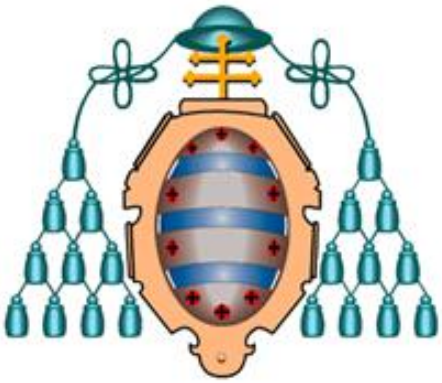


UNIVERSIDAD DE OVIEDO

INSTITUTO UNIVERSITARIO DE ONCOLOGIA

MÁSTER EN BIOMEDICINA Y ONCOLOGÍA MOLECULAR

Trabajo Fin de Máster

"APROXIMACIONES BIOINFORMÁTICAS PARA LA IDENTIFICACIÓN DE ALTERACIONES EN *SPLICING* EN CÁNCER"

Curso 2014-2015

Alumno: Ángel Álvarez Eguiluz

A mi tutor, el Dr. Xose Antón Suárez Puente,

Por permitir que me sumergiera en un mar de conceptos y estrategias llamados a determinar el futuro de la investigación biomédica

Por tratar de enseñarme a hacer buenísima ciencia como la que tú haces

GRACIAS

A Jesús Gutiérrez Abril,

Por tu ayuda incansable y los gratos momentos compartidos

GRACIAS

Al resto de compañeros del laboratorio,

Por hacer que cada día no fuera uno más

GRACIAS

ABREVIATURAS

BLAT	***B**LAST-**L**ike **A**lignment **T**ool*
BP	***B**iological **P**rocess*
COSMIC	***C**atalogue **O**f **S**omatic **M**utations **I**n **C**ancer*
DAVID	***D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery*
ES	***E**nrichment **S**core*
FC	***F**old **C**hange*
FISH	**H**ibridación **F**luorescente ***I**n **S**itu*
gWb	**g**e**W**ork**b**ench
ICGC	***I**nternational **C**ancer **G**enome **C**onsortium*
IgHV	***I**mmuno**g**lobulin **H**eavy chain **V**ariable*
LLC	**L**eucemia **L**infática **C**rónica
MF	***M**olecular **F**unction*
Mut	***Mut**ation* o **Mut**ado
PIR	***P**rotein **I**nformation **R**esource*
RNA-Seq	***RNA Seq**uencing*
SMD	**S**índrome **M**ielo**D**isplásico
SNP	***S**ingle **N**ucleotide **P**olymorphism*
snRNP	***s**mall **n**uclear **R**ibo**N**ucleic **P**rotein*
UCSC	***U**niversity of **C**alifornia, **S**anta **C**ruz*
URL	***U**niform **R**esource **L**ocator*
WT	***W**ild **T**ype*

ÍNDICE

1. RESUMEN

Aunque el *splicing* es un mecanismo pleiotrópico necesario para la función celular, la implicación directa de genes de *splicing* en el proceso de transformación neoplásica se conoce desde hace años gracias al desarrollo de las técnicas de secuenciación masiva. Uno de estos genes, conocido como *SF3B1*, es el gen más comúnmente mutado en síndromes mielodisplásicos (SMD) y el segundo en leucemia linfática crónica (LLC), reforzando su importancia en la carcinogénesis. Sin embargo, aún se desconoce el mecanismo por el que las mutaciones en *SF3B1* contribuyen a este proceso. En el presente trabajo se ha estudiado si la alteración de *SF3B1* provoca cambios específicos en los perfiles de expresión génica de pacientes con cáncer. El análisis de 159 casos de LLC, 269 de cáncer de páncreas y 253 de cáncer de mama, permitió identificar genes cuya expresión está alterada debido a mutaciones en dicho gen. Sin embargo, ninguno de ellos era recurrente en los tres tipos tumorales estudiados, ni siquiera cuando se enfrentaban los cánceres dos a dos. En este último caso, la realización de análisis menos restrictivos señaló la presencia de varias decenas de genes comunes que podrían ser dianas directas de la actividad mutante del factor SF3B1. Además, la mayor parte de los genes que son recurrentes en LLC y cáncer de páncreas (20 de los 24) tienen un cambio en la expresión en el mismo sentido en ambos tipos tumorales hacia una expresión reprimida. Por otra parte, se ha estudiado si la presencia de mutaciones somáticas en *SF3B1* provoca la activación de nuevos sitios aceptores de *splicing* alternativos mediante el análisis cualitativo de k-mers de RNA-Seq. Esta aproximación permitió identificar un nuevo evento de *splicing*, en el gen *KCTD17*. Un estudio posterior de la relación de eventos crípticos y canónicos de *splicing* que se producen en *ATM* reveló que la alteración de *SF3B1* sólo provoca un aumento de la formación del *splicing* aberrante que se da en condiciones normales. Estos datos sugieren que para determinar el efecto de las mutaciones en *SF3B1* sobre el splicing es necesario el empleo de estrategias que evalúen los fenómenos de *splicing* de manera cuantitativa.

2. INTRODUCCIÓN

El cáncer es responsable de una de cada siete muertes a nivel mundial (1). Abarca más de 100 enfermedades diferentes con factores de riesgo y epidemiología múltiples que se originan en la mayoría de órganos del cuerpo humano. Todas ellas comparten una patogénesis común (2). Surgen mediante el curso de un proceso evolutivo Darwiniano como consecuencia de cambios adquiridos somáticamente en el DNA genómico celular. Esto no significa que todas las alteraciones somáticas presentes en un genoma tumoral hayan estado implicadas en la carcinogénesis. De hecho, sólo un subgrupo minoritario confiere propiedades oncogénicas a la célula cancerosa (3). Son las denominadas mutaciones conductoras o '*driver*', capaces de atribuir una pequeña ventaja selectiva para el crecimiento del tumor. Están en contraposición con las mutaciones pasajeras o '*passengers*', también presentes en el genoma de las células neoplásicas pero que no contribuyen al desarrollo tumoral. Además, las mutaciones *driver* han sido seleccionadas positivamente en las células que dan origen al cáncer, de modo que su identificación proporcionaría conocimientos sobre la biología tumoral y delataría nuevas dianas farmacológicas además de *test* diagnósticos (4) (Figura 1).

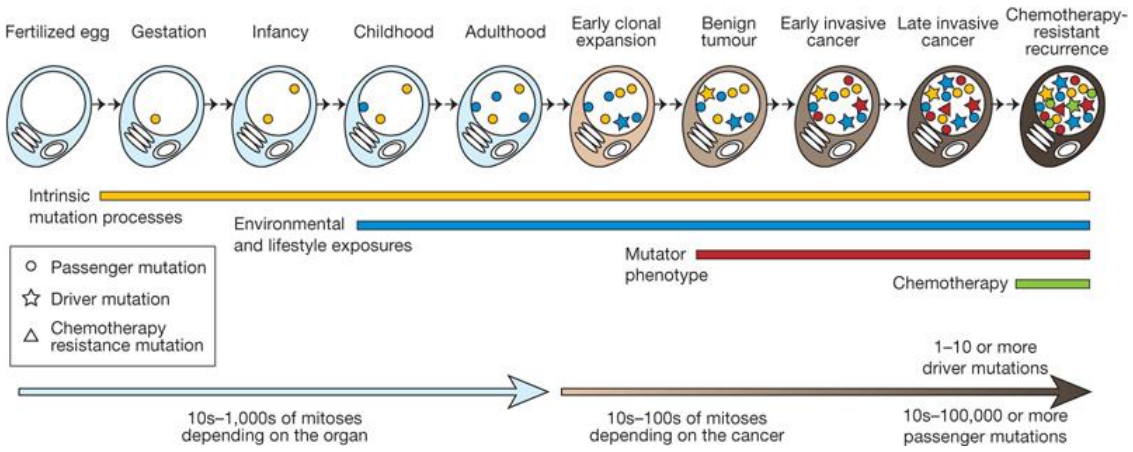


***Figura 1*. Linaje de divisiones celulares mitóticas desde el cigoto hasta la célula tumoral.** Se especifica la proporción de mutaciones somáticas que se van acumulando en función del estadio del desarrollo así como los procesos que inducen a ellas. La aparición de mutaciones somáticas conductoras, que confieren una ventaja de crecimiento (estrellas), permiten el desarrollo y la progresión tumoral (Stratton y cols., 2009).

La llegada al mercado de la tecnología de secuenciación masiva en paralelo ha renovado plenamente nuestra forma de interpretar la investigación biomédica. La principal ventaja respecto a los métodos convencionales viene dada por su capacidad de producir grandes volúmenes de datos relativos a secuencias biológicas de manera rápida, económica y precisa, lo que las hace útiles para muchas aplicaciones (5). Por ejemplo, el análisis preliminar de genomas tumorales completos ha revelado que la carga mutacional en cáncer es abundante y dispar (6–11). Con el objetivo de maximizar la eficiencia científica en la difícil tarea de comprender, prevenir y tratar este complejo grupo de enfermedades, en octubre del año 2007 se sentaron las bases del ICGC, un consorcio instaurado para poner en marcha y coordinar un gran número de proyectos de investigación cuyo propósito común radica en descifrar íntegramente las variantes genómicas, transcriptómicas y epigenómicas que contribuyen a la carga patológica de los 50 tipos y/o subtipos de cáncer de mayor relevancia clínica y social a nivel mundial (4) (Figura 2).

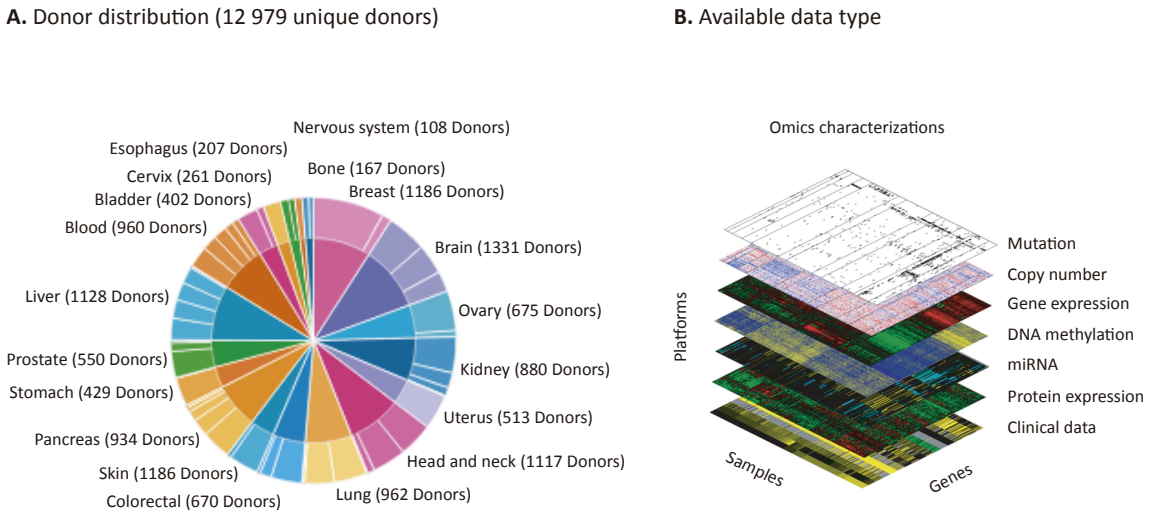


***Figura 2*. Consorcio Internacional del Genoma del Cáncer.** (A) Sitios primarios de los cánceres estudiados, especificando el número de donantes en cada caso. (B) Tipos de datos disponibles (adaptado de TCGA Research Network y cols., 2013).

Uno de los proyectos que forman parte del ICGC desde su fundación es el proyecto genoma de la leucemia linfática crónica (LLC), la leucemia más común en adultos en los países occidentales. A pesar de producirse por la transformación tumoral de linfocitos

B, la presentación y evolución de la enfermedad son muy heterogéneas (12). Mientras que un grupo de pacientes siguen un curso asintomático sin necesidad de tratamiento durante un largo periodo de tiempo, en otros progresa rápidamente con una enorme agresividad que se traduce en un pronóstico muy pobre (13). Este comportamiento diferencial ha sido asociado principalmente a la existencia de dos grandes subtipos moleculares de la enfermedad caracterizados respectivamente por un número alto y bajo de mutaciones somáticas en la región variable de los genes de las inmunoglobulinas (12). La introducción de técnicas de secuenciación masiva, así como el análisis de variaciones estructurales mediante hibridación '*in situ*' fluorescente (FISH) o *arrays* de genotipado, ha demostrado que la versatilidad clínica en la progresión de la enfermedad también se relaciona con la presencia de alteraciones genéticas concretas que incluyen anomalías citogenéticas y mutaciones específicas en genes como *TP53* y *NOTCH1* (12–14).

