Identificación de vías moleculares alteradas por mutaciones en KRAS y/o supervivencia en cáncer de colon

Lara Rodríguez Outeiriño
Máster en Bioinformática y Bioestadística
TFM - Área 32: Análisis de datos cómicos

Jeroni Luna Cornadó
David Merino Arranz

05 de junio de 2018

Licencias alternativas (elegir alguna de las siguientes y sustituir la de la página anterior)

A) Creative Commons:

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

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

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

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

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

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

B) GNU Free Documentation License (GNU FDL)

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	*"Identificación de vías moleculares alteradas por mutaciones en kras y/o supervivencia en cáncer de colon"*
Nombre del autor:	*Lara Rodríguez Outeiriño*
Nombre del consultor/a:	*Jeroni Luna Cornadó*
Nombre del PRA:	*David Merino Arranz*
Fecha de entrega (mm/aaaa):	06/2018
Titulación::	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*TFM - Área 32: Análisis de datos cómicos*
Idioma del trabajo:	*Castellano*
Palabras clave	*Cáncer de colon adenocarcinoma (COAD), K-RAS, Micro-ARNs, supervivencia*

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

Finalidad: En este trabajo se muestra el estudio comparativo de expresión diferencial de microRNAs (miRs) y ARN mensajeros (ARNm) en pacientes con adenocarcinoma de colon (COAD) subdividiendo a la muestra por dos criterios: estado mutacional de KRAS y baja o alta supervivencia. También se muestran las interacciones entre miRs sobreexpresados y ARNm infraexpresados en pacientes con COAD indicando genes que puedan ejercer alguna función por alteraciones en KRAS o en supervivencia. Los resultados servirán para encontrar posibles predictores moleculares que ayuden a determinar la relación del tumor con mutaciones en KRAS y la tasa de supervivencia.

Metodología: Obtenemos la información de los conteos de genes del repositorio The Genome Cancer Atlas (TCGA), así como los datos clínicos. Trabajamos con el lenguaje de programación R en RStudio, sirviéndonos del repositorio Bioconductor para los análisis bioestadísticos. Para el análisis comparativo de expresión diferencial, búsqueda de interacciones miRs-ARNm y análisis de enriquecimiento GO trabajamos con el paquete miRComb.

Resultados: Obtenemos los diferentes perfiles de expresión génica en las muestras de estudio, así como la funcionalidad GO de estos genes. Encontramos interacciones significativas de 6 parálogos del miR hsa-let-7 con varios ARNm en la comparación entre pacientes con alta y baja supervivencia, pero no se encuentra ningún significado biológico de dicha interacción.

Conclusiones: Esta memoria presenta resultados preliminares de las comparaciones entre pacientes con COAD de los perfiles de expresión génica, así como de enriquecimiento biológico de los genes diferencialmente expresados. Finalmente se proponen alternativas a los análisis computados para refinar los resultados obtenidos.

Abstract (in English, 250 words or less):

Purpose: In this study we show a comparative study of miRs's and mRNA's differentially expressed genes in patients with colon adenocarcinome (COAD), subdividing the sample by two criteria: KRAS mutational status and survival's rate. The interactions between overexpressed miRs and underexpressed mRNA in the disease are also shown. This can indicate genes that may play a role in COAD. Results can be used to find possible molecular predictors that would help us determine the relationship of tumours with KRAS mutation and/or survival rate.

Methods: Using The Genome Cancer Atlas (TCGA) genomic information, miRs and mRNA counts were extracted, as well as clinical data. For computing bio statistical analysis, we use Bioconductor repository which use R statistical programming language in RStudio. Differential expression comparative searching for miRs-mRNA interaction and GO enrichment analysis was computed with the package miRComb.

Results: Gene expression profiles were obtained by characterization of different samples' groups, as well as GO functional information. Six `has-let-7` miR paralogous have significant interactions with a few mRNA in survival analysis. However, this interaction doesn't show biological signification.

Conclusion: This report presents preliminary results of compared COAD gene expression profiles. In addition, we show biological enrichment of differentially expressed genes. Finally, we propose alternatives analysis to improve the results.

Índice

Lista de figuras

Gráficas de la memoria

Tablas de la memoria

1. Introducción

1.1 Contexto y justificación del Trabajo

El cáncer es un conjunto de enfermedades que se caracterizan por la división descontrolada y anormal de las células. Es la segunda causa de muerte tras las enfermedades cardiovasculares. El origen se debe tanto a factores externos (estilo de vida, alimentación, actividad física) como internos (condicionantes genéticos, hormonas o patologías inmunes) (American Cancer Society, 2017). Estos factores de riesgo pueden iniciar o favorecer el crecimiento del cáncer, así mismo, estos condicionantes son claves para la elección de un correcto tratamiento personalizado (American Cancer Society, 2017).

En este estudio nos centraremos en el cáncer colorectal, el tercer tipo de cáncer más común en hombres y mujeres en EEUU, causa de un 9% de los fallecimientos totales por cáncer (American Cancer Society, 2017; Muzny et al., 2012). Este tipo engloba a aquellos cáncer que se desarrollan en la parte final del aparato digestivo, el colon o el recto. Diversos estudios engloban al cáncer de colon y de recto cómo un único tipo de cáncer a escala genómica, por expresión de exosomas, por número de copias de DNA, por la metilación de promotores, por la expresión de ARNm y de miRs que sólo se distinguen por la localización anatómica (American Cancer Society, 2017; Muzny et al., 2012). En concreto, el tipo más común son los COAD con un 90% de los casos (Society, 2012). Este tipo de cancer se originan de las células glandulares intestinales que producen la mucosidad para lubricar el interior del colon y recto (American Cancer Society, 2017; Muzny et al., 2012).

En la mayoría de cánceres se encuentra sobreexpresado el factor de crecimiento epidérmico (EGFR), su activación favorece las rutas metabólicas implicadas en proliferación, angiogénesis, migración y adhesión celular (Tan & Du, 2012). EGFR está sobreexpresado en el 80% de los casos de cáncer colorectal, siendo una diana candidata para bloquear las células tumorales (Tan & Du, 2012). Por tanto, para el tratamiento del cáncer colorrectal se utiliza principalmente dos anticuerpos de unión al EGFR: cetuximab (Erbitux) o panitumumab (Vectibix) ((American Cancer Society, 2017; Muzny et al., 2012). Sin embargo, las mutaciones en esta vía de señalización suponen un problema en la eficacia de las terapias. Si observamos las diferentes mutaciones de las proteínas en la cascada de señalización de EGFR encontramos que una proteína, K-RAS aparece mutada en el 30-45% de los casos de cáncer colorrectal (Knickelbein & Zhang, 2015; Muzny et al., 2012; Roa, Sanchez, Majlis, & Schalper, 2013; Tan & Du, 2012).

KRAS es una proteína de unión al GTP/GDP anclada a la membrana, participa en la transducción de señales y es el principal responsable de la activación de la vía EGFR. Se han encontrado hasta 5000 tipos de mutaciones diferentes en el oncogen KRAS que dejan a la proteína en un modo de activación constante, promoviendo diversas rutas pro-proliferativas de las células tumorales (Tan & Du, 2012). Por tanto, el uso de terapias que bloqueen la vía EGFR en paciente con mutaciones KRAS son ineficaces y pueden llegar a ser tóxicas para el individuo (Tan & Du, 2012). En estos pacientes, la mutación en KRAS también se relacionan con un peor pronóstico de la enfermedad (Tortola, Silvia., Marcuello, Eugenio *et al,* 1999).

Surge la necesidad de nuevas dianas terapéuticas sobre las que actuar en pacientes con mutación en KRAS. Para elucidar nuevas dianas podemos realizar análisis diferenciales entre pacientes con mutación *versus* sin mutación en KRAS. En este contexto aparecen los miRs, pequeñas moléculas (19-25 nucleótidos) de ARN no codificante capaces de unirse a las regiones 3' no traducidas (3'UTR) de los ARNm interfiriendo así en su estabilidad y/o traducción (Jonas & Izaurralde, 2015).

Constituyen, por tanto, una nueva vía de control de la expresión génica a nivel post-transcripcional.

En este trabajo trataremos de identificar expresiones diferenciales de estos miRs entre los dos grupos de pacientes. Para ello nos valdremos de un nuevo paquete de Bioconductor, 'miRComb', capaz de integrar la información de miRs desregulados con dianas en ARNm desregulados y generar un informe con los resultados obtenidos (Vila-Casadesús, Gironella, & Lozano, 2016). A su vez, analizaremos como se encuentra distribuida la supervivencia de los pacientes según el estado mutacional de KRAS y analizaremos posibles rutas diferencialmente expresadas según el pronóstico de supervivencia. Los datos obtenidos servirán para crear perfiles de expresión de miRs y de ARNms dianas entre los dos grupos oncológicos. El análisis también servirá de base para nuevas líneas de investigación que usen miRs como terapias para el cáncer colorrectal.

1.2 Objetivos del Trabajo

Objetivos generales

1- Identificar miRs desregulados según el estado mutacional de KRAS en adenocarcinoma de colon.

2- Identificar ARNm desregulados según el estado mutacional de KRAS en adenocarcinoma de colon.

3- Integrar la información obtenida de miRs sobreexpresados con diana en ARNm infraexpresados según el estado mutacional de KRAS en adenocarcinoma de colon mediante el uso del paquete de Bioconductor '*miRComb*'.

4- Caracterización de la supervivencia en los pacientes con adenocarcinoma de colon según el estado mutacional de KRAS.

5- Identificar miRs sobreexpresados con diana en ARNm infraexpresados según supervivencia alta o baja en pacientes con adenocarcinoma de colon mediante el uso del paquete de Bioconductor '*miRComb*'.

6- Análisis e integración de los resultados obtenidos.

Objetivos específicos

1.1. Selección y descarga de datos de expresión de miRs en adenocarcinoma de colon con la librería TCGAbiolinks.
1.2. Crear una matriz de datos de expresión de miRs en adenocarcinoma de colon.

2.1. Selección y descarga de datos de expresión de ARNm en adenocarcinoma de colon con la librería TCGAbiolinks.
2.2. Crear una matriz de datos de expresión de ARNm en adenocarcinoma de colon.

3.1. Creación de una matriz fenotípica de los pacientes del estudio según el estado mutacional de KRAS
3.2. Integración de la expresión de ARNm y miRs según el estado mutacional de KRAS mediante el uso de miRComb.
3.3. Seleccionar las interacciones miRs-ARNm con correlación negativa y significativamente desregulados en adenocarcinoma de colon. Análisis de enriquecimiento biológico.

4.1. Análisis de la supervivencia en los pacientes con COAD según el estado mutacional de KRAS

5.1. Creación de una matriz fenotípica de los pacientes del estudio según la supervivencia, alta o baja.
5.2. Integración de la expresión de ARNm y miRs según la supervivencia mediante el uso de miRComb.
5.3. Seleccionar las interacciones miRs-ARNm con correlación negativa y significativamente desregulados en adenocarcinoma de colon. Análisis de enriquecimiento biológico.

6.1. Integración de toda la información obtenida.
6.2. Análisis de los resultados.

1.3 Enfoque y método seguido

El enfoque en este trabajo se ha orientado a la búsqueda de vías moleculares alteradas por mutación de KRAS o por supervivencia en COAD. Para alcanzar los objetivos marcados, proponemos la siguiente metodología a seguir:

Revisión bibliográfica sobre la temática escogida. Para comenzar el estudio es esencial tener conocimientos básicos sobre el cáncer de colon y saber cual es la novedad que aporta nuestro proyecto al campo de la biomedicina.

Extraer los datos del repositorio de TCGA (The Cancer Genome Atlas). Se elegirán aquellos pacientes que dispongan de datos de expresión del ARNm así como de miRs. Los datos se pre-procesarán y normalizarán para los pasos posteriores.

Uso del lenguaje de programación R y de RStudio como entorno de trabajo. Usaremos la versión: “Version 1.1.453 – © 2009-2018 RStudio, Inc.” Este entorno de trabajo permite la reproducibilidad de la investigación al tratarse de un software libre y de código abierto. A parte, R funciona en cualquier plataforma, permite cargar y crear diversos paquetes y funciones para compartir con la comunidad científica.

Usaremos **Bioconductor**, un repositorio de paquetes de R enfocado al análisis de datos biológicos. “Existen diferentes paquetes en R/Bioconductor capaces de elucidar interacciones de miRs-ARNm como RmiR, CORNA, miRNApath, microRNA o MultiMiR, pero ninguno de ellos permite realizar un análisis completo como el planteado en este estudio” (Vila-Casadesús et al., 2016). En su lugar usaremos un paquete de R diseñado para tal fin, **miRComb**. Este paquete nos va a permitir combinar los datos de expresión de ARNm y miRs con la información de hibridaciones de bases de datos ya existentes para descubrir dianas putativas de miRs-ARNm en una enfermedad concreta (Vila-Casadesús et al., 2016).

El *pipeline* del trabajo será semejante al del trabajo realizado en el artículo de María Vila-Casadesús y colaboradores mostrado en la siguiente ilustración (Vila-Casadesús et al., 2016):

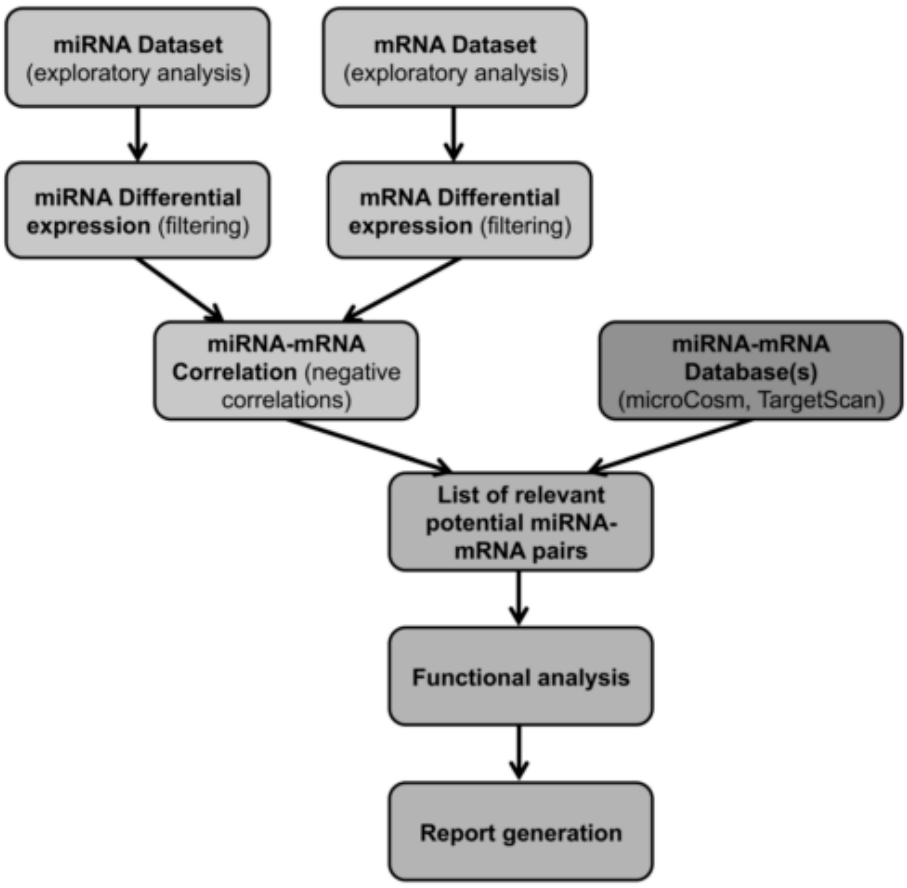


Fig 1. Flow diagram showing the main steps of an analysis using the *miRComb* package.

doi:10.1371/journal.pone.0151127.g001

Se obtendrán dos matrices de datos de expresión de miRs y de ARNm en pacientes con COAD. Sobre esta base de datos se realizará un análisis de expresión diferencial dividiendo las bases de datos en dos grupos según el estado mutacional de la proteína KRAS en los pacientes.

Posteriormente integraremos la información obtenida de las posibles dianas en los ARNm de los miRs gracias a diferentes bases de datos que usan diferentes algoritmos para predecir estas uniones. En este proyecto usaremos la base de datos MicroScan por usar el algoritmo miRanda, el cual necesita que la complementariedad de la secuencia diana sea perfectamente complementaria en el extremo 5' del miRs, esencial para una conformación estable; esta base de datos sólo acepta como dianas válidas sitios altamente conservados entre especies, indicativo de que se trata de una unión funcional a lo largo de la evolución. Por otra parte usaremos TargetScan, una de las bases de datos más actualizadas y extensas. TargetScan se basa en la complementariedad de la secuencia y diferencia entre sitios altamente conservados o no. MiRComb permite usar todas las bases de datos que queramos y cambiar la configuración de selección de las mismas, de este modo, para asemejar ambas bases de datos, sólo seleccionaremos los sitios altamente conservados de la base de datos TargetScan (Vila-Casadesús et al., 2016).

