Universitat Oberta de Catalunya

uoc.edu

Exploración del transcriptoma de la leucemia linfoblástica aguda de células T (T-ALL) y el papel de la expresión de las RBPs en el fenotipo de la enfermedad

Cristina Muntañola Valero

Máster Universitario en Bioinformática y Bioestadística

Área 1, Subárea 10: genómica comparativa

Nombre Consultor/a: Yolanda Guillén Montalbán

Nombre Profesor/a responsable de la asignatura: David Merino

Fecha Entrega 2/06/2022

FICHA DEL TRABAJO FINAL

Título del trabajo	*Exploración del transcriptoma de la leucemia linfoblástica aguda de células T (T-ALL) y el papel de la expresión de las RBPs en el fenotipo de la enfermedad*
Nombre del autor	*Cristina Muntañola Valero*
Nombre del consultor/a	*Yolanda Guillén Montalbán*
Nombre del PRA	*David Merino*
Fecha de entrega	*Máster Universitario en Bioinformática y Bioestadística UOC-UB*
Titulación	*Area 1, subárea 10: genómica comparativa*
Área del Trabajo Final	*Español*
Idioma del trabajo	*15*
Número de créditos	*Leucemia, ómicas, proteínas, empalme, ARN*
Palabras clave	*Exploración del transcriptoma de la leucemia linfoblástica aguda de células T (T-ALL) y el papel de la expresión de las RBPs en el fenotipo de la enfermedad*

Resumen del Trabajo

La leucemia linfoblástica aguda de células T (T-ALL) es una enfermedad caracterizada por el sobrecrecimiento de linfocitos T, que suele presentar un mal pronóstico. Esto se debe en parte a la falta de conocimiento de los mecanismos moleculares detrás de la enfermedad. Dado que se ha visto que los procesos de empalme de ARNm mediados por RBPs son importantes en otros procesos cancerígenos, diseñamos un estudio basado en un análisis de la expresión de los genes asociados con T-ALL para esclarecer el papel de las RBPs en la enfermedad. Para ello recolectamos, unificamos y procesamos datos transcriptómicos de T-ALL procedentes de bases de datos públicas que servirán como catálogo de referencia para para futuros estudios. Se desarrolló un análisis bioinformático del perfil de expresión génica de distintos grupos de pacientes con leucemia (fenotipos) mediante el cual 1) identificamos genes diferencialmente expresados entre varias condiciones de T-ALL, y 2) anotamos todos aquellos correspondientes a RBPs usando como referencia estudios anteriores. Al final comprendimos que, a pesar de la gran cantidad de RBPs desreguladas en T-ALL, de muchas de ellas no se conoce su papel bioquímico o celular y se necesitarán más estudios para conocer en profundidad su papel en la enfermedad. El código para llegar

a nuestros resultados se da a conocer, para entender los resultados, generar conocimiento, y para nuevas búsquedas de funciones biológicas en nuestras bases de datos de pacientes con T-ALL.

Abstract

T-cell acute lymphoblastic leukemia (T-ALL) is a disease characterized by the overgrowth of T lymphocytes, whose patients usually has a poor prognosis. This is due in part to a lack of understanding of the molecular mechanisms behind the disease. Since RBP-mediated mRNA splicing processes have been shown to be important in other cancer processes, we designed a study based on an expression analysis of T-ALL-associated genes to elucidate the role of RBPs in the disease. For this goal we collect, unify and process T-ALL transcriptomic data from public databases that will serve as a reference catalog for future studies. A bioinformatic analysis of the gene expression profile of different groups of leukemia patients (phenotypes) was developed by which we 1) identified differentially expressed genes among several T-ALL conditions, and 2) annotated all those corresponding to RBPs using as reference studies. previous. In the end, we understood that, despite the large number of deregulated RBPs in T-ALL, the biochemical or cellular role of many of them is not known and more studies will be needed to understand their role in the disease in depth. The code to came at our results is disclosed, to understand the results, generate knowledge, and for new searches for biological functions in our T-ALL patient databases.

Índice

Lista de figuras

Lista de tablas

1 Resumen

Antecedentes. La leucemia linfoblástica aguda (ALL) de células T (T-ALL) es una neoplasia maligna hematopoyética común cuyo comienzo y desarrollo no se entienden al completo a nivel bioquímico y celular. Los pacientes que la padecen se pueden clasificar en grupos con respecto a análisis inmunofenotípicos, y para algunos de ellos no existe un tratamiento claro, de forma que tienen un mal pronóstico clínico.

Existen muchos procesos por los que se cree que la enfermedad aparece y se desarrolla, tanto a nivel genético, como transcripcional y postranscripcional. Una parte del proceso de maduración del pre-ARNm se denomina empalme, y está mediado por unas proteínas llamadas RNA binding proteins (RBPs). Recientemente se han vinculado estos eventos de empalme alternativo de pre-ARNm con cánceres y neoplasias.

Así, nuestra hipótesis es que los procesos mediados por estas proteínas son claves en la enfermedad. Pensamos que un estudio sobre la comprensión de las RBPs facilitará el entendimiento de la enfermedad, ayudará al desarrollo de terapias más efectivas y, en última instancia, aumentará la esperanza de vida de los pacientes con T-ALL.

Método. A raíz de ello, nuestros propósitos son determinar las RBPs que presentan una desregulación génica entre grupos de individuos con T-ALL, y también esclarecer y dar a conocer la influencia de estas RBPs en las vías metabólicas del desarrollo de T-ALL.

Nuestro enfoque metodológico se basó en un estudio sobre el perfil de expresión genético de pacientes con T-ALL, a través de comparaciones entre grupos de pacientes. Se analizaron datos de microarrays y de RNA-sequencing (RNA-seq) a través de un análisis bioinformático. Generamos resultados de listas de genes expresados diferencialmente anotadas. A partir de ellas se siguió con otro tipo de enriquecimiento a partir de cuatro bases de datos de tres artículos genéticos. Por úlitmo, a partir de los últimos resultados y de los anteriores, se seleccionaron los genes anotados como proteínas implicadas en procesos de empalme de ARNm.

Resultados. Como resultado obtuvimos una serie de tablas con listas de genes y proteínas. Determinamos, a través de un análisis visual, qué genes, implicadas en procesos de empalme, eran los que aparecían como desregulados de forma más recurrente entre comparaciones de grupos con leucemia. Entonces, pudimos comprobar que existían muchas proteínas implicadas en procesos de empalme alternativo de ARN asociados con T-ALL.

Todos los resultados se encuentran disponibles para su descarga y uso, ya que quisiéramos que se generara otras investigaciones a partir de nuestro estudio. Se ha proporcionado también un documento para la interrogación de nuestras bases de datos globales, de forma que se ofrece la posibilidad de buscar otro tipo de funciones en los genes y proteínas expresados diferencialmente.

Conclusiones. Algunas de las proteínas expresadas diferencialmente según nuestras comparaciones ya se habían estudiado. Ciertas proteínas se ha visto que eran importantes para procesos relacionados con otros cánceres, y otras en neoplasias hematopoyéticas. Sin embargo, todavía no se ha estudiado muchas sobre ellas, y no se sabe qué papel ejercen en la T-ALL. Por ello, requerimos de más estudios que nos ayuden

a entender el papel de los genes y RBPs relacionados con procesos de empalme alternativo a la hora de determinar los perfiles de los grupos de pacientes con T-ALL.

2 Introducción

2.1 Contexto y justificación del trabajo

La leucemia linfoblástica aguda (ALL) es una neoplasia maligna que se caracteriza por la proliferación masiva de células linfoides (1). Es muy común, tanto en niños como en adultos (1,2), y es la neoplasia maligna más frecuente en el caso de los niños a pesar de que su patogenia sigue sin ser clara (3).

En concreto, la leucemia linfoblástica aguda de células T (T-ALL) conlleva el aumento descontrolado de células T, inmaduras o maduras dependiendo del fenotipo (4). Este tipo de ALL se ha atribuido a distintas alteraciones genéticas (5). Existen varias clasificaciones sobre los tipos de T-ALL, según diferentes criterios. La T-ALL se suele dividir según las firmas de expresión de los timocitos en pro-T, pre-T o immadura, cortical y madura (6).

Hay distintos factores asociados a la generación y/o el mantenimiento de T-ALL, no solo a nivel genético, sino también a nivel transcripcional y postranscripcional. En concreto, ciertas desregulaciones génicas que se dan durante la maduración de células sanguíneas, proceso denominado hematopoyesis, pueden dar lugar a distintas leucemias (entre ellas, T-ALL). A pesar de saber esto, los mecanismos específicos y los procesos moleculares subyacentes no están del todo claros (7). Para comprender la etiología de la enfermedad es necesario investigar su transcriptoma.

Aproximadamente un 20% de los pacientes con T-ALL vuelven a recaer tras recuperarse (8). Para estos pacientes no existe un tratamiento y un pronóstico favorable, y presentan tasas de supervivencia bajas con respecto a los inmunofenotipos (6,8). Hoy en día no existe tratamiento claro para ellos, como es el caso de otras leucemias. Se cree que, a partir de la comprensión de los procesos genéticos que influyen en la T-ALL, se identificarán dianas terapéuticas que harán posible el diseño de fármacos para favorecer una buena y más favorable recuperación (8). Por la falta de conocimiento mencionada anteriormente, vemos necesario un estudio comparativo entre enfermos con T-ALL e individuos sanos.

En los últimos años se ha reportado que la desregulación del empalme alternativo de ARN tiene un impacto en el desarrollo del cáncer (9,10). El empalme alternativo (AS, del inglés alternative splicing) es un proceso por el cual un trascrito primario de ARNm da lugar a diferentes isoformas y con ello proteínas, que pueden desempeñar distintas funciones (11).

Como comentábamos, el proceso por el que la enfermedad vuelve a aparecer cuando se erradica no está del todo claro. Se hipotetiza que el empalme alternativo juega un papel importante en T-ALL (12). Hay artículos que ya documentan la importancia del empalme alternativo en la expresión de moléculas y sus importantes consecuencias en la recaída en la enfermedad. Por ejemplo, en Yi et al., 2019 los autores indican que mediante empalme alternativo se generan ciertas isoformas disfuncionales de la

proteína Ikaros en pacientes con ALL. Este es un represor del gen FUT4, que puede activar ciertos mecanismos que generan invasión de células cancerígenas. Los autores creen que esto puede ayudar a entender la recaída en la enfermedad. De la misma forma, nuestro trabajo se centra en el estudio del desbalance de los genes implicados en empalme alternativo en individuos sanos o con varios tipos de T-ALL con respecto a otros.

Las proteínas de unión a ARN (RNA binding proteins o RBPs, en inglés) son moléculas clave para la regulación genética (7,13). Entre otros procesos moleculares esenciales, se encargan de orquestar el empalme alternativo.

Una determinada RBP puede dar lugar a uno o varios procesos de empalme alternativo. Esto puede acarrear procesos celulares aberrantes y ciertas enfermedades. Las RBPs no solo son importantes en el empalme alternativo, sino que también están implicadas en muchos de los procesos de transcripción y en cambios químicos y estructurales de ARN (13). En este estudio nos centraremos en la importancia de la desregulación de la expresión de las RBPs y, en concreto, del posible desajuste global de los eventos de splicing producidos por esta.

En el caso de T-ALL, creemos que la desregulación de la expresión de las RBPs tiene consecuencias importantes. Pensamos que una desregulación causada por la expresión diferencial de las RBPs puede desencadenar en procesos de splicing diferentes a los fisiológicos. De esta manera, el proceso de desregulación del empalme alternativo, dirigido por la función aberrante de ciertas RBPs, puede contribuir a la proliferación de las células tumorales, y pensamos que este podría ser un mecanismo clave para la comprensión de la enfermedad (14). De hecho, ya se ha visto que las células normales y las leucémicas se pueden diferenciar a partir de las diferentes expresiones de las RBPs (15).

Para determinar la existencia de un desbalance de splicing entre grupos de individuos se puede estudiar la desregulación de los genes que generan RBPs, que es lo que estudiaremos en concreto. Si, por ejemplo, determinamos dos grupos con expresión diferencial, referentes a pacientes que han recaído en la T-ALL, y pacientes que no han vuelto a tener cáncer, ello podrá suponer una asociación entre una expresión diferencial genética mediada por ciertas RBPs y la progresión de la enfermedad. Además, sabremos qué RBPs concretas están involucradas en los procesos.

2.2 Objetivos del trabajo

Objetivos generales

Los dos objetivos generales del estudio son:

1. Determinar las RBPs que presentan una desregulación génica entre grupos de individuos con T-ALL.
2. Esclarecer y dar a conocer la influencia de estas RBPs en las vías metabólicas del desarrollo de T-ALL.

Objetivos específicos

Los objetivos concretos, a partir de los generales, serán:

1.1. Elaborar una base de datos a partir de datos públicos de transcriptomas de microarrays y RNA-sequency. Esta se formará a partir de datos de otras bases de pacientes con T-ALL descargados de fuentes fiables (como GEO) de transcriptomas de células T aisladas (maduras y/o progenitores). Esta base global estará disponible a partir de un objeto R.

1.2. A partir de un análisis en R, otro objetivo específico será determinar los genes diferencialmente expresados, seleccionarlos, y generar anotaciones de los. Esto se debe a que queremos comprender qué genes están desregulados y en qué grupos existe una desregulación. Por ejemplo, podemos tener genes diferencialmente expresados entre grupos T-ALL vs controles, o entres distintos tipos de T-ALL (ETP T-ALL vs. non ETP T-ALL, vs. TAL1 mutated, etc). Entonces, debemos de identificar los genes que se expresan de manera diferente, y luego anotar su localización y estructuras producidas a partir de ellos.

2.1. Identificar la fracción de genes desregulados correspondientes a RNA-binding proteins (RBPs). Esto se puede realizar a partir de un "Gene Enrichment Analysis", en R, continuación del análisis. Deseamos, en última instancia, aportar conocimiento sobre las RBPs involucradas en las leucemias T-ALL, por lo que debemos de estudiar la función biológica de los genes desregulados seleccionados (vía metabólica, proceso molecular y/ o celular, etc) y, de ellos, nos interesaran los que den lugar a RBPs involucradas en procesos de maduración del ARNm.

2.3 Enfoque y método seguido

Como se sabe que ocurre, las RBPs son muy importantes en los procesos de splicing, y se ha visto su diferente intervención en pacientes con T-ALL y sin ella, como comentamos al principio. Por ello, queremos averiguar qué RBPs están implicadas en el empalme alternativo en pacientes con este inmunofenotipo, y cuál es el papel concreto que juegan en las células aberrantes.

Si bien es cierto que los métodos informáticos son muy importantes, existen otros enfoques, como el de los análisis de laboratorio. Existen pruebas que miden las interacciones entre proteínas y ARN, y, en concreto, se pueden estudiar qué RBPs se unen al ARN. Podemos estudiar los sitios de unión, cómo es el ensamblaje y qué produce biológicamente una unión de RBP al ARN, a partir del entrecruzamiento e inmunoprecipitación (CLIP) (16,17). Para este método se requieren proteínas que puedan unirse al ARN, de forma que este método no nos valdría, ya que quisiéramos estudiar todas las RBPs que podamos. De cualquier forma, quisiéramos estudiar la unión a cualquier región del ADN, y si quisiéramos hacerlo mediante este tipo de técnicas, nos costaría mucho tiempo y dinero. Además, deberemos de hacer otro estudio, a parte, sobre qué zonas del ADN delimitar para su purificación, y su posterior transcripción a ARN. Un análisis bioinformático de todas las regiones del genoma secuenciadas y con cierta calidad asegurada, que son volcadas en bases públicas como GEO, será menos

costoso, tanto en tiempo, como en materiales, como en dinero, y su alcance será mayor en el mismo tiempo.

Otros autores, como (8), nos indican otro tipo de pruebas de laboratorio para identificar RBPs que se han usado históricamente, pero, como comentan en el capítulo, tienen pegas que no contempla el análisis bioinformático, como un bajo rendimiento. Más tarde se han realizado otros experimentos de laboratorio para su identificación, pero requieren de otra validación y tienen poca sensibilidad, aunque su metodología se cree podría mejorarse y se ha mejorado (18). El enfoque de (8) nos permite seleccionar un dominio de las proteínas que se unirá al ARN para buscar esta secuencia en el genoma humano. Después usaron las bases de GO para la clasificación de las proteínas encontradas, por lo que necesitaron una posterior búsqueda y clasificación de las nuevas proteínas seleccionadas para identificar sus funciones. Nuestra metodología permite agrupar todos los datos de estudios transcriptómicos volcados en las bases de GO para analizarlos. No es necesaria ni la experimentación ni la posterior definición de las funciones de cada proteína identificada, sino que usamos una metodología convergente de bases de datos que conlleva solo el análisis bioinformático de todas las RBPs anotadas hasta la fecha.

En cuanto a los análisis bioinformáticos, se han descrito varias formas de identificar RBPs, algunas mediante el uso de machine learning (19–21).

El estudio que se realizó constó de un proceso de búsqueda de datos de microarrays y RNA-seq de pacientes con T-ALL y sus subtipos, más el análisis transcriptómico. A pesar de que el análisis de datos de RNA-Seq aportaría información no solo sobre la desregulación de las RBPs, sino sobre los propios procesos de empalme alternativo (mediante la cuantificación de las isoformas directamente), hemos decidido que en este TFM no analizaremos datos RNA-Seq a nivel de exón. Esto es así porque 1) se requiere tener una infraestructura computacional con una magnitud suficiente para analizar este tipo de datos y 2) existen muy pocas cohortes de T-ALL con datos de RNASeq, y con unos tamaños de población muy bajos.