Asimismo, los datos de secuenciación masiva de cientos de tumores han revelado que entre los genes más recurrentemente mutados en LLC se encuentra *SF3B1* (14). Este gen codifica un factor que participa en el mantenimiento de la fidelidad durante el *splicing* en el sitio aceptor. Además de estar mutado en LLC, el gen *SF3B1* también está alterado con bastante frecuencia en síndromes mielodisplásicos (SMD) (15). En ambos casos, todas las mutaciones observadas son sustituciones puntuales en heterocigosis (ninguna *sin sentido*) agrupadas en la secuencia génica entre los exones 12 y 18 sin comprometer sus sitios de *splicing* canónicos (14,15). Posteriores estudios se hicieran eco de estas observaciones e identificaron algunas de las alteraciones recurrentes de *SF3B1* en varios tipos de tumores sólidos, incluyendo mama (16) y páncreas (17), fortaleciendo la importancia de este factor en la aparición y desarrollo del cáncer (14).

Pese a que el *splicing* es un mecanismo pleiotrópico necesario para la función celular, este hallazgo, junto con la descripción simultánea de mutaciones menos recurrentes en otros componentes del espliceosoma, ha revelado el papel oncogénico que detentan los factores de *splicing* en la transformación maligna de las células (14). La incidencia de mutaciones en *SF3B1* en LLC varía según el estadio de la enfermedad al diagnóstico, siendo especialmente comunes en etapas avanzadas. Esto sugiere que la actividad mutacional en dicho gen tiende a ocurrir durante la evolución clonal de la

neoplasia (18). Además, al contrario de lo que sucede en SMD, los pacientes de LLC que son portadores de mutaciones somáticas en *SF3B1* presentan una enfermedad con características biológicas adversas que derivan en una progresión patológica más agresiva y supervivencia global reducida (14).

Como se ha mencionado anteriormente, *SF3B1* codifica un componente central de la snRNP espliceosomal U2 (19), implicada en la unión de ésta al punto de ramificación cerca de los sitios 3' de *splicing* (20). La proteína SF3B1 interacciona con las secuencias de RNA en la vecindad del punto de ramificación. Al mismo tiempo, se une al factor de reconocimiento temprano del sitio 3' de *splicing* U2AF65 y a la proteína de unión al punto de ramificación SF3B14. En consonancia con el papel que juega en la expresión génica, su secuencia aminoacídica retiene un nivel de conservación filogenética elevado (14) (Figura 3A). Estructuralmente, tiene dos regiones bien definidas; la región hidrofílica N-terminal, que contiene varios motivos de unión a proteína, y la región C-terminal, formada por 22 repeticiones HEAT no idénticas donde caen todas las alteraciones puntuales descritas hasta la fecha. Una en particular, la sustitución p.Lys700Glu, representa el ≈50% de la totalidad, y varios residuos aminoacídicos más de esta región son puntos calientes de mutación (14,15) (Figura 3B).

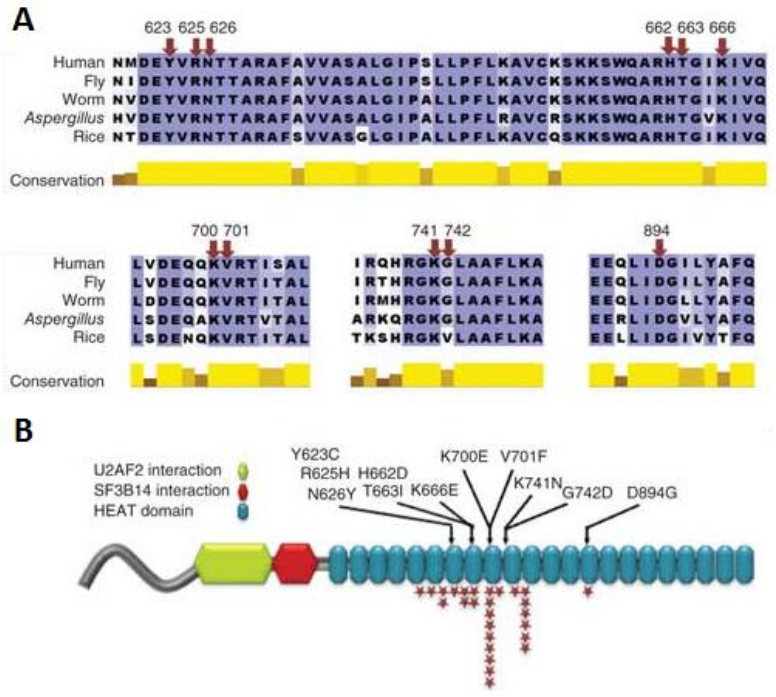


***Figura 3*. Impacto estructural de las alteraciones de SF3B1.** (A) Alineamientos en torno a los residuos alterados (flechas) de la secuencia proteica del dominio C-terminal de SF3B1 en varias especies evolutivamente distantes. (B)

Representación esquemática de la proteína SF3B1 humana en la que se resaltan sus dominios estructurales primarios. Se muestran las localizaciones de las diferentes alteraciones somáticas así como la frecuencia de cada una de ellas (adaptado de Quesada y cols., 2012).

Sin embargo, los mecanismos moleculares por los que *SF3B1* contribuye a la oncogénesis aún son desconocidos. Aunque el estudio de los perfiles de expresión génica con *microarrays* proporciona una visión global de los elementos expresados y de las interconexiones que moldean la biología de las células normales en su transformación y progresión hacia células cancerosas malignas, la secuenciación masiva del transcriptoma (RNA-Seq) está suplantando rápidamente a los *microarrays* como la tecnología *gold-standard* en todo sentido, dado su amplio rango dinámico así como la capacidad de cuantificación digital y la posibilidad de detectar y cuantificar formas de *splicing* conocidas o nuevas (21). Puesto que *SF3B1* garantiza la fidelidad del punto de ramificación 3', en el presente trabajo nos planteamos la hipótesis de que el efecto esperado de su disfuncionalidad vendría dado por la activación críptica de sitios 3' de *splicing* que conllevarían el procesamiento inadecuado del pre-mRNA y como consecuencia, cambios en el perfil transcriptómico.

3. OBJETIVOS

Por todo ello, para comprobar si la alteración de *SF3B1* en cáncer provoca cambios a nivel transcriptómico y con el objetivo de definir sus genes diana en caso de que así sea, se planteó la posibilidad de identificar cambios en los perfiles de expresión génica asociados a mutaciones puntuales en este gen en distintos tipos tumorales. Además, dado que se desconoce el mecanismo por el cual las mutaciones en *SF3B1* contribuyen a la oncogénesis, también planteamos la posibilidad de analizar datos de RNA-Seq de pacientes con cáncer para tratar de identificar nuevos eventos de *splicing* asociados a la actividad mutante de SF3B1. Los objetivos concretos del proyecto son los siguientes:

1. Análisis cuantitativo de cambios en los perfiles de expresión génica asociados a mutaciones en *SF3B1* en cánceres de sangre, páncreas o mama.
2. Identificación de nuevos eventos de *splicing* que pudieran producirse por la acumulación de mutaciones en el factor SF3B1 mediante análisis cualitativo de secuencias transcriptómicas en casos de LLC.
3. Análisis cuantitativo del patrón de *splicing* de *ATM* en LLC según el genotipo de *SF3B1*.

4. MATERIAL Y MÉTODOS

4.1.- OBTENCIÓN DE MUESTRAS

Se usaron las herramientas que ofrece la base de datos del ICGC para visualización, consulta y descarga de información transcriptómica y genética (en aquellos casos en los que esté disponible) relativa a pacientes con distintos tipos de cáncer:

- Tipo tumoral: Cáncer pancreático. Subtipo tumoral: Adenocarcinoma ductal. País: Australia.
- Número total de donantes a fecha de acceso: 431. Se dispuso públicamente de matrices de expresión génica normalizadas para 269. Entre ellos, 238 tenían caracterizado también el perfil mutacional de su enfermedad particular. Número total de descargas: 269 (8 corresponden a portadores de mutaciones somáticas puntuales en *SF3B1*).
- Método de normalización de los datos: Transformación Log2.

Por igual,

- Tipo tumoral: Cáncer de mama. Subtipo tumoral: Carcinoma ductal y lobular. País: Estados Unidos.
- Número total de donantes a fecha de acceso: 1045. Entre tantos, para 529 se dispone públicamente de matrices de expresión génica normalizadas y un total de 509 cuentan a su vez con la información genética correspondiente a su enfermedad. Número total de descargas: 253 (8 para pacientes con mutaciones puntuales en *SF3B1*).
- Método de normalización de los datos: Desconocido.

Para LLC se emplearon los datos de expresión presentes en nuestro laboratorio:

- Tipo tumoral: Leucemia linfática crónica. Subtipo tumoral: No mutados para la región IgHV.
- Número total de donantes de matrices de expresión génica normalizadas: 159 (16 de ellos tienen mutaciones puntuales en *SF3B1*).
- Método de normalización de los datos: Transformación Log2.

4.2.- PROCESAMIENTO DE DATOS

Los datos descargados fueron preprocesados mediante el empleo de rutinas de trabajo (*scripts*) escritos en lenguaje Perl (22) y ejecutadas en servidores Linux con distintos objetivos. En primer lugar, se procedió a reunir toda la información transcriptómica de la que se disponía para los pacientes afectados con el mismo tipo de tumor en una única matriz de expresión génica (una tabla de doble entrada cuyas filas y columnas albergan los múltiples genes y pacientes respectivamente mientras que el número recogido en cada celda define el valor de expresión normalizado de un gen particular en una muestra también particular). Sobre ésta, los identificadores de las sondas fueron sustituidos por los símbolos oficiales de los genes correspondientes y se distinguieron las réplicas numéricamente. Por último, se filtraron los genes cuyo promedio de expresión (independientemente de la condición genotípica muestral) fuera mayor o igual a 50, un punto de corte arbitrario e indicativo de expresión basal que había sido establecido y verificado antes por nuestro laboratorio con datos de LLC normalizados mediante la transformación Log2. Ahora bien, las matrices de las que se dispuso para los pacientes de cáncer de mama habían sido objeto de un método de normalización alternativo. Esto permite explicar el enorme contraste que se observó en la magnitud de sus valores de expresión génica, incluso tras reemplazar los susodichos por el resultado de la exponencial $2^{(\text{valor de expresión de cada gen en cada paciente})}$ y se determinó que el uso de un protocolo de filtrado análogo al descrito fuera inútil. Dado que nos habíamos propuesto comparar estos datos con los del resto de tipos tumorales en estudio, seleccionamos un ≈33% de las sondas (concretamente aquéllas

con una media de los valores entre pacientes más alta) ya que así se consigue un *set* muestral parecido al que resulta de filtrar las matrices de LLC y cáncer de páncreas.

4.3.- ANÁLISIS CUALITATIVO DE LECTURAS TRANSCRIPTÓMICAS

Se recopilaron datos crudos de RNA-Seq (pertenecientes a nuestro laboratorio) bajo archivos *.fastq* que almacenan lecturas del transcriptoma en muestras de dos grupos de 10 pacientes con LLC *unmut* para la región IgHV en los que se había definido con anterioridad si presentan o no mutaciones somáticas en *SF3B1*. Las secuencias fueron seguidamente k-merizadas haciendo uso de BFCounter (v0.3, Agosto 2011), un algoritmo que efectúa a su vez el recuento rápido y eficiente de los k-mers generados para dar lugar a archivos *.dump* (23). Después, se diseñaron en Perl y ejecutaron en Linux los códigos fuente de una serie de programas computacionales destinados a aplicar filtros. Los k-mers resultantes fueron ensamblados utilizando el programa BCalm (24) para tratar de reconstruir posibles fragmentos génicos. Fue entonces cuando se escribió un programa en lenguaje Perl que tras ser ejecutado en servidores Linux cambió a formato *.fa* el *output* obtenido previamente (*.bcalm*). A continuación, las secuencias fueron alineadas al genoma de referencia utilizando BLAT (v34, Abril 2007) instalado localmente (25). Por último, se programaron en código Perl nuevas herramientas de trabajo bioinformáticas que fueron ejecutadas en Linux con dos propósitos; cruzar las secuencias alineadas previamente con el locus cromosómico de todo gen del genoma (esta aproximación permitió determinar la correspondencia génica) y comprobar que la longitud de las regiones genómicas donde se han producido los alineamientos se ajusta “fielmente” a los tamaños de las secuencias alineadas. De no ser así, esta estrategia delataría todas las que pudieran señalar nuevos eventos de *splicing*. Para confirmar la existencia de tales procesos se utilizó la versión gráfica de BLAT.