Si los resultados obtenidos no son satisfactorios se analizará la distribución y dependencia de la variable supervivencia de los pacientes respecto al estado mutacional

de KRAS. Por último se valorarán las rutas metabólicas miRs-ARNm diferencialmente expresadas en pacientes de COAD según la supervivencia sea alta o baja.

1.4 Planificación del Trabajo

Objetivo 1: Identificar miRs desregulados según el estado mutacional de KRAS en adenocarcinoma de colon.

1.1.1. Instalación de la librería TCGAbiolinks en el entorno de Rstudio.
1.1.2. Selección de los estudios en pacientes con COAD en el repositorio TCGA. Los estudios deberán tener datos disponibles sobre expresión de miRs y ARNm.
1.1.3. Obtención de los datos de expresión de miRs de pacientes con COAD mediante la librería TCGAbiolinks.
1.1.4. Procesado de los datos: preprocesado, filtrado y normalización.

Objetivo 2: Identificar ARNm desregulados según el estado mutacional de KRAS en adenocarcinoma de colon.

2.1.1. Obtención de los datos de expresión de ARNm de pacientes con COAD mediante la librería TCGAbiolinks.
2.1.2. Procesado de los datos: preprocesado, filtrado y normalización.

Objetivo 3: Integrar la información obtenida de miRs sobreexpresados con diana en ARNm infraexpresados según el estado mutacional de KRAS en adenocarcinoma de colon mediante el uso del paquete de Bioconductor 'miRComb'.

3.1. Creación de la matriz fenotípica según el estado mutacional de KRAS.

3.2.1. Para cada grupo de pacientes integrar la información con miRComb de miRs y ARNm expresados diferencialmente obtenidos en los objetivos 1 y 2, con las posibles dianas putativas de MicroCosm y TargetScan.
3.2.2. Para cada grupo de pacientes realizar un análisis de correlación negativa de las interacciones miRs-ARNm seleccionadas con miRComb.

3.3.1. Análisis estadístico comparativo de las interacciones miRs-ARNm dependiendo del estado mutacional de KRAS en COAD.

3.3.2. Representación de los resultados, tablas, graficas, informe. Análisis de enriquecimiento biológico.

Objetivo 4: Caracterización de la supervivencia en los pacientes con adenocarcinoma de colon según el estado mutacional de KRAS.

4.1. Obtención de los datos clínicos de supervivencia en la base de datos TCGA de pacientes con COAD.

4.2. Análisis de correlación de la supervivencia según el estado mutacional de KRAS.

Objetivo 5: Identificar miRs sobreexpresados con diana en ARNm infraexpresados según supervivencia alta o baja en pacientes con adenocarcinoma de colon mediante el uso del paquete de Bioconductor 'miRComb'.

5.1. Creación de la matriz fenotípica según el supervivencia alta o baja y ajuste de las nuevas matrices de datos de expresión de miRs y ARNm.

5.2.1. Para cada grupo de pacientes integrar la información con miRComb de miRs y ARNm expresados diferencialmente obtenidos en los objetivos 1 y 2, con las posibles dianas putativas de MicroCosm y TargetScan.

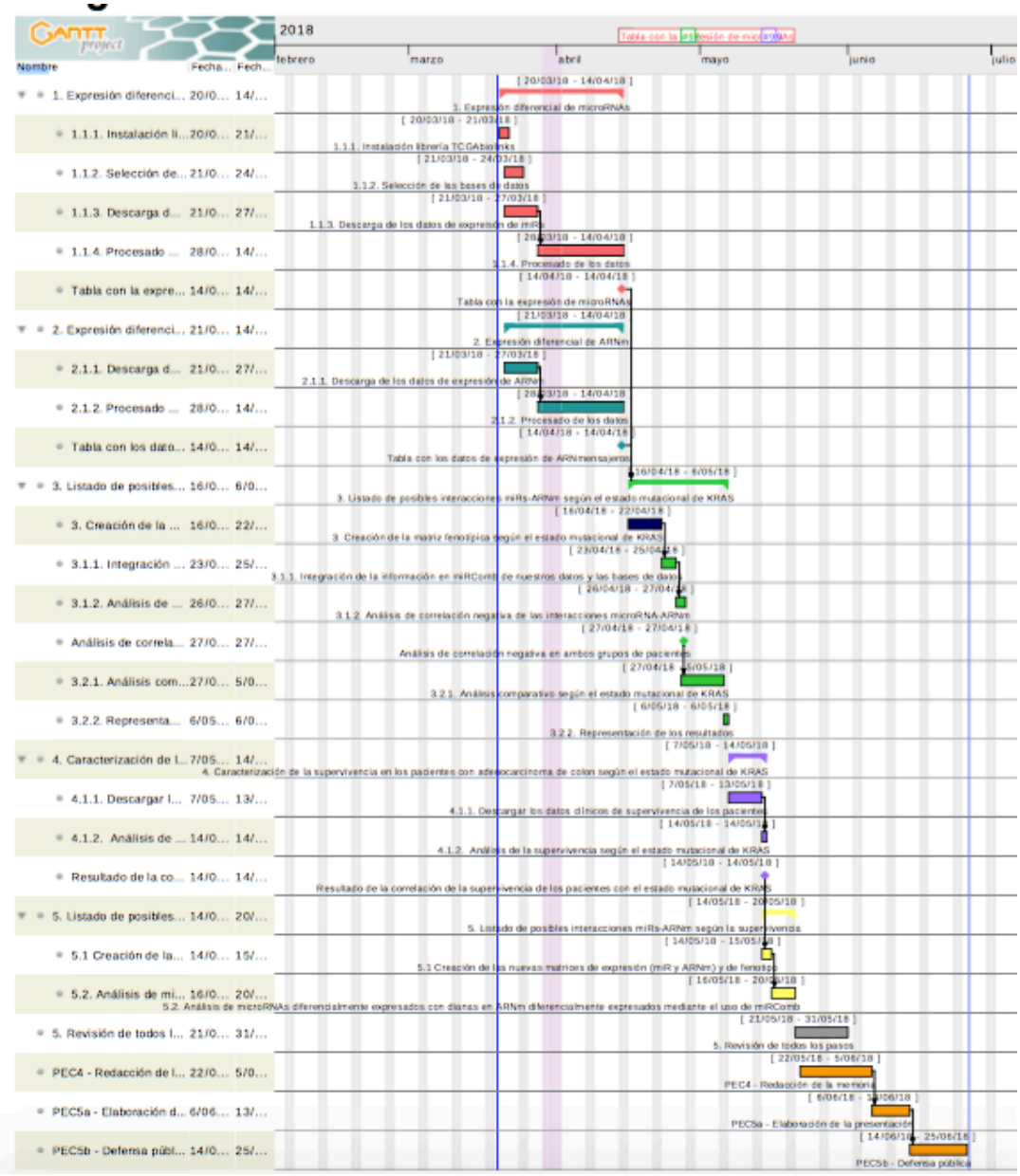
5.2.2. Para cada grupo de pacientes realizar un análisis de correlación negativa de las interacciones miRs-ARNm seleccionadas con miRComb.
5.3.1. Análisis estadístico comparativo de las interacciones miRs-ARNm dependiendo de la supervivencia de los pacientes con COAD.
5.3.2. Representación de los resultados, tablas, graficas, informe. Análisis de enriquecimiento biológico.

Objetivo 6: Análisis e integración de los resultados obtenidos.

6.1. Integración de los datos obtenidos
6.2. Justificación y reflexión de los resultados

Calendario

Para la realización del calendario nos hemos ayudado del software gratuito GanttProyect, que permite organizar las tareas y marcar los hitos necesarios para la realización del proyecto en un diagrama de Grantt:



Hitos

Durante la realización del proyecto existen varios hitos claves a lo largo del calendario:

- **Creación de la tabla con los niveles de expresión de miRs**. Para poder continuar el trabajo será esencial disponer de una tabla de datos de miRs sobre los que trabajar. En esta tabla figurarán los datos seleccionados, pre-procesados y filtrados.

- **Creación de tabla con los niveles de expresión de ARNm.** Al igual que el caso anterior, será esencial una tabla con los datos procesados de la expresión de ARNm.

- **Análisis de correlación negativa miRs-ARNm en cada grupo de pacientes según el estado mutacional de KRAS**. Esta información es crucial para seleccionar los genes e interacciones con miRs para los resultados finales. Estos datos serán integrados con las bases de datos de TargetScan y MicroCosm para continuar el análisis.

- **Análisis de correlación negativa miRs-ARNm en cada grupo de pacientes según la supervivencia.** Tarea esencial para analizar las rutas diferencialmente desreguladas en pacientes con una supervivencia alta o baja.

- **Análisis de enriquecimiento biológico.** Obtención de las funciones que describen a los genes diferencialmente expresados según el *gene ontology* (GO) en las comparaciones de muestras según el estado mutacional de KRAS y los niveles de supervivencia.

1.5 Breve sumario de productos obtenidos

Estos resultados se muestran parcialmente en la memoria como tablas para facilitar la lectura de la misma.

- Matriz de los niveles de expresión de miRs y ARNm filtrados y ajustados en pacientes con COAD para el análisis según el estado mutacional de KRAS. Archivos 'Expr-miRs-KRAS.txt' y 'Expr-ARNm-KRAS.txt' respectivamente.

- MiRs y ARNm diferencialmente expresados especificando FC y p-valores según el estado mutacional de KRAS. Archivos 'DifExpr-miRs-KRAS.txt' y 'DifExpr-ARNm-KRAS.txt' respectivamente.

- Análisis de la supervivencia en los pacientes de COAD según el estado mutacional de KRAS.

- Matriz de los niveles de expresión de miRs y ARNm filtrados y ajustados en pacientes con COAD para el análisis según el tipo de supervivencia. Archivos 'Expr-miRs-Superv.txt' y 'Expr-ARNm-Superv.txt' respectivamente.

- MiRs y ARNm diferencialmente expresados especificando FC y p-valores según la supervivencia alta o baja. Archivos 'DifExpr-miRs-Superv.txt' y 'DifExpr-ARNm-Superv.txt' respectivamente.

- MiRs con dianas putativas en genes diferencialmente expresados según el tipo de supervivencia de pacientes con COAD. Archivos 'Top10-miRs-frecARNm-Superv.txt' y 'Top10-miRs-namesARNm-Superv.txt'.

- Análisis de enriquecimiento biológico de GO de los genes diferencialmente expresados en pacientes con COAD según el estado mutacional de KRAS. Archivo 'GO-mutKRAS.txt'.

- Análisis de enriquecimiento biológico de GO de los genes diferencialmente expresados según el tipo de supervivencia en pacientes con COAD. Archivo 'GO-supervivencia.txt'.

1.6 Breve descripción de los otros capítulos de la memoria

- **Material y Métodos.** En este apartado detallamos en profundidad los recursos utilizados, estrategias seguidas, así como las alternativas propuestas.

- **Resultados y Discusión**. En este apartado detallamos y argumentamos los resultados obtenidos a lo largo del análisis.

2. Materiales y Métodos

En este apartado detallamos en profundidad los recursos utilizados, estrategias seguidas, así como las alternativas propuestas.

2.1. Uso del lenguaje de programación R, RStudio como entorno de trabajo y del repositorio Bioconductor.

Usaremos la versión de R '3.5.0'. R es un entorno de software libre para computación estadística. El lenguaje R fue creado en 1997 por Robert Gentleman y Ross Ihaka en la Universidad de Auckland. R puede considerarse como un conjunto integrado de softwares para la manipulación de datos, cálculo y gráficos. Una de las ventajas del lenguaje de programación en R es que el usuario posee el control total para modificar los estadísticos, gráficos y fórmulas. Funciona en cualquier plataforma, permite cargar y crear diversos paquetes y funciones para compartir con la comunidad científica. Es una herramienta que se puede extender con paquetes que facilitan fórmulas para el análisis y procesado de los datos. Toda la información sobre la descarga, el uso y funcionamiento de R se puede encontrar en la página de R-Project, https://www.r-project.org.

Para el uso del lenguaje de programación R trabajamos en un entorno de desarrollo integrado, RStudio: "Version 1.1.453 – © 2009-2018 RStudio, Inc". Este incluye una consola, un editor de sintaxis que soporta diferentes códigos de programación (R, Python, etc.), así como herramientas para gráficos y estadísticos de todo tipo. RStudio es accesible y gratuito permitiendo que los códigos sean reproducibles. Toda la información sobre el uso y funcionamiento de RStudio se puede encontrar en la página: https://www.rstudio.com.

Finalmente para esta memoria usaremos la herramienta Bioconductor, un repositorio de paquetes de R enfocado al análisis de datos biológicos (EdgeR, DESeq, TCGABiolinks, GO.db, etc). Es de código y desarrollo abierto. Toda la información sobre el uso y funcionamiento de Bioconductor se puede encontrar en la página https://bioconductor.org.

2.2. Obtención de datos de expresión génica y genotipo del *National Cancer Institute-GDC Data Portal*.

Obtenemos los datos de expresión desde el repositorio del Atlas del Genoma del Cáncer (The Cancer Genome Atlas, TCGA), una colaboración entre el Instituto Nacional de Cancer (National Cancer Institute, NCI) y el Instituto Nacional de Investigación del Genoma Humano (National Human Genome Research Institute (NHGRI)) creado para comprender y aumentar el conocimiento a nivel multidisciplinar de los cambios genómicos en 33 tipos de cáncer. Esta información se encuentra accesible de forma pública permitiendo una mejora en prevención, diagnóstico y tratamiento del cáncer.

Para la descarga de datos disponemos del portal GDC https://portal.gdc.cancer.gov. *GDC Data Portal* es una plataforma dónde se almacenan todos los datos de pacientes con cáncer, permitiendo a los investigadores el acceso para el estudio bioinformático de los mismos. Podemos visualizarlos directamente en la página web, descargarlos o acceder a ellos desde nuestro entorno R mediante la librería TCGABiolinks. Esta librería se encuentra dentro de la herramienta Bioconductor y ha sido creada para facilitar el acceso, descarga y análisis de los datos del TCGA (Colaprico et al., 2016). Usaremos las funciones *GDCquery*, *GDCdownload* y *GDCprepare* de la librería TCGABiolinks para la obtención de los archivos de expresión de los genes de interés en pacientes con

COAD. Los datos de expresión se descargan en nuestro directorio de R correspondiendo a un archivo por paciente.

Para el acceso a los datos de expresión de miRs optamos por seleccionar aquellos que correspondan al proyecto 'TCGA-COAD', categoría *Transcriptome Profiling* y tipo *miRNA Expression Quantification*; obteniendo un total de 462 muestras con información de expresión de 1881 miRs. Para cada miR tenemos 4 columnas informativas que corresponden al identificador, los conteos crudos de expresión, los conteos normalizadas (*per million*) y el mapeo del gen. Para la realización de este análisis seleccionamos la información de los conteos crudos como dato de expresión de los diferentes miRs.

Para el acceso a los datos de expresión de ARNm optamos por seleccionar aquellos que correspondan al proyecto 'TCGA-COAD', categoría *Transcriptome Profiling*, tipo *Gene Expression Quantification* y tipo de análisis seguido *HTSeq - FPKM-UQ*; obteniendo un total de 506 muestras con información de expresión de 56830 ARNm. Para cada ARNm tenemos 2 columnas informativas, una con el identificador del gen y otra con los datos de expresión ya normalizados en FPKM-UQ (*Fragments Per Kilobase of transcript per Million mapped reads upper quartile)* una versión modificada de la fórmula de FPKM en la cual se usan el percentil 75º de los conteos crudos como denominador en lugar de el número total de conteos codificantes de proteínas. Lo ideal en este paso sería seleccionar los conteos obtenidos por un análisis de tipo *HTSeq-Count* sin normalizar, realizando nosotros el filtraje y ajuste del modelo lineal, pero cuando intentamos acceder a dicha información, salta un error del servidor web con el portal de acceso impidiendo el acceso a dichos datos (Error: Error in getURL(url, fromJSON, timeout(600), simplifyDataFrame = TRUE): 'getURL()' failed: URL: https://gdc-api.nci.nih.gov/files/?pretty=true&expand=cases.samples.porti... error: SSLRead() return error -9806). Se realiza el análisis con los conteos en FPKM-UQ considerando que los resultados obtenidos son una aproximación al análisis ideal.

Seleccionamos a aquellos pacientes para los cuales existe información de expresión de ARNm y de miRs, reduciéndose el número de muestras a 441.

La información para la creación de la matriz fenotípica del estado mutacional de KRAS se obtuvo del portal GDC. Seleccionamos las mutaciones del oncogen KRAS en COAD con frecuencias mayores a 7/400 pacientes. Estas implican a los codones 12 y 13 del gen, en concreto: G12D, G12V, G12C, G12A, G12S y G13D (Phipps et al., 2013; Roa et al., 2013; Tan & Du, 2012). Integramos la información fenotípica para crear los dos grupos poblacionales a comparar en el primer análisis, obteniendo 131 pacientes con mutación en KRAS y 310 con genotipo salvaje (*Wild Type, WT*). La n resultante de ambos grupos no es homogénea, pero los grupos son grandes, por lo que no interfiere en el análisis estadístico.

