Sweave. Journal of Applied Econometrics, 27(1), 167-172.

Reddy, K. B. (2015). MicroRNA (miRNA) in cancer. Cancer Cell International, 15, 38.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, 77.

Robinson, M.D., McCarthy, D.J., Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139-140.

Robinson, M.D., Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol, 11(3), R25.

Srinivasan S, Patric IRP, Somasundaram K (2011) A Ten-microRNA Expression Signature Predicts Survival in Glioblastoma. PLoS ONE 6(3): e17438.

Stahlhut Espinosa, C. E., & Slack, F. J. (2006). The Role of MicroRNAs in Cancer. The Yale Journal of Biology and Medicine, 79(3-4), 131–140.

Therneau, T. (2014): A Package for Survival Analysis in S, R package version 2.37-7.

Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology, 19(1A), A68–A77.

van Rossum, G.(2003) The Python Language Reference Manual. Network Theory Ltd.

Wall, L., Christiansen, T., Orwant, J. (2004). Programming perl. O'Reilly Media, Inc.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., & Li, T. (2009). miRecords: an integrated resource for microRNA–target interactions. Nucleic Acids Research, 37(Database issue), D105–D110.

Zhang J, Chong CCN, Chen GG, Lai PBS (2015) A Seven-microRNA Expression Signature Predicts Survival in Hepatocellular Carcinoma. PLoS ONE 10(6): e0128628

7. MATERIAL SUPLEMENTARIO

Cancer	miRNA	logFC	logCPM	PValue	FDR
COAD	hsa-mir-508	2,8652284	6,16574061	2,53E-12	8,11E-10
COAD	hsa-mir-451a	-0,60486601	8,25520277	0,00022504	0,03600668
COAD	hsa-mir-3065	0,54079611	3,81016677	0,00124276	0,13256053
LUAD	hsa-mir-3653	-0,98123608	2,85418485	4,23E-08	1,42E-05
LUAD	hsa-mir-628	-0,84751263	4,6476788	4,52E-06	0,00075951
LUAD	hsa-mir-204	1,01040728	2,57933461	2,56E-05	0,0028641
LUAD	hsa-mir-942	0,46188547	2,74237762	0,00043425	0,02996014
LUAD	hsa-mir-153-2	-0,77345238	4,40569928	0,00044584	0,02996014
LUAD	hsa-mir-486-2	0,68922321	6,23644748	0,00076767	0,03760859
LUAD	hsa-mir-16-1	-0,31169042	8,16627382	0,00105749	0,03760859
LUAD	hsa-mir-431	0,93354273	4,11813424	0,00108197	0,03760859
LUAD	hsa-mir-486-1	0,66672549	6,23391549	0,001108	0,03760859
LUAD	hsa-mir-16-2	-0,30928312	8,19188232	0,0011193	0,03760859
LUAD	hsa-mir-29c	-0,46370313	11,8475832	0,00143073	0,04183212
LUAD	hsa-mir-29a	-0,3972274	13,7401119	0,00156531	0,04183212
LUAD	hsa-mir-3913-1	-0,56389487	1,53052337	0,0016185	0,04183212
LUAD	hsa-mir-30e	-0,26533969	13,56389	0,00193544	0,04645055
PAAD	hsa-mir-323b	-2,7537747	4,9577784	4,69E-11	1,82E-08
PAAD	hsa-mir-1180	-1,89746805	4,98907384	1,49E-09	2,56E-07
PAAD	hsa-mir-1251	-3,33464928	3,1624596	2E-09	2,56E-07
PAAD	hsa-mir-539	-1,82382045	5,07116086	5,64E-09	5,47E-07
PAAD	hsa-mir-323a	-1,92896327	4,23599592	1,32E-08	1,02E-06
PAAD	hsa-mir-7-2	-2,64464642	4,06620026	1,99E-08	1,286E-06
PAAD	hsa-mir-487b	-1,67962793	5,06999434	2,86E-08	1,43E-06
PAAD	hsa-mir-7-3	-2,62360894	4,00866351	2,94E-08	1,43E-06
PAAD	hsa-mir-642a	-1,75578999	3,67633831	3,39E-08	1,461E-06
PAAD	hsa-mir-7-1	-1,3152569	5,0489225	5,67E-08	2,20E-06
PAAD	hsa-mir-431	-1,5890185	5,20505955	1,38E-07	4,88E-06
PAAD	hsa-mir-889	-1,5852639	6,60671885	1,90E-07	0,00000571
PAAD	hsa-mir-433	-1,6531423	2,74657228	1,91E-07	0,00000571
PAAD	hsa-mir-153-2	-1,48809172	6,06611015	2,58E-07	6,88E-06
PAAD	hsa-mir-137	-2,87238675	3,312768	2,66E-07	6,88E-06
PAAD	hsa-mir-153-1	-1,95749663	2,92600941	3,13E-07	7,45E-06
PAAD	hsa-mir-592	-1,6858461	1,78421336	3,27E-07	7,45E-06
PAAD	hsa-mir-369	-1,40215445	5,52513906	4,34E-07	9,35E-06
PAAD	hsa-mir-541	-1,66280133	1,4201338	5,34E-07	1,09E-05
PAAD	hsa-mir-487a	-1,22883437	2,13563149	8,81E-07	1,71E-05
PAAD	hsa-mir-3200	-1,68789396	1,57403833	9,38E-07	1,73E-05
PAAD	hsa-mir-410	-1,42475181	6,05017468	1,18E-06	2,08E-05
PAAD	hsa-mir-129-2	-2,07473689	6,6597948	1,47E-06	2,38E-05
PAAD	hsa-mir-485	-1,24474455	3,69632518	1,47E-06	2,38E-05
PAAD	hsa-mir-1301	-0,9302719	2,91348567	1,86E-06	2,86E-05