4.4.- ESTADÍSTICA

Se utilizó el lenguaje de programación R (v3.1.3, Marzo 2015) para llevar a cabo un análisis estadístico y gráfico de los valores de expresión génica en pacientes con distintos tipos de tumores y genotipo mutante o no mutante para *SF3B1*. Para la identificación de genes con cambios de expresión entre grupos se empleó un Test-T de Student. En caso de realizarse comparaciones múltiples, los efectos se corrigieron mediante el método de Bonferroni o de Benjamini-Hochberg (26).

4.5.- ANÁLISIS INTEGRADO CON geWorkbench

geWorkbench (gWb) es un programa bioinformático que ofrece una colección completa y extensible de herramientas para gestión, análisis, visualización y anotación de datos biomédicos (27). Se utilizó con varias intenciones (gWb, v2.5.1, Diciembre 2014). Por un lado, se realizó un análisis básico *t* de Student para identificar los genes con expresión diferencial estadísticamente significativa entre dos colectivos de pacientes que albergan o no portadores de mutaciones somáticas en *SF3B1*. Los efectos provocados por la realización de *test* múltiples fueron corregidos con los métodos explícitos 'Sólo *alpha*', 'Bonferroni estándar' y 'Bonferroni ajustado'. Por otro lado, se desarrolló un análisis funcional no supervisado (*Clustering* jerárquico) de genes para localizar grupos coregulados o funcionalmente relacionados y de muestras, para descubrir posibles subclases tumorales. Las comparaciones entre los objetos a equiparar se hicieron en base al cálculo de tres índices de semejanza; la distancia euclídea y los coeficientes de correlación de Pearson o por rangos de Spearman. Puesto que la plataforma trabaja asociando repetidamente los dos *clusters* de elementos más próximos, se fue recalculando las distancias máxima, promedio o mínima entre el grupo recién integrado y el resto a la hora de construir el árbol jerárquico.

4.6.- ANÁLISIS DE ENRIQUECIMIENTO FUNCIONAL

El análisis de enriquecimiento constituye una estrategia útil para identificar posibles relaciones funcionales entre genes. Si un proceso concreto es anormal bajo unas condiciones específicas, los genes que están relacionados desde el punto de vista funcional adquieren mayor potencial (enriquecidos) de ser seleccionados por las herramientas de *screening* como un grupo relevante. Dado que el resultado está basado en una agrupación génica y no a nivel particular, se aumenta la probabilidad de identificar los mecanismos celulares alterados en la investigación realizada (28,29). En base a esto, se utilizaron los recursos bioinformáticos que ofrece la base de datos DAVID para averiguar el significado biológico de las grandes listas génicas obtenidas antes y después de filtrar respecto a COSMIC.

4.7.- URLs

Los datos se obtuvieron de las siguientes páginas web:

- BioMart *Project* (Ensembl *release* 79, Marzo 2015), *http://ensembl.org/index.html*
- COSMIC *database* (*The Cancer Gene Census* v72, Marzo 2015), *http://cancer.sanger.ac.uk/cosmic*
- DAVID *database* (DAVID *Bioinformatics Resources* 6.7, Enero 2010), *http://david.abcc.ncifcrf.gov*
- ICGC *database* (ICGC *Data Portal* 3.8.5.0, *Data Release* 18 Febrero 2015), *http:// dcc.icgc.org*
- UCSC *Genome Browser database* (vhg19, Febrero 2009), *http://genome.ucsc.edu*

5. RESULTADOS

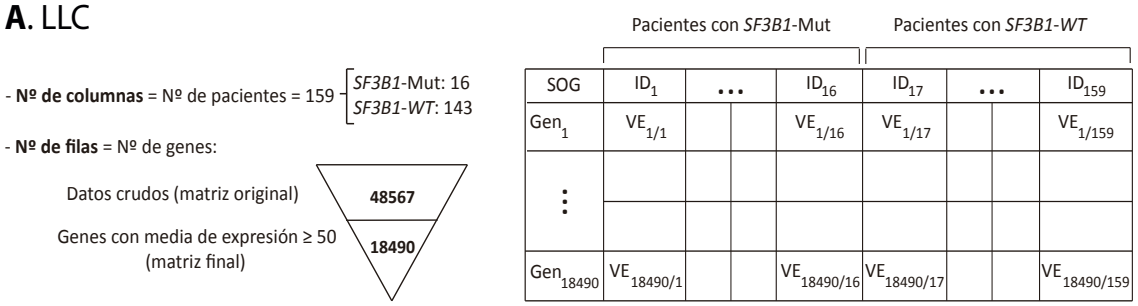
5.1.- ANÁLISIS CUANTITATIVO DE CAMBIOS EN LOS PERFILES DE EXPRESIÓN GÉNICA ASOCIADOS A MUTACIONES EN *SF3B1* EN CÁNCER

SF3B1 es el gen más comúnmente mutado en síndromes mielodisplásicos (20–28% de todos los casos) (19) y el segundo en leucemia linfática crónica, con una frecuencia total en torno al 10%. No obstante, si se consideran únicamente los pacientes afectados por el subtipo molecular de la enfermedad que se caracteriza por un número bajo de mutaciones somáticas en la región IgHV, su recurrencia en éstos aumenta hasta el 20-21% (14). La incidencia de alteraciones en *SF3B1* fue descrita también en otros cánceres hematológicos, incluyendo la leucemia mieloide aguda (≈5% del total de casos estudiados) y el mieloma múltiple (≈3%). Además, se ha estimado que entre un 1-5% de pacientes con varios tipos de tumores sólidos (como pueden ser los de hígado, páncreas, piel y mama entre otros ejemplos) son portadores de mutaciones somáticas en este gen (15).

Dado que se desconoce el mecanismo por el cual las mutaciones en *SF3B1* contribuyen a la transformación tumoral, en primer lugar se decidió analizar posibles cambios en los perfiles de expresión génica asociados a mutaciones en este gen. Para ello, fue necesario recopilar la información genética y transcriptómica de individuos con genotipo mutante (casos) y *WT* (controles) para *SF3B1* en los cánceres a estudiar. Con este objetivo se accedió a la base de datos del ICGC, en la que se depositan todos los datos sobre mutaciones, transcriptómica y epigenómica generados bajo este proyecto. Sin embargo, a pesar de disponer de datos de expresión normalizados para pacientes de un total de siete tipos de tumores diferentes, sólo identificamos tres proyectos: PACA-AU (páncreas), BRCA-US (mama) y CLL-ES (sangre), que dispusieran de un número “suficiente” de casos para constituir un grupo relativamente homogéneo que permitiera prever que el estudio proporcionaría resultados fiables y reproducibles. No obstante, la imposibilidad de averiguar el subtipo molecular de la enfermedad que

padece cada uno de los pacientes de LLC, determinó que finalmente se decidiera trabajar con los datos de expresión génica ya normalizados con los que contaba nuestro laboratorio para esta neoplasia hematolinfoide. En concreto, se analizarían aquellos que fueran no mutados para la región IgHV, independientemente del estado mutacional de *SF3B1*, puesto que así se reduce la variabilidad genética entre pacientes (*sets* más homogéneos) y se evitan posibles interferencias en los resultados.

Una vez descargados los datos de expresión y mutaciones génicas, se desarrolló un *script* en lenguaje Perl para aglutinar la información transcriptómica con la que se contaba para los distintos pacientes en una única matriz de expresión génica característica de cada tipo tumoral. Dicha matriz consistió en una tabla de doble entrada en la que las filas representan los genes cuya se expresión se había analizado mientras que las columnas corresponden a los pacientes estudiados. El valor asociado a cada celda define el valor de expresión normalizado de un gen particular en una muestra concreta (Tabla 1). En segundo lugar, se escribió un programa en formato Perl para filtrar estos *sets* de datos con el objetivo de seleccionar los genes con valor medio entre pacientes mayor o igual a 50 (Apéndice), por ser indicativo de expresión en datos normalizados mediante la transformación Log2. Ahora bien, las matrices individuales de las que se dispuso para cáncer de mama (las que había en la base de datos del ICGC) fueron normalizadas con un método alternativo, lo que explica el enorme contraste que existe entre sus valores de expresión génica y los del resto. Para tratar de hacer comparaciones con los otros tipos de cáncer estudiados, dichos valores fueron reemplazados por el resultado de la exponencial $2^{(\text{valor de expresión de cada gen en cada paciente})}$ y se eligió el ≈33% de las sondas con mayor promedio. Las dimensiones de las matrices generadas son especificadas en la siguiente figura según el tipo de tumor (Figura 4).



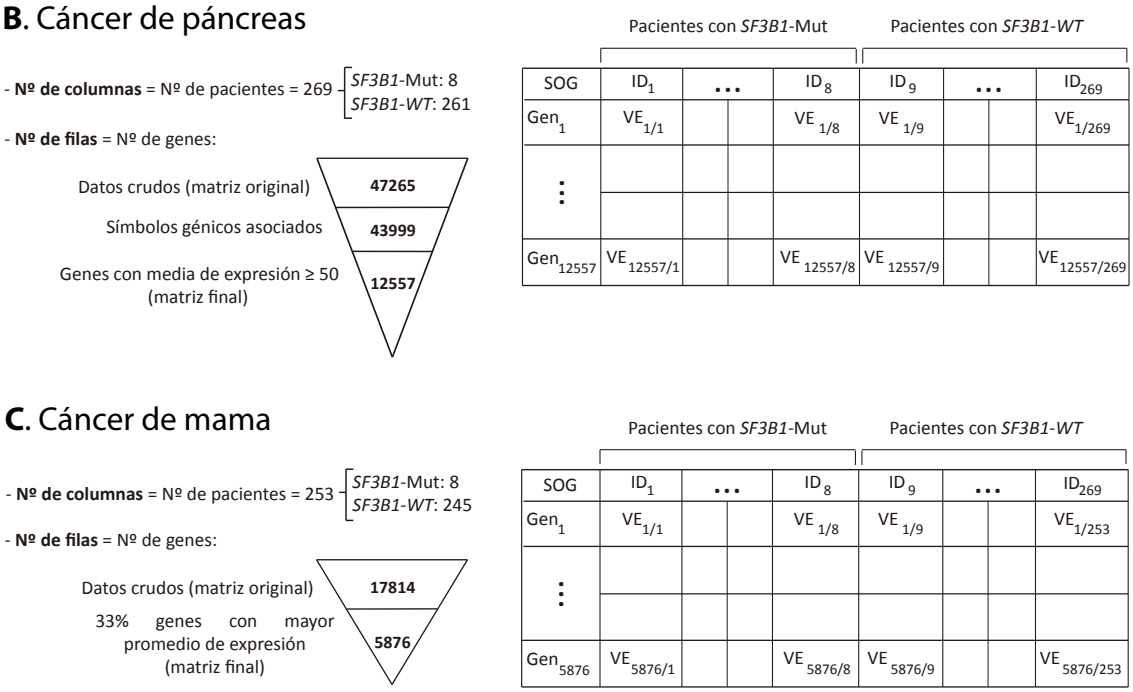


***Figura 4*. Dimensiones y estructura de las matrices de expresión génica obtenidas para cada tipo tumoral (A. LLC; B. Cáncer de páncreas; C. Cáncer de mama) de acuerdo con el tipo de procesamiento al que fueron sometidos los datos crudos.** LLC indica **L**eucemia **L**infática **C**rónica; *SF3B1*-Mut *SF3B1* **Mut**ado, *SF3B1-WT SF3B1 Wild Type*; SOG **S**ímbolo **O**ficial del **G**en; ID_X **ID**entificador del paciente **X**; $VE_{Y/Z}$ **V**alor de **E**xpresión del gen **Y** en el paciente **Z**.

5.1.1. - ANÁLISIS DE EXPRESIÓN DIFERENCIAL CON geWorkbench

Tras cargar cada matriz a título individual, se definieron los dos grupos de pacientes a comparar, 'casos' y 'controles' dependiendo de si presentan o no mutaciones en *SF3B1*. A continuación, se llevó a cabo una comparación estadística mediante una *t* de Student para tratar de identificar genes que presentasen cambios significativos en sus perfiles de expresión entre los tumores con mutaciones en *SF3B1* y los que no tenían mutado este gen de *splicing*.