2.3. Análisis de supervivencia y obtención del fenotipo supervivencia del GDC-Data Porta.

La supervivencia de un paciente se mide en días desde un acontecimiento inicial hasta la actualidad o hasta el fallecimiento, por tanto, usaremos como datos de supervivencia los días desde que el paciente está bajo seguimiento (en caso de estar vivo; *days_to_last_follow_up*) o los días desde el inicio del seguimiento hasta la muerte (en caso de fallecimiento; *days_to_death*). Obtenemos los datos del repositorio de información clínica del portal-GDC seleccionando las muestras del proyecto TCGA-COAD.

Analizamos el tipo de distribución de la variable supervivencia en las muestras con COAD según el estado mutacional de KRAS; comprobamos si existen diferencias significativas entre ambos grupos. Se analiza la distribución normal de la variable discreta supervivencia mediante un qqplot y test de Shapiro-Wilk. Optamos por un test no paramétrico de Wilcoxon para determinar si existe igualdad de distribución de densidad en la supervivencia de los pacientes según el estado mutacional de KRAS, considerando una igualdad de densidades al estadístico con un p-valor>0.05.

Para el análisis de independencia de la variable supervivencia y la variable estado mutacional de KRAS, ordenamos las muestras según los días de supervivencia. Dividimos en tres conjuntos equitativos de 146 individuos cada uno, escogemos el conjunto de mayor supervivencia y el de menor supervivencia, un total de 292 muestras para el análisis. Con esta selección aseguramos una correcta diferenciación de los datos de supervivencia, equiparamos los grupos muestrales y eliminamos muestras intermedias que puedan modificar el análisis. Realizamos un test de independencia de variables de Chi^2, hipótesis nula la independencia de variables e hipótesis alternativa de dependencia de variables.

2.4. Filtrado y ajuste de los datos de expresión génica (RefSeq).

Con el filtrado de datos eliminamos genes con poca evidencia de expresión que podrían interferir con las aproximaciones estadísticas. También reducimos las tasas de falso descubrimiento, aumentando la potencia para detectar genes diferencialmente expresados (Phipson et al., 2016).

Realizamos el proceso de filtrado con las *counts per million* (cpm), obtenidos mediante la función de mismo nombre del paquete de Biocondcutor '*edgeR'*. Las cpm son medidas descriptivas para ver la expresión real de un gen tras una normalización del tamaño de las librerías de los diferentes pacientes del estudio, es decir, las lecturas por kilobase y por millón. La obtención de las cpm debe hacerse siempre a partir de los conteos crudos, partimos de la base que la expresión de ARNm ya está normalizada por lo que este análisis no será del todo correcto pero, nos dará una aproximación a los resultados reales (Portal GDC Guide User). En la guía *GDC Data User's Guide NCI* se especifica trabajar con todo el set de genes cuando analizamos los conteos con FPKM-UQ, si procedemos de este modo, no obtenemos diferencias significativas en el análisis de expresión diferencial por la gran cantidad de genes a testar (no se detectan varianzas en la muestra) (Portal GDC Guide User). Nos vemos obligados a limpiar los datos de conteos de ARNm aún estando expresados en FPKM-UQ.

Seleccionamos aquellos genes cuyo cpm>1 en al menos un 'n' igual al grupo de estudio más pequeño. En este caso nuestro tamaño de muestra será n=7 para el análisis del estado mutacional de KRAS; estando formado por siete pacientes el subgrupo con la mutación en KRAS-G12S. Para el análisis de supervivencia la n será de 146 muestras, asegurándonos la expresión en al menos uno de los grupos, alta o baja supervivencia. En la tabla1 se muestra la disminución de miRs y ARNm tras el filtrado por cpm>1.

	Número de genes sin filtrar	Número de genes tras filtrar para el análisis mutacional de KRAS	Número de genes tras filtrar para el análisis de supervivencia
miRs	1881	802	350
ARNm	56830	33569	17079

Tabla 1 Número de miRs y ARNm antes y despues de filtrar.

Para poder comparar los datos de miRs entre las diferentes muestras y entre los genes dentro de una misma muestra llevamos a cabo un ajuste con Voom-limma (Ritchie et

al., 2015; Robinson & Oshlack, 2010; Smyth, Ritchie, & Thorne, 2015). Los conteos son variables discretas, para poder usar el modelo lineal de limma hay que ajustar la expresión de los genes como una variable continua. El ajuste con Voom permite esta transformación de forma que evita que los genes con bajos conteos tengan varianzas más grandes que los genes con altos conteos (Ritchie et al., 2015). Para los datos de expresión de ARNm usaremos una transformación log2 de los conteos obtenidos tras el filtrado por cpm (se desaconseja el ajuste voom con conteos expresados en FPKM-UQ) (Portal GDC Guide User).

2.5. Análisis interacciones miRs-ARNm con el paquete miRComb.

"Existen diferentes paquetes en R/Bioconductor capaces de elucidar interacciones de miRs-ARNm como RmiR, CORNA, miRNApath, microRNA o MultiMiR, pero ninguno de ellos permite realizar un análisis completo como el planteado en este estudio" (Vila-Casadesús et al., 2016). El paquete de R diseñado para tal fin es **miRComb**, http://mircomb.sourceforge.net. Permitir combinar los datos de expresión de ARNm y miRs con la información de hibridaciones de bases de datos ya existentes para descubrir dianas putativas de miRs-ARNm y rutas metabólicas alteradas en una enfermedad concreta (Vila-Casadesús et al., 2016).

Para trabajar con el paquete miRComb necesitamos varias matrices de datos para crear el objeto inicial sobre el cual trabajar, *corObject*. Partimos de las matrices de expresión génica de miRs y ARNm anterior con los conteos filtrados, ajustados y valores transformados a log2. Las columnas de dichas matrices corresponderán a las muestras, mientras que las filas corresponderán a los genes de interés. La información fenotípica la añadiremos en forma de *data frame*, dónde las filas correspondan a las muestras en el mismo orden que las matrices de conteos, y las columnas identifiquen a que grupo pertenece cada muestra (Vila-casades, 2015).

El análisis de expresión diferencial se lleva a cabo dentro del propio objeto *corObject* con el procedimiento de limma. Este análisis permite encontrar genes que se encuentren diferencialmente expresados entre los dos grupos de estudio y que esta diferencia sea significativamente estadística. El estadístico de limma realiza un contraste de hipótesis para cada gen moderando los errores estándar a lo largo de todos el set de genes. Ajustamos el modelo de limma para múltiples análisis por el método de Benjamini y Hochberg's (BH). El método de BH asume la independencia entre genes y controla la tasa de falsos descubrimientos. Seleccionamos aquellos genes diferencialmente expresado con un p-valor ajustado menor a 0.05 (adj-p.val < 0.05) preferiblemente o con un p-valor menor a 0.05 (p.val < 0.05).

Para determinar la relación entre los miRs y ARNm elegidos tras el análisis de expresión diferencial realizamos un test de correlación negativa de Pearson. Computamos un total de 135 miRs (adj.pval<0.05) y 3116 ARNm (adj.pval<0.05) en el análisis de expresión diferencial según el estado mutacional de KRAS; y de 350 miRs (pval<0.05) y 17079 ARNm (pval<0.05) en el análisis de expresión diferencial según la supervivencia. El test de Pearson asume una relación lineal entre miRs y ARNm. De los resultados obtenidos seleccionamos las correlaciones negativas, asumiendo que los miRs con dianas en un ARNm disminuyen su expresión.

Finalmente añadimos la información de las dianas putativas miRs-ARNm de dos bases de datos, MicroCosm y TargetScan. Obtenemos la información resumida de todos los análisis realizados, las puntuaciones obtenidas de las dianas y del análisis de correlación. Encontramos un resumen de los dos análisis realizados en miRComb en la tabla 2.

A

```
corObject with:
 miRNA slot with 441 samples and 802 probesets
 mRNA slot with 441 samples and 33569 probesets
Computations done:
- Differential expression mRNA:  limma method used
                                 estadoKRAS comparison used
- Differential expression miRNA:  limma method used
                                  estadoKRAS comparison used
- Correlation:  "pearson" method used
                "correlation" function used
                441 samples used
                135 miRNAs used
                   adj.pval < 0.05
                3116 mRNAs used
                   adj.pval < 0.05
- Database:  "microCosm_v5_18" database used
 - Database:  "targetScan_v6.2_18" database used
```

B

```
corObject with:
 miRNA slot with 292 samples and 350 probesets
 mRNA slot with 292 samples and 17079 probesets
Computations done:
- Differential expression mRNA:  limma method used
                                 Superv comparison used
- Differential expression miRNA:  limma method used
                                  Superv comparison used
- Correlation:  "pearson" method used
                "correlation" function used
                292 samples used
                350 miRNAs used
                   TRUE
                17079 mRNAs used
                   TRUE
- Database:  "microCosm_v5_18" database used
 - Database:  "targetScan_v6.2_18" database used
```

Tabla 2 Resumen del análisis del estado mutacional (A) y de supervivencia (B) por miRComb

2.6. Análisis de enriquecimiento biológico GO

Para el análisis de enriquecimiento biológico por GO podemos usar el propio paquete de miRComb, pero no obtenemos resultado alguno a pesar de tener ARNm diferencialmente expresados en nuestros resultados. Por ello, procedemos a extraer los datos de los genes diferencialmente expresados del objeto *corObject* de miRComb, así como su FC, p.val, adj-p.val, etc. Computamos el análisis con los paquetes orgDb específicos de especie de Bioconductor, en concreto org.Hs.eg.db para humanos. Estos paquetes contienen la información de identificadores de los genes para reconducirte a otras bases de datos y extraer información de dichos genes: GO, *symbol, genename*, ID de *ensemble*, *refseq*, *path*… Para integrar la información de nuestros genes de interés y las bases de datos usamos la función '*select*' del paquete *AnnotationDbi*. Para la creación de las tablas GO seleccionamos los genes diferencialmente expresados con un adj-p.val o p.val menor a 0.05, ordenándolos de mayor a menor significación estadística.

3. Resultados y Discusión

En este apartado detallamos y argumentamos los resultados obtenidos a lo largo del análisis.

3.1. Datos de expresión diferencial en pacientes con COAD según el estado mutacional de KRAS

El repositorio de TCGA posee la información necesaria de la cuantificación de expresión de genes, tanto de miRs como ARNm. Obtenemos la información del proyecto 'TCGA-COAD' con un total de 462 archivos correspondientes a cada uno de los pacientes con los niveles de expresión de 1881 miRs. Para los niveles de expresión de ARNm descargamos los conteos ya normalizados HTSeq-FPKM-UQ obteniendo los niveles de expresión de 56830 genes en 506 pacientes (tabla 3).

Show 5 entries Search:

data_type	cases	project
All	All	All
miRNA Expression Quantification	TCGA-DM-A1D7-01A-11H-A154-13	TCGA-COAD
miRNA Expression Quantification	TCGA-A6-3810-01A-01T-1021-13	TCGA-COAD
miRNA Expression Quantification	TCGA-AD-6548-01A-11H-1838-13	TCGA-COAD
miRNA Expression Quantification	TCGA-AU-3779-01A-01T-1722-13	TCGA-COAD
miRNA Expression Quantification	TCGA-DM-A28M-01A-12H-A16S-13	TCGA-COAD

Showing 1 to 5 of 462 entries Previous 1 2 3 4 5 … 93 Next

Show 5 entries Search:

data_type	cases	project
All	All	All
Gene Expression Quantification	TCGA-DM-A1D7-01A-11R-A155-07	TCGA-COAD
Gene Expression Quantification	TCGA-A6-3810-01B-04R-A277-07	TCGA-COAD
Gene Expression Quantification	TCGA-A6-3810-01A-01R-1022-07	TCGA-COAD
Gene Expression Quantification	TCGA-A6-3810-01A-01R-A278-07	TCGA-COAD
Gene Expression Quantification	TCGA-AD-6548-01A-11R-1839-07	TCGA-COAD

Showing 1 to 5 of 506 entries Previous 1 2 3 4 5 … 102 Next

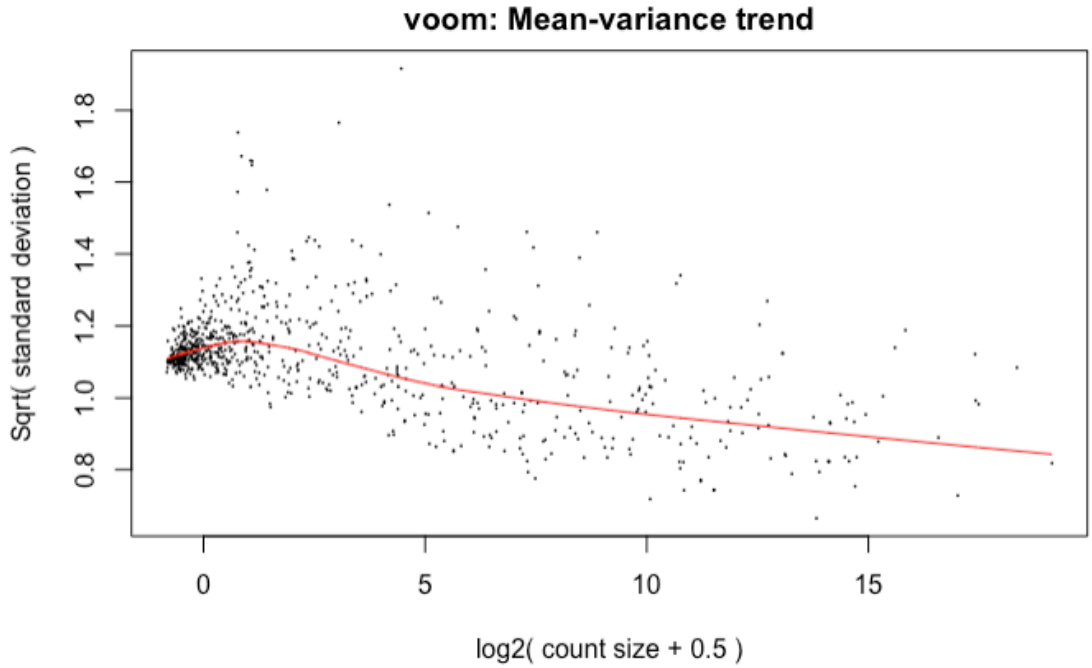
Tabla 3 Datos de expresión de miRs y ARNm descargados del repositorio TCGA-COAD

Para el análisis nos quedamos con los pacientes comunes a ambas bases de datos, información de los niveles de expresión de ARNm y de miRs, reduciéndose el grupo muestra a un total de 441.

Debemos pre-procesar los datos para eliminar genes que no se expresan en la mayoría de pacientes. De esta forma evitamos interferencias para los posteriores análisis estadísticos ya que una alta cantidad de genes con baja expresión podría enmascarar a genes que sí están diferencialmente expresados. Seleccionamos de ambas listas todos aquellos genes cuyas cpm>1 en al menos 7 muestras. Reducimos los genes a 802 miRs y 33569 ARNm en los 441 pacientes.

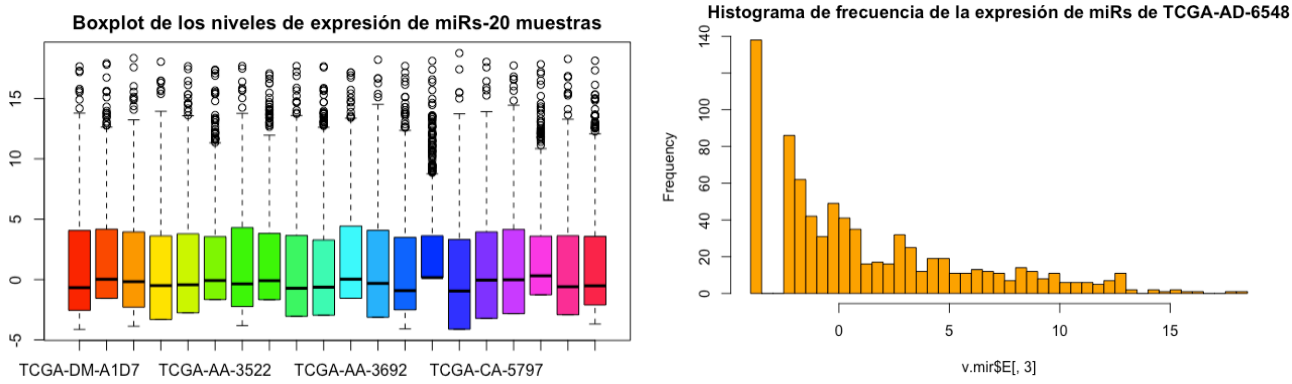
Realizamos un ajuste de los datos de expresión de miRs para poder trabajar posteriormente con ellos en un modelo lineal de limma. Computamos el ajuste voom, reduciendo la varianza de los genes con bajas expresiones como vemos en la gráfica 1.