El proceso de análisis bioinformático contendrá un primer paso referido al volcado de las bases de datos con R. Posteriormente, se realizará un análisis de metadatos a partir de un preprocesado de los datos, análisis de perfiles de expresión y genes diferencialmente expresados, una anotación funcional, un análisis de enriquecimiento funcional y una selección de los genes desregulados correspondientes a las RBPs que influyen en procesos de splicing. Todo ello nos dará una idea de si existen diferencias en las RBPs entre grupos de pacientes, y si estas diferencias pueden influir en la T-ALL.

Los grupos a estudiar dependerán de las muestras. Obtendremos los datos de la base de datos de GEO, de los diferentes estudios (22). Estos datos de microarrays y RNA-seq serán referentes a gene expression data series (GSE), y sus datos asociados. Los GSE usados eran los mismos que usaron Bigas et al., 2020. Además, usamos el GSE33469 y GSE33470 mencionado en el artículo de Vlierberghe et al., 2011, el GSE32215 usado por Piovan et al., 2013, y el GSE10609 de van Vlierberghe et al., 2008. Es importante tener en cuenta las peculiaridades de cada serie para el análisis informático, indicadas en la web de NCBI:

-GSE8879: perfil de expresión génica de T-ALL atípica. Extracción de linfoblastos de niños en el momento del diagnóstico. Uso de array U133A de Affymetrix. Última fecha de actualización 2018, EE.UU.

-GSE10609: perfiles de expresión por array. Estudio de un subgrupo de T-ALL. 67 pacientes, 92 muestras. Array de Affymetrix Human Genome U133 Plus 2.0. Última fecha de actualización 2019, Países Bajos.

-GSE14618: generación de perfiles de expresión por array. Chip Affymetrix U133 Plus 2.0. 50 niños recién diagnosticados. Última fecha de actualización 2019, EE.UU. Conlleva el GPL570 y el GPL 96.

-GSE26713: generación de perfiles de expresión por array, se identificaron dos subtipos de T-ALL. Chip Affymetrix U133 Plus 2.0. 117 muestras de pacientes pediátricos. Última fecha de actualización 2019, Países Bajos.

-GSE28703: generación de perfiles de expresión por array, T-ALL temprana (ETP ALL). Affymetrix HT HG-U133+ PM Array Plate. Secuenciación de genoma de ADN tumoral y normal, 12 niños con ETP-ALL, 52 ETP con tumores y 42 muestras de T-ALL no ETP. Última fecha de actualización 2018, EE.UU.

-GSE32215: array Affymetrix Human Genome U133 Plus 2.0. 228 muestras de T-ALL, con 117 muestras reanálisis de GSE26713. Última fecha de actualización 2019, Países Bajos.

-GSE33469: generación de perfiles de expresión por array. Illumina HumanHT-12 V4.0 expression beadchip. 57 muestras de T-ALL de adultos. Última fecha de actualización 2018, EE.UU.

-GSE33470: perfil de expresión de poblaciones de timocitos. 3 pacientes pediátricos (de 7 días a 6 meses de edad, es decir, 7 poblaciones de células T). Illumina HumanHT-12 V4.0 expression beadchip. Última fecha de actualización 2018, EE.UU.

-GSE37389: generación de perfiles de expresión por array. Agilent-014850 Whole Human Genome Microarray 4x44K G4112F. Última fecha de actualización 2019, India.

-GSE56488: generación de perfiles de expresión por array. Affymetrix Human Gene 2.0 ST Array. Muestras de T-ALL. Última fecha de actualización 2015, Francia.

-GSE62156: generación de perfiles de expresión por array. Affymetrix Human Genome U133 Plus 2.0 Array. 64 muestras de T-ALL. Última fecha de actualización 2019, Bélgica.

-GSE110633: generación de perfiles de expresión por array. Cohorte primaria de pacientes con T-ALL. Illumina NextSeq 500. Última fecha de actualización 2019, Bélgica.

-GSE110636: generación de perfiles de expresión por array. Cohorte primaria de pacientes con T-ALL. Illumina NextSeq 500. Última fecha de actualización 2020, Bélgica.

Si determinamos una expresión diferencial de RBPs entre grupos de individuos de cada serie, estudiaremos su asociación con los eventos de splicing. Seleccionaremos solo aquellas RBPs relacionadas con el empalme alternativo. A partir de ellas y de la bibliografía se determinará si las rutas concretas de las RBPs influyen en eventos relacionados con el cáncer y, en concreto, con T-ALL.

En los documentos tipo .Rmd del archivo de Github se especifica la metodología seguida en los análisis de forma extensa, y en el apartado de metodología se comentan los tipos de análisis seguidos, a modo de resumen.

2.4 Planificación del trabajo

Como comentamos, uno de los objetivos de nuestro trabajo era determinar si los genes codificantes de proteínas de unión a ARN (RBPs), que intervienen en el proceso de empalme alternativo, tienen un papel relevante en T-ALL mediante la comparación del transcriptoma de distintos grupos de pacientes. Otro objetivo general se basaba en ofrecer la información resultante de este estudio a través de la aplicación R, además de información sobre datos de los pacientes y datos referentes a procedimientos de laboratorio, si es conveniente.

El primer paso general fue referente a la recopilación de los datos. Luego, se volcaron los datos en R, donde se siguió con el preprocesado de los datos y la exploración de estos. Se completaron los análisis relacionados con la selección de genes diferencialmente expresados de ambos tipos de datos (microarray y RNA-seq), el análisis de enriquecimiento funcional de datos de microarrays, la anotación de los datos, y de forma similar para los datos de RNA-Seq.

En cuanto a los nuevos análisis, pensamos en generar diferentes gráficos a partir de los análisis realizados, y en cruzar nuestros resultados con bases de datos de otros estudios centrados en RBPs RBPs (ver Huang et al., 2018; Sebestyén et al., 2016; Wang et al., 2019).

Primeramente se han realizado unos “volcano plots” y “heatmaps”, típicos de los análisis de expresión génica, para ciertas comparaciones resultado de los análisis de genes diferencialmente expresados, indicados en el HTML de figuras de GitHub. El primero de los gráficos representan los p valores con respecto a un eje x del “fold change”, medida del cambio de la expresión de los genes. El segundo tipo de grafico es interesante para generar y visualizar agrupaciones entre genes y muestras en función de la similitud de sus patrones de expresión. Conjuntamente con ellos, también se ha determinado el gráfico que usaremos para representar los resultados de las vías del análisis de enriquecimiento funcional de los genes.

A parte, también se comentó la anotación de los resultados con la base de datos de GeneOntology para determinar el significado celular, la función molecular o la función biológica de los genes.

Tras la nueva anotación, posterior al análisis de enriquecimiento funcional, intentamos interrogar un subset de genes involucrado en el procesamiento de RNA (RBPs). Así podemos ver cómo podemos realizar este tipo de preguntas a cualquiera de los datos resultantes de cada comparación entre grupos con leucemia.

Como últimos pasos, estamos seleccionando los genes anotados como implicados en el empalme alternativo, se recopiló más información sobre las RBPs implicadas en el splicing, y se pasó a relacionar esta información y a escribir la memoria.

2.4.1 Tareas

En la tabla 1 se puede visualizar el ultimo diagrama de Gantt con la lista de tareas y sus correspondientes nuevas fechas referente a la PEC3. Hemos modificado la lista de tareas a realizar dentro de la selección de genes diferencialmente expresados, y hemos incluido otras tres tareas.

Reescribimos las tareas referentes a cada nuevo apartado modificado. Esto concuerda con la tabla 1. Hemos variado el resultado de la tarea T3.1, hemos incluido otras tres tareas, y hemos modificado la duración de todas, de forma que concuerde con la fecha de finalización de nuestro trabajo. Las tareas modificadas se marcan con un subrayado:

T1.1: importado de datos a R a partir de datos de microarrays y de RNA-seq.

- Recopilación de datos de GSE de los datasets de GEO usados por Bigas et al., 2020.
- Buscar otros datos del mismo tipo, de GSE de microarrays o de RNA-seq de pacientes con T-ALL, en la web de GEO.
- Volcado de los datos en R o RStudio.
- Datos de RNA-seq: tras la lectura, se crearon las matrices de contajes a partir de los datos crudos de lecturas por gen.
- Cada uno de los data frames contiene la información que contiene cada GSE, más los tipos de pacientes y los subtipos de T-ALL que tienen. También contienen información referente a los datos técnicos de los laboratorios y de los pacientes indicados en páginas de referencia.

Resultado: archivo .gz de cada GSE descargado de la web de GEO.

Duración: desde el 27 de febrero hasta el 5 de marzo. 6 días.

T2.1: análisis de preprocesado 1 de los datos genéticos en R.

- Datos de microarrays: exploración de los datos, control de calidad y estudio de posibles covariables experimentales y efecto batch
- Datos de RNA-seq: estandarización de los contajes, filtrado de genes poco expresados, uso de clases específicas para manejar los datos, normalización y transformación logarítmica

Resultado: documento de R con la información de la tarea T1.1, más estos análisis informáticos.

Duración: desde el 5 de marzo hasta el 23 de marzo. 18 días.

T2.2: preprocesado 2 y exploración.

- Datos de microarrays. Preprocesado 2: normalización, filtraje y anotación.
- Datos de RNA-seq. Preprocesado 2: normalización y transformación logarítmica.
- Datos de RNA-seq: exploración de los datos. Dentro encontramos el análisis de la distribución de los contajes y el análisis de similaridad entre muestras a partir de varios métodos.

Resultado: documento de R con la información de la tarea T1.1 y T2.1, más estos análisis. Además, un archivo.Rda de datos normalizados, filtrados y anotados por cada

serie de microarrays. Para los datos de RNA-seq, guardamos los datos de los genes filtrados y un objeto normalizado en el mismo tipo de objeto, por cada serie.

Duración: desde el 24 de marzo al 11 de abril, es decir, 18 días.

T3.1: finalización del análisis ómico.

- Análisis de expresión diferencial y posterior selección de genes diferencialmente expresados. Dentro de la selección de genes diferencialmente expresados debimos de crear las matrices de diseño y de contrastes, hicimos una estimación del modelo y selección de genes, a parte de comparaciones múltiples, y realizamos varios métodos de visualización.
- Incluimos la representación de los genes diferencialmente expresados a través de "volcano plots", y de "heatmaps" que los relacionen con las muestras.
- Datos de RNA-seq: anotación.
- Análisis de enriquecimiento funcional ("Gene Enrichment Analysis").

Resultado: documento de R con todo el análisis hasta aquí y gráficos. Archivo de las RBPs seleccionadas.

Duración: desde mirad de abril (día 12) hasta finales de mes (día 26). En total, 14 días.

T4.1: selección de las RBPs.

- Enriquecimiento de las bases de datos resultado de cada análisis de enriquecimiento funcional a partir de otros datasets de tres autores. Mediante esto podemos anotar informacion sobre RBPs que se nos haya escapado por el anotado convencional de los datasets.
- Interrogación de un subset a modo de ejemplo sobre genes involucrados en el splicing. Escribiremos un código modelo en R, que se puede ejecutar según se indica para la búsqueda de genes de splicing en una comparación entre pacientes con leucemia. Este mismo se puede modificar para la búsqueda de otras proteínas o de funciones biológicas concretas.
- Nueva selección de los genes, pero en este caso para determinar aquellos que generen proteínas relacionadas con eventos de splicing de ARN.

Resultado: documentos de R con la anotación nueva a partir de otros datasets. Otro documento de R para su carga y búsqueda de proteínas o funciones concretas. Ademas, un documento mas sobre como se seleccionarán las RBPs de todos nuestros datasets, que es lo que usaremos posteriormente.

Duración: últimos días de abril (desde el 27), hasta el 19 de mayo, 22 días. La etapa que más tiempo lleva es el enriquecimiento, por la gran carga computacional.

T4.2: papel de las RBPs en T-ALL y redacción de la memoria.

- Recopilación de información a partir de bibliografía de cada RBP seleccionada.
- Determinación del papel de cada RBP en el metabolismo celular y, en concreto, en la producción de células anormales y el posible papel en la leucemia. Esto se realizará a partir de nuestro análisis y del estudio de cada RBP de otros autores.

- Integración de toda esta informacion recopilada y redacción de la memoria final del trabajo.

Resultado: artículos para leer, y, a parte, la memoria escrita.

Duración: del 20 de mayo hasta, como mucho, el 4 de abril. Aunque esperamos acabar antes para poder realizar el siguiente paso.

T5.1: elaborar una shiny app en R. Tarea no indispensable.

- Se intentará desarrollar una shiny app en R con la información obtenida de nuestros análisis.

Resultado: aplicación web interactiva en R con la información.

Duración: si los apartados anteriores se han realizado en menor tiempo del previsto, este apartado ocupará el tiempo restante hasta el 6 de junio, que se entrega el trabajo. Será, por tanto, un trabajo de una o dos semanas.

Todos los objetivos anteriormente mencionados han sido completados parcialmente, menos la shinny app de R.

Todas las nuevas tareas son importantes para la comprensión de nuestros objetivos, que eran determinar las RBPs que presentan una desregulación génica entre grupos de individuos y esclarecer y dar a conocer la influencia de estas RBPs.

Si no anotamos todas las RBPs correctamente, es decir, sin contrastar con otros artículos, por ejemplo, dejándonos información, no podremos optimizar el alcance de estos objetivos, ya que perderíamos información valiosa.

2.4.2 Calendario

Podemos consultar el calendario en la tabla 1. En la figura 2 también mostramos el orden de los pasos seguidos para el análisis transcriptómico.

2.4.3 Hitos

- Importar correctamente ambos tipos de datos en R.
- Conseguir el archivo.Rda de datos de microarrays y, a parte, el referente a los datos de RNA-seq.
- Entrega de la PEC2 - Desarrollo del trabajo - Fase 1. Observar el desarrollo del trabajo, si se han cumplido los plazos del calendario, etcétera.
- Entrega de la PEC3 - Desarrollo del trabajo - Fase 2. Lo mismo que el apartado anterior.
- Completar el análisis ómico en R entero, teniendo un documento de R completo.

- Obtención de varias RBPs relacionadas con el splicing seleccionadas (no una o ninguna), escritas en un documento para su entendimiento y consulta.
- Documentar la asociación, según la bibliografía, de esas RBPs y la enfermedad.
- Haber terminado de escribir bien la memoria.
- Entregar la memoria.
- Elaborar la presentación.
- Defender el trabajo.

Tabla 1. Diagrama de Gantt corregido de las tareas. Las tareas en gris claro son referentes a los datos de microarrays, las de color gris intermedio para ambos tipos de datos, y las de color oscuro solo para datos de RNA-seq. El color lila hace referencia a una tarea no imprescindible, que no se realizó.

Actividad	Febrero	Marzo	Abril	Mayo	Junio
T1.1. Recopilación de datos	27-28	1-2			
T1.1. Recopilación de otros datos		3-4			
T1.1. Volcado en R de cada GSE		4-5			
T1.1. Creación matriz de contajes		5-6			
T2.1. Preprocesado de datos 1: -Exploración -Control de calidad -Análisis de batch		7-23			
T2.1. Preprocesado de datos 1: -Estandarización -Filtraje -Clases específicas		7-23			
T2.2. Preprocesado de datos 2: -Normalización -Filtraje -Anotación		24-31	1-5		
T2.2. Preprocesado de datos 2: -Normalización -Transformación logarítmica		24-31	1-5		
T2.2. Exploración de los datos			6-11		
T3.1. Selección de genes diferencialmente expresados			12-16		
T3.1. Anotación			17-21		
T3.1. Análisis de significación biológica			22-26		
T4.1. Enriquecimiento de funciones en cada comparativa			27-30	1-4	
T4.1. Interrogación de un subset de genes involucrado en splicing				5-12	
T4.1. Selección de las RPBs				13-19	
T4.2. Recopilación de información del papel de cada RBP				20-25	
T4.2. Papel RBPs y redactar el trabajo				26-31	1-4
T5.1. Desarrollar una shinny app					?-6

2.5 Breve sumario de contribuciones y productos obtenidos

- Plan de trabajo.
- Obtuvimos datasets de varios tipos de datos.
- Documentos “Prodecimiento creacion datasets e info”, “Expres diferenc y signif, “Figuras”, los tres nuevos archivos de R de las anotaciones nuevas de los GSE a partir de los tres artículos, y el archivo “Selección RBPs”. Todos de tipo .Rmd. Además, un documento de R para la interrogación a modo de ejemplo de un subset a partir de genes involucrados en el splicing de ARN, “Ejemplo_interrogracion”.
- Listado sobre las RBPs, seleccionadas por su expresión diferencial, influyentes en el procesado del ARNm e involucradas en T-ALL. Este es el documento de Word “Tabla genes y proteinas RNA splicing”.
- Todos los Excel de las comparaciones, generados a partir de la anotación de genes con bases de datos de otros autores. Estos son “comparison_GSE8879_Transcriptome”, “comparison_GSE8879_Targeting2”, “comparison_GSE8879_Targeting5”, y “comparison_GSE8879_Large” para GSE8879. Se han incluido estos, y no todos los resultados, porque eran demasiado grandes, en GitHub. Estos hacen referencia a la comparación entre grupos de GSE8879. En las tablas solo se indican los genes que aparecían en los tres artículos y en nuestras tablas, ademas de la informacion anotada. Todos nuestros genes al completo los tenemos en varias tablas generadas anteriormente, resultado de los análisis de enriquecimiento funcional.
- Gráficos explicativos del tipo “volcano plots”, “heatmaps” y “dotplots” del análisis de enriquecimiento de todas las series GSE.
- Memoria.
- Presentación virtual

2.6 Breve descripción de los otros capítulos de la memoria

En la introducción se hace un repaso sobre los conceptos clave relacionados con la T-ALL para la comprensión del problema. En el primer apartado dentro de este se comentan las peculiaridades de la enfermedad, por qué es importante estudiarla, el pronóstico médico de cada tipo de paciente, qué es el empalme alternativo de ARNm y las RBPs, y nuestro enfoque para la resolución del problema planteado. En el segundo apartado se comentan los objetivos generales y específicos del trabajo. A partir de estos, en el tercer apartado se comenta el enfoque seguido para conseguir los objetivos y la metodología a grandes rasgos. En el cuarto apartado se hace una descripción de los recursos, las tareas a realizar, la planificación temporal de las tareas y los hitos del trabajo realizado.