Debido al elevado número de genes analizados (>20.000), para un nivel de significación concreto (por ejemplo, $P<0.05$), la probabilidad de identificar genes con diferencias de expresión entre casos mutados y no mutados para *SF3B1* con una significación estadística inferior a 0.05 es muy elevada, esperándose encontrar unos 1.000 genes sólo debido al azar. Por esta razón, es necesario llevar a cabo una corrección para *test* múltiples (Tabla 1).

TIPO DE TUMOR	MÉTODO DE CORRECCIÓN	LONGITUD DE LA LISTA GÉNICA
LLC	Bonferroni ajustado	21
	Bonferroni estándar	21
	Just alpha	870
Cáncer de páncreas	Bonferroni ajustado	78
	Bonferroni estándar	78
	Just alpha	709
Cáncer de mama	Bonferroni ajustado	18
	Bonferroni estándar	18
	Just alpha	294

***Tabla 1*. Número de genes expresados diferencialmente en cualquier dirección (P<0.01) según el método de corrección empleado y el tipo tumoral.** LLC significa **L**eucemia **L**infática **C**rónica.

Los métodos 'Bonferroni estándar' y 'Bonferroni ajustado' están basados en la independencia de los *test* realizados. Además son fuertemente restrictivos (30), por eso el tamaño de las listas génicas obtenidas es muy limitado. En cambio la prueba 'Sólo *alpha*' (Figura 5) no es exigente a la hora de ajustar los resultados de múltiples comparaciones. En base a ellos, se puede concluir que el método Bonferroni protege excesivamente contra la posibilidad de identificar falsos positivos a costa de disminuir gravemente la potencia del análisis.

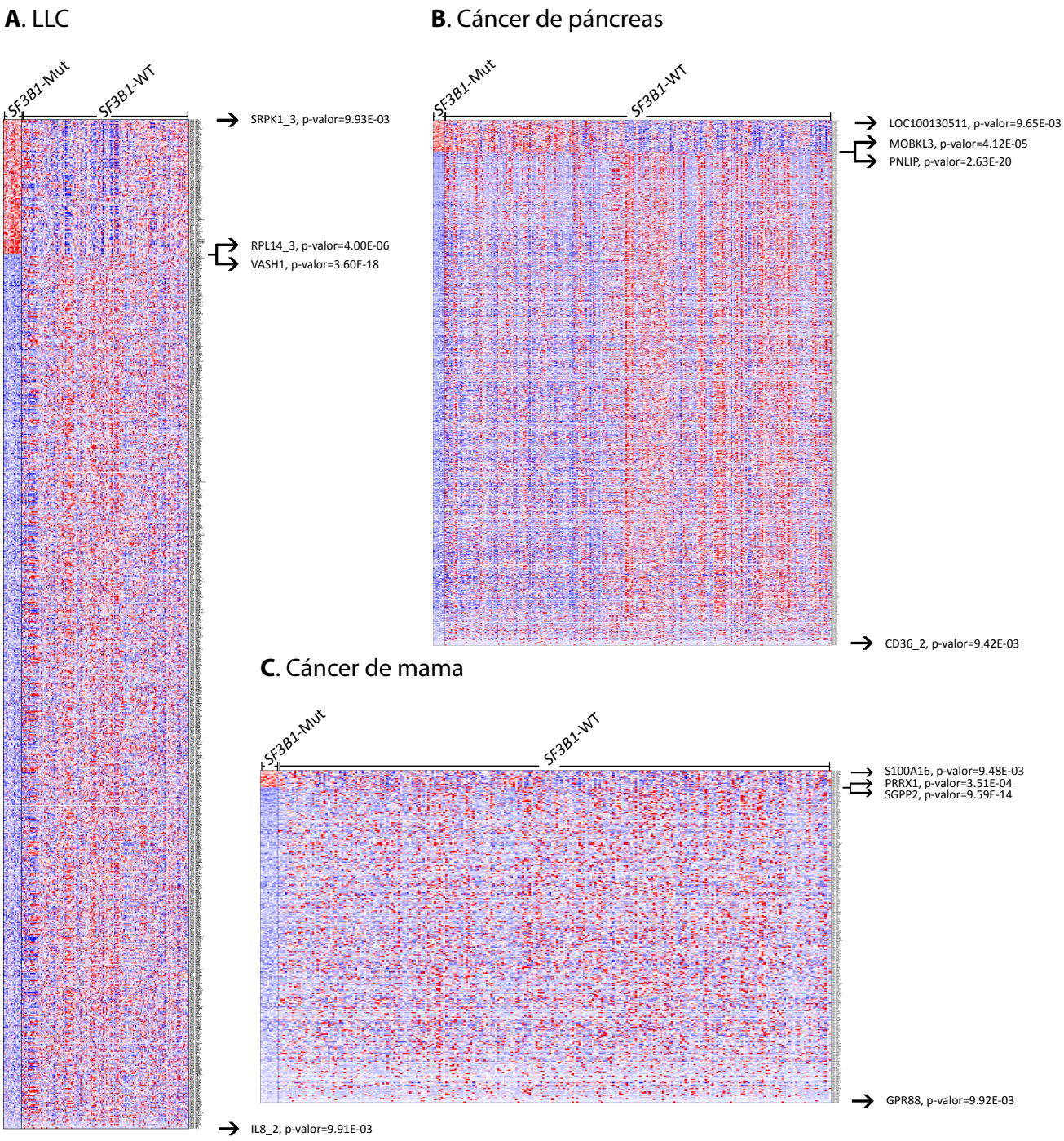


***Figura 5*. Representación gráfica de los valores de expresión de las sondas con diferencias estadísticamente significativas entre los dos grupos muestrales estudiados.** Los valores de expresión se muestran en una matriz bidimensional cuyas filas y columnas representan los múltiples genes y pacientes respectivamente. El espectro de colores indica la magnitud de estos valores en relación a su media total (Blanco – Valor medio; Rojo – Valores mayores que la media; Azul – Valores menores que la media), por lo que sólo se pueden hacer comparaciones directas dentro de la misma sonda. A la derecha se enumeran los genes con mayor y menor significación estadística así como sus respectivos p-valores. Con LLC se denota la **L**eucemia **L**infática **C**rónica y con *SF3B1*-Mut y *SF3B1-WT* *SF3B1* **Mut**ado y *SF3B1 **W**ild **T**ype* respectivamente.

5.1.2.- ANÁLISIS DE LOS GENES EXPRESADOS DIFERENCIALMENTE QUE SEAN RECURRENTES EN LOS TRES TIPOS DE TUMORES ESTUDIADOS

A pesar del gran poder de este protocolo para estudiar cambios en los perfiles de expresión génica, no permite deducir si la alteración de *SF3B1* repercute indirectamente en los genes que están expresados diferencialmente entre los estados mutado y no mutado por estar sujetos a la actividad de alguno en particular que es diana de SF3B1 o si, por el contrario, son afectados de manera directa, en cuyo caso se esperaría que fueran comunes en los tres tipos tumorales. Para solucionar esta cuestión, se cruzaron todas las listas génicas que resultan de gWb según la prueba de ajuste. Dado que no existe ningún gen recurrente en los tres cánceres, se procedió al estudio dos a dos. Este enfoque permitió identificar 24 genes comunes a LLC y cáncer de páncreas, 4 a LLC y cáncer de mama y 11 a cáncer de páncreas y mama (Tabla 2). Todos ellos están incluidos únicamente en las grandes listas que derivan de gWb sin corregir los resultados de comparaciones múltiples ('Sólo *alpha*') en sus matrices correspondientes.

Tabla 2. **Genes expresados diferencialmente en dos colectivos de pacientes (con y sin *SF3B1* mutado) que son recurrentes en: A. LLC y cáncer de páncreas, B. LLC y cáncer de mama o en C. Cáncer de páncreas y mama.** Se indican también sus p-valores correspondientes en función del tipo tumoral mientras que en azul se especifican aquéllos que recoge la base de datos COSMIC. Con $\overline{X}_{SF3B1\text{-Mut}}$ y $\overline{X}_{SF3B1\text{-}WT}$ se denota respectivamente el valor de expresión medio para cada gen entre pacientes con y sin mutaciones somáticas en *SF3B1* mientras que *FC* significa ***F**old **C**hange*.

A. LLC *vs*. Cáncer de páncreas

	LLC				Cáncer de páncreas			
Gen	$\overline{X}_{SF3B1\text{-Mut}}$	$\overline{X}_{SF3B1\text{-WT}}$	**p-valor**	***FC***	$\overline{X}_{SF3B1\text{-Mut}}$	$\overline{X}_{SF3B1\text{-WT}}$	**p-valor**	***FC***
AMT	47.68	75.11	1.10E-03	-0.66	74.97	136.82	1.48E-03	-0.87
ANGEL1	57.45	75.53	3.89E-03	-0.40	70.92	102.22	5.56E-03	-0.53
ARHGAP9	1210.09	1497.02	2.14E-03	-0.31	59.07	111.72	8.48E-04	-0.92
DTX1	112.26	217.24	7.04E-06	-0.95	22.30	51.81	4.30E-05	-1.22
ERP27	77.68	105.66	6.54E-03	-0.44	93.46	555.96	1.34E-06	-2.57
GNB4	39.21	204.66	1.52E-08	-2.38	36.74	65.95	2.56E-03	-0.84
IKBKB	78.18	100.38	8.05E-03	-0.36	119.66	169.00	1.25E-03	-0.50
IL10RA	893.21	1240.32	5.44E-03	-0.47	34.65	62.19	1.27E-03	-0.84
LASS5	426.04	496.22	3.18E-03	-0.22	132.71	184.51	2.22E-04	-0.48
LTBP3	41.48	50.75	6.22E-03	-0.29	307.95	497.56	5.13E-04	-0.69
LZTR1	40.95	51.02	8.17E-05	-0.32	434.99	570.88	1.60E-03	-0.39
MTMR10	83.98	119.11	6.42E-03	-0.50	183.42	239.17	6.71E-03	-0.38
PRKCH	73.49	135.44	3.48E-03	-0.88	243.03	325.40	9.31E-03	-0.42
RFNG	63.45	84.47	3.55E-04	-0.41	411.29	516.20	4.09E-03	-0.33
RING1	236.76	147.25	2.72E-03	0.69	1008.65	1218.87	2.86E-03	-0.27
RPL18A	16973.81	16292.33	9.10E-03	0.06	11189.48	12520.55	7.31E-03	-0.16
SNRPN	1124.96	1317.27	4.69E-03	-0.23	359.51	552.91	8.64E-04	-0.62
TCF4	6673.51	5567.60	1.19E-03	0.26	205.47	302.08	9.67E-03	-0.56
TMEM175	106.61	145.50	1.15E-04	-0.45	241.80	303.46	6.60E-03	-0.32
TPR	662.83	792.76	1.69E-03	-0.26	47.97	129.72	3.13E-03	-1.44
TSPAN33	332.98	475.84	1.96E-04	-0.52	61.26	123.88	2.55E-10	-1.02
TTLL3	210.76	289.05	7.37E-03	-0.46	20.95	52.46	1.60E-03	-1.32
VASH1	26.05	157.27	3.60E-18	-2.59	51.51	101.85	6.33E-04	-0.98
ZMIZ2	107.34	144.13	8.63E-04	-0.43	98.48	145.34	8.35E-03	-0.56

B. LLC *vs*. Cáncer de mama

	LLC				Cáncer de mama			
Gen	$\overline{X}_{SF3B1\text{-Mut}}$	$\overline{X}_{SF3B1\text{-WT}}$	**p-valor**	***FC***	$\overline{X}_{SF3B1\text{-Mut}}$	$\overline{X}_{SF3B1\text{-WT}}$	**p-valor**	***FC***
C1orf9	51.03	66.21	9.96E-04	-0.38	1.80	2.43	7.96E-03	-0.44
C12orf51	88.73	109.71	1.23E-03	-0.31	1.21	1.73	7.77E-03	-0.52
HIST1H4J	146.87	252.81	2.07E-03	-0.78	1.55	2.66	5.99E-03	-0.78
MEX3C	48.11	65.23	1.64E-04	-0.44	1.07	1.39	1.25E-03	-0.38