Tras el ajuste obtenemos un objeto Elist con los datos de los conteos ajustados y en escala logarítmica de base 2 listos para el análisis con miRComb.



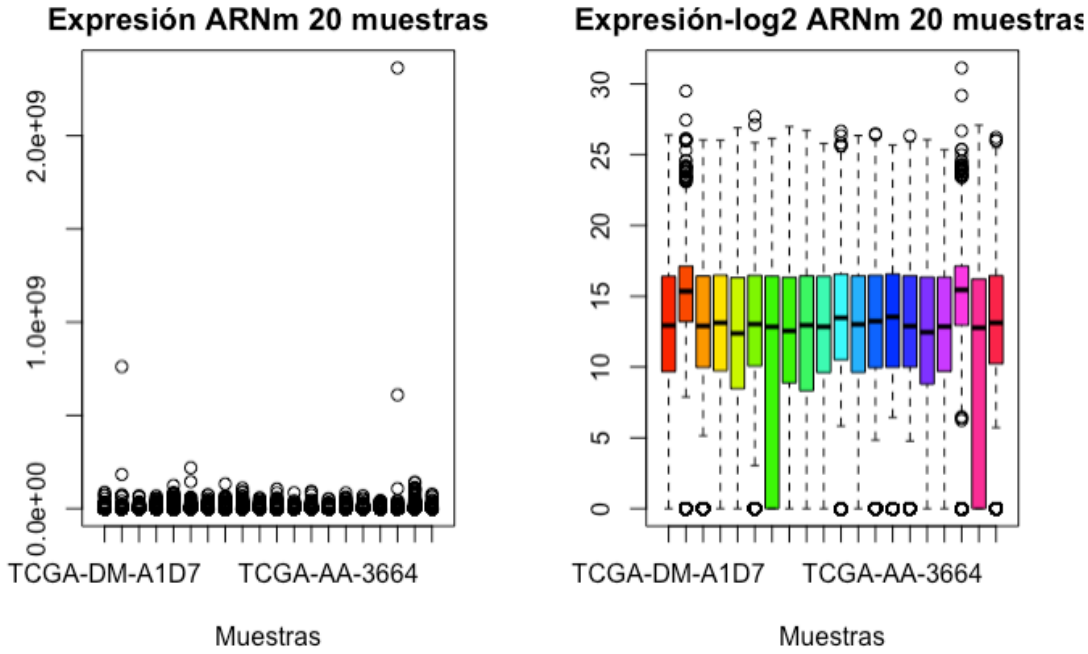
Gráfica 1 Ajuste Voom - Desviación estándar de la expresión de miRs para el análisis del estado mutacional de KRAS

Podemos visualizar en la gráfica 2 la distribución de la expresión de los miRs en varias muestras tras el ajuste realizado (datos en base log2). Observamos como la mayoría de miRs se expresan muy poco o nada y, unos pocos se expresan a niveles más altos (rango del boxplot de Q3 al máximo), causando que los datos se distribuyan en su mayoría en en Q1 del boxplot. Esta distribución la podemos observan mejor en un histograma de frecuencias para una de las muestra de TCGA-COAD (gráfica 2).



Gráfica 2 Visualización de la distribución de la expresión de miRs para el análisis del estado mutacional de KRAS. Boxplot con los niveles de expresión de 20 muestras. Histograma de frecuencia con los niveles de expresión de los miRs en la muestra TCGA-AD-6548

Realizamos una transformación a escala log2 de los datos de expresión de ARNm filtrados para poder visualizar correctamente los conteos, gráfica 3.



Gráfica 3 Distribución de la expresión de ARNm para el análisis del estado mutacional de KRAS. Distribución antes y después de la transformación logarítmica en base 2.

3.2. Análisis de expresión diferencial en pacientes con COAD según el estado mutacional de KRAS - miRComb

Comparamos la expresión de genes de pacientes de COAD con alteraciones en el oncogen KRAS y los que tienen un estado salvaje del gen. Como ya hemos visto, el gen KRAS aparece mutado en el 30-45% de los casos (Markman, 2012; Muzny et al., 2012; Roa et al., 2013; Tan & Du, 2012) y concretamente el 90% de las mutaciones se concentran en los codones 12 y 13 (Liu, Jakubowski, & Hunt, 2011; Roa et al., 2013). Por ello hemos seleccionado las mutaciones de los codones 12 y 13 con una frecuencia de ocurrencia superior a 1,75% en los datos de cohorte de TCGA (frecuencias de las mutaciones en el repositorio de TCGA mostradas en la tabla 4).

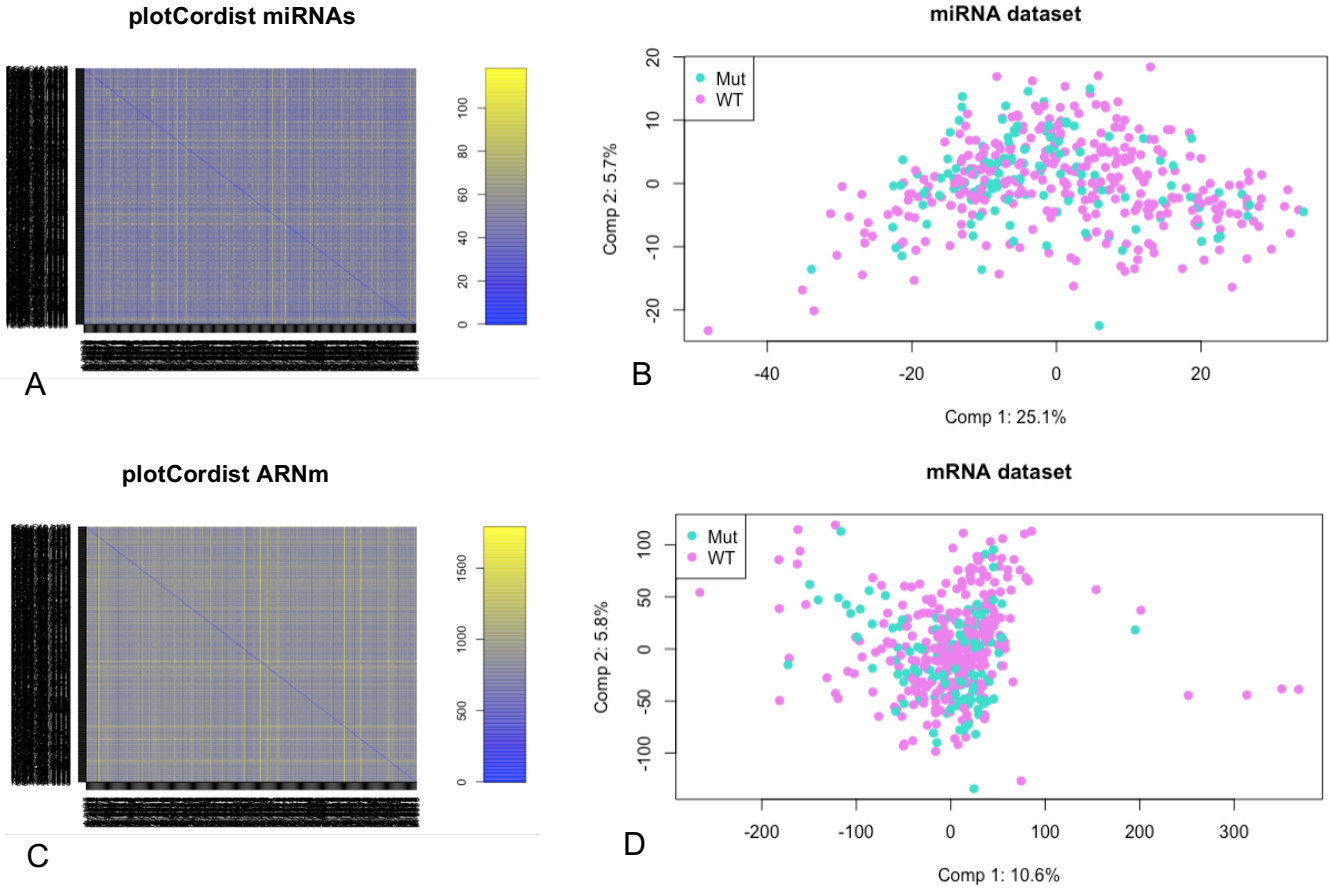
Mutación	Tipo de mutación	Consecuencia	Casos en la cohorte de TCGA-COAD	Porcentaje de casos en la cohorte de TCGA-COAD
G12D	Sustitución	*Missense*	49/400	12,25%
G12V	Sustitución	*Missense*	33/400	8,25%
G12C	Sustitución	*Missense*	10/400	2,5%
G12A	Sustitución	*Missense*	8/400	2%
G12S	Sustitución	*Missense*	7/400	1,75%
G13D	Sustitución	*Missense*	31/400	7,75%

Tabla 4 Mutaciones más frecuentes de KRAS en COAD según la cohorte de TCGA. Link: https://portal.gdc.cancer.gov/exploration?filters=%7B%22op%22%3A%22and%22%2C%22content%22%3A%5B%7B%22op%22%3A%22in%22%2C%22content%22%3A%7B%22field%22%3A%22cases.primary_si

Tras la descarga de los datos clínicos de todos los pacientes con alguna de las mutaciones seleccionadas, obtenemos un total de 138 pacientes de los cuales 131

corresponden a pacientes con información para la expresión de miRs y ARNm. Esta cifra supone que un 29,7% de nuestra muestra poblacional posee mutaciones en el oncogen KRAS, acercándose a los porcentaje observados de 30-45% en diversos estudios en todo el mundo (Muzny et al., 2012; Roa et al., 2013; Tan & Du, 2012).

Procedemos a realizar el análisis de expresión diferencial en miRComb. Creamos un objeto tipo clase *corObject* en el que iremos añadiendo todos los resultados obtenidos: matriz de expresión de miRs (Expr-miRs-KRAS.txt) y ARNm (Expr-ARNm-KRAS.txt), y tabla fenotípica con el estado mutacional de KRAS. Observamos los datos antes de procesarlos en un gráfico *Cordist* 4.A y 4.C dónde se muestran las distancias/correlación entre los genes de cada muestra, vemos como en ambas gráficas se forma una diagonal por coincidencia de las muestras entre los ejes X e Y (debido a la cantidad de muestras del estudio no se aprecian los nombres). Mediante el gráfico 4.B y 4.D de análisis de componentes principal (PCA) observamos los dos componentes principales no correlacionados que miden la varianza original entre las dos poblaciones de estudio. Así, vemos como el primer componente de la expresión de miRs explica un 25,1% de la varianza entre los dos grupos poblacionales (gráfica 4.B) y un 10,6% de la varianza en cuanto a la expresión de ARNm (gráfica 4.D). No observamos una tendencia al agrupamiento de las muestras según el estado mutado o WT de KRAS en ninguno de los PCA, indicativo de que las poblaciones no se diferencian en terminos de expresión genica.



Gráfica 4 Distribución de las muestras según el estado muatcional de KRAS. (A) Grafico corDist de miRs. (B) PCA de expresión de miRs. (C) Grafico corDist de ARNm. (D) PCA de expresión de ARNm

Realizamos un análisis de expresión diferencial mediante un modelo lineal de limma. Limma comprueba con un test de hipótesis si hay o no diferencia significativa entre los

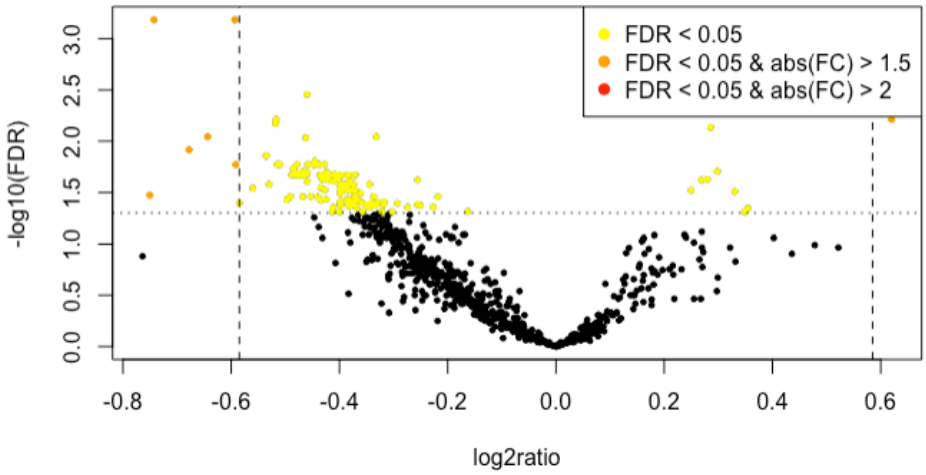
valores de expresión de cada gen según el estado mutacional de KRAS. Obtenemos un p-valor y el *fold change* (FC) en log2 para cada gen. Dado que estamos realizando múltiples test debemos ajustar el modelo con el método de Benjamini y Hochberg's para controlar la tasa de falsos descubrimientos, obteniendo así el valor p-ajustado de los genes por el cual filtraremos los resultados finales. Tablas adjuntas con los datos de miRs en el archivo 'DifExpr-miRs-KRAS.txt' y de ARNm en el archivo 'DifExpr-ARNm-KRAS.txt'.

Observamos lo resultados en un gráfico Volcano 5.A y 6.A dónde se representa la dimensión biológica en el eje de las abscisas (FC*)* y la estadística en el eje de las ordenadas por el logaritmo negativo del factor de corrección múltiple (FDR - *False Discovery Rate*). Los genes con un FDR por debajo del límite se resaltan en amarillo (FDR < 0.05). Los elementos con FDR por debajo del límite y por encima de un FC=1 o 1.5 se resaltan en naranja. Los elementos con FDR por debajo del límite y por encima de FC-2 se resaltan en rojo.

Vemos como hay muy pocos miRs expresados diferencialmente y que estos varían su FDR/FC muy poco entre las dos condiciones del estudio (gráfica 5.A). En las tablas de la gráfica 5.B y 5.C sólo encontramos 11 miRs sobre-expresados en el análisis, con un FC muy bajo, y 124 miRs regulados a la baja, con un logFC que no supera 1.5. Estos datos son indicativos de la poca diferencia de expresión de miRs entre las dos condiciones del estudio. Pre-visualizamos la falta de un perfil genético característico en la expresión de miRs realizando un heatmap, gráfica 5.D, dónde las muestras no parecen tener una expresión definida/única según el estado mutacional de KRAS.

En la gráfica 6.A del Volcano de expresión ARNm encontramos más genes diferencialmente expresados con unos FC/FDR mayores y un p-valor más bajo que en los resultados de expresión de miRs. Mostramos 20 genes regulados a la baja (gráfica 6.B) o sobre-expresados (gráfica 6.C) con niveles elevados de logFC, es decir, aquellos en los que existe mayor diferencia en los niveles de expresión del gen entre las dos condiciones. En este caso obtenemos un total de 396 genes sobre-expresados y de 308 genes regulados a la baja con un FC>2 y p-valor ajustado < 0.05. Graficamos un *heatmap 6.D* con los 20 genes de expresión más diferente entre las dos condiciones y observamos un cierto perfil de expresión génica (semejanza a nivel de expresión de ARNm en las muestras con KRAS WT y semejanza en las muestras con KRAS mutado).