El siguiente apartado hace referencia al estado del arte, en el que explicamos cómo ha sido la investigación en los últimos años para la comprensión de la T-ALL y qué se sabe de la enfermedad. También se comenta por qué es importante nuestro estudio y la hipótesis sobre el papel de las RBPs en la T-ALL.

En el apartado de metodología, el cuarto, contamos cómo se ha realizado el estudio, a través de un recorrido de los análisis bioinformáticos, del enriquecimiento de las

funciones de las proteínas a partir de otras bases de datos, y por último de la selección de genes que están implicado en el empalme alternativo de ARNm en la leucemia linfocítica aguda de células T.

En el quinto apartado, referente a los resultados, se comentan los productos obtenidos más importantes acordes con la metodología. Todos ellos se comentan en la siguiente parte.

En la discusión se hace referencia a los resultados. Se plantea si se han conseguido los objetivos, se responde a la pregunta de si la hipótesis era veraz y se hace referencia a los resultados de otros autores.

En la última sección se describen los resultados más importantes del trabajo, se planea si se han conseguido los resultados y el por qué de ambas cuestiones, en referencia a resultados y discusión. Luego, comentamos posibles caminos de investigación futuros para hacer otro tipo de trabajos, los cuales nos indiquen otras RBPs importantes en la T-ALL. En el último de estos subcapítulos hacemos un análisis sobre si la metodología y la planificación a lo largo de los análisis ha seguido la planificación, y si ha habido que introducir algún cambio en ella.

3 Estado del arte

Como ya comentamos, la leucemia linfocítica de células T es una enfermedad cuyos pacientes muchas veces tienen un futuro no favorable en la superación de esta condición, sobre todo los que han sufrido una recaída. Esto se debe, entre otras cuestiones, a que hoy en día todavía no sabemos las bases genéticas ni comprendemos del todo cómo se dan los procesos de traducción en la patogenia de esta enfermedad (26). Entonces, se necesita una investigación más profunda del entorno genético para poder desarrollar tratamientos efectivos, a favor de combatir esta enfermedad tan agresiva (8,26).

Recientemente se ha entendido que las proteínas de unión a ARN, importantes reguladoras del entorno genético, son las que se encargan de la producción de proteínas normales o irregulares, lo cual puede asociarse con procesos tumorales. Se empieza a comprender, en definitiva, que la regulación génica postranscripcional está íntimamente relacionada con la aparición y progresión de tumores (27).

Como es de esperar, las RBPs también son cruciales en las leucemias agudas. Algunas de las diferentes isoformas de ARNm generan tumores, al estar mal empalmadas, en la aparción de la leucemia (28). Ya se ha demostrado que las mutaciones en las secuencias que dan lugar a las RBPs generan desregulaciones en el empalme de ARN de la leucemia mieloide y linfoide (29). En el caso de la leucemia mieloide aguda, existe una intrincada red de RBPs implicadas en procesos de empalme de ARN, las cuales se encuentran reguladas a la alza en pacientes con esta enfermedad (10).

A pesar de todo este conocimiento, no debemos olvidar que el estudio de las vías metabólicas y la genética detrás de las RBPs, los procesos de empalme alternativo y su relación con la T-ALL todavía no se conoce al completo.

Debido a la importancia de estos procesos, se cree que esclarecer los mecanismos por los que las RBPs regulan la progresión del cáncer dará lugar a una base teórica notable que ayudará a tratar el cáncer de forma más eficaz (27). Por ejemplo, en el caso de la leucemia mieloide aguda, al eliminar una RBP llamada RBM39 in vitro se produjeron eventos de empalme anormales de otras RBPs implicadas en el mantenimiento de la leucemia. A partir de este descubrimiento, se cree que se pueden producir avances en el tratamiento de la leucemia mieloide aguda al eliminar RBM39 (10). Este es solo un ejemplo de cómo podemos construir la base teórica de tratamientos focalizados en la supresión o continuación de procesos de empalme alternativo de ARN mensajero.

Existen muchos mecanismos por los que se da un mantenimiento de las células cancerosas en la T-ALL, pero nuestra hipótesis se basa en que los cambios en la expresión de las RBPs son claves en estos procesos (26). Si existen diferencias de expresión de RBPs entre grupos de pacientes, esto nos puede indicar que en la prolongación de la enfermedad, o en la recaída en ella, están implicadas las RBPs, de forma que podemos entender en qué fases de la enfermedad son importantes. De ser así, es probable que dichas RBPs desreguldas puedan estar controlando eventos de splicing que afectan a genes que son claves para el crecimiento y/o mantenimiento de células leucémicas en los diferentes inmunofenotipos.

La bioinformática es una disciplina en auge que nos ayuda a estudiar perfiles de expresión génica, características clínicas y patologías en conjunto. Por ejemplo, se pueden estudiar las vías metabólicas y genéticas y su papel en la formación de tumores (30). A partir de este tipo de análisis ómico bioinformático, se ha pensado que los procesos relacionados con cáncer tienen que ver con la sobreexpresión y desregulación de la maquinaria metabólica. En esta línea, se han visto similitudes a nivel de expresión génica entre varios tipos de cáncer (54). De esta forma, pensamos que un análisis bioinformático de conjuntos de datos GSE, los cuales se han usado en las metodologías mencionadas en este párrafo, nos conducirá al entendimiento del papel de las RBPs y el empalme alternativo en la T-ALL.

4 Metodología

Aquí comentaremos los pasos seguidos para alcanzar nuestros objetivos y plantearnos la veracidad de nuestra hipótesis. Los pasos seguidos se desarrollan en documentos accesibles desde nuestro repositorio de GitHub https://github.com/CristinaMuntanola/TFM_T_ALL.git. A lo largo del texto comentaremos los nombres de los archivos subidos al repositorio para acceder a cierto tipo de contenido.

4.1 Análisis bioinformático ómico de datos

En la figura 1 representamos el análisis seguido. Aquí comentaremos los pasos seguidos de forma general. El workflow del análisis detallado se encuentra en los archivos de extensión .Rmd que se pueden descargar del Github.

Los siguientes pasos hacen referencia al código del archivo "Procedimiento creación datasets e info". Este es el primero de los archivos creados, como se puede ver en la figura 1. El flujo de análisis dentro de este documento se puede estudiar a partir de la figura 2.

Se analizaron dos tipos de series de datos de GEO de expresión génica (GSE, o "GEO Series" en inglés), tanto de microarrays como de RNA-seq. Los GSE usados se indican en la tabla 2 de Bigas et al., 2020. Además, se usaron los datos indicados por (23), que son GSE33469 y GSE33470, correspondientes a datos de pacientes con leucemia temprana. También usamos el GSE32215 de pacientes con leucemia (sin especificar) (24). También adquirimos los datos de GSE10609 de (25). Todos los datos se recopilaron del repositorio GEO (Gene Expression Omnibus) la web de NCBI.

Primeramente se leyeron y volcaron los datos en R. Los demás pasos son distintos, debido a que los datos son de distinta índole.

Para los datos de microarrays se continuó con un preprocesado de los datos. Dentro de este se hicieron una exploración de los datos, un control de calidad, un análisis de posibles efectos batch, una normalización, el filtraje de los datos y su anotación, como workflow de análisis común que se suele realizar.

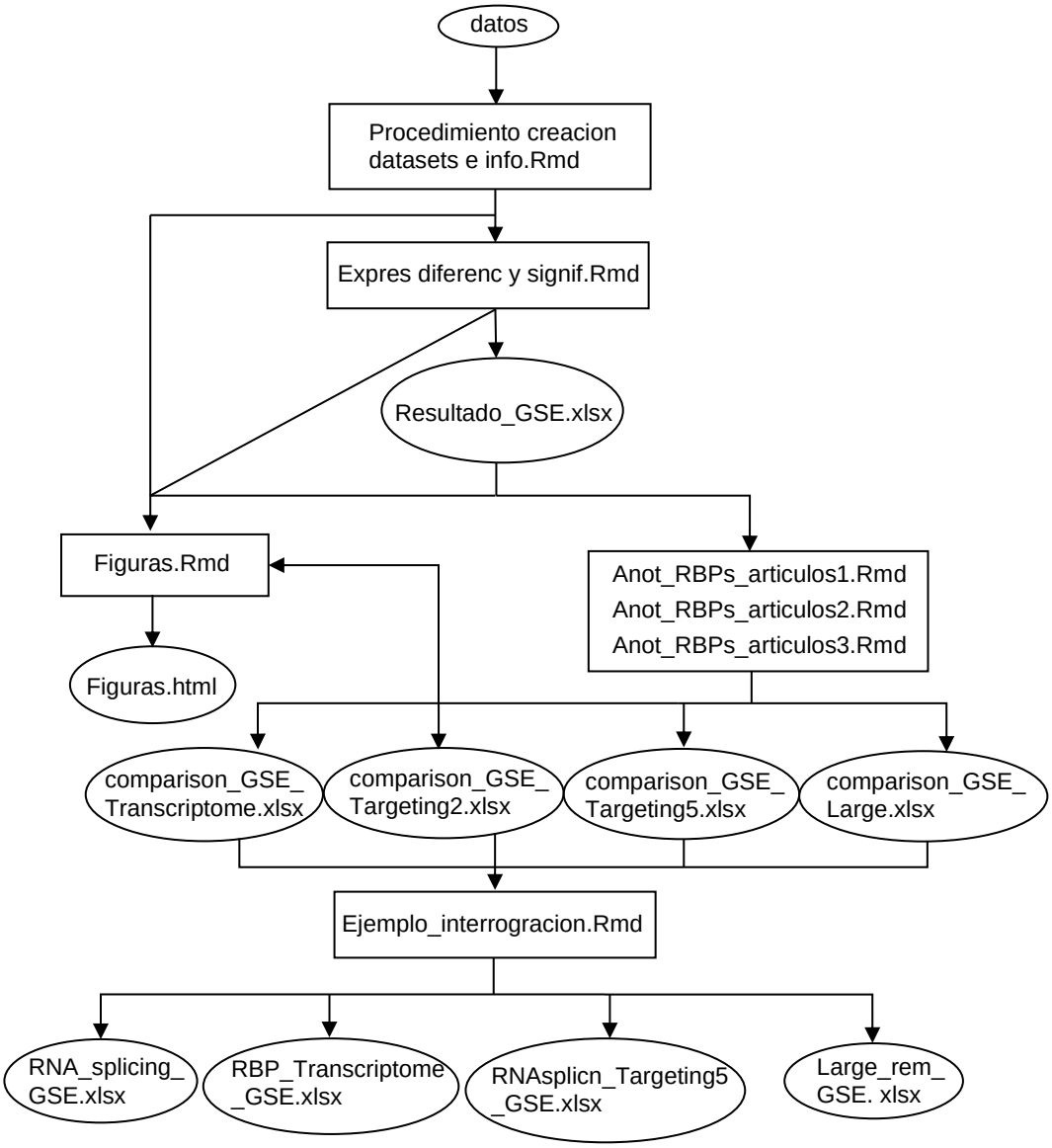


Figura 1. Diagrama de flujo de trabajo seguido para conseguir todos nuestros resultados. En los óvalos se muestran los resultados obtenidos, a partir de los documentos en los que se desarrolla el código, indicados con cuadrados.

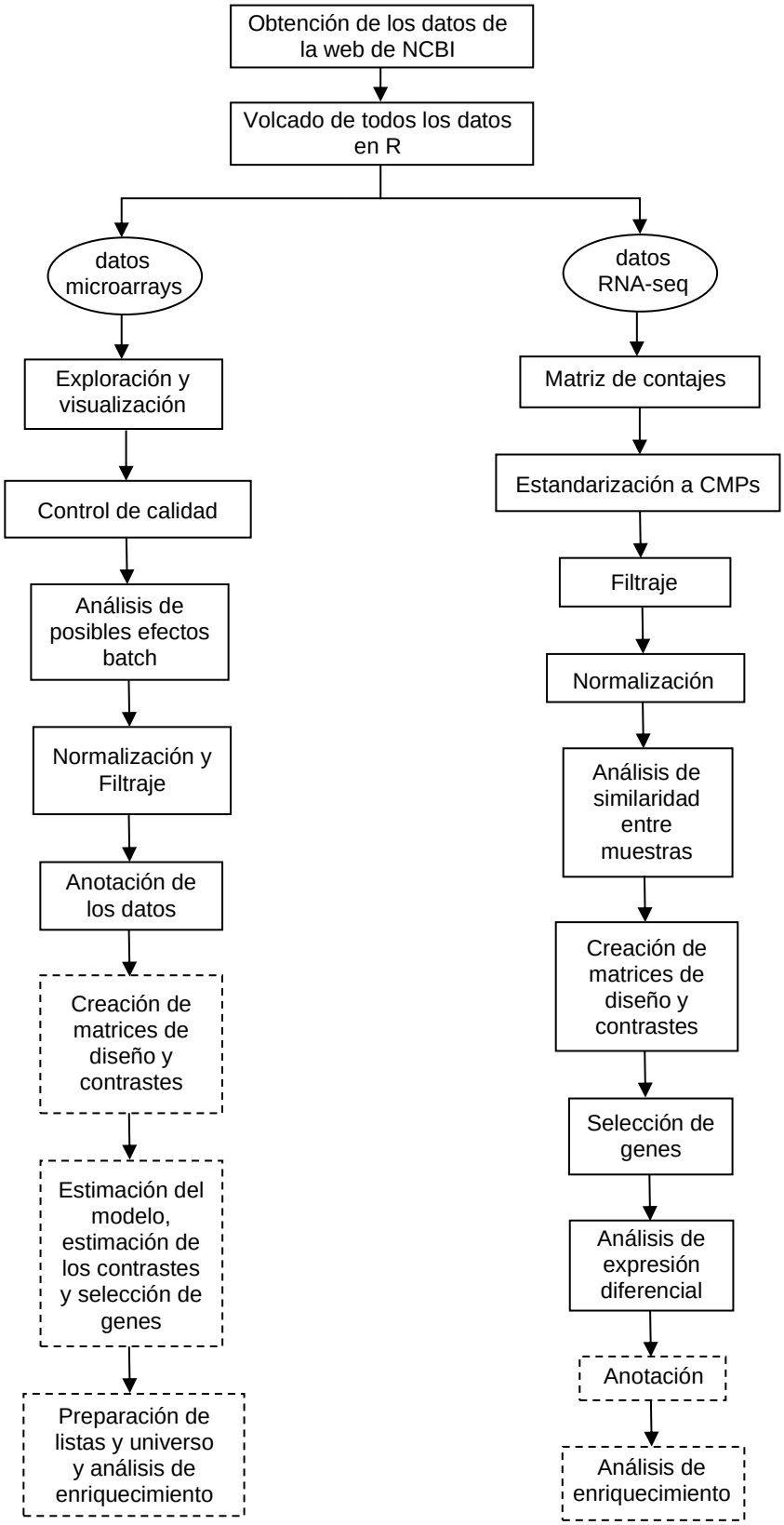


Figura 2. Diagrama de flujo de trabajo seguido para el análisis bioinformático ómico. Los datos se analizaron en el documento "Prodecimiento creacion datasets e info.Rmd", en línea contínua, y "Expres diferenc y signif.Rmd", en línea discontínua.

En la tabla 2 se indican los grupos creados a partir de cada serie en el apartado de exploración, según lo indicado por los creadores de cada una, para los datos de microarrays y RNA-seq. Debemos de tener en cuenta, para la interpretación de los resultados posteriores, si los grupos de pacientes están constituidos por más de cinco muestras. Aquellos grupos con menos o de cinco muestras son el grupo LMO1 de GSE26713 (solo un individuo), MYB de la misma serie (consta de dos individuos), todos los del GSE33470 (solo tienen tres individuos por grupo), TCR de γδ (con cinco individuos) y casi todos los grupos de GSE62156. De esta última serie, CALM-AF10 poseía un individuo, HOX otro, "HOXA *of unknown mechanism*" solamente uno, immature HOXA13-t estaba formado por dos muestras, MLL-t por un individuo, TLX1 por cinco, y desconocido tenía solo uno. Solamente Immature y TLX3 de esta serie, y todos los demás grupos de todas las series, tenían suficiente cantidad de individuos.