PAAD	hsa-mir-129-1	-2,06486219	6,500929	1,92E-06	2,86E-05
PAAD	hsa-mir-496	-1,22655861	3,13419945	2,09E-06	3,01E-05
PAAD	hsa-mir-95	-1,24487153	3,31789946	2,61E-06	3,61E-05
PAAD	hsa-mir-376a-1	-1,17894223	2,23949651	2,99E-06	4,00E-05
PAAD	hsa-mir-598	-1,02201863	5,04291912	5,019E-06	6,49E-05
PAAD	hsa-mir-409	-1,06119567	6,18727594	6,62E-06	8,29E-05
PAAD	hsa-mir-5683	-1,7382486	1,97941056	1,44E-05	0,00017422
PAAD	hsa-mir-543	-1,0668493	1,99378062	2,01E-05	0,00023687
PAAD	hsa-mir-381	-0,96945706	7,12879801	2,15E-05	0,00024544
PAAD	hsa-mir-328	-0,77535905	4,97609444	2,6131E-05	0,00028878
PAAD	hsa-mir-1224	-1,77071938	4,35600985	2,68E-05	0,00028878
PAAD	hsa-mir-376a-2	-1,16644698	1,2008593	3,01E-05	0,00030915
PAAD	hsa-mir-744	-0,79347202	4,80873984	3,03E-05	0,00030915
PAAD	hsa-mir-874	-0,90799413	5,75260686	3,13E-05	0,00031184
PAAD	hsa-mir-376c	-0,87481847	4,08733665	5,65E-05	0,00054786
PAAD	hsa-mir-495	-0,79514213	4,22257291	6,97E-05	0,00065161
PAAD	hsa-mir-412	-1,33618582	3,91848774	7,0535E-05	0,00065161
PAAD	hsa-mir-204	-1,32587954	5,00470966	7,33E-05	0,00066156
PAAD	hsa-mir-655	-0,83022074	2,69400006	8,09E-05	0,00071345
PAAD	hsa-mir-188	-0,6146966	1,632206	9,84E-05	0,00084871
PAAD	hsa-mir-370	-0,93313773	4,7601558	0,0001032	0,0008705
PAAD	hsa-mir-136	-0,85087379	6,87666749	0,00012056	0,00099523
PRAD	hsa-mir-508	2,73224784	3,75915627	1,06E-13	3,25E-11
PRAD	hsa-mir-654	-0,42859906	3,94748645	3,08E-05	0,00471338
PRAD	hsa-mir-411	-0,35620238	2,70175845	0,00036034	0,03675424
PRAD	hsa-mir-379	-0,33374059	8,76574345	0,00065417	0,04111269
PRAD	hsa-mir-370	-0,28838349	2,30899091	0,00068683	0,04111269
PRAD	hsa-mir-95	-0,47228796	0,79529485	0,00080613	0,04111269
PRAD	hsa-mir-134	-0,30135093	6,79989586	0,00104543	0,0445218
PRAD	hsa-mir-152	-0,29956019	7,82155575	0,00116397	0,0445218
PRAD	hsa-mir-299	-0,36693054	1,32257909	0,00145011	0,04930374
READ	hsa-mir-508	2,96904873	5,89992649	2,20E-08	7,99E-06
READ	hsa-mir-509-3	2,13440334	2,91325107	9,79E-07	0,00017778
READ	hsa-mir-509-1	2,30863309	2,78418748	2,78E-06	0,00033609
READ	hsa-mir-509-2	2,06274063	2,77853919	1,07E-05	0,00097141
READ	hsa-mir-195	-0,71330492	4,89682577	2,24E-05	0,00146219
READ	hsa-mir-31	2,04645406	4,85412771	2,42E-05	0,00146219
READ	hsa-mir-26a-1	-0,48234378	9,81146844	0,00013917	0,00641179
READ	hsa-mir-26a-2	-0,48636039	9,80629054	0,00014131	0,00641179
STAD	hsa-mir-3607	0,94028396	5,64069822	1,27E-05	0,00226918
STAD	hsa-mir-15b	-0,38378238	7,94817967	1,84E-05	0,00226918
STAD	hsa-mir-628	0,65097229	3,71399035	2,18E-05	0,00226918
STAD	hsa-mir-942	-0,44993375	3,56288277	0,00022657	0,01767273