A **VolcanoPlot miRNAs**



B

Table: miRs downregulated FC>1 y adj.pval<0.05 según el estado mutacional de KRAS

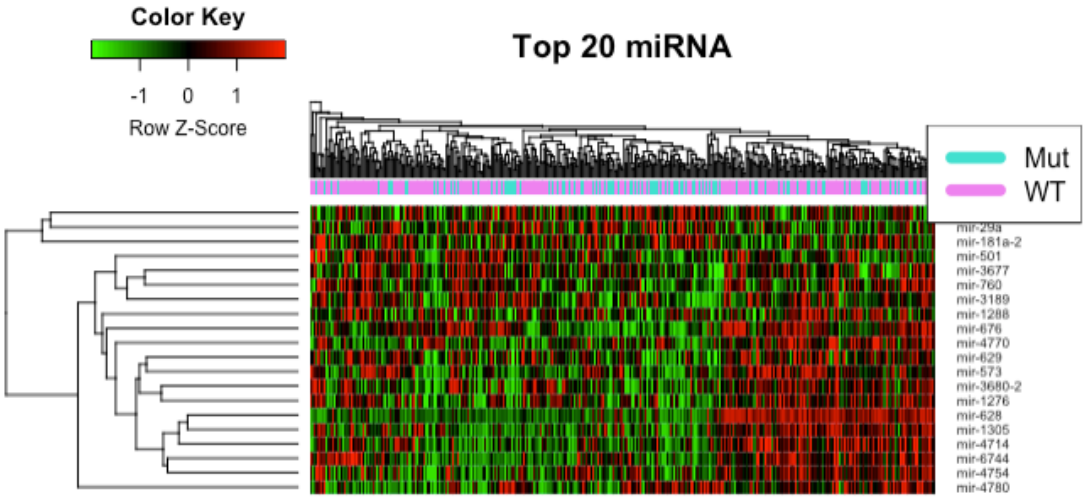
	FC	logratio	meanExp	pval	adj.pval
hsa-mir-374c	-1.6820	-0.7502	-0.8280	0.0035	0.0335
hsa-mir-760	-1.6733	-0.7427	-0.1580	0.0000	0.0007
hsa-mir-676	-1.5997	-0.6778	-0.7551	0.0002	0.0121
hsa-mir-628	-1.5621	-0.6435	3.7648	0.0001	0.0091
hsa-mir-1288	-1.5086	-0.5933	-0.7747	0.0000	0.0007
hsa-mir-573	-1.5070	-0.5917	-1.3136	0.0004	0.0168
hsa-mir-4326	-1.4998	-0.5848	1.9648	0.0054	0.0402
hsa-mir-548f-1	-1.4745	-0.5602	-0.5402	0.0026	0.0287
hsa-mir-3189	-1.4493	-0.5354	-0.9613	0.0002	0.0139
hsa-mir-551a	-1.4440	-0.5301	-1.1731	0.0021	0.0261
hsa-mir-3680-2	-1.4320	-0.5180	-1.7436	0.0001	0.0068
hsa-mir-1276	-1.4310	-0.5170	-1.7310	0.0001	0.0061
hsa-mir-6744	-1.4288	-0.5148	-2.0239	0.0005	0.0168
hsa-mir-4638	-1.4242	-0.5101	-0.4601	0.0005	0.0168
hsa-mir-7702	-1.4117	-0.4974	-0.4762	0.0044	0.0371

C

Table: miRs upregulated FC>1 y adj.pval<0.05 según el estado mutacional de KRAS

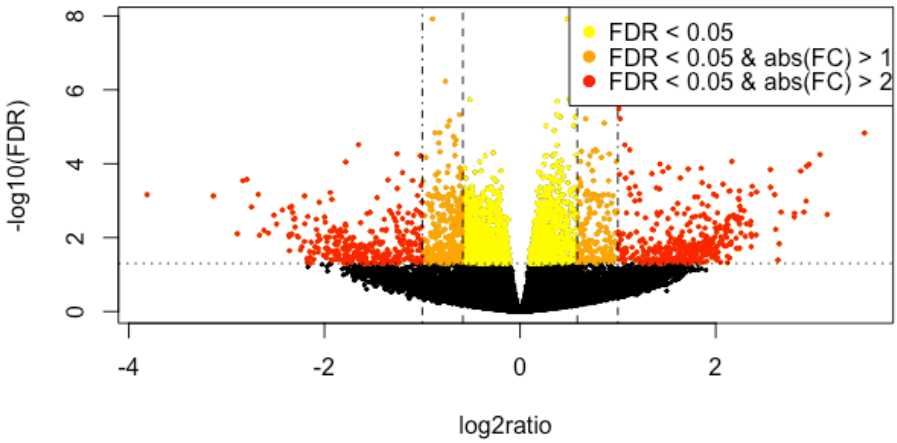
	FC	logratio	meanExp	pval	adj.pval
hsa-mir-4780	1.5371	0.6202	-1.5267	0.0000	0.0061
hsa-mir-10a	1.3785	0.4631	15.8922	0.0000	0.0016
hsa-let-7f-2	1.2789	0.3549	11.4777	0.0071	0.0450
hsa-let-7f-1	1.2735	0.3488	11.4603	0.0082	0.0487
hsa-mir-96	1.2572	0.3303	4.0615	0.0031	0.0309
hsa-mir-181a-2	1.2516	0.3237	9.8000	0.0000	0.0029
hsa-mir-181b-2	1.2296	0.2982	6.3429	0.0007	0.0195
hsa-mir-29a	1.2191	0.2858	12.9691	0.0001	0.0073
hsa-mir-34a	1.2139	0.2796	6.6755	0.0017	0.0235
hsa-mir-27a	1.2051	0.2692	10.5626	0.0018	0.0238
hsa-mir-181b-1	1.1888	0.2495	6.5474	0.0030	0.0301

D **Heatmap miRNAs**



Gráfica 5 Análisis de expresión diferencial de miRs según el estado mutacional de KRAS. (A) VolcanoPlot de la expresión diferencial de miRs. (B) Tabla de miRs regulados a la baja, con FC>1 y con un un adj.p-valor <0.05. (C) Tabla de miRs regulados a la alta, con FC>1 y con un un adj.p-valor <0.05. (D) HeatMap de la expresión de 20miRs en las 441 muestras de TCGA-COAD.

A **VolcanoPlot ARNm**



B

Table: mRNA downregulated FC>2 y adj.pval<0.05 según el estado mutacional de KRAS

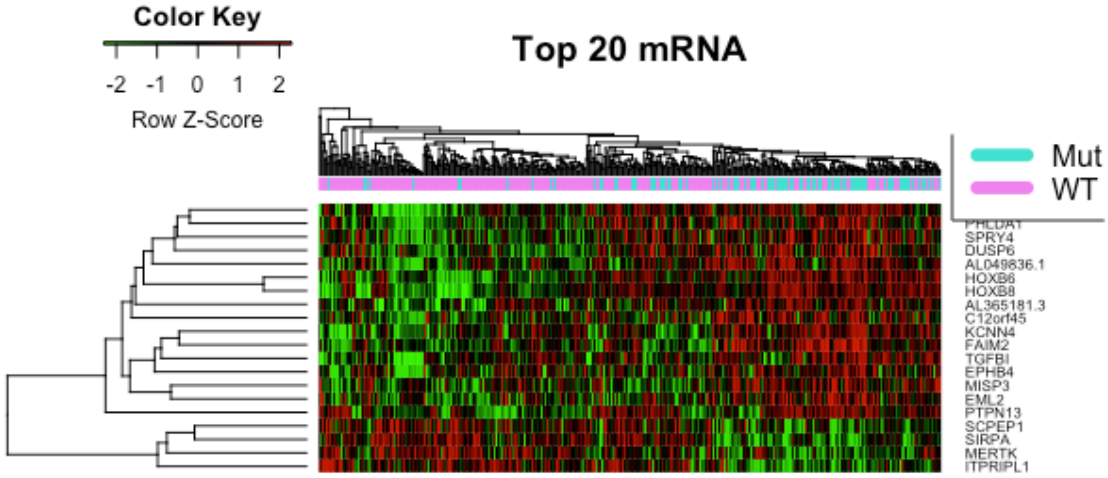
	FC	logratio	meanExp	pval	adj.pval
IGHD3.3	-14.0555	-3.8131	7.9344	0.0000	0.0007
RNA5SP149	-8.7886	-3.1356	3.6848	0.0000	0.0007
IGLJ3	-7.4193	-2.8913	9.8026	0.0002	0.0078
COPS8P2	-7.1334	-2.8346	7.1344	0.0000	0.0003
ZNF663P	-6.9290	-2.7927	7.8012	0.0000	0.0003
VTRNA1.1	-6.7056	-2.7454	3.4655	0.0000	0.0015
AL121949.1	-6.3962	-2.6772	5.0690	0.0000	0.0007
IGLJ2	-6.3386	-2.6642	12.3881	0.0003	0.0086
RNA5S9	-6.1393	-2.6181	3.7251	0.0002	0.0064
INSL5	-5.9568	-2.5745	6.7510	0.0002	0.0075
HZAFZP4	-5.7205	-2.5161	8.0392	0.0000	0.0025
TMIGD1	-5.6143	-2.4891	10.5764	0.0001	0.0043
CLC	-5.3883	-2.4298	12.4030	0.0000	0.0018
MYL6P2	-5.1364	-2.3608	7.7975	0.0003	0.0084
RNA5SP370	-5.1319	-2.3595	3.7656	0.0003	0.0093
AC007991.2	-5.1272	-2.3582	6.6127	0.0013	0.0226
AC017013.1	-5.1164	-2.3551	4.8910	0.0000	0.0015
AC245128.3	-5.0874	-2.3469	9.4668	0.0000	0.0015
AL020994.1	-5.0406	-2.3336	3.5996	0.0000	0.0014
COX5AP2	-5.0380	-2.3329	7.4831	0.0002	0.0079

C

Table: mRNA upregulated FC>2 y adj.pval<0.05 según el estado mutacional de KRAS

	FC	logratio	meanExp	pval	adj.pval
AC011611.2	11.4772	3.5207	7.4281	0.0000	0.0000
RF02246	8.8151	3.1400	6.5818	0.0000	0.0024
FOXI1	8.3782	3.0666	7.2621	0.0000	0.0001
AC005256.1	7.7518	2.9545	10.9118	0.0000	0.0001
FEZF1	7.6281	2.9313	10.9070	0.0000	0.0001
LINC02512	7.5941	2.9249	6.2231	0.0000	0.0010
MIR135B	7.4740	2.9019	6.3890	0.0000	0.0020
GLYATL1P4	7.3063	2.8691	7.4042	0.0000	0.0002
SPRR2A	6.9941	2.8061	8.2873	0.0000	0.0022
RF02247	6.9798	2.8032	5.7767	0.0000	0.0028
LINC02178	6.3565	2.6682	4.9364	0.0000	0.0020
SPRR1A	6.2726	2.6491	8.5883	0.0006	0.0146
RPL41P1	6.2211	2.6372	16.4678	0.0033	0.0404
RF00156.11	6.1206	2.6137	7.0895	0.0002	0.0060
AL451050.1	6.0739	2.6026	8.2354	0.0000	0.0007
AC140479.1	5.9004	2.5608	3.5568	0.0000	0.0004
FEZF1.AS1	5.8937	2.5592	12.3025	0.0000	0.0001
AC011611.3	5.3428	2.4176	9.4436	0.0000	0.0006
AC004009.1	5.3276	2.4135	5.6900	0.0003	0.0085
KRT6A	5.1815	2.3734	11.4433	0.0000	0.0004

D **Heatmap ARNm**



Gráfica 6 Análisis de expresión diferencial de ARNm según el estado mutacional de KRAS. (A) VolcanoPlot de la expresión diferencial de ARNm. (B) Tabla de ARNm regulados a la baja, con FC>2 y con un adj.p-valor <0.05. (C) Tabla de ARNm regulados a la alta, con FC>2 y con un un adj.p-valor <0.05. (D) HeatMap de la expresión de 20ARNm en las 441 muestras de TCGA-COAD.

Seleccionamos los genes diferencialmente expresados con un p-valor ajustado menor al 0.05 y procedemos a realizar el análisis de correlación negativa con el test de Pearson para miRs-ARNm. El valor nos indicará la relación negativa de los niveles de expresión entre miRs y ARNm asociado a un p-valor del estadístico (tabla 5 mostrando algunos valores de la correlación). Sólo seleccionamos las correlaciones negativas, partimos de la premisa que la mayoría de miRs regulan a la baja a sus ARNm dianas. Cuanto más se acerque a '-1' el valor de la correlación, más fuerza tendrá la relación entre el miR y el ARNm testado.

Table: Correlación negativa miRs-ARNm según estado mutacional de KRAS

	FGR	CFH	MYH16	SNX11	CYP26B1	ICA1	TFPI	CD38	RBM6	HSPB6
hsa-let-7f-1	-0.107	0.011	0.234	-0.220	-0.062	0.144	0.040	-0.181	0.235	-0.049
hsa-let-7f-2	-0.109	0.010	0.236	-0.221	-0.062	0.148	0.039	-0.182	0.235	-0.051
hsa-mir-10a	-0.214	-0.270	0.013	-0.039	-0.151	0.174	-0.246	-0.272	0.118	-0.195
hsa-mir-1197	0.076	0.063	-0.210	0.124	0.069	-0.103	0.049	0.079	-0.294	0.060
hsa-mir-1226	-0.117	-0.170	-0.105	0.000	-0.136	-0.100	-0.033	-0.053	0.002	-0.086
hsa-mir-1227	-0.096	-0.174	-0.152	0.149	-0.089	-0.063	-0.025	0.023	-0.211	-0.126
hsa-mir-1228	0.036	-0.054	-0.128	0.068	-0.059	-0.171	0.050	0.092	-0.087	-0.066
hsa-mir-1229	-0.091	-0.192	-0.143	-0.002	-0.143	-0.161	-0.068	-0.012	-0.019	-0.108
hsa-mir-1255a	-0.032	-0.183	-0.098	0.068	-0.069	-0.106	-0.037	-0.060	-0.148	-0.202
hsa-mir-1276	-0.089	-0.186	-0.115	0.088	-0.091	-0.127	-0.060	-0.063	-0.169	-0.160
hsa-mir-1286	-0.015	-0.120	-0.226	0.167	-0.046	-0.148	-0.036	0.057	-0.342	-0.161
hsa-mir-1288	0.094	0.027	-0.105	0.042	0.102	-0.190	0.022	0.094	-0.146	-0.079
hsa-mir-1305	0.096	0.024	-0.224	0.187	0.079	-0.154	0.057	0.093	-0.270	-0.096
hsa-mir-130b	-0.229	-0.371	-0.186	0.089	-0.232	-0.031	-0.170	-0.094	-0.088	-0.258
hsa-mir-146b	0.533	0.368	-0.075	0.101	0.206	-0.263	0.278	0.386	-0.168	-0.012
hsa-mir-181a-2	-0.208	-0.055	0.120	-0.212	-0.114	0.037	-0.078	-0.103	0.155	-0.160
hsa-mir-181b-1	-0.139	-0.022	0.072	-0.169	-0.111	0.019	-0.026	-0.105	0.188	-0.117
hsa-mir-181b-2	-0.175	-0.028	0.134	-0.221	-0.117	0.029	-0.052	-0.125	0.236	-0.133
hsa-mir-186	-0.169	-0.308	-0.161	0.033	-0.219	-0.030	-0.134	-0.016	-0.048	-0.297
hsa-mir-18b	-0.129	-0.203	-0.108	0.069	-0.109	-0.031	-0.014	-0.069	-0.091	-0.031

Tabla 5 Muestra de la correlación negativa de miRs-ARNm según el estado mutacional de KRAS

Finalmente integramos la información de las bases de datos sobre las dianas de los miRs. Usamos la información de dos bases de datos: microCosm (v5.18) y TargetScan(v6.2.18). Podemos ordenar la tabla resultante por aquellas interacciones que tengan una puntuación más alta: con dianas en más bases de datos, menor p-valor, mayor FC, etc. (tabla 6). No tenemos ninguna coincidencia de miRs diferencialmente sobreexpresados que tengan diana en alguno de los ARNm diferencialmente infraexpresados, el algoritmo de miRComb no detecta que ninguno de los ARNm seleccionados aparezca en ninguna base de datos como dianas de los miRs seleccionados.

Table: Resumen de puntuaciones en las interacciones miRs-ARNm

miRNA	mRNA	cor	pval	logratio.miRNA	meanExp.miRNA	logratio.mRNA	meanExp.mRNA	dat.microCosm_v5_18	dat.targetScan_v6.2_18	dat.sum	score
-											
hsa-mir-374c	IGHD3.3	0.121	0.994	-0.750	-0.828	-3.813	7.934	0	0	0	-5.721
hsa-mir-760	IGHD3.3	0.016	0.628	-0.743	-0.158	-3.813	7.934	0	0	0	-5.664
hsa-mir-676	IGHD3.3	0.102	0.984	-0.678	-0.755	-3.813	7.934	0	0	0	-5.169
hsa-mir-628	IGHD3.3	0.195	1.000	-0.644	3.765	-3.813	7.934	0	0	0	-4.908
hsa-mir-374c	RNA5SP149	0.458	1.000	-0.750	-0.828	-3.136	3.685	0	0	0	-4.705
hsa-mir-760	RNA5SP149	0.114	0.992	-0.743	-0.158	-3.136	3.685	0	0	0	-4.658
hsa-mir-1288	IGHD3.3	0.055	0.875	-0.593	-0.775	-3.813	7.934	0	0	0	-4.524
hsa-mir-573	IGHD3.3	0.020	0.661	-0.592	-1.314	-3.813	7.934	0	0	0	-4.512
hsa-mir-4326	IGHD3.3	-0.008	0.436	-0.585	1.965	-3.813	7.934	0	0	0	-4.460
hsa-mir-4780	AC011611.2	0.160	1.000	0.620	-1.527	3.521	7.428	0	0	0	-4.367
hsa-mir-374c	IGLJ3	0.513	1.000	-0.750	-0.828	-2.891	9.803	0	0	0	-4.338
hsa-mir-760	IGLJ3	0.079	0.951	-0.743	-0.158	-2.891	9.803	0	0	0	-4.295
hsa-mir-548f-1	IGHD3.3	-0.104	0.015	-0.560	-0.540	-3.813	7.934	0	0	0	-4.272
hsa-mir-374c	COPS8P2	0.369	1.000	-0.750	-0.828	-2.835	7.134	0	0	0	-4.253
hsa-mir-676	RNA5SP149	0.154	0.999	-0.678	-0.755	-3.136	3.685	0	0	0	-4.251
hsa-mir-760	COPS8P2	0.069	0.926	-0.743	-0.158	-2.835	7.134	0	0	0	-4.210
hsa-mir-374c	ZNF663P	-0.083	0.041	-0.750	-0.828	-2.793	7.801	0	0	0	-4.190
hsa-mir-760	ZNF663P	-0.004	0.465	-0.743	-0.158	-2.793	7.801	0	0	0	-4.148
hsa-mir-374c	VTRNA1.1	0.262	1.000	-0.750	-0.828	-2.745	3.465	0	0	0	-4.119
hsa-mir-3189	IGHD3.3	-0.098	0.019	-0.535	-0.961	-3.813	7.934	0	0	0	-4.083

Tabla 6 Resumen de las puntuaciones miRs-ARNm en el análisis de expresión diferencial según el estado mutacional de KRAS

3.3. Enriquecimiento biológico GO de genes diferencialmente expresados según el estado mutacional de KRAS en pacientes con COAD

Realizamos el análisis de significación biológica de los 3116 ARNm diferencialmente expresados según el estado mutacional de KRAS (adj.p-val>0.05). Ordenamos ascendentemente los genes por el p-valor ajustado y obtenemos la información de GO y symbol de los ARNm (archivo adjunto en anexos 'GO-mutKRAS.txt'). Encontramos 2341 genes implicados en muerte celular, apoptosis, factores de crecimiento, receptores celulares… así como un gen de asociación a cáncer de colon (COLCA1) o el oncogen p53. Formará parte del trabajo del biólogo investigador el posterior análisis detallado y significación biológica de estos resultados.