Para la normalización y el filtraje no especifico o general de genes poco expresados, anotamos cada serie según lo indicado por los autores de cada estudio. Se anotaron en R los resultados del apartado anterior con las siguientes bases de datos, para cada uno de las siguientes series:

-hgu133a.db: usado para anotar GSE8879.

-hgu133plus2.db: con ella anotamos GSE10609, GSE14618, GSE26713, GSE32215, GSE62156.

-hthgu133pluspm.db: para GSE28703.

-illuminaHumanv4.db: GSE33469, GSE33470.

-hgug4112a.db: GSE37389.

-hugene20sttranscriptcluster.db: GSE56488.

-org.Hs.eg.db: GSE110633 y GSE110636.

Se anotaron en todos los casos, a partir de los identificadores de Ensembl, el identificador tanto de Entrez, como el símbolo del gen, el símbolo de la proteína resultante según UniProt, el nombre del gen, y la ruta metabólica en la que están implicados los genes.

Aplicamos posteriormente un filtraje "estándar" que retuvo el 50% de los genes con mayor variabilidad de entre aquellos que están correctamente anotados, eliminando aquellos que no tienen identificador en la base de datos Entrez.

Tabla 2. Grupos de pacientes para cada gene expression data series.

Serie	*Grupos*
GSE8879	Leucemia precursora temprana de células T (ETP o early T-cell precursor leukemia) y sin ETP diagnosticada
GSE10609	Anomalías citogenéticas moleculares "TAL1", "HOX11, "HOX11L2", "HOXA", "LMO2", y desconocido.
GSE14618	Perfiles de expresión de pacientes con leucemia recién diagnosticada "NR" (no response (induction failure)), "F" (failure (relapse)), y "C" (Complete Continous Remission)
GSE26713	"BM", "HOXA", "LMO2", "TAL1", "TLX3", "unknown", "TLX1", "TAL2/LMO1", "MYB", "TAL2" y "LMO1"
GSE28703	Pacientes con T-ALL temprana ETP, y no temprana
GSE32215	"Adult T-ALL", “pediatric T-ALL" o desconocido.
GSE33469	Un solo grupo
GSE33470	"CD3+CD4+", "CD3+CD8+", "CD34+CD1a-", "CD34+CD1a+", "CD4ISP", "DPCD3-"
GSE37389	Receptor de las células T (TCR) alpha/beta o αβ, y TCR gamma/delta o γδ
GSE56488	Cinética del injerto corta y cinética larga
GSE62156	“CALM-AF10”, “HOX”, “HOXA of unknown mechanism”, “immature”, “immature, HOXA13-t”, “MLL-t”, “TAL-R”, “TLX1”, “TLX3” y desconocido
GSE110633	“HOXA”(HOXA-overexpressing T-ALL), “IMM”(Immature T-ALL), “TAL” (TAL T-ALL), “TLX” (T-cell Leukemia Homeobox)
GSE110636	“HOXA”(HOXA-overexpressing T-ALL), “IMM”(Immature T-ALL), “TAL” (TAL T-ALL), “TLX” (T-cell Leukemia Homeobox)

A continuación se comentarán los pasos indicados en el archivo "Expresión diferencial y signif" del GitHub. Los pasos se muestran en la figura 2.

A partir de los resultados del preprocesado de los datos de microarrays se realizó la selección de aquellos genes que se expresaban diferencialmente entre los grupos de pacientes comparados. Los grupos enfrentados eran los ya comentados para cada serie de GEO, de manera que se realizaron todas las comparaciones posibles entre grupos.

Antes de nada se crearon las matrices de diseño de las comparaciones entre grupos, y la matriz de contrastes. Las matrices de contrastes se definieron de forma diferente si se comparaban dos grupos o más:

-Comparaciones entre dos grupos: como se hace de forma tradicional, la matriz de contrastes definió este tipo de comparaciones como un grupo menos el otro. Entonces, uno de los grupos se representaba como 1, y el otro como -1.

-Comparaciones entre más de dos grupos: existían grupos a comparar con características inmunofenotípicas similares. Por ejemplo, para GSE26713 podríamos hacer comparaciones entre un grupo referente a TAL y MYB, aunque los individuos con TAL estén diferenciados en dos grupos, TAL1 y TAL2. Se pensó en comparaciones según el promedio de la expresión diferencial de dos grupos con características similares, con respecto a otro grupo. Por ejemplo, se realizó una comparación de la siguiente manera: (TAL1 + TAL2)/2 – MYB. La adicción simple entre dos grupos nos llevaría a error, debido a que sumaríamos la expresión genética de dos grupos, por lo que probablemente resultarían muchos genes, y no sería adecuado para entender el efecto de TAL con respecto a MYB, en este caso. Aquí hacemos un promedio de las expresiones de dos grupos relacionados, en este caso, con leucemia TAL, lo cual no produce que, si un gen se expresa en ambas condiciones, no tenga una expresión desorbitada con respecto al otro grupo de la comparación, en este caso MYB.

La aproximación utilizada aquí a posteriori para evaluar la expresión diferencial hace referencia a la utilización de un modelo lineal general (55). Tras definir las matrices y la estimación del modelo, se estimaron los contrastes y se realizaron las pruebas de significación para determinar qué genes estaban diferencialmente expresados. Como en la mayoría de series GSE no existían más que un par de grupos a comparar, es decir, no existían una gran cantidad de comparaciones múltiples, no se realizaron ajustes de los p-valores entre comparaciones distinto al realizado entre genes. De esta forma se determinaron si los genes se encontraban regulados a la alza, regulados a la baja o no regulados a partir de un cutoff de 0.05 para los p-valores.

Se usó una ampliación del método tradicional para análisis de significación, de forma que se usaron modelos de Bayes empíricos para combinar la información de toda la matriz de datos y de cada gen individual, y obtener así estimaciones de error mejoradas. El objeto creado tiene, al igual que el análisis tradicional, estadísticos como el fold-change, t-moderados, o p-valores ajustados que se utilizan para ordenar los genes de más a menos diferencialmente expresados.

Posterior a este análisis de significación, se realizaron "volcano plots" para ciertas comparaciones de cada GSE, típicos de los análisis genéticos. Se realizaron a partir del objeto creado según la moderación empírica de Bayes de los errores, es decir, para el conjunto de genes de una serie, no de cada comparación.

También se crearon “heatmaps”, otras figuras que se suelen usar para visualizar los perfiles de expresión de cada gen, agrupados por este tipo de perfiles. De esta forma, los genes diferencialmente expresados en alguna de las comparaciones se representan de forma enfrentada con respecto a cada muestra o paciente.

Como paso final del análisis de las series referentes a microarrays, se realizó un análisis de enriquecimiento funcional de las listas de genes diferencialmente expresados utilizando como referencia los procesos biológicos descritos en GO, para la interpretación de los resultados. De esta forma podemos entender si las listas tienen una cantidad superior a la esperada de genes asociadas a funciones o procesos biológicos determinados relevantes para este estudio. Para comenzar se prepararon las listas de genes y los “universos”, y posteriormente se realizó un test de Fisher (hipergeométrico) para todas las categorías GO de la ontología seleccionada.

El último tipo de gráfico estudiado se suele denominar “dotplot”, empleados a partir de los análisis de enriquecimiento funcional. Estos gráficos nos indican las rutas metabólicas o procesos biológicos en los que están implicados nuestros genes para cada comparación enriquecida. Se suelen representar los valores de p relacionados con el enriquecimiento y la proporción de genes que están implicados en cada ruta. Cuantos más genes estén diferencialmente expresados en una ruta metabólica, mayor será el punto, por lo que representan unos gráficos muy importantes.

Los “volcano plots”, “heatmaps” y “doptplots” se representan en el HTML “Figuras” en el GitHub.

Debemos de recordar que para algunos grupos existían pocos individuos que comparar (menos o 5 individuos por grupo). Por ello, no podemos aceptar los resultados de las comparaciones de BM – MYB, BM – LMO1, BM – TAL2_LMO1, BM – TAL2, HOXA – TAL2_LMO1, LMO2 – TAL2_LMO1, (LMO1 + LMO2) – MYB, (LMO1 + LMO2) - TAL2, LMO2 - TAL2_LMO1, MYB - TLX3, TAL1 - TAL2_LMO1, (TAL1 + TAL2) - TAL2_LMO1, (TAL1 + TAL2) – MYB, (TLX1 + TLX3) - TAL2, (TLX1 + TLX3) – MYB y (TLX1 + TLX3) - TAL2_LMO1 de GSE26713. Tampoco hay suficiente poder estadístico, dado el bajo número de muestras por grupo, para extraer conclusiones con seguridad de GSE33470. Lo mismo ocurre con los resultados de las comparaciones de GSE62156, excepto de las comparaciones Immature + Immature_HOXA13_t - (TLX1 + TLX3) y (Immature + Immature_HOXA13_t) - TLX3.

A la par que el análisis de datos de microarrays, se realizó el análisis de datos de RNA-seq del GSE110633 y GSE110636. También se volcaron los datos en R, pero en este caso eran de tipo “ReadsPerGene.out.tab”, un tipo de formato que se puede analizar transformándolos en matrices de contajes (“CPMs” o “counts per million”), que fue el principio de nuestro análisis, para poder comparar las muestras en número. Primeramente se introdujeron las lecturas por millón, luego se transformaron en una matriz de contajes, y posteriormente se estandarizaron en forma de CPMs. Esto se realiza para que los datos de los contajes estén balanceados en cuanto a la cantidad de secuencias.

Sobre estos datos de RNA-seq se realizaron los mismos tipos de análisis que para los datos de microarrays, pero con un orden distinto: se realizó un preprocesado de los datos, una exploración, un análisis de expresión diferencial, una anotación de los

resultados y un análisis de enriquecimiento. Se puede estudiar el orden según la figura 2.

En cuanto al preprocesado de los datos, posterior a la estandarización de los contajes, hicimos un filtraje de genes poco expresados, después una normalización y por último una transformación logarítmica.

Antes de la selección de genes los valores normalizados se sometieron a un filtraje no específico para eliminar los genes de baja señal (los genes cuya media de la señal en cada grupo no supera un umbral mínimo) y los genes de baja variabilidad (los genes cuyo rango intercuartil entre todas las muestras no superó un umbral mínimo). Realizamos un filtraje estándar que retuvo el 50% de los genes con mayor variabilidad de entre aquellos que están correctamente anotados según la base de datos Entrez.

Realizamos la normalización a través del método TMM, con un cálculo de factores de normalización para definir el tamaño efectivo de las muestras. También ejecutamos una transformación logarítmica a partir de los resultados normalizados para trabajar en estas unidades. El resultado del análisis se representó mediante “heatmaps”.

Sobre la exploración de los datos, se estudiaron diversos gráficos exploratorios que indicaran una buena calidad de los datos o si existen problemas en las muestras. Estudiamos la distribución de los contajes y se hizo un análisis de similitud entre muestras. Dentro de este último se estudiaron las matrices de distancias entre pares de muestras, y también se realizaron un estudio de agrupamiento jerárquico entre muestras y un PCA para visualización en dimensión reducida de las agrupaciones y comportamientos atípicos de las muestras junto con las demás.

El penúltimo de los pasos se refiere a un análisis de expresión diferencial para la selección de genes entre grupos. Se realizó a través de dos objetos que trabajan con modelos lineales generalizados (GLMs), y en última instancia se regularizaron los estimadores de los errores a través de los modelos de Bayes, igual que para los datos de microarrays. Tras ello, se representaron los resultados a través de “volcano plots”. Se eligieron los genes con un p valor menor a 0.05 y un logaritmo de fold-change superior a 2. Estos valores son los que se suelen definir en este tipo de análisis.

Luego se realizó la anotación de los identificadores genéticos y también un análisis de enriquecimiento, a partir de los identificadores de Ensembl. También se representaron los resultados en gráficos, de forma en que todos los archivos mencionados anteriormente también se encuentran los análisis de estas dos series GSE110633 y GSE110636.

De todos los resultados de las comparaciones de cada serie se generaron documentos tipo Excel (.xlsx) de los genes diferencialmente expresados y su descripción y papel en procesos biológicos. Debido a la cantidad de información tan elevada, y a que existen muchos genes que no son de interés en este estudio, en el apartado 4.3 se comentará cómo se han seleccionado los genes que estén implicados en procesos de empalme y que generen RBPs.

4.2 Enriquecimiento funcional a partir de otras bases de datos

Debido a que es posible que en el proceso de anotación de las bases de datos de las comparaciones no se indicara si existe una relación entre un gen particular y procesos relacionados con eventos de empalme de ARN, de forma que podríamos interpretar que ese gen no nos interesa para el estudio, se realizó una anotación a partir de otras bases de datos.

Tomamos las bases de datos de tres autores (9,10,31) para anotar genes que dieran lugar a proteínas implicadas en procesos de splicing de ARN mensajero.

Usamos el dataset llamado Dataset_S01 (XLSX) de Huang et al., 2018, las tablas Table S2 y Table S5 de Wang et al., 2019, y el archivo Supp Tables.xlsx de Sebestyén et al., 2016 para el enriquecimiento con nuevas anotaciones. La elección de las tablas es referente a que, de alguna forma, tienen anotados en cada gen o proteína si la funcionalidad de estos en el genoma es referente a eventos de splicing.

Estas tablas se cargaron en R, al igual que cada uno de los resultados de nuestras comparaciones entre grupos de pacientes. Después se leyeron ambas tablas, y si tenían algún gen o proteína común anotada de la misma forma, se guardaban objetos con el nombre de los genes, el de las proteínas (si se hacía referencia a ellas), la clasificación del tipo de RBP según los autores y la función biológica de cada gen. Con ello se generaron nuevas tablas de anotaciones.

Los pasos para el enriquecimiento funcional a partir de otras secuencias se pueden encontrar en los documentos “Anot_RBPs_articulos1”, “Anot_RBPs_articulos2” y “Anot_RBPs_articulos3” de la página de GitHub.

Además, en Figuras de GitHub podemos ver algunos diagramas de Venn. Estos se crearon a partir de dos medidas, referentes a cantidades de genes. El círculo azul hace referencia a la cantidad de genes expresados diferencialmente que contiene una tabla de nuestros resultados, por ejemplo, la referente a la serie GSE8879. El círculo gris muestra la cantidad de genes que contenía una de las cuatro tablas usadas de otros autores: o bien una de Huang et al., 2018 (31), o las dos indicadas de Wang et al., 2019 (10), u otra de Sebestyén et al., 2016 (9). En la intersección entre ambos círculos de las figuras se ha indicado la cantidad de genes una de nuestras bases de datos que han sido anotadas a partir de este análisis de enriquecimiento. Es decir, en la zona de unión de ambas tablas se indican los genes de nuestros resultados entre comparaciones que aparecen en las listas de referencia. Así, podemos comprender cuántos genes han resultado de nuestros análisis y tambien se encuentran en resultados de otros autores, a parte de favorecer la comprensión de las funciones de nuestras proteínas.

Debemos de tener en cuenta que este proceso no generó tablas con todo los genes anotados por nosotros, referentes a procesos de splicing, y los demás genes nuevamente anotados que indicaron otros autores que tenían que ver con este tipo de procesos. Las tablas anteriormente generadas conllevan toda la información de los genes expresados diferencialmente. Aquí desarrollamos un análisis para la creación de otras nuevas solamente con los genes anotados como codificadores de RBPs.

Como comentábamos al principio, existen RBPs que no están involucradas en el proceso de empalme de ARN mensajero. Si los genes se relacionaron con RBPs, podemos seleccionarlos, aunque no se indique si presentan un papel importante en la maduración del pre-ARN. Entonces, un apartado de la metodología, explicado posteriormente, hace referencia a la búsqueda de información del papel de cada RNA binding protein para la comprensión del papel de estas proteínas a nivel celular y bioquímico, y en última instancia, en la enfermedad.

4.3 Selección de genes involucrados en procesos de splicing de RNA

En las tablas generadas a partir de la anotación y el enriquecimiento funcional que se suele realizar se ha indicado la funcionalidad de muchos de nuestros genes diferencialmente expresados. Debido a que tenemos demasiada informacion de muchos tipos de genes, creímos necesario desarrollar un método para obtener la informacion de aquellas RBPs que tengan relación con procesos de splicing de ARN mensajero.

Además, esto se ha utilizado a modo de ejemplo, como interrogación de una de las comparaciones enriquecidas por nosotros (no en referencia a los tres autores) para encontrar genes involucrados en procesos concretos que desee buscar cualquier usuario. En nuestro caso, queremos ver, para todas las comparaciones de todas las series GSE, si los genes se relacionan con procesos de maduración de ARN. Así, podemos buscar los genes que den lugar a RBPs.

Para correr el código y generar tablas a partir de interrogaciones con otros tipos de cuestiones se puede descargar el archivo “Ejemplo_interrogación” del GitHub a parte de las tablas anotadas.

Cualquier usuario que tenga descargados nuestros resultados puede cambiar ciertas partes del código para buscar cualquier tipo de función anotada de los genes. En el código se comentan las partes del análisis y qué debemos cambiar si queremos realizar otro tipo de búsqueda. El usuario puede generar tablas de resultados enriquecidos según su criterio y además tablas de Excel con estos resultados, de tipo .xlsx. El código y cómo cambiarlo para generar otras tablas a partir de funciones concretas se comentan paso a paso en el documento.

Realizamos una búsqueda de “splicing” y “binding protein”. De esta manera, aquellos genes y proteínas involucrados en procesos de splicing de ARNm, o simplemente RBPs se obtuvieron. Posteriormente creamos una tabla en Word (ver Anexo) donde se indican tanto las series, como las comparaciones, como los genes y rutas y el proceso o tipo de proteína que tienen relacion con los genes, y que son relevantes para contestar a nuestras cuestiones.