C. Cáncer de páncreas *vs*. Cáncer de mama

	Cáncer de páncreas				Cáncer de mama			
Gen	$\overline{X}_{SF3B1\text{-Mut}}$	$\overline{X}_{SF3B1\text{-WT}}$	**p-valor**	***FC***	$\overline{X}_{SF3B1\text{-Mut}}$	$\overline{X}_{SF3B1\text{-WT}}$	**p-valor**	***FC***
ARHGEF17	53.40	79.11	3.18E-03	-0.57	1.97	1.58	9.68E-03	0.31
CDC2L6	250.72	350.73	5.45E-04	-0.48	1.04	1.46	5.88E-04	-0.49
ISLR	595.01	1068.14	4.97E-04	-0.84	11.53	13.16	1.45E-04	-0.19
KCNJ8	101.03	170.68	7.03E-03	-0.76	2.56	3.79	8.16E-03	-0.56
LRRC32	525.21	967.54	4.22E-05	-0.88	9.27	7.25	1.92E-03	0.36
PAPLN	22.41	68.75	2.71E-07	-1.62	2.05	1.65	4.74E-03	0.31
PDCL3	319.25	265.54	7.01E-03	0.27	1.18	1.51	1.27E-03	-0.36
PTP4A3	70.66	109.49	3.03E-03	-0.63	1.07	1.54	5.30E-03	-0.53
RASGRP3	97.95	142.53	5.51E-03	-0.54	10.58	14.19	5.72E-05	-0.42
SLC5A1	124.29	273.26	6.31E-04	-1.14	4.46	11.22	9.19E-04	-1.33
SLC16A2	36.66	57.02	2.64E-03	-0.64	1.49	2.31	5.83E-03	-0.63

A continuación, se procedió a analizar con R si el sentido de cambio de la expresión de los 24 genes comunes a LLC y cáncer de páncreas debido a la presencia de mutaciones somáticas en *SF3B1* (también evaluado) era el mismo en ambos tipos tumorales. En base a ellos, y mediante diagramas de cajas para cada gen, se representaron las distribuciones y los estadísticos de resumen [valores máximo y mínimo y cuartiles Q1, Q2 (mediana) y Q3] de sus valores de expresión correspondientes según la clase genotípica a la que pertenecen (*WT* o *SF3B1* mutado). La siguiente figura (Figura 6) resume el análisis conjunto de expresión génica diferencial en pacientes de LLC y cáncer de páncreas de acuerdo con el estado mutacional de *SF3B1.* En los ejes X e Y se representan, mediante escala logarítmica, los *fold changes* de cada gen en los cánceres de páncreas y sangre respectivamente. El tamaño de las burbujas refleja la abundancia relativa (en términos de $\overline{X}_{total}$) de los distintos genes en función del tipo tumoral.

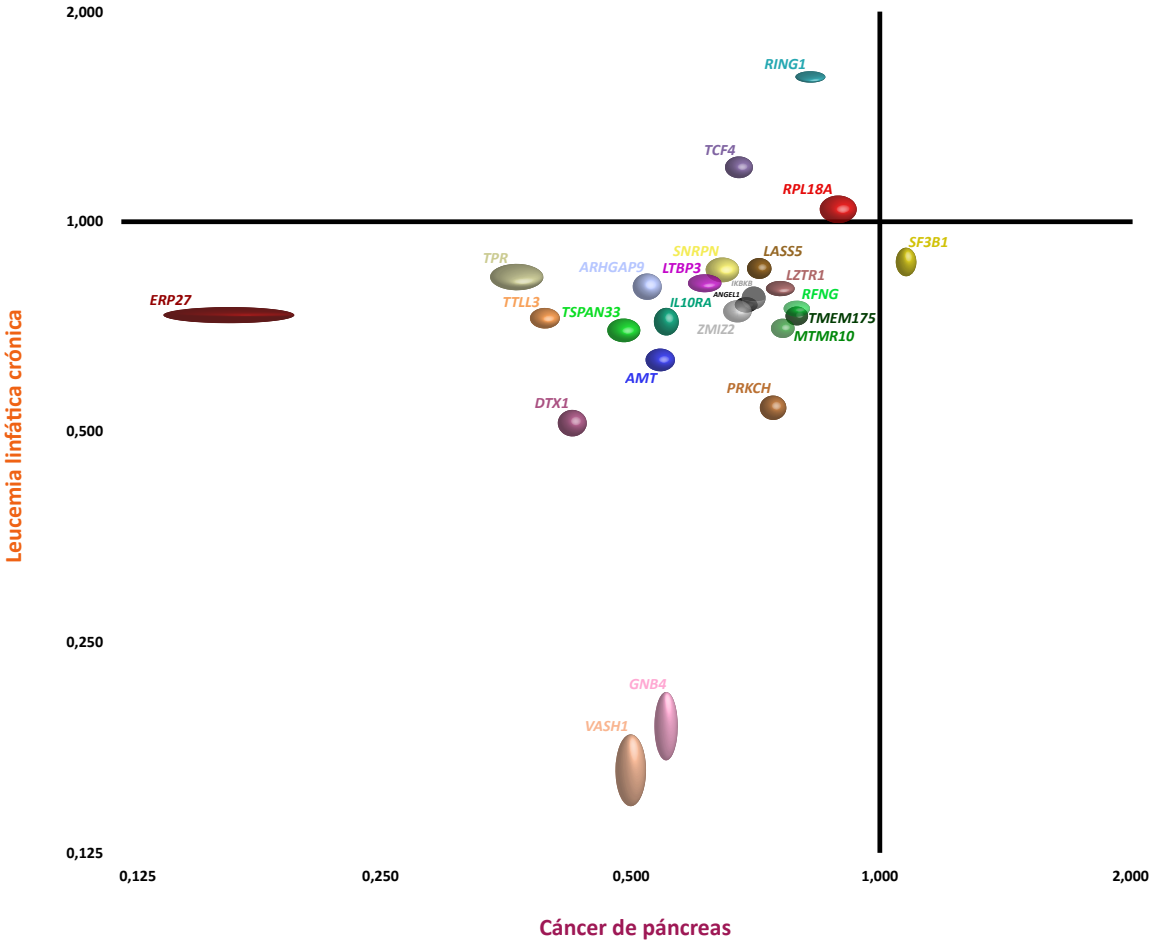


Figura 6. **Representación gráfica del análisis conjunto de expresión génica diferencial asociada a *SF3B1* en casos de leucemia linfática crónica y cáncer de páncreas.**

Como se puede observar, de los 24 genes con expresión diferencial entre tumores mutados y no mutados para *SF3B1* y comunes a LLC y cáncer de páncreas, la mayor

parte (20 de los 24) tienen un cambio de expresión en el mismo sentido en ambos tipos tumorales, y en los dos casos coincide con una expresión reprimida.

5.1.3.- ANÁLISIS DE ENRIQUECIMIENTO FUNCIONAL CON DAVID

Hasta ahora, se ha utilizado la información transcriptómica disponible para tratar de determinar cambios en los perfiles de expresión génica entre dos grupos de pacientes (portadores y no portadores de mutaciones somáticas en *SF3B1*) con cánceres de sangre, páncreas o mama. Sin embargo, la tarea se complica cuando las diferencias a nivel de genes particulares son sutiles. El análisis de enriquecimiento se revela como una estrategia de alto rendimiento para superar este obstáculo, pues se aleja del enfoque a nivel individual para centrarse en grupos de genes que están relacionados funcionalmente. Por consiguiente, permite identificar mecanismos celulares alterados en la investigación realizada. Con este objetivo, las listas génicas fueron cargadas de manera independiente en la base de datos DAVID. Sólo se obtuvieron resultados significativos (P<0.01) con datos de LLC sometidos al análisis estadístico previo sin corregir los efectos provocados por la ejecución de *test* múltiples ('Sólo *alpha*') (Tabla 3).

ES	**Mecanismo celular** (y categoría)	**Total**	**%**	**p-valor**	**Genes**
3.75	Regulación de la transcripción (GOTERM_BP_FAT)	128	20.65	7.18E-06	*EIF2C1, IL16, PASK, ZXDC, CBX4, BBX, CNOT2, FOXO1, RORA, LASS5, RAB1A, RCBTB1, GABPB1, EPC2, SMARCD2, MED28, PHTF2, MIER3, PSIP1, BCL7A, MTERFD3, ZNF180, RBL2, TAF4B, ZNF507, ARID1B, ZNF37A, KDM2B, ZNF238, RAB18, PARP14, ZNF384, EDF1, ZNF434, ZNF589, ZNF587, SUDS3, SUV420H1, ZNF586, SIVA1, UBE2V1, ATF1, NR2C1, PLAGL1, TRIM66, KRAS, HNRNPD, BTF3, TCF4, RUNX3, KLF7, KLF8, IKZF1, SMAD5, SMAD3, SNW1, MALT1, ZNF626, ZNF836, DDX5, ZNF585A, ABCG1, GCFC1, SAFB2, NOTCH2, DMRTC1, HIVEP3, JAZF1, ZNF318, TMPO, IKBKB, ZNF117, KLF3, BMI1, BACH2, ZNF532, EZH1, ZNF10, ZNF673, ZBTB38, NFATC2IP, MCM7, ELOF1, LRRFIP1, KDM5A, SERTAD2, ZNF548, SSBP2, SP100, ZNF142, RING1, ZNF684, IRF2BP2, IFI16, ZNF140, UBE2N, HNRPDL, BRWD1, ZMIZ2, BTG1, C19ORF2, CAND1, RBM39, VOPP1, WASL, NSD1, ZNF484, BTAF1, BCLAF1, CARHSP1, ZNF468, BDP1, NR3C2, CTCFL, ARID2, TSC22D2, ZSCAN21, LANCL2, MAP3K1, PRDM10, BCL3, ATRX, ZNF671, GMCL1, DR1, JAK3, TBL1X, ZBTB1*

ES	**Mecanismo celular** (y categoría)	**Total**	**%**	**p-valor**	**Genes**
3.75	Transcripción (SP_PIR_KEYWORDS)	101	16.29	1.40E-05	*EIF2C1, IL16, CBX4, ZXDC, BBX, CNOT2, FOXO1, RORA, RCBTB1, KDM1B, GABPB1, EPC2, SMARCD2, MED28, MIER3, PHTF2, PSIP1, ZNF180, MTERFD3, RBL2, TAF4B, ZNF507, ARID1B, ZNF37A, KDM2B, ZNF238, PARP14, ZNF384, ZNF434, EDF1, ZNF589, ZNF587, SUDS3, SUV420H1, ZNF586, ATF1, NR2C1, PLAGL1, HNRNPD, BTF3, TCF4, RUNX3, KLF7, KLF8, IKZF1, SMAD5, SMAD3, ZNF836, ZNF626, ZNF585A, SAFB2, GCFC1, NOTCH2, DMRTC1, HIVEP3, JAZF1, ZNF318, ZNF117, KLF3, BMI1, BACH2, ZNF532, EZH1, ZNF10, ZBTB38, MCM7, ELOF1, LRRFIP1, KDM5A, SERTAD2, ZNF548, ZNF142, RING1, ZNF684, IRF2BP2, IFI16, ZNF140, HNRPDL, BRWD1, ZMIZ2, CAND1, VOPP1, WASL, RBM39, NSD1, ZNF484, BCLAF1, ZNF468, BDP1, NR3C2, CTCFL, POLR2C, ARID2, ZSCAN21, PRDM10, BCL3, PHF11, ZNF671, DR1, TBL1X, ZBTB1*
3.75	Regulación de la transcripción (SP_PIR_KEYWORDS)	99	15.97	1.58E-05	*EIF2C1, IL16, CBX4, ZXDC, BBX, CNOT2, FOXO1, RORA, RCBTB1, KDM1B, GABPB1, EPC2, SMARCD2, MED28, MIER3, PHTF2, PSIP1, ZNF180, MTERFD3, RBL2, TAF4B, ZNF507, ARID1B, ZNF37A, KDM2B, ZNF238, PARP14, ZNF384, ZNF434, EDF1, ZNF587, SUDS3, SUV420H1, ZNF586, ATF1, NR2C1, PLAGL1, HNRNPD, BTF3, TCF4, RUNX3, KLF7, KLF8, IKZF1, SMAD5, SMAD3, ZNF836, ZNF626, ZNF585A, SAFB2, GCFC1, NOTCH2, DMRTC1, HIVEP3, JAZF1, ZNF318, ZNF117, KLF3, BMI1, BACH2, ZNF532, EZH1, ZNF10, ZBTB38, MCM7, ELOF1, LRRFIP1, KDM5A, SERTAD2, ZNF548, RING1, ZNF142, ZNF684, IRF2BP2, IFI16, ZNF140, HNRPDL, BRWD1, ZMIZ2, CAND1, VOPP1, WASL, RBM39, NSD1, ZNF484, BCLAF1, ZNF468, BDP1, NR3C2, CTCFL, ARID2, ZSCAN21, PRDM10, BCL3, PHF11, ZNF671, DR1, TBL1X, ZBTB1*
3.75	Transcripción (GOTERM_BP_FAT)	102	16.45	1.56E-04	*EIF2C1, IL16, ZXDC, CBX4, BBX, CNOT2, FOXO1, RORA, RCBTB1, GABPB1, EPC2, SMARCD2, MED28, PHTF2, MIER3, PSIP1, ZNF180, MTERFD3, RBL2, TAF4B, ZNF507, ARID1B, ZNF37A, KDM2B, ZNF238, PARP14, ZNF384, ZNF434, EDF1, ZNF589, ZNF587, SUDS3, SUV420H1, ZNF586, DIDO1, ATF1, NR2C1, PLAGL1, HNRNPD, BTF3, TCF4, RUNX3, KLF7, KLF8, IKZF1, SMAD5, SMAD3, ZNF836, ZNF626, ZNF585A, SAFB2, GCFC1, NOTCH2, DMRTC1, HIVEP3, JAZF1, ZNF318, ZNF117, KLF3, BMI1, BACH2, ZNF532, EZH1, ZNF10, ZBTB38, MCM7, ELOF1, LRRFIP1, KDM5A, SERTAD2, ZNF548, ZNF142, RING1, ZNF684, IRF2BP2, IFI16, ZNF140, HNRPDL, BRWD1, ZMIZ2, CAND1, VOPP1, WASL, RBM39, NSD1, ZNF484, BCLAF1, ZNF468, BDP1, NR3C2, CTCFL, POLR2C, ARID2, ZSCAN21, PRDM10, BCL3, DTX1, PWP1, ZNF671, DR1, TBL1X, ZBTB1*
1.63	Procesamiento del RNA (GOTERM_BP_FAT)	33	5.32	2.16E-03	*RNMT, RPL14, POLR2C, SF3B1, RPL7, CD2BP2, PCBP1, PCBP2, HNRNPD, WDR12, RPL5, SCNM1, RPL11, LUC7L3, NSA2, TXNL4B, SRPK2, KRR1, SNRPN, MAGOH, SMAD3, SNW1, RPS6, DDX5, HNRNPA1, SRPK1, HNRPDL, RSL1D1, SNRNP48, PAPD4, CELF1, RBM39, SNRNP25*