3.4. Distribución de la supervivencia en pacientes con COAD

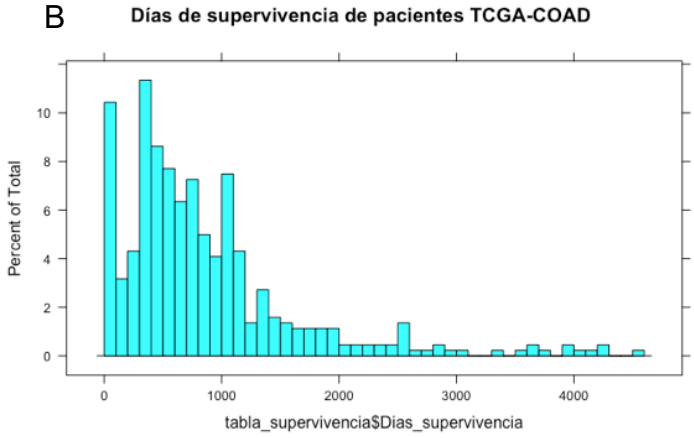
No hemos encontrado miRs diferencialmente sobreexpresados que tengan una correlación negativa con ARNm diferencialmente infraexpresados según el estado mutacional del oncogen KRAS. Esta ausencia de resultados podría deberse a que los pacientes no experimentan un cambio en la supervivencia si poseen alteraciones en el gen KRAS, sin alteraciones diferencialmente significativas en los niveles de expresión génica. Para elucidar el papel de la supervivencia en pacientes con COAD recurrimos a los datos clínicos seleccionando los días de seguimiento desde la diagnosis hasta la actualidad o ocurrencia de un evento si el paciente está vivo; o los días hasta la muerte si el paciente ya ha fallecido (tabla de la gráfica 7.A). Podemos ver en el histograma de la gráfica 7.B la distribución de la variable supervivencia en la población muestral de 441 pacientes del proyecto TCGA-COAD.

A

Información clínica de supervivencia TCGA-COAD

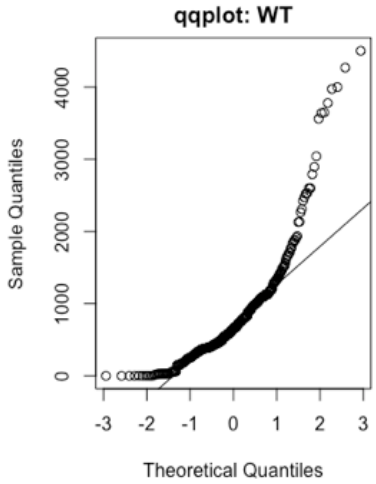
pacientesID	Estado_vital	Dias_seguimiento	Dias_muerte
TCGA-5M-AATE	alive	1200.0	--
TCGA-CA-6718	dead	3.0	306.0
TCGA-AD-5900	alive	370.0	--
TCGA-AA-3855	alive	975.0	--
TCGA-G4-6317	alive	1095.0	--
TCGA-AZ-4308	alive	3324.0	--
TCGA-A6-A565	dead	--	494.0
TCGA-AA-3989	dead	0.0	242.0
TCGA-D5-6920	alive	377.0	--
TCGA-AA-3833	alive	485.0	--
TCGA-A6-4105	dead	--	442.0
TCGA-WS-AB45	alive	2130.0	--
TCGA-AA-A00Q	alive	1278.0	--
TCGA-AA-3877	alive	943.0	--
TCGA-DM-A1D6	dead	1518.0	1518.0
TCGA-D5-6535	alive	460.0	--

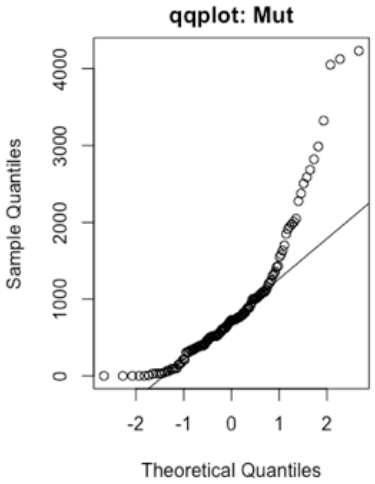
B



Gráfica 7 Variable supervivencia en TCGA-COAD. (A) Tabla con los datos clínicos. (B) Histograma de distribución de la variable supervivencia en las 441 muestras de TCGA-COAD.

Para determinar el estadístico a usar hacemos un análisis para observar la distribución normal de los datos de supervivencia en los dos grupos a comparar, WT y KRAS-mutado. En la gráfica 8 observamos como la distribución de la variable supervivencia se alejan mucho de una recta de normalidad ideal. Confirmamos la ausencia de normalidad con un test de Shapiro-Wilks, p-valor de 2.2e-16 y de 1.774e-11 < 0.05, rechazando la hipótesis nula de normalidad.






```
        Shapiro-Wilk normality test

data:  Dias_supervivencia[estadoKRAS == 0]
W = 0.8048, p-value < 2.2e-16


        Shapiro-Wilk normality test

data:  Dias_supervivencia[estadoKRAS == 1]
W = 0.81718, p-value = 1.774e-11
```

Gráfica 8 Análisis de la distribución normal de la variable supervivencia. QQplots de la variable en la subpoblación con KRAS mutada y con KRAS WT. Análisis de Shapiro-Wllks para las dos poblaciones.

Realizamos un test no paramétrico, prueba de Wilcoxon, para variables de dos grupos independientes categóricos. La hipótesis nula será "igualdad de distribución de densidad en los dos grupos" y la alternativa de "diferencia de distribución de densidad". Con un p-valor de 0.5692 (> 0.05) no podemos rechazar la hipótesis nula (tabla 7). Por tanto no hay evidencia significativa suficiente para determinar cambios en la supervivencia de los pacientes según el estado mutacional de KRAS. Este resultado confirma nuestras sospechas de que la ausencia de resultados positivos puede deberse a que los pacientes no presentan diferencias de supervivencia según el estado del oncogen KRAS.

```
        Wilcoxon rank sum test with continuity correction

data:  Dias_supervivencia by estadoKRAS
W = 19418, p-value = 0.5692
alternative hypothesis: true location shift is not equal to 0
```

Tabla 7 Análisis estadistico de la distribución de la variable supervivencia dependiendo del estado mutacional de KRAS

En los pasos posteriores pasaremos a analizar la expresión génica diferencial de miRs-ARNm según la supervivencia en los pacientes con COAD sea alta o baja. Para ello ordenamos las muestras según los días de supervivencia y dividimos en tres conjuntos equitativos de 146 muestras, escogemos el primero (mayor supervivencia) y el último (menor supervivencia) para generar las dos muestras a contrastar. Eliminamos pacientes intermedios que puedan modificar el análisis y aseguramos una correcta diferenciación de los datos de supervivencia.

Podemos analizar si existe relación entre la variable 'poseer el oncogen KRAS mutado' y la variable tener una 'alta' o 'baja' supervivencia; esta vez, la supervivencia se analizará como variable dicotómica (baja o alta). Tras computar el test de independencia de Chi^2 y con un p-valor=0,6039 (>0.05) determinamos que las variables son independientes (tabla 8), por lo que una variable no varía entre los distintos niveles de la otra. Confirmamos que la supervivencia no depende del estado mutacional de KRAS.

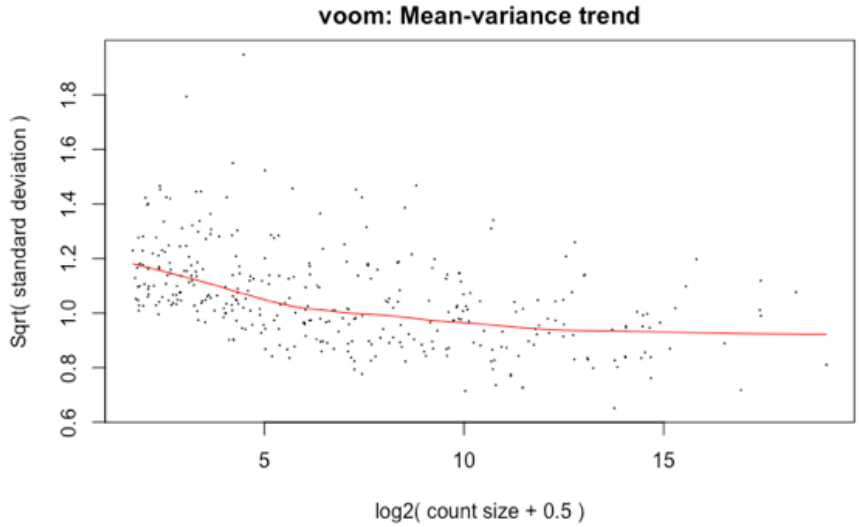
```
        Pearson's Chi-squared test with Yates' continuity correction

data:  Superv and grupo
X-squared = 0.26933, df = 1, p-value = 0.6038
```

Tabla 8 Análisis estadístico de dependencia entre la variable mutación/WT oncogen KRAS y alta/baja supervivencia

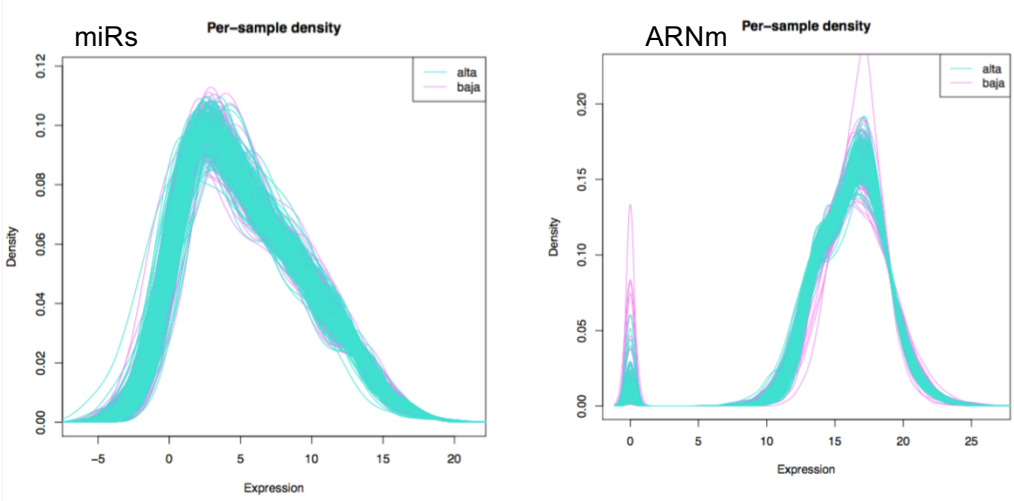
3.5. Análisis de expresión diferencial en pacientes con COAD según la supervivencia - miRComb

Para elucidar si existen diferencias en la expresión de rutas metabólicas entre los pacientes con alta supervivencia y baja supervivencia, computaremos el mismo análisis de miRComb pero dividiendo a la muestra de pacientes de TCGA-COAD por la variable dicotómica supervivencia. Debemos reajustar las matrices de expresión de miRs y ARNm a las 292 muestras con las que trabajaremos ahora (146 con alta y 146 con baja supervivencia). Para evitar interferencias en los posteriores análisis estadísticos eliminamos de ambas listas todos aquellos genes cuyas cpm sean menores a 1 en al menos 146 muestras. Reducimos los genes a 305 miRs y 17079 ARNm en los 292 pacientes (tabla 1).

Como en el análisis anterior realizamos un ajuste voom de los datos de expresión de miRs para poder trabajar posteriormente con ellos en un modelo lineal de limma (gráfica 9). Para los datos de expresión de ARNm realizamos una transformación a escala log2 de los datos de expresión.

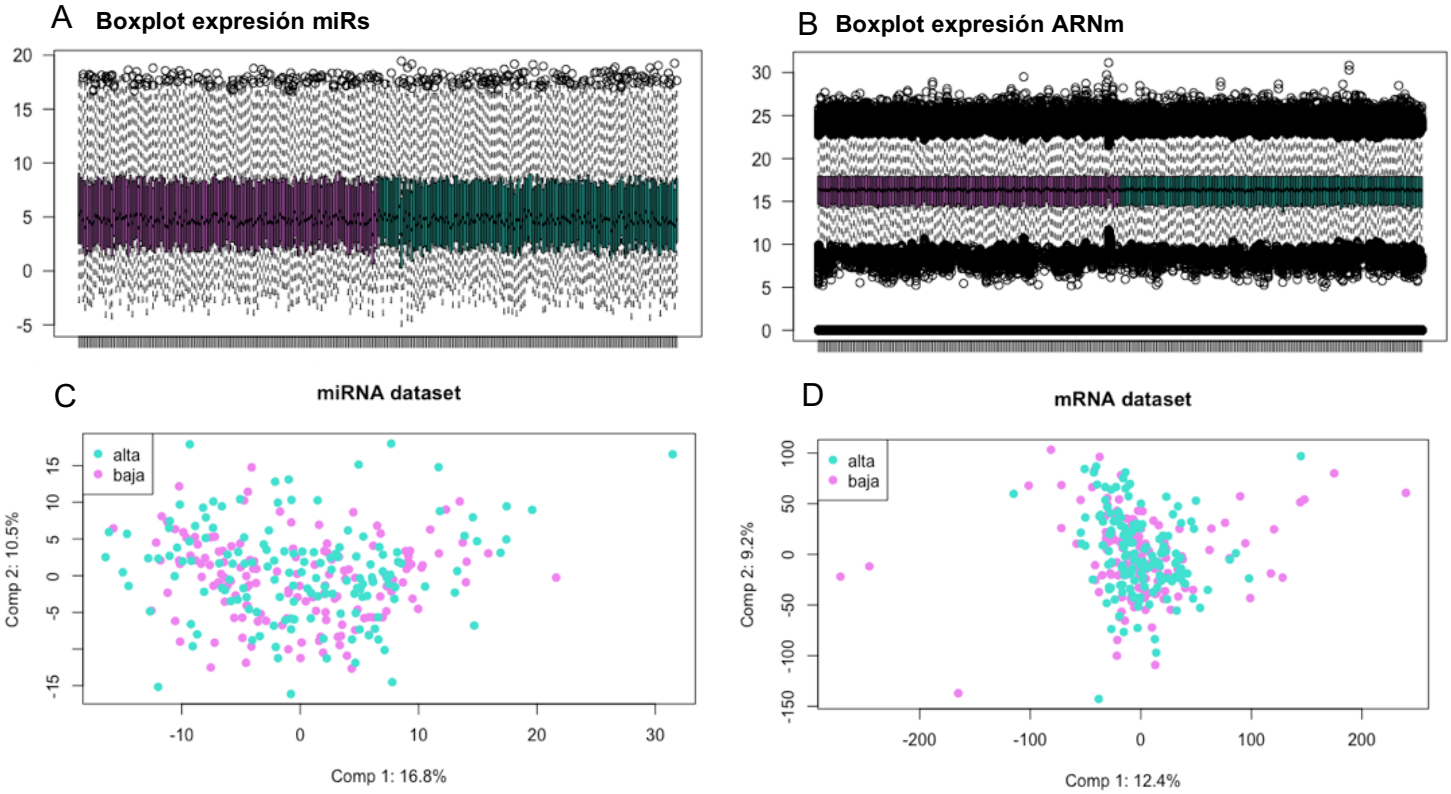


Gráfica 9 Ajuste Voom - Desviación estándar de la expresión de miRs para el análisis de alta/baja supervivencia



Gráfica 10 Distribución de la densidad de expresión de miRs y ARNm en el anàlisis según alta/baja supervivencia.