5 Resultados

A continuación se mencionan los resultados más importantes del preprocesado y exploración de los datos de microarrays:

-GSE8879: un histograma y boxplot muestra que las distribuciones de los datos son similares en formas pero no en posición, por lo que probablemente se necesite un centrado de los datos. Esto quiere decir que cada barra, de cada muestra, del histograma es igual de "alta", pero se encuentra alejada del resto de barras de las demás muestras. Es posible que, tras un análisis de las fechas de análisis de las muestras en el laboratorio, exista efecto batch entre las muestras, aunque parece que se distribuyen aleatoriamente tras un análisis a través de un PCA y un agrupamiento jerárquico de las muestras.

-GSE10609: se realizaron los mismos análisis anteriores, que indicaban que los datos necesitaban centrarse. Un agrupamiento jerárquico y un PCA mostraron que la muestra numero 9, del tipo HOX11, era distinta al resto. Posiblemente exista efecto batch.

-GSE14618: vimos que los datos de los GSM o muestras de GPL570 eran diferentes y requerían centrado, es decir, algún tipo de metodología para que se parecieran más entre sí con respecto a un centro. Además, C23, F13 y C22 eran los que más definían la primera componente principal (C23 y F13) y la segunda (C22) de un PCA hecho sobre los datos de GPL570. El clúster jerárquico indicó que, de estos tres, F13 y C22 eran los más diferentes al resto de muestras. Aunque no supimos determinar el efecto batch según otro método no relacionado con el día de extracción de la muestra, por la falta de información proporcionada por los autores, pensamos que, como todas las muestras se analizaron el 28 de enero, no existe efecto batch relacionado con el día de los análisis de laboratorio. Sobre GPL96 descubrimos que el GSM365050 o F5 era el que más variación aportaba a la base de datos, y ademas también GSM365078 o C15 era muy distinto al resto de muestras, según un PCA, aunque un clúster agrupó todas las muestras, indicando que el más diferente al resto era solamente C15.

-GSE26713: un PCA reveló que las muestras del día 02/22/05 se parecían más entre sí, y menos que al resto de muestras.

-GSE28703: el clúster y el PCA revelaron que la muestra más diferente al resto era SJTALL157 antes de la normalización. No detectamos batch.

-GSE32215: contiene muestras diferentes entre sí, en general, por lo que requerirá de centrado. Las muestras de T-ALL de adulto se parecen mas entre si que con respecto a un gran grupo de muestras del grupo pediátrico, según el PCA. Para este ultimo grupo se diferencia una nube de muestras semejantes al otro grupo, y otra nube muy diferente, a lo largo de la segunda componente principal. Las muestras 141, 144, 39, 219, 194 y 156 definen la componente principal primera, todos pediátricos. La segunda componente la determinan el 159 (adulto), 141, 166 y 144, sobre todo, siendo estos últimos pacientes pediátricos.

-GSE33469: los datos muestran distribuciones muy similares tanto en forma como en posición. No parece que exista ninguna muestra malograda. El PCA nos indicó que las muestras que definen mayormente la segunda componente principal son las muestras 44 y 57, y el clúster indicó que la muestra 12 era la que más se diferenciaba del resto,

aunque no hicimos un análisis más profundo debido a que no parecen atípicos según el resto de análisis.

-GSE33470: de nuevo, el histograma y boxplot nos arrojaron información sobre que las muestras son muy parecidas. El PCA nos hizo ver que las muestras de CD4ISP se agrupan juntas y se asocian negativamente con la primera componente principal. El resto de muestras en este eje parecen distribuirse aleatoriamente. A pesar de ello, a partir de la segunda componente principal interpretamos que debemos te tener cuidado con las muestras "CD34+CD1a+ rep2" y "CD34+CD1a- rep1", ya que definen mucha parte de la variabilidad de esta. El clúster de distancias euclídeas no indicó que estas muestras fueran diferentes al resto.

-GSE37389: el histograma nos indicó que haría falta una normalización, ya que las muestras eran dispares en cuanto a su densidad. El boxplot realizado para las muestras también nos dio la misma información aunque no sobresalía la disparidad de ninguna muestra en concreto. A través del PCA, en cambio, pensamos que "G4: TCR-GD013" y "A3: TCR-AB003", y el agrupamiento corroboró nuestras sospechas sobre la primera de estas dos muestras.

-GSE56488: las muestras se parecían en forma y distribución, pero había una distinta según el histograma. El clúster nos hizo ver que las muestras 5 y 9 se parecían más entre sí con respecto al resto. Por eso pensamos en que realizar un centrado de los datos sería importante. No pensamos que haya efecto batch debido al momento de realización de los análisis de laboratorio por las similitudes. La muestra 13 es distinta al resto según la primera componente del PCA, lo cual no parece tener que ver con la hora.

-GSE62156: según el histograma y el boxplot parece que existen dos grupos de muestras con tamaños diferentes, aunque no corresponden a grupos concretos, por lo que parece que necesitamos un proceso de normalización general. El PCA mostró los mismos resultados, dos nubes de puntos que definían en la parte positiva del eje de la primera componente, y otra nube a lo largo de la parte negativa. Las muestras JSR_044 y JSR_064 generan la mayor parte de la variabilidad del segundo eje. De nuevo, el clúster nos arroja la misma información de las dos nubes de muestras.

Para las muestras de RNA-seq se hizo un estudio de otro tipo:

-GSE110633: los contajes normalizados se parecen entre sí, según un boxplot. Podríamos entender que las muestras TLX28, TAL16, TLX35 y TAL40 son las que tienen valores de CPM atípicos, aunque el log de los recuentos por millón de lecturas asignadas no parece mostrar diferencia entre muestras.

-GSE110636: en gráficos que podemos realizar a partir de los contajes normalizados vemos que hay muestras, como la GSM3004630/TALL_JS_26/TLX11, que tienen valores relativamente bajos con respecto al resto (cerca de 12500), y muestras con una gran cantidad de contajes, como GSM3004621/TALL_JS_3/TLX2, que tiene cerca de 25100 contajes tras la normalización. Si realizamos el logaritmo de estos contajes se parecen mucho más. El agrupamiento jerárquico entre muestras mostró que el logaritmo en base dos de los contajes de GSM3004629/TALL_JS_23/HOXA10, GSM3004640/TALL_JS_58/TLX21, GSM3004636/TALL_JS_44/TAL17 y GSM3004637/TALL_JS_51/TAL18 se parecen más entre sí que con respecto al resto de muestras. A pesar de ello, realizamos un escalado multidimensional de estos, y su

representación no indicó que hubiera ciertas muestras que fueran diferentes al resto, y no distinguimos una distribución clara de los grupos en general.

Teniendo en cuenta estas consideraciones, realizamos el resto de pasos del análisis. Realizamos una serie de volcano plots para la representación de los genes diferencialmente expresados entre comparaciones de cada GSE. Estos se pueden estudiar en el HTML de Figuras del Anexo. Estas conllevan la representación de los genes para el conjunto del GSE, de forma que todos los genes expresados diferencialmente entre grupos se indican en ellos.

A modo de ejemplo se estudiará el volcano plot de la figura 3, referente al GSE8879 y a su comparación entre grupos ETP – no ETP. Aquí consideraremos genes significativamente expresados a aquellos con un p-valor de 0.05, y con un valor del log de fold change inferior a -1 o superior a 1. Entonces, debemos buscar aquellos que se distribuyan en el eje vertical a partir de 1.3, aproximadamente. Los genes con mayor expresión para el grupo primero (ETP) se muestran en el lado derecho. Los que son de gran interés por sus valores altos del logaritmo en base dos del fold change y bajo p-valor son GSS, MEF2C y MYH10. Por el contrario, aquellos genes más expresados en el segundo grupo (no ETP) son TGFB1, ANGPT1, CDC25B, ANGPT1, RASGRP1, MYCN, CD28 y TCF7. A pesar de esto, estos genes no son los que se suelen expresar diferencialmente, según la tabla “Tabla genes y proteinas RNA splicing.docx”. Por ello, si queremos estudiar qué genes particulares se expresan más en ciertos grupos, y menos en otros, debemos de estudiar estos diagramas, y no solamente entender qué proteínas son las que se encuentran mayoritariamente en todas nuestras tablas de resultados.

Como decíamos, los genes más diferencialmente expresados se pueden examinar a partir de los volcano plots del Anexo, y los heatmaps se pueden revisar para estudiar los diferentes perfiles de expresión. Comentaremos los resultados más importantes de los heatmaps, de aquellas muestras cuya expresión se diferencia del resto, ya que los volcano plots indican genes diferencialmente expresados que luego comentaremos.

-GSE8879: de nueve muestras, cinco se parecen más entre sí con respecto a sus perfiles de expresión que con respecto al resto de muestras, que corresponden a pacientes con ETP. Entre ellas, se encuentran las muestras 4, 15 y 47, mientras que la 22 y 23, que también hacen referencia a ETP T-ALL. De las muestras que tienen mayores valores de expresión génica son los pacientes con ETP, en general. Todo ello se puede estudiar a partir de la figura 4, a modo de ejemplo de cómo son el resto de los heatmaps.

-GSE10609: ciertas muestras, como la 25, 23, 13 y 49, se parecen entre sí más que con respecto al resto. Estas son de pacientes con HOX11, HOX11L2, HOX11, y unknown. Entonces, pensamos que los pacientes con HOX11 T-ALL tienen altos valores de expresión génica que con respecto al resto de muestras.

-GSE14618: observamos a partir del heatmap del GPL570 que varios NR se parecen más entre sí que con respecto al resto (entre ellos, NR1 y NR4). Es decir, “no response” se parecen más entre sí, en general, que con respecto a F (“failure, relapse”) y C (“complete continous remission”). GPL96 a su vez presenta NR4 como muy diferente con respecto al resto de muestras, con niveles de expresión muy por encima del resto. Estas muestras no son las que se consideraban diferentes según los análisis del preprocesado de los datos de esta serie.

-GSE26713: no parece que haga distinción entre grupos. No hay ciertas muestras se parezcan más entre ellas con respecto al resto y correspondan a un grupo concreto.

-GSE28703: las muestras rojas, de pacientes con "early T-ALL", se parecen más entre sí que con respecto a la mayoría de pacientes sin ETP. Algunas muestras azules, que representan pacientes sin ETP, como SJTAL050, SJTAL041, SJTAL071, SJTAL075 Y SJTAL019, se diferencial del resto de muestras azules y se asemejan a los pacientes con T-ALL temprana.

-GSE33470: las muestras CD34+CD1a+ rep1, CD34+CD1a- rep1, CD34+CD1a- rep3, CD4ISP rep1, CD4ISP rep3, CD4ISP rep2 se parecen más entre sí. Además, las muestras de CD4ISP se parecen más entre sí dentro de este grupo, estando las muestras de este tipo todas juntas.

-GSE62156: las muestras del grupo TAL-R, que son las de la derecha, se parecen más entre sí que con el resto, a excepción de la 149.

-GSE110633: en este caso los perfiles de expresión son parecidos entre muestras. No parece que haya grupos que se relacionan más entre si que con respecto al resto.

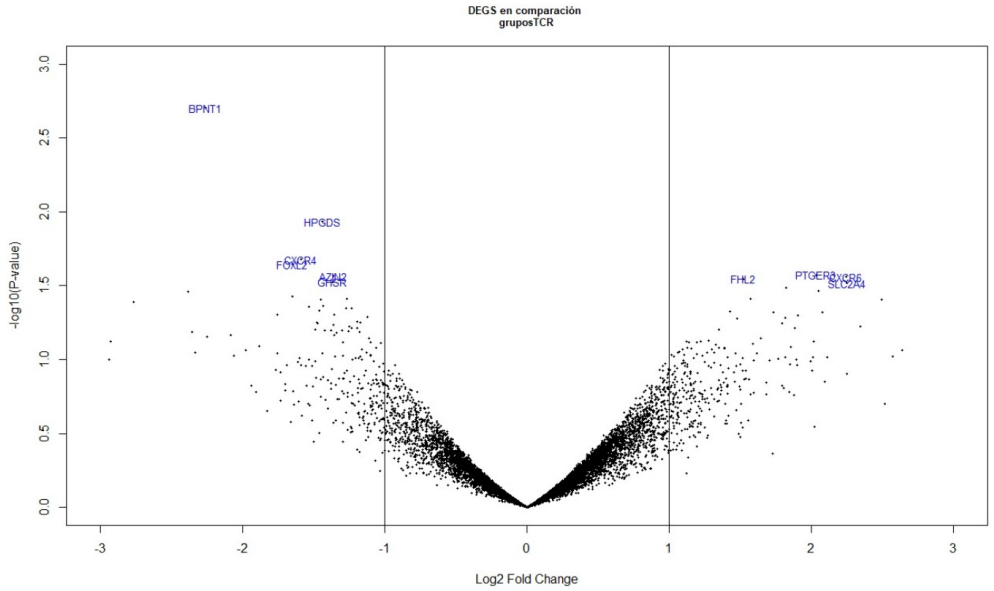
-GSE110636: cuatro individuos se parecen más que con el resto. Estas son las muestras GSM3004640/TALL_JS_58 (grupo TLX), GSM3004629/TALL_JS_23 (grupo HOXA), GSM3004636/TALL_JS_44 (grupo TAL) y GSM3004637/TALL_JS_51 (grupo TAL).

Las dos últimas son del mismo grupo.

Podemos determinar una relación entre ciertas muestras que poseían determinadas distinciones con respecto del resto según los análisis exploratorios y los perfiles de expresión para ciertos GSE. Por ejemplo, en GSE33470 vimos que CD34+CD1a- rep1, junto con otras muestras, se parecía en cuanto a los perfiles de expresión menos que al resto, y hemos visto en la exploración que solo esta, de ese grupo, que definía gran parte de la variabilidad de la segunda componente principal de un PCA con todas las muestras. También hemos visto que las muestras GSM3004629/TALL_JS_23/HOXA10, GSM3004640/TALL_JS_58/TLX21, GSM3004636/TALL_JS_44/TAL17 y GSM3004637/TALL_JS_51/TAL18 se parecían entre sí según el agrupamiento jerárquico del logaritmo en base dos de los contajes, y también presentaban perfiles de expresión similares. Sin embargo, no encontramos para estas similaridades un efecto batch debido ni al día de realización de los análisis de laboratorio, ni de la hora, ni del centro en el que se realizaron.

Además, se realizaron dotplots, gráficos muy importantes para entender las funciones de la mayoría de los genes expresados diferencialmente entre grupos comparados. Uno de nuestros resultados se puede observar en la figura 5. La mayoría de genes expresados diferencialmente para GSE8879 intervienen en procesos de activación de leucocitos, en la activación de linfocitos, en la respuesta inflamatoria, en la regulación de una señal de transducción de GTP, en una ruta antigénica y en otra ruta relacionada con la señalización de un receptor de células T. Estos resultados son clave para comprender los procesos globales que se desarrollan detrás de la enfermedad, y se pueden estudiar todos en el HTML de figuras del GitHub. A pesar de ello, si quisiéramos estudiar qué función concreta realiza un gen particular, o si da lugar a RBPs, podemos consultar las tablas de nuestros resultados, en los que se indican los genes particulares y también estas rutas en las que están implicados.

Figura 3. Volcano plot de los datos de GSE8879 entre el grupo de individuos con linfoma linfoblástico agudo precursor de células T tempranas (ETP) y otro grupo con el resto de individuos.



Algunas de las proteínas más importantes en nuestros resultados, que aparecen más de 80 veces en nuestra tabla word genes, son las del tipo POLR2, las variantes de hnRNP, aquellas que constituyen los grupos NCBP, POLR2A y también POLR2B, SRSF9 y las SRSF, la familia de genes RBM, las ribonucleoproteínas nucleares pequeñas o SNRNP, las proteínas B' asociadas a ribonucleoproteínas nucleares pequeñas SNRPB, tambien las SNRPD, nucleoporinas NUP, ribonucleoproteínas nucleares pequeñas, los factores de especificidad de escisión y poliadenilación CPSF, los antígenos de superficie SRS para virus, las LSM2 y también el resto del conjunto de proteínas del tipo LSm, y las DEAD-box o DDX.

Por otro lado, no descubrimos genes o proteínas propias de un grupo concreto. En algunos casos en las comparaciones entre grupos, existen muchas proteínas comunes, pero ninguna se relaciona con un grupo particular. Entonces, no parecen existir proteínas que definan inmunofenotipos a partir de nuestros resultados.

A parte, como ya hemos comentado, realizamos una serie de anotaciones funcionales de las proteínas resultado a partir de tres artículos. Luego, generamos una serie de diagramas de Venn. Todas estas figuras se recogen, de nuevo, en el HTML de Figuras de GitHub. En general, los genes anotados a partir de nuestras bases de datos son muy pocos, en comparación con las nuestras y con las tablas del resto de autores. Todas las tablas se pueden consultar en la carpeta Tablas del GitHub.

Figura 4. Heatmap de los datos de GSE8879 entre el grupo de individuos con linfoma linfoblástico agudo precursor de células T tempranas (ETP) y otro grupo con el resto de individuos. Las muestras en rojo hacen referencia al grupo "diagnostic leukemic blasts of early T-cell precursor acute lymphoblastic leukemia (ETP)", y en amarillo se muestran las que no son del tipo ETP.