ES	**Mecanismo celular** (y categoría)	**Total**	**%**	**p-valor**	**Genes**
1.63	*Splicing* del RNA (GOTERM_BP_FAT)	20	3.23	4.13E-03	*TXNL4B, SRPK2, SNRPN, MAGOH, SNW1, DDX5, HNRNPA1, POLR2C, SRPK1, SF3B1, SNRNP48, CD2BP2, PCBP1, PCBP2, HNRNPD, CELF1, SCNM1, RBM39, SNRNP25, LUC7L3*
1.63	Procesamiento del mRNA (GOTERM_BP_FAT)	21	3.39	7.21E-03	*TXNL4B, SRPK2, RNMT, MAGOH, SNW1, DDX5, HNRNPA1, POLR2C, SRPK1, SF3B1, SNRNP48, PAPD4, CD2BP2, PCBP1, PCBP2, HNRNPD, CELF1, SCNM1, RBM39, SNRNP25, LUC7L3*

***Tabla 3*. Análisis de enriquecimiento con DAVID para genes expresados diferencialmente entre pacientes de leucemia linfática crónica con genotipos *SF3B1*-Mutado y *SF3B1-Wild Type* (método de corrección, 'Sólo *alpha*').** Se seleccionaron los mecanismos celulares cuya alteración podría estar relacionada de manera directa con la actividad mutante del factor de *splicing* SF3B1. Se especifican también sus valores de enriquecimiento (*ES*) y p-valores correspondientes así como los grupos de genes relacionados funcionalmente.

De manera adicional, y con el objetivo de tratar de identificar posibles genes del cáncer cuya expresión pudiera verse afectada por *SF3B1*, llevamos a cabo un filtrado de las tres listas génicas que produjo gWb (una por método de corrección aplicado) con la base de datos COSMIC. Los genes resultantes se analizaron con DAVID. De forma análoga a lo que sucede en el caso anterior, el resultado sólo fue significativo tras filtrar aquella que surge de analizar estadísticamente la matriz de LLC con la prueba 'Sólo *alpha*' (Tabla 4).

ES	**Mecanismo celular** (y categoría)	**Total**	**%**	**p-valor**	**Genes**
1.61	Actividad reguladora de la transcripción (GOTERM_MF_FAT)	10	32.26	3.37E-03	*NOTCH2, IKZF1, BTG1, JAZF1, ZNF384, BCL3, KDM5A, DDX5, NSD1, ATF1*
1.61	Regulación positiva de la transcripción (GOTERM_BP_FAT)	6	19.36	7.43E-03	*IKZF1, BCL3, JAK3, KDM5A, DDX5, NSD1*
1.61	Regulación positiva de la expresión génica (GOTERM_BP_FAT)	6	19.36	8.40E-03	*IKZF1, BCL3, JAK3, KDM5A, DDX5, NSD1*
1.15	Regulación de la transcripción (GOTERM_BP_FAT)	17	54.84	3.08E-05	*IKZF1, MALT1, DDX5, ARID2, ATF1, ATRX, NOTCH2, KRAS, BTG1, JAZF1, PSIP1, ZNF384, BCL3, JAK3, KDM5A, BCL7A, NSD1*
1.15	Regulación de la transcripción (SP_PIR_KEYWORDS)	10	32.26	2.87E-03	*NOTCH2, IKZF1, JAZF1, ZNF384, PSIP1, BCL3, KDM5A, NSD1, ARID2, ATF1*
1.15	Transcripción (SP_PIR_KEYWORDS)	10	32.26	3.33E-03	*NOTCH2, IKZF1, JAZF1, ZNF384, PSIP1, BCL3, KDM5A, NSD1, ARID2, ATF1*

ES	**Mecanismo celular** (y categoría)	**Total**	**%**	**p-valor**	**Genes**
1.15	Actividad reguladora de la transcripción (GOTERM_BP_FAT)	10	32.26	3.37E-03	*NOTCH2, IKZF1, BTG1, JAZF1, ZNF384, BCL3, KDM5A, DDX5, NSD1, ATF1*

***Tabla 4*. Análisis de enriquecimiento funcional con DAVID después de filtrar con COSMIC los genes con perfiles de expresión diferenciales entre estados mutado y no mutado de *SF3B1* para casos de leucemia linfática crónica (sin corrección, 'Sólo *alpha*').** Sólo se incluyen los que podrían estar alterados con mayor probabilidad debido a la presencia de mutaciones en *SF3B1*. Para cada uno de ellos se especifica el valor de enriquecimiento (*ES*) y p-valor correspondientes así como los genes relacionados en términos funcionales.

5.1.4.- ANÁLISIS FUNCIONAL NO SUPERVISADO CON geWorkbench: *CLUSTERING* JERÁRQUICO

Con el objetivo de tratar de identificar si los tumores con mutaciones en *SF3B1* podrían tener perfiles de expresión génica similares entre ellos, se procedió a realizar un análisis de *clusterización* no supervisado. Antes de nada, las tres matrices iniciales (una por tipo de cáncer) fueron analizadas con R para seleccionar los 1000 genes cuya expresión tiene una mayor variabilidad entre los dos grupos de pacientes estudiados (mutantes y *WT* para *SF3B*1). Aunque gWb no requiere este filtro previo cuando se trata de analizar diferencias entre los perfiles de expresión génica de estas mismas poblaciones muestrales, es un requisito indispensable para hacer un *clustering* jerárquico, pues éste tiene unos requerimientos de memoria muy altos, por lo que la eliminación de los genes que muestran menor variabilidad permite disminuir los requerimientos computacionales. En concreto, se realizó una *clusterización* de muestras por cada combinación posible de los distintos métodos (distancias máxima, promedio y mínima entre *clusters*) y métricas (distancia euclídea y los coeficientes de correlación de Pearson o por rangos de Spearman) de *clustering* que ofrece el programa. Con la siguiente figura (Figura 7) se expone, a modo de ejemplo, el resultado de unos de los análisis de *clusterización* realizados.

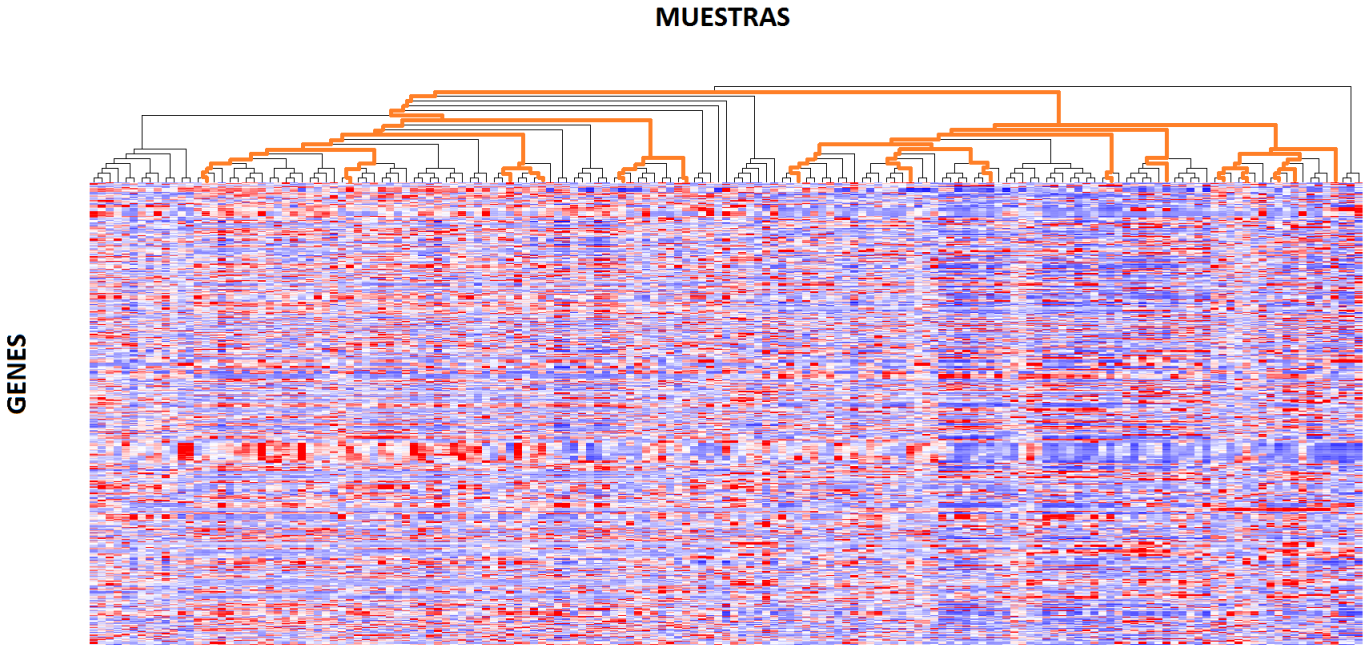


***Figura 7*. Representación gráfica del resultado de un análisis de *clusterización* de muestras.** En concreto, éste fue realizado para el *set* de datos de leucemia linfática crónica (método y métrica de *clustering*; Distancia promedio y coeficiente de correlación de Pearson respectivamente). En naranja se especifican las líneas correspondientes a los casos con *SF3B1* mutado.

Aunque este tipo de análisis permitía el agrupamiento claro de algunos tipos de muestras, en ningún caso se obtuvieron resultados concluyentes que guarden relación con el genotipo de *SF3B1*. Un análisis detallado reveló que la principal causa de agrupación de muestras no se debía a las mutaciones en *SF3B1*, sino a la presencia de mutaciones cromosómicas específicas (trisomía del cromosoma 12, pérdida de 13q14, …). Estos resultados sugieren que aunque es posible que las mutaciones en *SF3B1* puedan provocar cambios en la expresión de otros genes, probablemente las diferencias a nivel transcriptómico no sean lo suficientemente fuertes como para determinar que las muestras *clustericen* en base a ellas.