Tabla S1: miRNAs significativos en el modelo de expresión diferencial.

Cancer	miRNA	log2FC	Survival Pvalue
COAD	hsa-miR-3065	0,54079612	1.12E-04
COAD	hsa-miR-508	2,8652284	8.26E-03
LUAD	hsa-miR-153-2	-0,77345238	1.36E-03
LUAD	hsa-miR-16-1	-0,31169042	2.48E-03
LUAD	hsa-miR-16-2	-0,30928312	2.63E-03
LUAD	hsa-miR-29a	-0,3972274	3.1E-02
LUAD	hsa-miR-29c	-0,46370313	1.57E-03
LUAD	hsa-miR-3653	-0,98123608	4.37E-04
LUAD	hsa-miR-628	-0,84751263	1.41E-02
PAAD	hsa-miR-1251	-3,33464928	2.70E-02
PAAD	hsa-miR-3065	-0,68492952	8.89E-04
PAAD	hsa-miR-323b	-2,7537747	1.72E-02
PAAD	hsa-miR-369	-1,40215445	4.98E-02
PAAD	hsa-miR-487b	-1,67962793	2.13E-02
PAAD	hsa-miR-539	-1,82382045	4.74E-02
PAAD	hsa-miR-592	-1,6858461	4.01E-03
PAAD	hsa-miR-642a	-1,75578999	1.24E-02
PAAD	hsa-miR-7-2	-2,64464642	1.28E-02
PAAD	hsa-miR-7-3	-2,62360894	4.38E-02
PRAD	hsa-miR-508	2,73224785	8.32-E03
READ	hsa-miR-509-3	2,13440334	2.2E-02
STAD	hsa-miR-15b	-0,38378238	2.92E-03
STAD	hsa-miR-3607	0,94028396	1.19E-03
STAD	hsa-miR-628	0,65097229	8.58E-04
STAD	hsa-miR-942	-0,44993375	1.48E-02

Tabla S2: miRNAs diferencialmente expresados implicados en la supervivencia general de los pacientes