Creamos otro objeto *corObject* para el análisis por el paquete miRComb y cargamos las matrices de conteos (tabla con los conteos de miRs archivo 'Expr-miRs-Superv.txt' y de ARNm 'Expr-miRs-Superv.txt'). Observamos los datos antes de procesarlos para ver la distribución de la densidad de expresión en todas las muestras (gráfica 10) o la distribución en cada muestra mediante boxplot (gráfica 11.A y 11.B). En el análisis de componentes principal vemos como el primer componente de la expresión de miRs explica un 16.8% de la varianza entre los dos grupos de estudio (gráfica 11.C) y un 12.4% de la varianza en cuanto a la expresión de ARNm (gráfica 11.D). Como en el análisis anterior apenas hay diferencias, no vemos una tendencia al agrupamiento de las muestras según el tipo de supervivencia, alta o baja.



Gráfica 11 Distribución de las muestras según baja/alta supervivencia. (A) Boxplot de la expresión de miRs. (B) Boxplot de la expresión de ARNm. (C) PCA de expresión de miRs. (D) PCA de expresión de ARNm

Realizamos el análisis de expresión diferencial de miRs y ARNm por el modelo lineal de limma, con un ajuste por el método de Benjamini y Hochberg's. Vemos como hay muy pocos miRs expresados diferencialmente y que estos varían su FC muy poco entre las dos poblaciones y como se ha distorsionado un p-valor ajustado tras el análisis con miRComb (Tablas de la gráfica 12.A y 12.B, archivo adjunto 'DifExpr-miRs-Superv.txt'). Algún algoritmo interno del paquete de miRComb está dando problemas y los p-valores ajustados están distorsionados, por lo que seleccionamos los genes según su p-valor. En el *heatmap* de la gráfica 12.C observamos como las muestras no parecen tener perfiles de expresión génica característico de una supervivencia alta o baja.

A

Table: miRs downregulated FC>1 y pval<0.05 según supervivencia

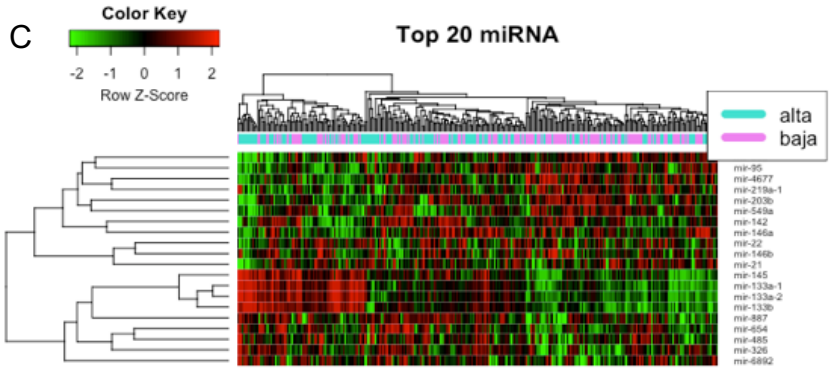
	FC	logratio	meanExp	pval	adj.pval
hsa-mir-549a	-1.3022	-0.3810	0.3384	0.0198	0.8073
hsa-mir-219a-1	-1.2388	-0.3090	1.7699	0.0097	0.7393
hsa-mir-95	-1.2224	-0.2897	3.5600	0.0435	0.8073
hsa-mir-4677	-1.2119	-0.2773	2.6760	0.0041	0.7393
hsa-mir-146a	-1.2090	-0.2738	7.5691	0.0327	0.8073
hsa-mir-20a	-1.2013	-0.2646	8.8518	0.0384	0.8073
hsa-mir-146b	-1.1839	-0.2435	8.8599	0.0082	0.7393
hsa-mir-21	-1.1249	-0.1698	17.5283	0.0299	0.8073
hsa-mir-22	-1.0904	-0.1249	15.3779	0.0443	0.8073

B

Table: miRs upregulated FC>1 y pval<0.05 según supervivencia

	FC	logratio	meanExp	pval	adj.pval
hsa-mir-133b	1.4919	0.5771	1.0902	0.0143	0.8073
hsa-mir-133a-1	1.4731	0.5589	2.7721	0.0050	0.7393
hsa-mir-133a-2	1.4089	0.4946	2.6406	0.0106	0.7393
hsa-mir-326	1.2578	0.3309	1.7534	0.0284	0.8073
hsa-mir-145	1.2251	0.2928	10.9303	0.0234	0.8073

C



Gráfica 12 Análisis de expresión diferencial de miRs según alta/baja supervivencia. (A) Tabla de miRs regulados a la baja, con FC>1 y con un p-valor <0.05. (B) Tabla de miRs regulados a la alta, con FC>1 y con un p-valor <0.05. (C) HeatMap de la expresión de 20miRs en las 292 muestras de TCGA-COAD.

Algo similar ocurre con la expresión diferencial de los ARNm. Observamos pocos genes expresados diferencialmente con unos p-valores elevados aunque menores a 0.05 (gráfica 13.A y 13.B, archivo adjunto 'DifExpr-ARNm-Superv.txt'). El *heatmap* de la gráfica 13.C tampoco parece indicar perfiles de expresión génica característicos de un tipo de supervivencia.

A

Table: mRNA upregulated FC>2 y pval<0.05 según supervivencia

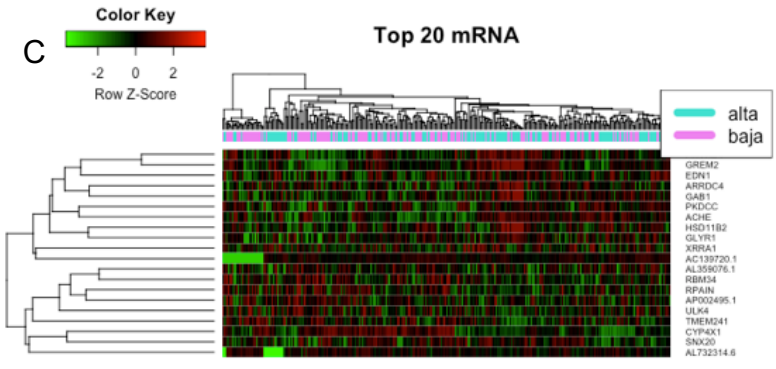
	FC	logratio	meanExp	pval	adj.pval
AL022718.1	3.3277	1.7345	13.7360	0.0263	1
RPS4Y1	3.2629	1.7062	15.3344	0.0161	1
RNU6.353P	3.0437	1.6058	9.0565	0.0355	1
DDX3Y	2.9860	1.5782	11.6993	0.0326	1
AL355309.1	2.9599	1.5655	12.8178	0.0476	1
SCARNA22	2.8839	1.5280	10.8123	0.0256	1
MIR6835	2.8462	1.5090	10.9037	0.0478	1
AC139720.1	2.8127	1.4919	11.8635	0.0016	1
MIR4635	2.7674	1.4685	11.2291	0.0432	1
AL589843.1	2.5296	1.3389	11.6497	0.0095	1
URAD	2.3994	1.2626	14.5817	0.0378	1
TRAV12.2	2.3961	1.2607	11.3774	0.0226	1
RF00561.15	2.3450	1.2296	12.2110	0.0426	1
SNORA59A	2.3433	1.2286	12.4239	0.0176	1
RF00410.1	2.3366	1.2244	13.1937	0.0130	1
TRAV2	2.3303	1.2205	10.9908	0.0305	1
MT1DP	2.3179	1.2128	12.0457	0.0343	1
AC006026.3	2.2770	1.1871	13.2555	0.0313	1
AC012360.2	2.1950	1.1342	11.8406	0.0149	1
UGT2A3	2.1190	1.0834	14.6571	0.0079	1
AC135507.1	2.1033	1.0727	11.7080	0.0049	1
AL133215.1	2.0570	1.0406	11.9113	0.0243	1
GREM2	2.0453	1.0323	13.0599	0.0008	1
C3orf85	2.0085	1.0061	13.6044	0.0091	1
AP006748.1	2.0035	1.0025	13.7451	0.0152	1

B

Table: mRNA downregulated FC>2 y pval<0.05 según supervivencia

	FC	logratio	meanExp	pval	adj.pval
MIR5581	-2.8870	-1.5296	10.9190	0.0453	1
AL442067.1	-2.4953	-1.3192	10.3596	0.0444	1
LINC02163	-2.4019	-1.2642	12.5716	0.0294	1
AL732314.6	-2.2809	-1.1896	12.1190	0.0006	1
REG1A	-2.1006	-1.0708	18.4441	0.0402	1

C



Gráfica 13 Análisis de expresión diferencial de ARNm según alta/baja supervivencia. (A) Tabla de ARNm regulados a la alta, con FC>2 y con un p-valor <0.05. (B) Tabla de ARNm regulados a la baja, con FC>2 y con un p-valor <0.05. (C) HeatMap de la expresión de 20ARNm en las 292 muestras de TCGA-COAD

Incorporamos el análisis de correlación negativa de miRs-ARNm con los p-valores correspondientes como se observa en la tabla 9.

Table: Correlación negativa miRs-ARNm según supervivencia

	TSPAN6	TNMD	DPM1	SCYL3	C1orf112	FGR	CFH	FUCA2	GCLC	NFYA
hsa-let-7a-1	-0.148	-0.064	-0.017	-0.054	-0.186	-0.082	0.107	-0.161	-0.020	0.090
hsa-let-7a-2	-0.148	-0.064	-0.017	-0.057	-0.187	-0.080	0.108	-0.156	-0.016	0.092
hsa-let-7a-3	-0.148	-0.062	-0.016	-0.054	-0.186	-0.080	0.109	-0.157	-0.018	0.091
hsa-let-7b	-0.098	-0.064	-0.127	0.109	0.019	0.200	0.281	-0.092	-0.125	-0.098
hsa-let-7c	-0.026	0.069	-0.071	0.030	-0.169	0.254	0.419	-0.091	-0.022	-0.014
hsa-let-7d	0.025	-0.045	-0.085	0.050	0.093	-0.186	-0.219	0.027	-0.022	-0.021
hsa-let-7e	-0.150	-0.066	-0.003	-0.063	-0.212	-0.007	0.217	-0.139	0.006	0.047
hsa-let-7f-1	-0.067	-0.043	0.099	-0.171	-0.184	-0.144	-0.010	-0.080	0.057	0.120
hsa-let-7f-2	-0.069	-0.046	0.101	-0.169	-0.183	-0.145	-0.011	-0.083	0.056	0.119
hsa-let-7g	0.196	0.039	0.135	0.031	0.130	-0.414	-0.421	-0.012	0.006	0.001
hsa-let-7i	-0.189	-0.070	-0.134	-0.152	-0.157	0.403	0.377	-0.009	-0.127	-0.137
hsa-mir-1-1	0.047	0.081	0.017	-0.026	-0.138	0.046	0.123	-0.098	0.030	0.061
hsa-mir-1-2	0.060	0.102	0.005	-0.052	-0.169	0.066	0.141	-0.104	0.001	0.023
hsa-mir-100	-0.016	0.132	-0.078	0.067	-0.228	0.357	0.465	-0.132	0.010	-0.008
hsa-mir-101-1	0.250	0.116	0.146	0.045	0.071	-0.039	-0.106	-0.008	0.104	0.057
hsa-mir-101-2	0.249	0.114	0.146	0.044	0.068	-0.043	-0.107	-0.006	0.107	0.059
hsa-mir-103a-1	-0.121	-0.058	0.048	-0.066	-0.138	-0.091	-0.053	-0.005	0.032	0.068
hsa-mir-103a-2	-0.120	-0.058	0.049	-0.066	-0.137	-0.090	-0.053	-0.004	0.032	0.067
hsa-mir-106a	0.310	0.057	0.106	0.098	0.343	-0.043	-0.182	0.147	0.052	-0.078
hsa-mir-106b	0.052	-0.077	0.138	-0.157	0.133	-0.227	-0.350	0.098	0.096	0.031

Tabla 9 Muestra de correlaciones negatives miRs-ARNm según alta/baja supervivencia

Finalmente introducimos la información de las bases de datos de microCosm y TargetScan. Ordenando los resultados por la suma de la puntuación de dianas en bases de datos, encontramos algunos miRs diferencialmente sobreexpresados con unión putativa en varios ARNm diferencialmente infraexpresados, tabla 10.

Table: Resumen de puntuaciones en las interacciones miRs-ARNm según supervivencia

miRNA	mRNA	cor	pval	logratio.miRNA	meanExp.miRNA	logratio.mRNA	meanExp.mRNA	dat.microCosm_v5_18	dat.targetScan_v6.2_18	dat.sum	score
hsa-let-7b	PRSS22	-0.082	0.081	-0.039	13.121	-0.274	17.449	1	1	2	-0.021
hsa-let-7b	PLEKHG6	-0.033	0.289	-0.039	13.121	0.211	18.332	1	1	2	0.017
hsa-let-7b	MAP4K3	-0.039	0.254	-0.039	13.121	0.057	17.059	1	1	2	0.004
hsa-let-7b	ZC3H3	-0.107	0.034	-0.039	13.121	0.007	17.776	1	1	2	0.001
hsa-let-7b	RUFY3	-0.025	0.338	-0.039	13.121	-0.027	15.599	1	1	2	-0.002
hsa-let-7b	PLEKHO1	0.175	0.999	-0.039	13.121	0.095	16.158	1	1	2	0.007
hsa-let-7b	PQLC2	0.000	0.503	-0.039	13.121	-0.017	16.828	1	1	2	-0.001
hsa-let-7b	CTNS	0.139	0.991	-0.039	13.121	-0.076	16.440	1	1	2	-0.006
hsa-let-7b	MGAT4A	0.017	0.615	-0.039	13.121	0.024	17.111	1	1	2	0.002
hsa-let-7b	SSH1	0.065	0.866	-0.039	13.121	0.014	16.111	1	1	2	0.001
hsa-let-7b	CHRD	0.168	0.998	-0.039	13.121	0.020	13.300	1	1	2	0.002
hsa-let-7b	DTX2	-0.106	0.035	-0.039	13.121	0.050	18.053	1	1	2	0.004
hsa-let-7b	CDC34	-0.126	0.016	-0.039	13.121	0.062	19.785	1	1	2	0.005
hsa-let-7b	KCTD17	0.016	0.606	-0.039	13.121	-0.080	17.449	1	1	2	-0.006
hsa-let-7b	ELF4	0.109	0.969	-0.039	13.121	0.138	18.490	1	1	2	0.011
hsa-let-7b	OLFM4	0.080	0.913	-0.039	13.121	0.136	21.406	1	1	2	0.011
hsa-let-7b	NME4	-0.155	0.004	-0.039	13.121	0.010	18.690	1	1	2	0.001
hsa-let-7b	BRF2	0.128	0.985	-0.039	13.121	-0.117	16.111	1	1	2	-0.009
hsa-let-7b	PLD3	0.087	0.931	-0.039	13.121	0.030	19.333	1	1	2	0.002
hsa-let-7b	AP1S1	-0.065	0.136	-0.039	13.121	-0.023	20.703	1	1	2	-0.002

Tabla 10 Resumen de las puntuaciones miRs-ARNm en el análisis de expresión diferencial según alta/baja supervivencia

Podemos visualizar cuantos miRs poseen dianas en genes del estudio y visualizar los nombres de las dianas en la tabla 11.