5.2.- ANÁLISIS CUALITATIVO DE SECUENCIAS TRANSCRIPTÓMICAS. CARACTERIZACIÓN DE NUEVOS EVENTOS DE *SPLICING* EN LLC ASOCIADOS A MUTACIONES EN *SF3B1*

El segundo objetivo de este trabajo consistió en tratar de identificar nuevos fenómenos de *splicing* que pudieran producirse por la acumulación de mutaciones en el factor SF3B1. Para ello, se decidió emplear una estrategia cualitativa basada en el análisis de k-mers. En el siguiente diagrama de flujo se muestra el protocolo experimental utilizado para tratar de identificar nuevos eventos de *splicing* asociados a mutaciones en *SF3B1* mediante comparación directa de lecturas transcriptómicas k-merizadas (Figura 8).

En primer lugar, asumimos que la alteración de *SF3B1* provoca cambios en el perfil de *splicing* de determinados genes que sólo son recurrentes entre pacientes con el genotipo mutante. Teniendo en cuenta esta premisa, la aparición de un fenómeno de *splicing* anormal se traducirá en la presencia de nuevos k-mers de 23 bases correspondientes a esta secuencia alrededor del *splicing*, que no debería aparecer en los casos control (Figura 9).

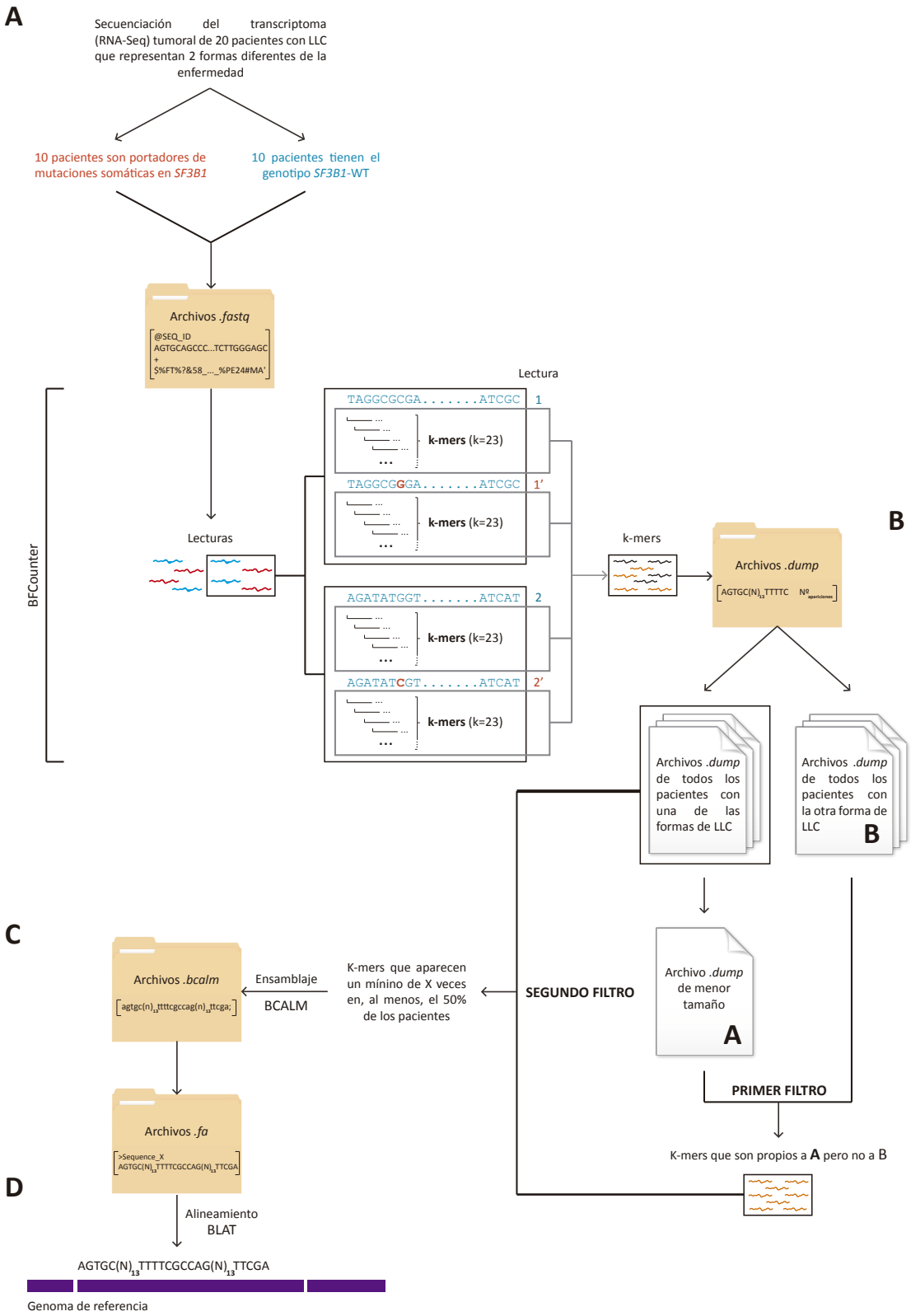


Figura 8. **Caracterización de patrones de *splicing* desconocidos asociados a *SF3B1* mutado en casos de leucemia linfática crónica (LLC).** (**A**) Los datos crudos de RNA-Seq (ficheros *.fastq*) fueron procesados empleando el programa BFCounter para obtener los k-mers de 23 bases. Se procesaron los datos de RNA-Seq de 10 pacientes de LLC con genotipo *SF3B1*-Mutado y 10 pacientes *SF3B1-WT*. En ambos casos el subtipo tumoral es no mutado para la región IgHV. (**B**) A continuación, se utilizó un *script* en Perl para filtrar los datos crudos y seleccionar los k-mers presentes en tumores con *SF3B1* mutado, pero no en casos control. (**C**) Los k-mers resultantes fueron ensamblados *de novo* con BCalm, generándose distintas secuencias en formato *.fasta*. (**D**) Las secuencias obtenidas fueron alineadas al genoma de referencia usando BLAT. Finalmente, se programaron en Perl y ejecutaron en Linux nuevas herramientas de trabajo bioinformáticas con el objetivo de determinar las posibles correspondencias génicas de estas secuencias así como para identificar aquéllas que, una vez cargadas en la versión gráfica de BLAT, pudieran delatar nuevos patrones de *splicing*. Con *SF3B1-WT* se denota *SF3B1* ***W****ild* ***T****ype,* X indica el Umbral de expresión basal y k el tamaño del k-mer.

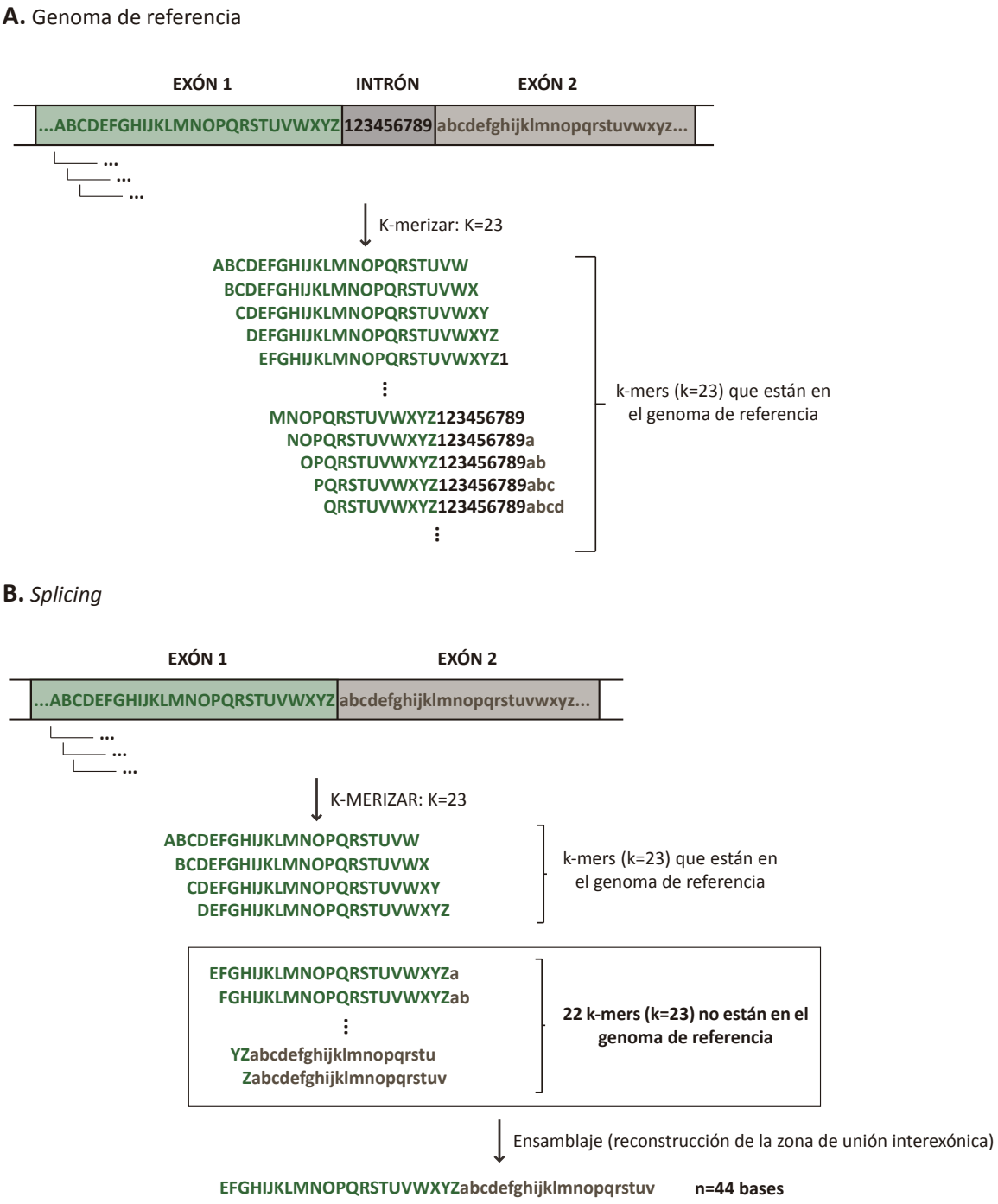


Figura 9. **Esquema de cómo un evento de *splicing* puede generar hasta k – 1 k-mers distintos.** k indica el tamaño del k-mer (en el ejemplo 23 bases). La presencia de un splicing anormal en una muestra provoca la aparición de k-1 k-mers únicos no presentes en una muestra normal (en este caso 23-1=22 k-mers). El ensamblaje *de novo* de estos 22 k-mers permite reconstruir una secuencia de n=2*k-2 bases (en este caso 44 bases).

Los datos de RNA-Seq se procesaron para extraer y contar todos los k-mers de 23 bases. Una vez extraídos, se buscaron aquéllos que estuvieran presentes en casos de LLC con *SF3B1* mutado y en ninguno de los *WT* (PRIMER FILTRO). Acto seguido se filtraron los k-mers que aparecían un mínimo de 4 veces en, al menos, el 50% de estos

pacientes (SEGUNDO FILTRO). Dado que la aproximación con k-mers no se había utilizado jamás para la identificación de nuevos eventos de *splicing* asociados a mutaciones en *SF3B1*, el establecimiento de un umbral de expresión fijado en 4 k-mers fue totalmente arbitrario. Era lo suficientemente bajo como para esperar que muchos superarían el filtro, pero al mismo tiempo aseguraba una cierta significatividad en el nivel de expresión génica (evitando efectos demasiado marginales). Sin embargo, el número de k-mers identificados fue bastante bajo (1316) y ninguno de ellos era recurrente en el 100% de los pacientes estudiados. Dado que un fenómeno de *splicing* nuevo puede generar hasta 22 k-mers distintos consecutivos, una vez filtrados los k-mers diferenciales se procedió a su ensamblaje mediante el programa BCALM. Esto resultó en la generación de 156 secuencias, de las que 84 pudieron ser alineadas al genoma de referencia. Entonces, se seleccionaron únicamente 36 por ser las que alinean en genes conocidos, fuera de zonas de DNA repetitivo. Para evitar resultados no asociados al genotipo de *SF3B1* se descartaron también aquéllas que podrían haber superado el primer filtro por contener *SNP*s (se asume que la heterogeneidad de la enfermedad entre pacientes está limitada a alteraciones en nuestro gen de estudio). Al final, un total de 7 secuencias tienen un tamaño que no coincide con el de la región cubierta por el alineamiento de modo que podrían ser indicativas de procesos de *splicing* desconocidos (Figura 10).