Ruta Reactome	cancer	Genes
Oxidative Stress Induced Senescence	COAD	HIST1H2BN;CBX4;TFDP2;HIST1H2AH;MDM2;MDM4;FOS;TP53
	LUAD	HIST1H2BK;BMI1;E2F3;CBX6;JUN;CBX4;H3F3B;CDKN2A;CBX2;MINK1;HIST1H2AJ;H2AFX;FOS;PHC3;HIST2H2BE;MOV10;HIST1H4B;CDK6;CDK4;AGO4;AGO1;MAPKAPK2;MDM2;TP53;HIST1H2BC;TNRC6B
	PRAD	HIST1H2BN;CBX4;TFDP2;HIST1H2AH;MDM2;MDM4
	PAAD	SUZ12;CBX6;PHC2;MINK1;FOS;BMI1;MAPK14;TNRC6C;MAPK9;TFDP1;AGO3;AGO1;E2F2;MDM4;HIST2H3D;HIST1H3B;TP53;HIST1H3D;TNRC6A;MAP3K5;TNRC6B
	STAD	SUZ12;MAP2K4;CBX6;CBX4;CBX2;MINK1;PHC3;HIST2H2BE;RNF2;HIST2H3A;CDK6;TFDP2;CDK4;AGO4;AGO1;MAPKAPK2;MAPK1;E2F3;TNRC6B
MET activates RAP1 and RAC1	COAD	DOCK7;RAPGEF1;CRKL
	PRAD	DOCK7;RAPGEF1;CRKL
	PAAD	RAP1B;GRB2;RAC1;CRKL
	STAD	RAPGEF1;GRB2;CRK;CRKL
Regulation of TP53 Degradation	COAD	CCNG1;MDM2;MDM4;TP53
	LUAD	USP7;CDKN2A;PPP2R5C;MTOR;CCNA2;PPP2R1B;PPP2R1A;AKT2;CDK2;AKT3;MDM2;CDK1;RICTOR;SGK1;TP53
	PRAD	CCNG1;MDM2;MDM4
Regulation of TP53 Expression and Degradation	COAD	CCNG1;MDM2;MDM4;TP53
	LUAD	USP7;CDKN2A;PPP2R5C;MTOR;CCNA2;PPP2R1B;PPP2R1A;AKT2;CDK2;AKT3;MDM2;CDK1;RICTOR;SGK1;TP53
	PRAD	CCNG1;MDM2;MDM4
Major pathway of rRNA processing in the nucleolus and cytosol	COAD	RBM28;HMX3;UTP6;RPL12;FCF1;NOL9;WDR12;RPS15A;CIRH1A;RPS19;RPL27A;BMS1;RPL37;EXOSC2;RPS24
	LUAD	RPL4;RBM28;RPL5;RPL30;RPL3;RPL31;RPLP1;WDR3;RPL10L;RPLP0;FCF1;RPL9;WDR43;RPL6;RRP9;RPS4X;RPS17;NOB1;RPL36;UTP14A;RBM6;RPS12;UTP15;RPS9;RPL21;RPS5;RPS6;RPL22;WDR75;RPS3A;RPSA;CSNK1E;DDX52;DIEXF;CIRH1A;SIK1;RPL10;RPL12;HEATR1;ISG20L2;DDX21;NOC4L;BMS1;RPL14;DHX37;RPS3;RPL13;RRP36;UTP20;RPS2;DCAF13;NOP14;UTP3;RIOK3;WDR18;BYSL;RCL1;RPS25;TBL3;RPS27;RPL27A;TSR1;NOL11;RPL17-C18orf32;RPS24
	PRAD	RBM28;HMX3;UTP6;CIRH1A;RPS19;RPL27A;BMS1;RPL12;NOL9;WDR12;EXOSC2
Cellular Senescence	LUAD	CDKN1A;CDKN1B;ANAPC16;HIST1H2BK;BMI1;CABIN1;RPS6KA3;CDC23;CDC27;E2F3;HIST1H1C;CBX6;JUN;CBX4;H3F3B;CDKN2A;UBE2C;CBX2;MINK1;HIST1H2AJ;H2AFX;HMGA1;HMGA2;FOS;PHC3;NFKB1;HIST2H2BE;CCNA2;MOV10;CDK6;HIST1H4B;CCNE2;SP1;CDK4;AGO4;CCNE1;AGO1;CDK2;MAPKAPK2;MDM2;TERF2IP;TP53;TNRC6B;HIST1H2BC
	PAAD	BMI1;RELA;LMNB1;MAPK9;MAPK7;CDC27;E2F2;HIST1H3B;HIST1H3D;MAP3K5;SUZ12;CBX6;PHC2;MINK1;MRE11A;FOS;MAPK14;TNRC6C;H1F0;RAD50;TFDP1;AGO3;SP1;CCNE1;AGO1;TERF2IP;MDM4;TPP1;HIST2H3D;TP53;TNRC6A;TNRC6B

	STAD	CDKN1A;RNF2;RPS6KA3;CDC27;MAPK1;E2F3;SUZ12;MAP2K4;CBX6;CBX4;CBX2;MINK1;HMGA1;TERF2;PHC3;HIST2H2BE;HIST2H3A;CDK6;TFDP2;CCNE2;CCNE1;CDK4;AGO4;AGO1;MAPKAPK2;TNRC6B
Gene and protein expression by JAK-STAT signaling after Interleukin-12 stimulation	LUAD	HSPA9;ARF1;RPLP0;FOXO3;CDC42;HNRNPDL;IFNG;HNRNPF;HNRNPA2B1;TCP1;PDCD4;SNRPA1;PAK2
	PAAD	RAP1B;RALA;CAPZA1;HNRNPF;HNRNPA2B1;SOD2;PAK2;LMNB1
	STAD	IFNG;HNRNPDL;HNRNPA2B1;RPLP0;TCP1;PDCD4;PAK2
Oncogene Induced Senescence	COAD	TFDP2;SP1;MDM2;MDM4;TP53
	LUAD	MOV10;CDK6;CDKN2A;CDK4;AGO4;SP1;AGO1;MDM2;E2F3;TP53;TNRC6B
	PAAD	TNRC6C;TFDP1;AGO3;SP1;AGO1;E2F2;MDM4;TP53;TNRC6A;TNRC6B
	STAD	CDK6;TFDP2;CDK4;AGO4;AGO1;MAPK1;E2F3;TNRC6B

Tabla S3: Relación de genes encontrados en las rutas metabólicas significantes