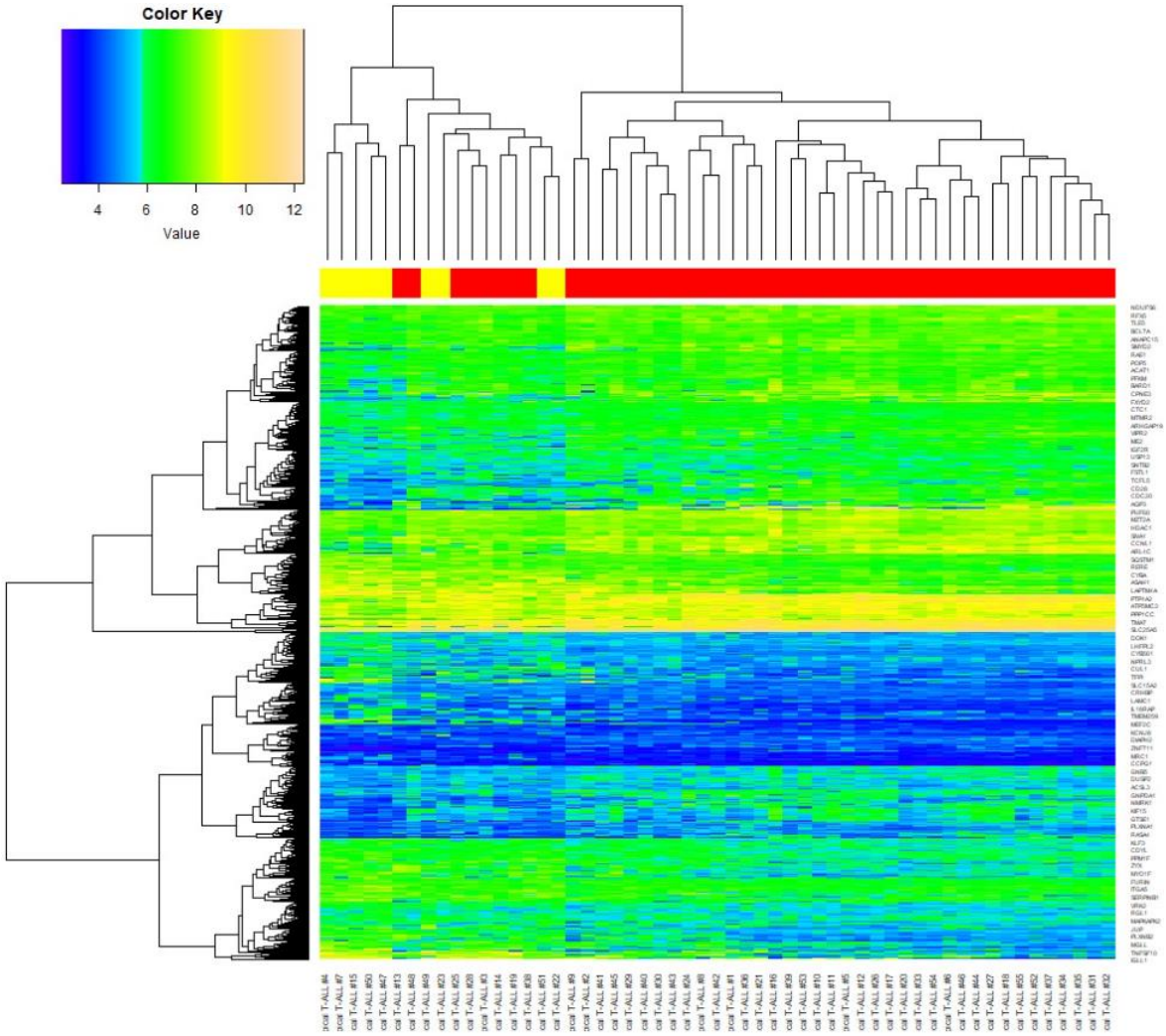


Figura 5. Doplot de los datos de GSE8879 entre el grupo de individuos con linfoma linfoblástico agudo precursor de células T tempranas (ETP) y otro grupo con el resto de individuos.

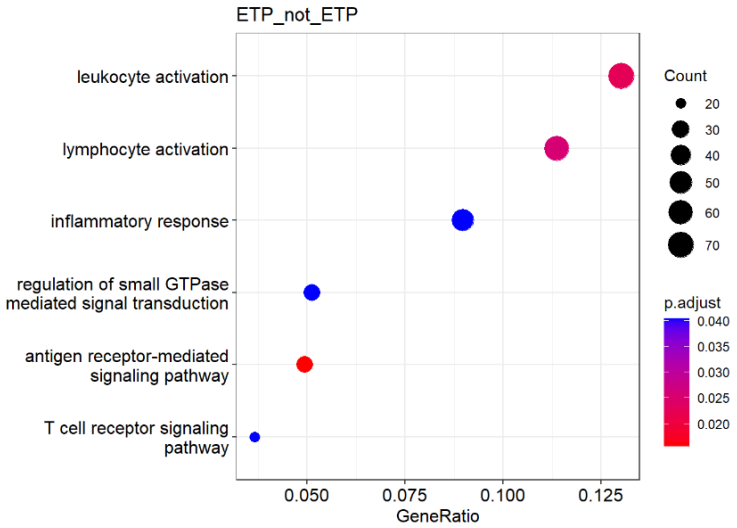


Figura 6. Análisis de componentes principales de los valores de expresión de las muestras crudas sin tratar de GSE33470.

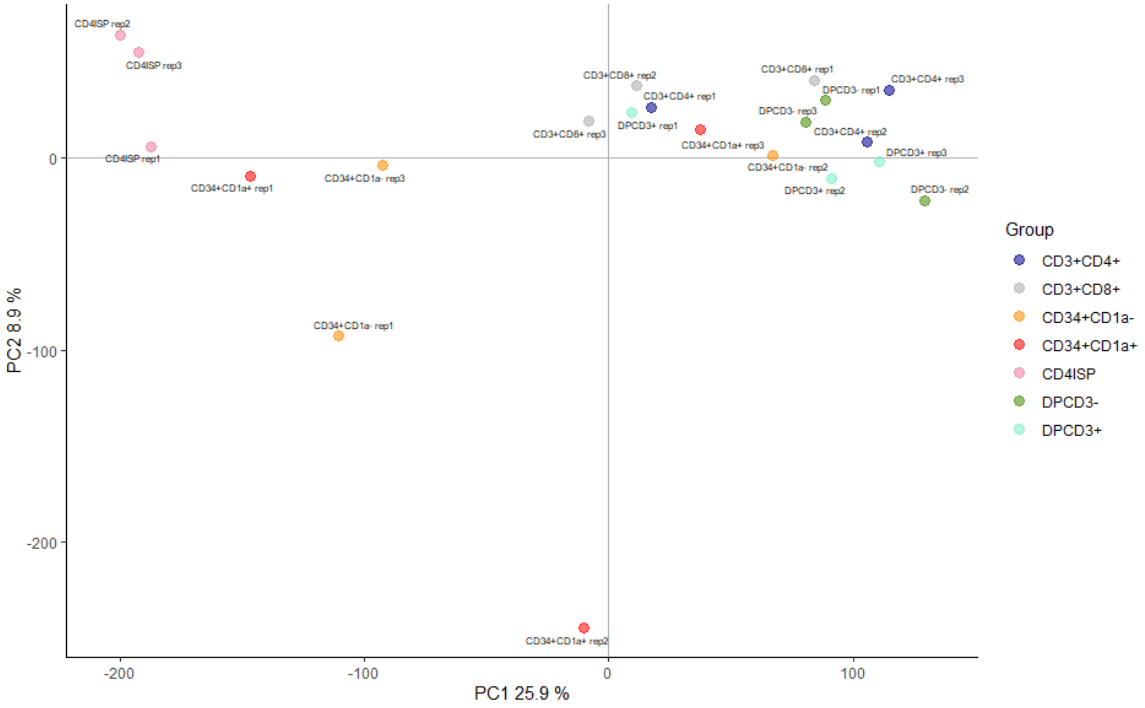
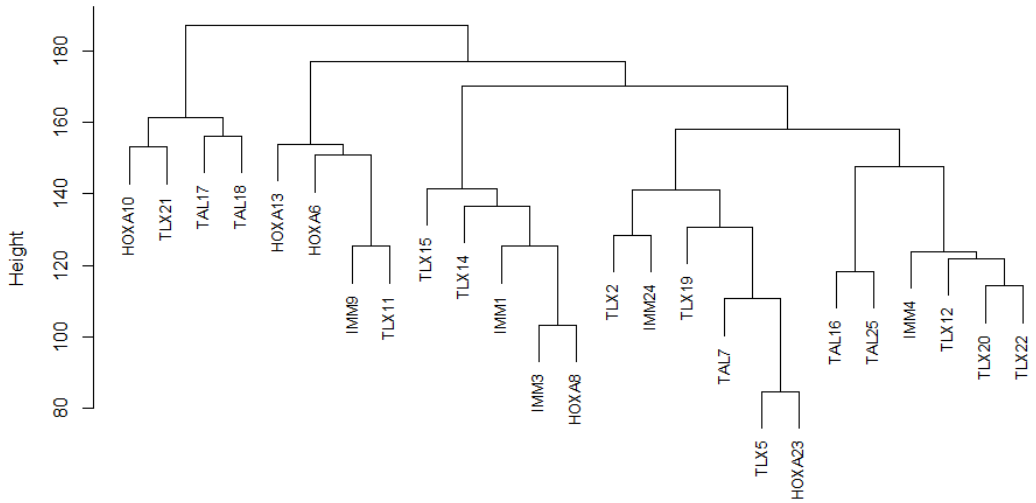


Figura 7. Agrupamiento jerárquico en función de la matriz de distancias euclídea del logaritmo en base dos de los contajes normalizados para las muestras de GSE110636.



6 Discusión

Recientemente se ha entendido que el empalme alternativo de ARNm está recurrentemente implicado en el desarrollo y progresión de tumores. No solo esto, sino que también puede suprimir el crecimiento tumoral, por lo que pensamos que entender cómo funciona en tejidos tumorales generará avances en tratamiento de la leucemia linfocítica de células T (32).

En nuestro estudio hemos determinado, a partir de la expresión génica de pacientes con T-ALL, miles de proteínas implicadas en el empalme de ARNm. Se ha visto que las diferentes regulaciones de ciertas RBPs son muy importantes en T-ALL, como por ejemplo TTP, importante supresora de tumores (28). Potencialmente, y debido a ambas razones, creemos que el estudio será una base teórica para otros de tipo molecular, que esperamos serán muy importantes para comprender la enfermedad.

En nuestra tabla se puede comprobar como ciertas proteínas, identificadas a partir de los genes expresados diferencialmente, son muy comunes. Algunas de ellas ya se han estudiado, y se ha visto que eran importantes en procesos tumorales. Por ejemplo, una de ellas es la generada a partir del gen RBFOX2, cuya funcion es tan importante que se pueden clasificar las líneas celulares cancerosas del tejido mesenquimatoso a través de sus firmas de empalme alternativo de la proteína FGFR2 (34).

En relación con lo anterior, también existen ciertos grupos de proteínas que hemos visto que eran muy importantes en nuestras bases de datos. Las ribonucleoproteínas nucleares heterogéneas (hnRNP) y las proteínas ricas en serina/arginina (SR) también son necesarias para los procesos de empalme alternativo de ARNm y, en última instancia, generan diferentes proteínas a partir de una sola molécula de ARN. Se ha visto que en varios tipos de cáncer estas proteínas generan otras relacionadas con procesos cancerígenos, y aparecen en demasía en nuestra tabla de comparaciones (33).

Podemos citar algunas RBPs estos dos últimos grupos de proteínas, como SRSF1, hnRNPH y hnRNP A2/B1. La regulación del empalme alternativo de MST1R, a través de estas tres proteínas, puede dar lugar a transiciones de estados epiteliales, ligado a su vez con el desarrollo de cáncer (34).

Como se puede entender, existen varios tipos de cáncer en los que los procesos mediados por RBPs son importantes. Otra proteína de interés del tipo hnRNP es hnRNPM, la cual ha sido reportada como importante generadora de metástasis del cáncer de mama (36), cáncer de próstata (37) y cancer gástrico (38) y cáncer de colon (39), entre otros. Cuando se sintetiza esta proteína, a través de una forma de regulación concreta, se activa el empalme alternativo, y con ello las células se dividen sin control (36).

Los cánceres de cuello uterino y la neoplasia cervical son otros tipos de neoplasias que también han sido asociados con estos grupos de proteínas. Estas enfermedades están relacionadas con infecciones víricas, en concreto, con el virus del papiloma humano. Se ha demostrado que estos grupos de proteínas ayudan, mediante el empalme alternativo, a la propagación vírica. Hoy en día estas vías moleculares y celulares son de gran interés para generar nuevas terapias efectivas (33).

Como vemos, existen RBPs relacionadas con procesos tumorales y víricos. Algunas de ellas están relacionadas con varios tipos de virus y con procesos hematológicos, como el del papiloma humano que comentábamos. A modo de ejemplo, otro virus, el del VIH-1, está relacionado con la DHX9 y NCBP1, y esta última con NCBP2, representadas en nuestra tabla Word anexo en múltiples comparaciones como importantes en procesos de empalme de ARNm (43). Junto con la última, ya se ha informado en varios estudios sobre que NUP98 es importante en neoplasias hematológicas malignas, y nuestro estudio puede corroborar que las tres lo son (44).

Como hemos visto en varios casos, se ha descubierto una relacion entre ciertos virus humanos con neoplasias y RBPs. Además, en nuestros resultados se expresan diferencialmente muchas proteínas SRS, antígenos de superficie relacionados con procesos de virulencia (45). Entonces, es posible que la respuesta a infecciones víricas sea importante en procesos relacionados con RBPs y T-ALL, pero no se ha estudiado el tema todavía.

Como también hemos comprendido, algunas de las proteínas de nuestros resultadas están relacionadas con otros tipos de cáncer, pero otras tambien se han determinado como importantes en la leucemogénesis. Por ejemplo, SRSF2 es un factor importante que interviene en el empalme de pre-ARNm en pacientes con cáncer, y se sabe que es clave en leucemias (32,40). Se conoce que en enfermedades neoplásicas una mutación en el gen que produce esta proteína tiene una especificidad mayor a diferentes secuencias de pre-ARNm. Entonces, se generan nuevos ARNm maduros, en un proceso de empalme alternativo, que se han unido de manera diferente con respecto a individuos sin este tipo de enfermedades (32).

Otro de los genes importantes tanto en otros cánceres como en la leucemia es NCBP2. Se ha descubierto que la proteína del gen NCBP2, que promueve el empalme del pre-ARNm y otra serie de regulaciones traduccionales, se encuentra en zonas hipóxicos tumorales y que está asociado a fibroblastos asociados al cáncer, de forma que puede producir metástasis. Los tratamientos para los pacientes con hipoxia tumural son poco eficaces (41). Este gen también se ha relacionado recientemente con la leucemia mieloide aguda en niños, y lo tenemos representado en nuestra tabla (42). No solo en la mieloide, sino que nosotros reportamos que podría ser importante en la leucemia linfoide.

De cualquier forma, existen ciertas proteínas importantes que se han denotado como importantes solamente en la leucemia. Se sabe que en leucemias linfocíticas y mieloides están alteradas las funciones de las proteínas de empalme SF3B1, SF1, PRPF40B, U2AF35 y ZRSR2 (40). Se piensa que las células con funciones diferentes de estas proteínas no prosperan, y además se relacionan con el comienzo de la leucemia y definen tanto el fenotipo clínico como el pronóstico del paciente (40). De esta manera, la comprensión de las vías metabólicas que están detrás del empalme alternativo de ARNm en pacientes con leucemia, con ambos tipos, parece muy importante como para olvidarnos de estudios centrados en este tipo de cuestiones.

A pesar de esto, debemos de saber diferenciar entre ambos tipos de leucemias. Sabemos que existen ciertas similitudes en la genética aberrante que producen RBPs entre la leucemia linfocítica y la mieloide, como hemos visto en algunos casos, como por ejemplo, en el de NUP98. Sin embargo, hay genes que están mutados en células

mieloides leucémicas, pero no en linfocíticas. Debido a ello, creemos que es de suma importancia la determinación de los genes generadores de RBPs diferencialmente expresados para cada grupo leucémico. Esto nos puede llevar a comprender mejor la leucemia T-ALL, y creemos que puede ser muy importante a la hora de generar terapias adecuadas. Aquí presentamos varios genes y proteínas candidatas para distinguir transcriptómicamente ambos tipos de leucemia.

Un ejemplo es el de la RBP MSI2, que aumenta la expresión de FLT3 si existe, en la leucemia mieloide. Se cree que el proceso de empalme de ARNm generado por MSI2 a partir de FLT3 regula el crecimiento de la leucemia (46). Sin embargo, esta proteína no está representada en nuestra base de datos, por lo que creemos que no está implicada en procesos relacionados con T-ALL. Como proteína importante en T-ALL citamos anteriormente la TTP, que puede desencadenar procesos tumorales, aunque esta no la encontramos en nuestra base de datos (28). Un gen que codifica una RBP importante en T-ALL, particularmente, es QKI, pero nosotros no la encontramos tampoco (52).