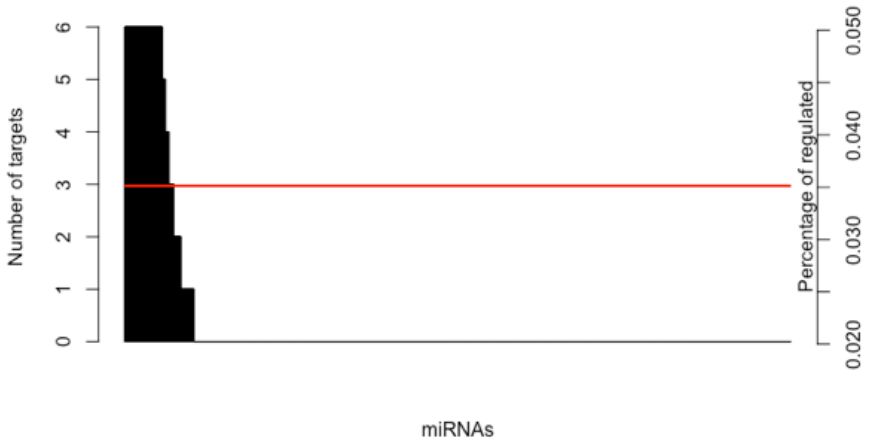
hsa-let-7i	hsa-let-7g	hsa-let-7d	hsa-let-7b	hsa-let-7e	hsa-let-7c	hsa-let-7a-1	hsa-let-7a-2	hsa-let-7a-3	hsa-let-7f-1
1281	1274	1268	1254	1239	1231	0	0	0	0

	freq <dbl>	names <fctr>	perc.regulated <dbl>
hsa-let-7i	1281	SCYL3, C1orf112, SEMA3F, LASP1, TFPI, PLXND1, PRSS22, SLC25A5, ZFX, LAMP2, YBX2, TTC22, SCIN, ADIPOR2, GAS7, ST7L, RHBDF1, TEAD3, E2F2, CYTH3, ADAM22, AASS, PLEKHG6, ST3GAL1, TRAF3IP3, ...	7.500439
hsa-let-7g	1274	SCYL3, SEMA3F, HS3ST1, LASP1, TFPI, PLXND1, PRSS22, SLC25A5, ZFX, LAMP2, YBX2, SCIN, ADIPOR2, GAS7, ST7L, TEAD3, E2F2, CYTH3, AASS, PLEKHG6, ST3GAL1, TRAF3IP3, STARD3NL, IDS, ZNF200, LYP...	8.343580
hsa-let-7d	1268	SCYL3, SEMA3F, TFPI, PLXND1, PRSS22, ZFX, LAMP2, YBX2, ADIPOR2, GAS7, ST7L, RHBDF1, TEAD3, E2F2, CYTH3, AASS, PLEKHG6, ST3GAL1, TRAF3IP3, STARD3NL, PHF7, ZNF200, LYPLA2, SYT7, PLAUR, M...	9.233562
hsa-let-7b	1254	SCYL3, SEMA3F, TFPI, PLXND1, PRSS22, LAMP2, YBX2, ADIPOR2, GAS7, ST7L, RHBDF1, TEAD3, E2F2, CYTH3, AASS, PLEKHG6, ST3GAL1, TRAF3IP3, STARD3NL, ZNF200, HFE, SYT7, PLAUR, MAP4K3, MBTPS2...	9.807366
hsa-let-7e	1239	SCYL3, C1orf112, SEMA3F, PLXND1, PRSS22, ZFX, LAMP2, YBX2, ADIPOR2, GAS7, ST7L, TEAD3, E2F2, CYTH3, AASS, PLEKHG6, ST3GAL1, TRAF3IP3, STARD3NL, PHF7, IDS, ZNF200, SYT7, PLAUR, MAP4K3, ...	10.217226
hsa-let-7c	1231	SCYL3, SEMA3F, TFPI, PLXND1, PRSS22, LAMP2, YBX2, ADIPOR2, GAS7, ST7L, RHBDF1, TEAD3, E2F2, CYTH3, AASS, PLEKHG6, ST3GAL1, TRAF3IP3, STARD3NL, PHF7, ZNF200, LYPLA2, SYT7, PLAUR, MAP4K...	10.305053
hsa-let-7a-1	0		10.305053

7 rows | 1-2 of 3 columns

Tabla 11 Top 10 miRs con dianas en ARNm. Nombre identificativo del miR y número de dianas putativas en ARNm. Resumen de miRs con dianas en ARNm tras el análisis de alta/baja supervivencia. (archivos Top10-miRs-frecARNm-Superv.txt y Top10-miRs-namesARNm-Superv.txt)

A la inversa, también podemos en la gráfica 14 como varía el número de dianas en los ARNm para estos 6 miRs diferencialmente expresados. Si nos fijamos el resultado obtenido corresponde a un único miRs 'hsa-let-7' y a sus diferentes isoformas/parálogos a lo largo del genoma humano, de ahí que no sorprenda que los genes diana sean muy parecidos, o que todos los genes resultantes posean dianas para todos o casi todos estos parálogos.

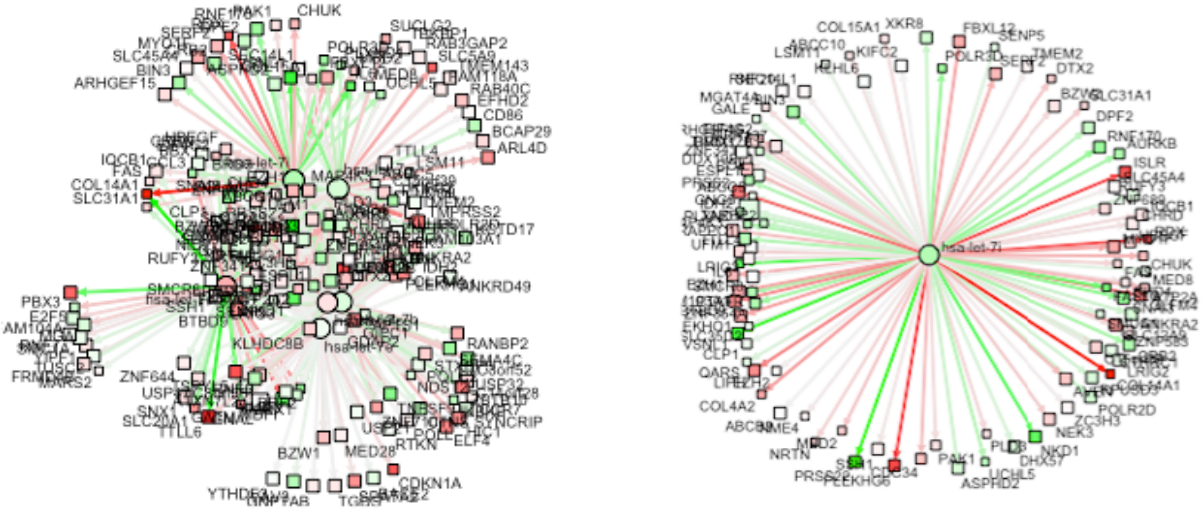


Gráfica 14 Número de dianas en los ARNm para miRs. La linea roja representa el porcentaje acumulado de dianas

Graficamos la red que se forma de los miRs resultantes con todas sus dianas de las interacciones con un p-valor<0.01 (gráfica 15.A) o las dianas de uno de los miRs 'hsa-let-7i' (gráfica 15.B). Cada interacción mostrada en el gráfico 15 es una correlación negativa y predicha en al menos las dos bases de datos (microCosm y TargetScan). Los círculos representan los miRs y los cuadrados los ARNm; el relleno rojo significa una regulación a la alza miR/ARNm, mientras que si es verde está regulado a la baja; las líneas rojas indican puntuaciones positivas y las verdes negativas; el tamaño de la flecha es proporcional al numero de ocurrencias en las bases de datos testadas.

A. NetWork miRs

B. NetWork 'hsa-let-7i'



Gráfica 15 Red de interacciones miRs-ARNm tras el análisis según alta/baja supervivencia en muestras de TCGA-COAD

3.6. Enriquecimiento biológico GO de genes diferencialmente expresados según la supervivencia en pacientes con COAD

Realizamos el análisis de significación biológica de los 480 ARNm computados en el análisis con miRComb con un p.val<0.05 (recordemos que los valores adj-p.val se han distorsionado). Ordenamos ascendentemente los genes por el p-valor y obtenemos la información de 470 genes con el GO y symbol (archivo adjunto en anexos 'GO-supervivencia.txt').

En los primeros puestos encontramos genes implicados en cáncer (proto-oncogenes), también recetores y genes de muerte celular. Forma parte del trabajo del biólogo investigador el posterior análisis detallado y significación biológica de estos resultados.

4. Conclusiones

- El análisis por miRComb de ARNm diferencialmente infraexpresados con dianas para miRs diferencialmente sobreexpresados según el estado mutacional de KRAS, en pacientes del repositorio TCGA-COAD, no muestra resultados positivos.

- El análisis de enriquecimiento biológico muestra 2341 genes diferencialmente expresados, según el estado mutacional de KRAS, que participan en rutas de muerte celular, cáncer de colon, señalización celular y oncogenes. Estos podrían ser predictores moleculares que ayuden a determinar la relación del tumor con mutaciones en KRAS.

- La supervivencia de los pacientes con COAD no parece ser dependiente del estado mutacional de KRAS.

- El análisis por miRComb de ARNm diferencialmente infraexpresados con dianas para miRs diferencialmente sobreexpresados según la supervivencia de los pacientes del repositorio TCGA-COAD, detecta seis isoformas/parálogos del hsa-let-7 como miR diferencialmente sobreexpresados con dianas en varios ARNm diferencialmente infraexpresados.

- El análisis de enriquecimiento biológico muestra 470 genes diferencialmente expresados, según la supervivencia de los pacientes, que participan en rutas de muerte celular, señalización celular y proto-oncogenes. Estos podrían ser predictores moleculares que ayuden a determinar la relación del tumor con la tasa de supervivencia.

El paquete miRComb no arroja resultados positivos o muy pocos resultados en los dos análisis ejecutados. Este paquete es relativamente nuevo, creación en 2015 (Vila-casades, 2015), por lo que todavía se encuentra en construcción y *debugging* (depuración de las funciones). Esto origina que algunas funciones, como la obtención final del informe (*mkReport*) no puedan ser ejecutadas desde RStudio.

Una de las principales limitaciones observadas en las funciones del paquete miRComb para el análisis del objeto *corObject* es que se encuentran muy acotadas, no dejando al usuario seleccionar los argumentos a su criterio. La falta de personalización de los análisis ante la ausencia de resultados, impide concretar el origen del error o posible fallo de los datos iniciales, como ha ocurrido con el análisis de enriquecimiento biológico de GO. Esto nos ha obligado al uso de otros paquetes (*AnnotationDbi*) para completar el análisis GO sin el uso de miRComb; y a cambiar la estrategia del trabajo ante la falta de resultados, analizando los datos de expresión génica según la variable supervivencia para completar la memoria.

Planteamos la posibilidad de que la ausencia de resultados se deba a la elección de los datos de conteos en FPKM-UQ de los ARNm, y las aproximaciones realizadas para el pre-filtrado y ajuste de las matrices. Como ya hemos comentado, se intenta la descarga de HTSeq-Count de los genes, pero la ejecución del código resulta en un error con la conexión de la web GCD *Data Portal*. Este error concuerda con la última actualización del portal, 21 de Mayo de 2018, saltando un aviso en la propia web de los cambios realizados y la nueva política de protección de datos. Se permite la descarga directa de los datos desde el servidor web al directorio de trabajo, pero se desconoce la posibilidad

de alguna función que permita la subida de dicha información al *enviroment* de RStudio. Como continuidad en la línea de trabajo se propone realizar el análisis con los datos de los conteos crudos de ARNm.

Para mejorar el análisis bioinformático planteamos en un futuro estudiar en detalle las funciones del paquete miRComb que nos puedan estar llevando a error, para ello proponemos computar de forma paralela el mismo o semejantes análisis con las librerías para expresión diferencial de genes *DESeq2*, *edgeR* o *limma*. Así mismo podemos realizar un análisis paralelo para verificar las dianas de miRs-ARNm en los genes seleccionados tras el análisis con paquetes específicos como RmiR, CORNA, miRNApath, microRNA o MultiMiR (Vila-Casadesús et al., 2016).

La validación y verificación de las interacciones miRs-ARNm se planea como otra posible línea de trabajo futuro. La validación del RNAseq se realiza por RTqPCR (*Quantitative Reverse Transcription Polymerase Chain Reaction)* de los genes, tanto miRs como ARNm. Posteriormente se procede a la verificación funcional de la diana en la 3'UTR del ARNm para el miR estudiado mediante experimentos de perdida y recuperación de la actividad luciferasa en plásmidos (Kuhn et al., 2008; Yi Jin, Zujian Chen, Xiqiang Liu, 2013).

En resumen, esta memoria presenta resultados preliminares del análisis de expresión diferencial de genes, miRs y ARNm, en pacientes de COAD según el estado mutacional de oncogén KRAS y de la supervivencia. También se muestra la línea de trabajo a seguir en los estudios de interacciones de miRs con ARNm con el paquete miRComb. Y finalmente se muestra un resultado preliminar de enriquecimiento biológico con las funciones de los genes alterados de forma significativa; estos podrían ser predictores moleculares que ayuden a determinar la relación del tumor con mutaciones en KRAS y la tasa de supervivencia.

5. Glosario

ARNm – ARN mensajero
BH - *Benjamini–Hochberg*
COAD - *Colon Adenocarcinoma*
cpm – *count per million*
GO – *Gene Ontology*
FC – *Fold Change*
FDR – *False Discovery Rate*
miRs - microRNAs
TCGA - *The Cancer Genome Atlas*

6. Bibliografía

American Cancer Society. (2017). Colorectal Cancer Facts & Figures 2017 - 2019. *Atlanta*, 1–40. https://doi.org/http://dx.doi.org/10.1016/S0140-6736(13)61649-9

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. https://doi.org/10.1093/nar/gkv1507

Jonas, S., & Izaurralde, E. (2015). Towards a molecular understanding of microRNA-mediated gene silencing. *Nature Reviews Genetics*, *16*(7), 421–433. https://doi.org/10.1038/nrg3965

Knickelbein, K., & Zhang, L. (2015). Mutant KRAS as a critical determinant of the therapeutic response of colorectal cancer. *Genes and Diseases*, *2*(1), 4–12. https://doi.org/10.1016/j.gendis.2014.10.002

Kuhn, D. E., Martin, M. M., Feldman, D. S., Terry, A. V., Nuovo, G. J., & Elton, T. S. (2008). Experimental validation of miRNA targets. *Methods*, *44*(1), 47–54. https://doi.org/10.1016/j.ymeth.2007.09.005

Liu, X., Jakubowski, M., & Hunt, J. L. (2011). KRAS gene mutation in colorectal cancer is correlated with increased proliferation and spontaneous apoptosis. *American Journal of Clinical Pathology*, *135*(2), 245–252. https://doi.org/10.1309/AJCP7FO2VAXIVSTP

Markman, M. (2012). Colorectal Cancer and KRAS / BRAF. *Medscape*, (July), 6–10.

Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., Thomson., E. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*(7407), 330–337. https://doi.org/10.1038/nature11252

Phipps, A. I., Buchanan, D. D., Makar, K. W., Win, A. K., Baron, J. A., Lindor, N. M., … Newcomb, P. A. (2013). KRAS-mutation status in relation to colorectal cancer survival: The joint impact of correlated tumour markers. *British Journal of Cancer*, *108*(8), 1757–1764. https://doi.org/10.1038/bjc.2013.118

Phipson, A. B., Trigos, A., Ritchie, M., Doyle, M., Dashnow, H., & Law, C. (2016). RNA-seq analysis in R Differential expression analysis, (November), 1–57.

Portal, GDC (web-pdf). GDC Data User ' s Guide. *https://docs.gdc.cancer.gov/Data_Portal/Users_Guide/Getting_Started/*.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Roa, I., Sanchez, T., Majlis, A., & Schalper, K. (2013). [KRAS gene mutation in colorectal cancer]. *Revista Medica de Chile*, *141*(9), 1166–1172. https://doi.org/10.4067/S0034-98872013000900009

Robinson, M., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. Retrieved from http://www.biomedcentral.com/content/pdf/gb-2010-11-3-r25.pd

Smyth, G. K., Ritchie, M., & Thorne, N. (2015). Linear Models for Microarray and RNA-Seq Data User ' s Guide. *R*, (September). https://doi.org/10.1093/nar/gkv007

Society, A. C. S. A. C. (2012). Colorectal Adenocarcinoma, *37*, 1–25.

Tan, C., & Du, X. (2012). KRAS mutation testing in metastatic colorectal cancer. *World Journal of Gastroenterology*, *18*(37), 5171–5180. https://doi.org/10.3748/wjg.v18.i37.5171

Vila-casades, M. (2015). miRComb - An R package for analyzing miRNA-mRNA interactions Contents, 1–17.

Vila-Casadesús, M., Gironella, M., & Lozano, J. J. (2016). MiRComb: An R package to analyse miRNA-mRNA interactions. Examples across five digestive cancers. *PLoS ONE*, *11*(3), 1–18. https://doi.org/10.1371/journal.pone.0151127

Yi Jin, Zujian Chen, Xiqiang Liu, and X. Z. (2013). Evaluating the MicroRNA Targeting Sites by Luciferase Reporter Gene Assay. *Methods Mol Biol.*, *936*(4), 117–127. https://doi.org/10.1007/978-1-62703-083-0

7. Anexos

- Código ejecutado en R, archivo adjunto en formato MarkDown 'Codigo_TFM_LaraRguezOute_Bioinfo.Rmd'.
- Resultados de tablas en archivos formato txt.
- Diagrama de Grant con la planificación del TFM
- Instalación de la librería miRComb a través del siguiente código R:

```{r}
install.packages(c("gplots","gtools","network","WriteXLS","Hmisc","glmnet","scatterplot3d", "VennDiagram","xtable","survival","pheatmap","mvoutlier","mclust"))
source("http://www.bioconductor.org/biocLite.R")
biocLite(c("RankProd","GOstats","limma","RamiGO","KEGG.db","circlize","ReactomePA","DESeq","DO.db"))

library(devtools)
if(!require(miRData)) install_github("mariavica/miRData", ref="master", build_vignettes = FALSE)
install_github("mariavica/miRComb", ref="master", build_vignettes = FALSE)
```