Sin embargo, no hemos sabido identificar, a partir de un análisis visual, grupos de proteínas relacionadas con ciertos subgrupos de T-ALL, de los grupos indicados por los autores que usamos para realizar las comparaciones. En realidad, son los factores de transcripción, proteínas que se unen al ADN, los que se usan para definir los subgrupos moleculares, como TAL1 y HOXA, por ejemplo (53). Así, parece que las RBPs no son importantes para la clasificación de los subgrupos, aunque seguimos pensando que son importantes para entender la enfermedad y generar procedimientos clínicos favorables para todos los pacientes.

Todos los genes comentados que dan lugar a RBPs, excepto donde se indica explícitamente, se encuentran expresados diferencialmente para muchas de nuestras comparaciones entre grupos de pacientes con leucemia (ver Tabla genes y proteinas RNA splicing.docx del repositorio del Github). Así, comprendemos que la desregulación del empalme alternativo se da de forma sustancial entre varios tipos de cáncer, y parece que también juega un papel importante en nuestra leucemia. Dado que los procesos de empalme alternativo son tan importantes, y que ya se conocen muchas rutas bioquímicas relacionadas con RBPs, creemos que el desarrollo de terapias dirigidas a la correcta generación de proteínas no aberrantes derivadas de procesos de empalme alternativo puede generar una mejora en el pronóstico médico de los pacientes con leucemia linfoblástica aguda de células T.

No obstante, aunque hemos determinado algunas proteínas comunes en cáncer, neoplasias y tambien ya estudiadas en T-ALL, la gran mayoría de proteínas son desconocidas en cuanto a su papel en la leucemogénesis y mantenimiento de leucemia aguda de células T. Tal como POLR2A, importante subunidad de la ARN polimerasa II. Esta se sabe que desencadenar el síndrome del neurodesarrollo con hipotonía profunda, pero de momento no hay literatura para determinar la relación de esta con la leucemia (47). Una vez más, señalamos la importancia de más estudios relacionados con los genes que más han aparecido en nuestra tabla de resultados.

A pesar de todas las proteínas que hemos determinado como importantes, debemos de tener en cuenta que ha habido comparaciones en las que uno de los grupos (o conjunto de grupos, si se hacía el promedio de la expresión de dos grupos) tenía menos o 5 individuos. Creemos que con tan poca cantidad de muestras las inferencias

realizadas a partir de los grupos no deberían ser tomadas como representativas de los grupos. Es decir, si por ejemplo estudiamos para un GSE la comparación entre dos condiciones, IMM y TLX3, y para el grupo TLX3 tenemos 3 muestras, es posible que la expresión diferencial entre grupos se deba a particularidades de las muestras concretas. En este ejemplo, podemos suponer que la expresión del grupo TLX3 es referente a los individuos, y no podemos asumir esta expresión para cualquier paciente con TLX3 ni los resultados de sus comparaciones con IMM. Por ello, requerimos de más estudios con suficiente cantidad de individuos para cada tipo de leucemia que nos ayuden a entender si nuestros resultados son extrapolables o no.

A parte, hemos visto que para GSE33470 y GSE110636 existían muestras que se parecían entre si, muestras que no parecían del mismo grupo. Estas eran GSM3004629/TALL_JS_23/HOXA10, GSM3004640/TALL_JS_58/TLX21, GSM3004636/TALL_JS_44/TAL17 y GSM3004637/TALL_JS_51/TAL18 para GSE110636, y también la muestra CD34+CD1a- rep1 de GSE33470 se alejaba de la mayoría de las muestras de cada serie en cuanto a los datos normalizados y filtrados, y en cuanto a los resultados de sus perfiles de expresión. No pudimos determinar estos patrones, ni con análisis de efectos batch ni mediante una búsqueda en los artículos de distinciones entre las muestras. Por ello, pensamos que estas peculiaridades se deben a los pacientes concretos y no a posibles efectos batch del trabamiento ni análisis de las muestras.

7 Conclusiones

7.1 Conclusiones

En este estudio hemos determinado los perfiles de expresión génica mediante comparaciones de muestras de pacientes con leucemia T-ALL, usando datos públicos. Se han determinado tanto los genes como las proteínas más importantes en la enfermedad, a partir de la comparación del transcriptoma de distintos grupos de pacientes. A raíz de esto, hemos comprendido que hay muchos genes desregulados que participan en procesos de empalme alternativo en la leucemia linfocítica de células T. Esperamos que este estudio suponga una fuente de información útil para otros trabajos centrados en T-ALL, ya que cualquier investigador podría obtener diferentes datos de nuestros resultados ya recopilados en una única base de datos, y usarlos como información adicional para sus análisis.

De tal manera, muchos genes se expresan diferencialmente entre pacientes con diferentes inmunofenotipos. Algunas de las RBPs generadas a partir de estos genes, que hemos determinado como importantes, ya se han estudiado en otras enfermedades y cánceres, como por ejemplo las del tipo hnRNP y SR, o RBFOX2 y NCBP2. Además, también se ha determinado que algunas de ellas son importantes en procesos relacionados con leucemias, tanto mieloides como linfoblásticas. Sin embargo, existen otras muchas proteínas de las no conocemos su función ni podemos definir su papel en T-ALL.

Hemos visto que ciertas RBPs importantes en T-ALL interactuaban con otras proteínas relativas al ciclo de vida de los virus. De forma admisible, algunos de los pacientes con T-ALL pueden estar condicionados por ciertos virus, por lo que pensamos que las consecuencias patológicas derivadas de estudios víricos en pacientes con T-ALL pueden ser cuantiosas.

Habida cuenta de que no se conoce del todo el trasfondo genético ni cómo se desencadena y desarrolla esta enfermedad, comprendemos que debemos de intentar determinar las características que hacen peculiar a la T-ALL. Aunque existan similitudes genéticas y proteómicas entre individuos con diferentes tipos de leucemias, pensamos que esta enfermedad se caracterizar a partir de ciertos perfiles bioquímicos y genéticos, y esto nos puede ayudar a entenderla mejor.

En ultima instancia, creemos que es importante determinar los perfiles genéticos y proteómicos relacionados con eventos de empalme alternativo de los grupos inmunofenotípicos de pacientes con T-ALL. Esto puede ayudar, según creemos, a desarrollar tratamientos especializados para cada tipo de paciente y, en ultima instancia, aumentar la probabilidad de supervivencia de estos.

Entonces, no debemos descuidar que todavía hace falta una investigación más profunda a partir de estudios bioquímicos, celulares y bioinformáticos. Creemos que es muy importante y sería muy beneficiosa la caracterización de las rutas bioquímicas y el papel celular para profundizar en la patogenia de la enfermedad.

7.2 Líneas de futuro

En nuestro estudio, hemos tomado datos de análisis transcriptómicos de pacientes con leucemia de diferentes laboratorios, y los hemos analizado.

Sin embargo, como ya hemos comentado en varias ocasiones, las contribuciones relacionadas con la genética o bioquímica de los procesos de empalme alternativo de ARN posiblemente nos ayuden a generar fármacos y procedimientos más efectivos a la hora de tratar la leucemia linfoblástica de células T. También han cobrado importancia en el diagnóstico, seguimiento y evaluación de pacientes con leucemia (49). Por todo ello, los estudios relacionados con estos temas son muy importantes, y creemos que debe de seguir analizándose la información ómica de pacientes con diferentes inmunofenotipos de leucemia.

Ya se conocen grupos de genes y proteínas muy implicadas en procesos neoplásicos hematopoyéticos, como los que generan las proteínas ricas en serina/arginina ya comentadas. En pos de la contribución al entendimiento de procesos genéticos, nuestro estudio indica otras regiones y proteínas que pueden estar implicadas en la T-ALL.

Se han comentado los genes y proteínas ya estudiados por algunos autores como importantes y que aparecen en nuestros análisis, pero hay muchas regiones en nuestros resultados de las que todavía no se entiende su papel. Pensamos que estudios bioquímicos de las rutas indicadas por nosotros, en las que están implicadas ciertos genes y RBPs que hemos determinado como importantes, esclarezcan un camino para el tratamiento de la enfermedad.

No solo esto, sino que también creemos que se pueden desarrollar dianas terapéuticas a partir de la comprensión de la biología celular en la leucemia, en conjunto con la genética y bioquímica. Se sabe que la genética anormal en pacientes con linfomas repercute en receptores celulares de las células T, que conduce a más desregulaciones. También se avanzó en la clínica de los linfomas a partir del número de copias celulares en interfase (50).

Además, requerimos de nuevos estudios de expresión diferencial de pacientes con T-ALL. De estos, esperamos que los datos se sigan proporcionando en portales abiertos. El estudio de la expresión diferencial de genes a partir de este tipo de informacion en la leucemia linfoblástica aguda es muy importante y se cree que proporcionará informacion sobre la patogenia y los procesos subyacentes a la leucemia (51).

Como hemos visto en este trabajo, se pueden determinar muchos genes que podrían significar una importante vía para la proliferación y mantenimiento de células leucémicas. De esta forma, se relacionan todas las fuentes de extracción y procesamiento de los datos de forma eficaz, a partir de los estudios genéticos comentados anteriormente. Creemos que, a partir de ellos, los estudios bioinformáticos como este representarán un avance en la comprensión de la enfermedad.

Así, esperamos que estudios bioinformáticos de esta índole se realicen para nuevas muestras de pacientes con T-ALL, y se podrían proporcionar en bases de datos de fácil acceso. Es cierto que aquí mostramos un método para la extracción de informacion a partir de nuestras bases de datos resultado, pero pensamos que la generación de una shinny app en R, por ejemplo, ayudaría el acceso a estas.

7.3 Seguimiento de la planificación

Hemos realizado una serie de cambios en el calendario desde el principio, desde la definición de tareas inicial. En él no constaban los siguientes pasos que se han realizado:

-La no creación de una base datos global, integración de los distintos GSE descargados de la web de GEO. Se realizaron los análisis de forma individual de cada una de las series, por comodidad y simplicidad. Si para un GSE, por ejemplo, tuviéramos datos de pacientes con leucemia/linfoma linfoblástico agudo de tipo CD7+, CD3e+ y CD43+, el análisis entre grupos sería simple. Sin embargo, si tuviéramos todos los GSE y, por tanto, todas las muestras en un mismo dataset, tendríamos demasiados grupos. Además, el análisis en conjunto de todas las series podría resultar en una pérdida de información de las características de cada muestra de cada GSE.

-Creación de la matriz de contajes a partir de los datos de las lecturas por gen de los datos de RNA-seq. Esto es de nuevo debido a la comodidad de trabajar con este tipo de medidas, en vez de con las lecturas por gen.

-Se han creado, dentro de un apartado que sí que estuvo definido en el calendario anterior, una serie de gráficos explicativos. Esto se ha incluido en los planes debido a la importancia de las representaciones gráficas de los resultados para la comprensión de estos.

-Se han enriquecido con nuevas anotaciones los datos resultantes de cada comparación tras la última etapa del análisis ómico, correspondiente al análisis de enriquecimiento funcional. Esto es debido a que, a pesar de la anotación previa general, es posible que ciertos genes no se hayan anotado, o no se haya detallado toda la informacion referente a RBPs, con la consecuente pérdida de informacion. En este caso, tomamos ciertos datasets de Huang et al., 2018; Sebestyén et al., 2016 y E. Wang et al., 2019 para llevar a cabo este paso, de forma que si ellos anotaron algún gen que diera lugar a RBPs, si nosotros tambien lo tenemos en las bases de datos de nuestras comparaciones, lo anotáramos a partir de sus bases. Usamos el dataset llamado Dataset_S01 (XLSX) de Huang et al., 2018, las tablas Table S2 y Table S5 de E. Wang et al., 2019, y el archivo Supp Tables.xlsx de Sebestyén et al., 2016. La elección de las tablas es referente a que, de alguna forma, tienen anotados en cada gen o proteína si la funcionalidad de estos en el genoma es referente a eventos de splicing, o simplemente si se tratan de RBPs. Si se trata de RBPs, podemos seleccionarlas, aunque no indiquen si presenta un papel importante en la maduración del pre-ARN, y posteriormente buscar si se tratan de proteínas importantes para nuestro estudio, en un apartado definido a partir de la modificación de las tareas, llamado “recopilación de informacion de cada RBPs” (ver Tabla 1). De esta manera, tomamos los resultados de nuestro estudio, con los genes ya anotados, e intentamos volver a anotar aquellos genes en común entre nuestro estudio y de los tres mencionados para indicar si tienen relacion con RBPs.

-Se realizó la interrogación de uno de una de las comparaciones enriquecidas en el apartado anterior para encontrar genes involucrados en procesos de maduración de ARN. Este paso se ha incluido para representar un ejemplo de cómo podemos preguntar a cualquiera de nuestras bases de datos de entre comparaciones de grupos de pacientes. Así, podemos buscar los genes que den lugar a RBPs.

-El siguiente paso hace referencia a la selección de las RBPs implicadas en el splicing de todos los datos, lo cual se puede generar a partir de la interrogación anterior. Se usó para tener listas de las RBPs resultantes entre los análisis de expresión diferencial de entre grupos de pacientes, mediante lo que se culminó el primer objetivo.

-Es importante incluir en la memoria final una tabla descriptiva sobre los datasets recopilados, y dentro de ellos los grupos a comparar. Esto nos ayudara a entender los genes diferencialmente expresados.

-Luego, se incluyó la búsqueda de información de cada RBPs. Esto es debido a que es una importante parte de comprensión del problema, y no hay que hacerlo al final en el poco tiempo que quede.

-La siguiente tarea sí se ha incluido a parte de las anteriores. Escribimos el papel de cada RBPs implicada en procesos de splicing que encontramos entre comparaciones y el resto de la memoria. Sin embargo, en esta no constaba la redacción del trabajo. Debido a que parte del tiempo se ha invertido en el análisis y la redacción de las PECs, creímos conveniente la inclusión de un apartado para dedicarlo entero a la redacción del trabajo.

En cualquier caso, la metodología se ha modificado en pos de un análisis más cómodo, de la generación de contenido enriquecido de funciones, de la comprensión de los resultados, de la no pérdida de información y de la generación de informacion accesible y estudiable por cualquier usuario. Por ello, creemos que la desviación de las tareas originales es justificable, y que no ha supuesto cambios relevantes que supediten el resto de los análisis, sino solo en una parte de la metodología de ciertos pasos. Todo el resto del análisis se ha realizado sin complicaciones y se han obtenido resultados favorables, por lo que pensamos que la modificación no ha sido en vano ni ha entorpecido el desarrollo del resto de tareas.

Sin embargo, ha habido un objetivo muy importante para dar a conocer nuestro trabajo, y generar por tanto otro tipo de cuestiones a partir de él, relacionado con la shinny app, que no se ha podido desarrollar. Esta no se ha podido realizar debido a la falta de tiempo, incluso cuando hemos tenido que modificar las fechas de las tareas porque las realizábamos antes de tiempo. Sin embargo, no es un apartado fundamental del análisis a partir del cual se siga con el estudio, sino un complemento que, como hemos comentado, no era totalmente necesario. Por ello, podemos pensar que no completar este apartado no ha trastocado nuestra metodología para determinar la veracidad de nuestra hipótesis.

8 Glosario

- Cáncer: neoplasia maligna que puede invadir otros tejidos y con capacidad diseminadora.
- Desregulación genética: desajuste en el proceso de control del momento, ubicación y/o nivel de expresión de los genes.
- Empalme/ayuste de ARNm: proceso de maduración del ARN mensajero por el cual una molécula de pre-ARN mensajero es cortada en sus regiones delimitantes de intrones y exones, con la posterior unión de los exones, para generar una molécula de ARNm madura.
- Empalme alternativo de ARNm: mecanismo molecular de procesado del pre-ARNm por el cual se generan diferentes tipos de ARNm maduro y, por tanto, suelen generarse varias proteínas, a partir de una secuencia génica.
- Hematopoyesis: mecanismo celular de proliferación y diferenciación de células sanguíneas a partir de células madre hematopoyéticas.
- Hematopoyético/hematopoyética: que hace referencia a células sanguíneas.
- Leucemia: tipo de neoplasia maligna hematopoyética con origen medular.
- Leucemia linfoblástica aguda de células T: neoplasia maligna hematopoyética de progresión rápida caracterizada por la proliferación de linfocitos T inmaduros.
- Neoplasia/tumor: parte de un tejido caracterizado por la proliferación anormal en demasía de células.
- Patogenia: disciplina referente al estudio de las causas y el desarrollo de las enfermedades.
- Postranscripcional: referente a los procesos que se dan después de la generación de moléculas de ARN.
- Proteína de unión a ARN/RBP: aquellas que se acoplan y son claves en el control postranscripcional del ARN
- Transcripcional: relativo a la transcripción genética.

9 Bibliografía

1. Terwilliger T, Abdul-Hay M. Acute lymphoblastic leukemia: a comprehensive review and 2017 update. Blood Cancer J [Internet]. 2017 Jun 30 [cited 2022 Jun 1];7(6):e577. Available from: https://pubmed.ncbi.nlm.nih.gov/28665419/
2. van der Sligte NE, Kampen KR, Elst A ter, Scherpen FJG, Meeuwsen-de Boer TGJ, Guryev V, et al. Essential role for cyclic-AMP responsive element binding protein 1 (CREB) in the survival of acute lymphoblastic leukemia. Oncotarget [Internet]. 2015 [cited 2022 Jun 1];6(17):14970–81. Available from: https://pubmed.ncbi.nlm.nih.gov/26008971/
3. Zhang H, Zhu L, He H, Zhu S, Zhang W, Liu X, et al. NF-kappa B mediated Up-regulation of CCCTC-binding factor in pediatric acute lymphoblastic leukemia. Molecular Cancer [Internet]. 2014 Jan 7 [cited 2022 Jun 1];13(1):1–14. Available from: https://molecular-cancer.biomedcentral.com/articles/10.1186/1476-4598-13-5
4. Bigas A, Guillén Y, Schoch L, Arambilet D. Revisiting β-Catenin Signaling in T-Cell Development and T-Cell Acute Lymphoblastic Leukemia. BioEssays [Internet]. 2020 Feb 1 [cited 2022 Jun 1];42(2):1900099. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/bies.201900099
5. Cordo' V, van der Zwet JCG, Canté-Barrett K, Pieters R, Meijerink JPP. T-cell Acute Lymphoblastic Leukemia: A Roadmap to Targeted Therapies. Blood Cancer Discov [Internet]. 2020 Jan [cited 2022 Jun 1];2(1):19–31. Available from: https://pubmed.ncbi.nlm.nih.gov/34661151/
6. Hunger SP, Raetz EA. How I treat relapsed acute lymphoblastic leukemia in the pediatric population. Blood [Internet]. 2020 Oct 15 [cited 2022 Jun 1];136(16):1803–12. Available from: https://ashpublications.org/blood/article/136/16/1803/461194/How-I-treat-relapsed-acute-lymphoblastic-leukemia
7. Schuschel K, Helwig M, Hüttelmaier S, Heckl D, Klusmann JH, Hoell JI. RNA-Binding Proteins in Acute Leukemias. International Journal of Molecular Sciences [Internet]. 2020 May 2 [cited 2022 Jun 1];21(10). Available from: /pmc/articles/PMC7279408/
8. Fattizzo B, Rosa J, Giannotta JA, Baldini L, Fracchiolla NS. The Physiopathology of T- Cell Acute Lymphoblastic Leukemia: Focus on Molecular Aspects. Front Oncol [Internet]. 2020 Feb 28 [cited 2022 Jun 1];10. Available from: https://pubmed.ncbi.nlm.nih.gov/32185137/
9. Sebestyén E, Singh B, Miñana B, Pagès A, Mateo F, Pujana MA, et al. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Research [Internet]. 2016 Jun 1 [cited 2022 Jun 1];26(6):732–44. Available from: https://pubmed.ncbi.nlm.nih.gov/27197215/
10. Wang E, Lu SX, Pastore A, Chen X, Imig J, Chun-Wei Lee S, et al. Targeting an RNA-Binding Protein Network in Acute Myeloid Leukemia. Cancer Cell [Internet]. 2019 Mar 18 [cited 2022 Jun 1];35(3):369-384.e7. Available from: https://pubmed.ncbi.nlm.nih.gov/30799057/

11. Ghigna C, Valacca C, Biamonti G. Alternative splicing and tumor progression. Curr Genomics [Internet]. 2008 Dec 13 [cited 2022 Jun 1];9(8):556–70. Available from: https://pubmed.ncbi.nlm.nih.gov/19516963/
12. Yi L, Hu Q, Zhou J, Liu Z, Li H. Alternative splicing of Ikaros regulates the FUT4/LeX-α5β1 integrin-FAK axis in acute lymphoblastic leukemia. Biochemical and Biophysical Research Communications [Internet]. 2019 Jan 22 [cited 2022 Jun 1];510(1):128–34. Available from: https://europepmc.org/article/med/30683310
13. González RA, Tejera Hernández B, Ocampo PH. Proteínas de unión con ARN, moléculas integradoras de la expresión genética. inventio.uaem.mx [Internet]. 2012 [cited 2022 Jun 1];4:1–5. Available from: http://inventio.uaem.mx/index.php/inventio/article/view/309
14. Zhang B, Babu KR, Lim CY, Kwok ZH, Li J, Zhou S, et al. A comprehensive expression landscape of RNA-binding proteins (RBPs) across 16 human cancer types. RNA Biology [Internet]. 2020 Feb 1 [cited 2022 Jun 1];17(2):211. Available from: /pmc/articles/PMC6973330/
15. Saha S, Murmu KC, Biswas M, Chakraborty S, Basu J, Madhulika S, et al. Transcriptomic Analysis Identifies RNA Binding Proteins as Putative Regulators of Myelopoiesis and Leukemia. Front Oncol [Internet]. 2019 Aug 6 [cited 2022 Jun 1];9. Available from: https://pubmed.ncbi.nlm.nih.gov/31448224/
16. Ule J, Hwang HW, Darnell RB. The Future of Cross-Linking and Immunoprecipitation (CLIP). Cold Spring Harbor Perspectives in Biology [Internet]. 2018 Aug 1 [cited 2022 Jun 1];10(8). Available from: /pmc/articles/PMC6071486/
17. Hafner ED, Techel F, Leinss S, Bühler Y. Mapping avalanches with satellites-Evaluation of performance and completeness. Cryosphere. 2021 Feb 24;15(2):983–1004.
18. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. Nature Reviews Genetics 2014 15:12 [Internet]. 2014 Nov 4 [cited 2022 Jun 1];15(12):829–45. Available from: https://www.nature.com/articles/nrg3813
19. Shi Y, Manley JL. The end of the message: multiple protein–RNA interactions define the mRNA polyadenylation site. Genes & Development [Internet]. 2015 May 5 [cited 2022 Jun 1];29(9):889. Available from: /pmc/articles/PMC4421977/
20. Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, et al. Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet [Internet]. 2018 Jan 1 [cited 2022 Jun 1];50(1):151–8. Available from: https://pubmed.ncbi.nlm.nih.gov/29229983/
21. Zhang X, Liu S, Tramontano A. RBPPred: predicting RNA-binding proteins from sequence using SVM. Bioinformatics [Internet]. 2017 Mar 15 [cited 2022 Jun 1];33(6):854–62. Available from: https://academic.oup.com/bioinformatics/article/33/6/854/2557689
22. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res [Internet]. 2002 Jan 1 [cited 2022 Jun 1];30(1):207–10. Available from: https://pubmed.ncbi.nlm.nih.gov/11752295/
23. Vlierberghe P van, Ambesi-Impiombato A, Perez-Garcia A, Haydu JE, Rigo I, Hadler M, et al. ETV6 mutations in early immature human T cell leukemias. J Exp

Med [Internet]. 2011 Dec 19 [cited 2022 Jun 1];208(13):2571–9. Available from: https://pubmed.ncbi.nlm.nih.gov/22162831/

24. Piovan E, Yu J, Tosello V, Herranz D, Ambesi-Impiombato A, DaSilva AC, et al. Direct reversal of glucocorticoid resistance by AKT inhibition in acute lymphoblastic leukemia. Cancer Cell [Internet]. 2013 Dec 9 [cited 2022 Jun 1];24(6):766–76. Available from: https://pubmed.ncbi.nlm.nih.gov/24291004/
25. van Vlierberghe P, van Grotel M, Tchinda J, Lee C, Beverloo HB, van der Spek PJ, et al. The recurrent SET-NUP214 fusion as a new HOXA activation mechanism in pediatric T-cell acute lymphoblastic leukemia. Blood [Internet]. 2008 May 1 [cited 2022 Jun 1];111(9):4668–80. Available from: https://pubmed.ncbi.nlm.nih.gov/18299449/
26. Efficace F, Stagno F, Iurlo A, Breccia M, Cottone F, Bonifacio M, et al. Health-related quality of life of newly diagnosed chronic myeloid leukemia patients treated with first-line dasatinib versus imatinib therapy. Leukemia [Internet]. 2020 Feb 1 [cited 2022 Jun 1];34(2):488–98. Available from: https://pubmed.ncbi.nlm.nih.gov/31477798/
27. Qin H, Ni H, Liu Y, Yuan Y, Xi T, Li X, et al. RNA-binding proteins in tumor progression. Journal of Hematology and Oncology [Internet]. 2020 Jul 11 [cited 2022 Jun 1];13(1):1–23. Available from: https://jhoonline.biomedcentral.com/articles/10.1186/s13045-020-00927-w
28. Elcheva IA, Spiegelman VS. Targeting RNA-binding proteins in acute and chronic leukemia. Leukemia 2020 35:2 [Internet]. 2020 Nov 4 [cited 2022 Jun 1];35(2):360–76. Available from: https://www.nature.com/articles/s41375-020-01066-4
29. Hodson DJ, Screen M, Turner M. RNA-binding proteins in hematopoiesis and hematological malignancy. Blood [Internet]. 2019 May 5 [cited 2022 Jun 1];133(22):2365. Available from: /pmc/articles/PMC6716123/
30. Deng JL, Xu YH, Wang G. Identification of Potential Crucial Genes and Key Pathways in Breast Cancer Using Bioinformatic Analysis. Front Genet [Internet]. 2019 [cited 2022 Jun 1];10(JUL). Available from: https://pubmed.ncbi.nlm.nih.gov/31428132/
31. Huang R, Han M, Meng L, Chen X. Transcriptome-wide discovery of coding and noncoding RNA-binding proteins. Proc Natl Acad Sci U S A [Internet]. 2018 Apr 24 [cited 2022 Jun 1];115(17):E3879–87. Available from: www.pnas.org/cgi/doi/10.1073/pnas.1718406115
32. Zhang Y, Qian J, Gu C, Yang Y. Alternative splicing and cancer: a systematic review. Signal Transduction and Targeted Therapy 2021 6:1 [Internet]. 2021 Feb 24 [cited 2022 Jun 1];6(1):1–14. Available from: https://www.nature.com/articles/s41392-021-00486-7
33. Cerasuolo A, Buonaguro L, Buonaguro FM, Tornesello ML. The Role of RNA Splicing Factors in Cancer: Regulation of Viral and Human Gene Expression in Human Papillomavirus-Related Cervical Cancer. Front Cell Dev Biol [Internet]. 2020 Jun 12 [cited 2022 Jun 1];8. Available from: https://pubmed.ncbi.nlm.nih.gov/32596243/
34. Venables JP, Brosseau JP, Gadea G, Klinck R, Prinos P, Beaulieu JF, et al. RBFOX2 is an important regulator of mesenchymal tissue-specific splicing in both

normal and cancer tissues. Mol Cell Biol [Internet]. 2013 Jan 15 [cited 2022 Jun 1];33(2):396–405. Available from: https://pubmed.ncbi.nlm.nih.gov/23149937/

35. Qi F, Li Y, Yang X, Wu YP, Lin LJ, Liu XM. Significance of alternative splicing in cancer cells. Chinese Medical Journal [Internet]. 2020 Jan 1 [cited 2022 Jun 1];133(2):221. Available from: /pmc/articles/PMC7028187/
36. Xu Y, Gao XD, Lee JH, Huang H, Tan H, Ahn J, et al. Cell type-restricted activity of hnRNPM promotes breast cancer metastasis via regulating alternative splicing. Genes & Development [Internet]. 2014 Jun 6 [cited 2022 Jun 1];28(11):1191. Available from: /pmc/articles/PMC4052765/
37. Ho JSY, di Tullio F, Schwarz M, Low D, Incarnato D, Gay F, et al. Hnrnpm controls circrna biogenesis and splicing fidelity to sustain cancer cell fitness. Elife. 2021 Jun 1;10.
38. Wang X, Li J, Bian X, Wu C, Hua J, Chang S, et al. CircURI1 interacts with hnRNPM to inhibit metastasis by modulating alternative splicing in gastric cancer. Proc Natl Acad Sci U S A. 2021 Aug 17;118(33).
39. Chen TM, Lai MC, Li YH, Chan YL, Wu CH, Wang YM, et al. hnRNPM induces translation switch under hypoxia to promote colon cancer development. EBioMedicine. 2019 Mar 1;41:299–309.
40. el Marabti E, Younis I. The cancer spliceome: Reprograming of alternative splicing in cancer. Frontiers in Molecular Biosciences. 2018 Sep 7;5(SEP).
41. Kugeratski FG, Atkinson SJ, Neilson LJ, Lilla S, Knight JRP, Serneels J, et al. Hypoxic cancer-associated fibroblasts increase NCBP2-AS2/HIAR to promote endothelial sprouting through enhanced VEGF signaling. Sci Signal [Internet]. 2019 Feb 5 [cited 2022 Jun 1];12(567). Available from: https://pubmed.ncbi.nlm.nih.gov/30723174/
42. Zhang H, Cheng L, Liu C. Regulatory Networks of Prognostic mRNAs in Pediatric Acute Myeloid Leukemia. J Healthc Eng [Internet]. 2022 [cited 2022 Jun 1];2022. Available from: https://pubmed.ncbi.nlm.nih.gov/35035819/
43. Singh G, Fritz SE, Seufzer B, Boris-Lawrie K. The mRNA encoding the JUND tumor suppressor detains nuclear RNA-binding proteins to assemble polysomes that are unaffected by mTOR. J Biol Chem [Internet]. 2020 May 28 [cited 2022 Jun 1];295(22):7763–73. Available from: https://pubmed.ncbi.nlm.nih.gov/32312751/
44. Stengel A, Shahswar R, Haferlach T, Walter W, Hutter S, Meggendorfer M, et al. Whole transcriptome sequencing detects a large number of novel fusion transcripts in patients with AML and MDS. Blood Adv. 2020 Nov 10;4(21):5393–401.
45. Jung C, Lee CYF, Grigg ME. The SRS superfamily of Toxoplasma surface proteins. Int J Parasitol [Internet]. 2004 Mar 9 [cited 2022 Jun 1];34(3):285–96. Available from: https://pubmed.ncbi.nlm.nih.gov/15003490/
46. Hattori A, McSkimming D, Kannan N, Ito T. RNA binding protein MSI2 positively regulates FLT3 expression in myeloid leukemia. Leuk Res [Internet]. 2017 Mar 1 [cited 2022 Jun 1];54:47–54. Available from: https://pubmed.ncbi.nlm.nih.gov/28107692/
47. Haijes HA, Koster MJE, Rehmann H, Li D, Hakonarson H, Cappuccio G, et al. De Novo Heterozygous POLR2A Variants Cause a Neurodevelopmental Syndrome with Profound Infantile-Onset Hypotonia. Am J Hum Genet [Internet]. 2019 Aug

1 [cited 2022 Jun 1];105(2):283–301. Available from: https://pubmed.ncbi.nlm.nih.gov/31353023/

48. Mannelli F. Immunophenotyping of Acute Leukemias – From Biology to Clinical Application. Flow Cytometry - Select Topics. 2016 Aug 24;
49. Alvares CJ, Blough R, Mazzella FM, Schumacher HR. Genetics of the chronic leukemias. Journal of Clinical Ligand Assay. 2001;24(3):201–9.
50. Bailey NG, Elenitoba-Johnson KSJ. Impact of Genetics on Mature Lymphoid Leukemias and Lymphomas. Cold Spring Harbor Perspectives in Medicine [Internet]. 2020 Jan 13 [cited 2022 Jun 1];10(11):a035444. Available from: http://perspectivesinmedicine.cshlp.org/content/early/2020/01/13/cshperspect.a035444
51. Zhang S, Zhang Q, Yin J, Wu X. Overlapped differentially expressed genes between acute lymphoblastic leukemia and chronic lymphocytic leukemia revealed potential key genes and pathways involved in leukemia. J Cell Biochem [Internet]. 2019 Sep 1 [cited 2022 Jun 1];120(9):15980–8. Available from: https://pubmed.ncbi.nlm.nih.gov/31081970/
52. Kang H, Sharma N, Nickl CK, Ness S, Devidas M, Wood BL, et al. New Insights into Deregulated Gene Expression Pathways in MLL- and AF10-Rearranged T-Lineage Acute Lymphoblastic Leukemia. Blood [Internet]. 2016 Dec 2 [cited 2022 Jun 6];128(22):2906–2906. Available from: https://ashpublications.org/blood/article/128/22/2906/113811/New-Insights-into-Deregulated-Gene-Expression
53. Steimlé T, Dourthe ME, Alcantara M, Touzart A, Simonin M, Mondesir J, et al. Clinico-biological features of T-cell acute lymphoblastic leukemia with fusion proteins. Blood Cancer Journal 2022 12:1 [Internet]. 2022 Jan 26 [cited 2022 Jun 6];12(1):1–8. Available from: https://www.nature.com/articles/s41408-022-00613-9
54. Feltes BC, Poloni J de F, Nunes IJG, Faria SS, Dorn M. Multi-Approach Bioinformatics Analysis of Curated Omics Data Provides a Gene Expression Panorama for Multiple Cancer Types. Front Genet [Internet]. 2020 Nov 23 [cited 2022 Jun 6];11. Available from: https://pubmed.ncbi.nlm.nih.gov/33329726/
55. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology [Internet]. 2004 Feb 12 [cited 2022 Jun 6];3(1). Available from: https://www.degruyter.com/document/doi/10.2202/1544-6115.1027/html

ANEXOS

Los documentos descritos aquí hacen referencia a lo que podemos encontrar en el reporitorio de GitHub https://github.com/CristinaMuntanola/TFM_T_ALL.git. En metodología se explican los pasos seguidos y el orden de los documentos:

- Anot_RBPs_articulos1.Rmd
- Anot_RBPs_articulos2.Rmd
- Anot_RBPs_articulos3.Rmd
- Ejemplo_interrogracion.Rmd
- TFM.docx (esta memoria)
- Expres diferenc y signif.Rmd
- Figuras.html
- Prodecimiento creacion datasets e info.Rmd
- Tabla genes y proteinas RNA splicing.docx