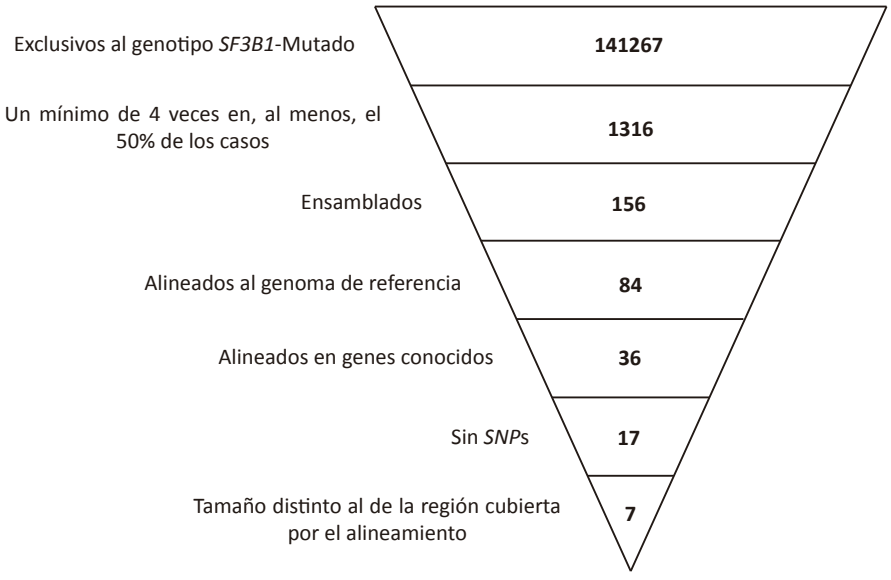


***Figura 10*. Esquema que muestra el proceso de filtrado usado para identificar k-mers candidatos en el estudio de nuevos eventos de *splicing* asociados a mutaciones en *SF3B1*.** Se asume que la alteración de este gen induce la expresión de ciertos genes silenciados en pacientes con el genotipo *Wild Type*. *SNP* significa ***S**ingle **N**ucleotide **P**olymorphism*.

En base a todos estos criterios, se identificaron un total de 13 genes que tan sólo se expresan en casos de LLC que son portadores de mutaciones somáticas puntuales en *SF3B1* (*PKD2*, *USP3*, *C10orf128*, *NRF1*, *KDM5B*, *FAM13B*, *AP1G1*, *RP13-297E16.4*, *RP11-25L3.3*, *ZC3H14*, *FCGR2A*, *ASRGL1* y *C1orf112*) además de un nuevo evento de *splicing*, concretamente entre los exones 3 y 4 del gen *KCTD17 (**K** **C**hannel **T**etramerization **D**omain containing **17**)* (Figura 11).

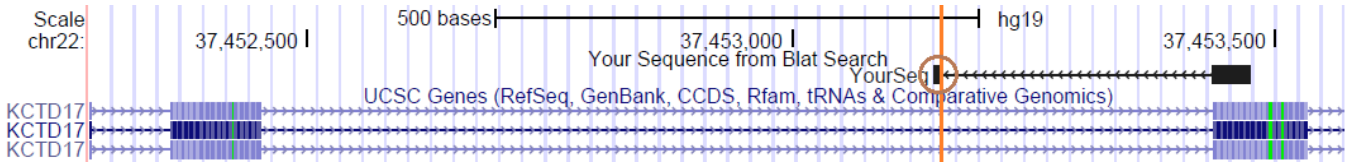


Figura 11. **Representación gráfica de nuevo evento de *splicing* para el gen *KCTD17*.** Se señala la presencia de un sitio críptico de *splicing* en 5'.

Como se ha dicho, la cantidad de k-mers que superaron el umbral de expresión basal fue muy inferior al esperado, por lo que se podrían estar sobrefiltrando los datos. Para descartar esta posibilidad, se repitió el experimento reduciéndolo de 4 a 2 k-mers. El número de secuencias candidatas al final fue considerablemente mayor pero ninguna sirvió para identificar patrones de *splicing* desconocidos ni genes que se expresen sólo en el grupo de pacientes con *SF3B1* mutado que sean especialmente importantes en el proceso oncogénico. Tampoco se obtuvieron resultados significativos tras realizar un análisis de enriquecimiento funcional en DAVID con las correspondencias génicas de las secuencias alineadas.

En segundo lugar, dado que la presencia de mutaciones somáticas puntuales en *SF3B1* podría traer como consecuencia la pérdida de eventos de *splicing* alternativos necesarios para la expresión de ciertos genes, se procedió a buscar los k-mers que aparecieran en pacientes de LLC con *SF3B1-WT* y no en los mutantes (PRIMER FILTRO). Por lo demás, los criterios de selección de secuencias permanecieron invariables y se desarrolló el protocolo con los dos umbrales de expresión basal (4 y 2 k-mers) fijados en la aproximación anterior. También se llevó a cabo el análisis de enriquecimiento funcional con DAVID de las correspondencias génicas que resultan después del alineamiento. No se obtuvieron resultados significativos en ningún caso.

5.3.- ANÁLISIS CUANTITATIVO DEL PATRÓN DE *SPLICING* DE *ATM* EN LLC SEGÚN EL GENOTIPO DE *SF3B1*

Antes de la realización del presente trabajo, un estudio ya había descrito cambios en el *splicing* de *ATM* asociados a la presencia del gen *SF3B1* mutado en casos de LLC (13). Según este artículo, la actividad mutante del factor SF3B1 provoca la activación de un sitio críptico 3' de *splicing*, cuyo uso genera una proteína truncada. Por eso, *ATM* fue un control positivo bueno para valorar la utilidad del enfoque cualitativo con k-mers en la búsqueda de nuevos eventos de *splicing*. En contraposición con lo esperado, esta aproximación no permitió identificar el sitio de *splicing* alternativo en 3' que se activa de manera específica y recurrente en *ATM* debido a mutaciones puntuales en *SF3B1*. Una posible explicación a esto sería que tanto el *splicing* canónico de *ATM* como el críptico tienen lugar independientemente de la condición genotípica de *SF3B1* (nunca silenciamiento total), de ahí que los k-mers correspondientes nunca superen el PRIMER FILTRO. Sin embargo, la presencia de *SF3B1* mutado repercutiría en un aumento significativo de la frecuencia con la que se usa el sitio 3' de *splicing* alternativo durante el procesamiento del pre-mRNA de *ATM*. Para verificar esta hipótesis, se procedió a analizar cuantitativamente el patrón de *splicing* de este gen en función del estado mutacional de *SF3B1* mediante comparación directa de lecturas transcriptómicas k-merizadas para casos de LLC. Con este propósito, se seleccionaron dos k-mers concretos (uno representa al *splicing* canónico y el otro al críptico según el sitio aceptor que contengan sus secuencias) y se contó el número de veces que aparecen ambos en dos grupos de 10 pacientes de LLC con genotipos *SF3B1*-Mut o *SF3B1-WT*.

El análisis de dicho *splicing* alternativo en casos con *SF3B1* mutado y casos *WT* reveló que este proceso se detectaba en ambos tipos de tumores, por lo tanto, no es un *splicing* generado específicamente por la presencia de mutaciones en *SF3B1*. Sin embargo, tal y como se puede observar en la Figura 12, el k-mer correspondiente al *splicing* críptico es significativamente (P<0.01) más común entre pacientes con mutaciones en *SF3B1* que en los *WT*. Tanto es así que en estos casos, la frecuencia con la que se activa el sitio 3' de *splicing* alternativo es casi el doble que en aquéllos que no

son portadores de mutaciones somáticas en *SF3B1.* Asumiendo que la heterogeneidad de la enfermedad entre los dos grupos pacientes está restringida al genotipo de *SF3B1,* se puede concluir que la activación del sitio críptico aceptor en *ATM* es específica a *SF3B1* mutado y recurrente entre pacientes, tal y como habían descrito Ferreira et al. meses antes (13).

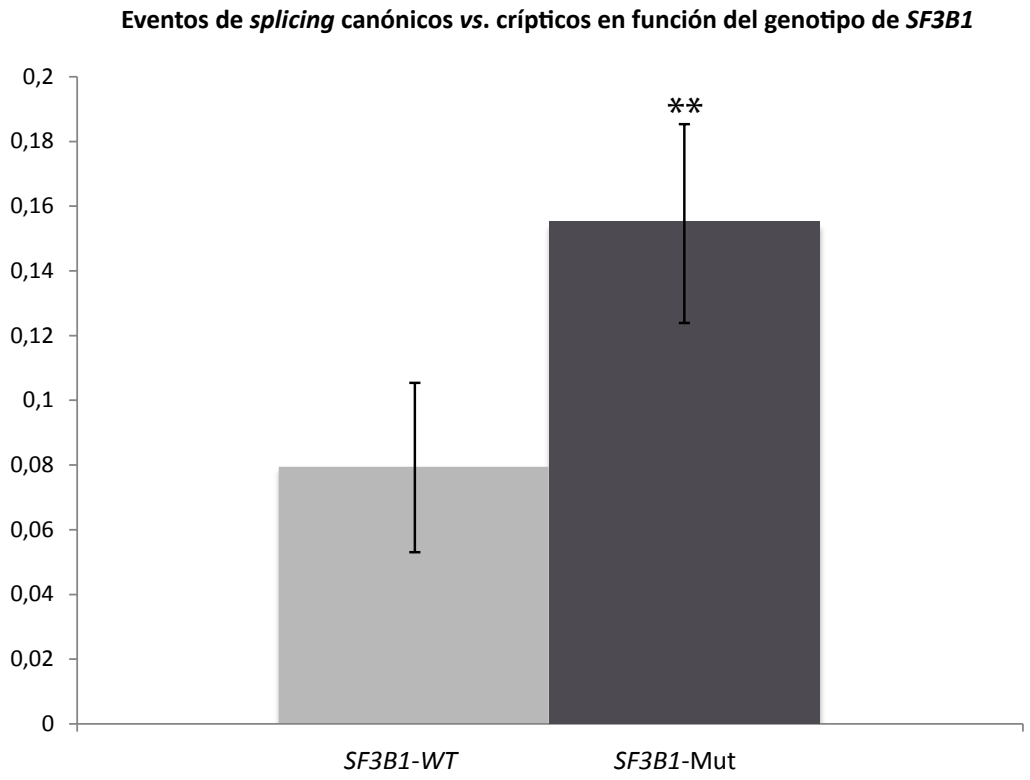


Figura 12. **Representación gráfica de la proporción media de eventos de *splicing* crípticos *versus* canónicos en casos de leucemia linfática crónica con *SF3B1* mutado** (*SF3B1*-Mut) **o *Wild Type*** (*SF3B1-WT*). (**, $P<0.01$).

6. DISCUSIÓN

La participación directa de genes de *splicing* en el proceso de transformación tumoral se conoce desde hace pocos años gracias al desarrollo de la tecnología de secuenciación masiva en paralelo. Dentro del enorme conjunto de genes que están alterados en cáncer, *SF3B1* es el más recurrente en síndromes mielodisplásicos y el segundo en LLC, así como en menor frecuencia en otros tipos de cáncer, incluyendo leucemias y tumores sólidos. Aunque aún se desconoce el mecanismo por el que las mutaciones en *SF3B1* contribuyen a la transformación tumoral, el hecho de que este gen codifique un componente central de la snRNP espliceosomal U2 que garantiza la fidelidad del punto de ramificación 3' incita a pensar que la actividad mutante del factor SF3B1 pueda provocar la activación de nuevos sitios aceptores de *splicing* que conllevarían un procesamiento inadecuado del pre-mRNA y como consecuencia, cambios en los perfiles de expresión génica. En el presente trabajo hemos tratado de comprobar si la alteración de dicho gen provoca cambios específicos en el perfil transcriptómico de pacientes con cáncer, o si estimula la activación de nuevos sitios de *splicing* alternativos que sean recurrentes.

Este análisis permitió identificar algunos genes cuya expresión aparece alterada en tumores con mutaciones en *SF3B1*. Aunque es posible que algunos de ellos puedan estar alterados por estas mutaciones, el reducido número de muestras con *SF3B1* alterado que hemos podido conseguir (8 casos en mama y 8 en páncreas), ha hecho imposible llevar a cabo una validación con una serie de pacientes distinta que permitiera confirmar o descartar estos hallazgos. Por esta razón, se procedió a realizar comparaciones entre los tres tipos de tumores estudiados. Este análisis reveló que no existía ningún gen común cuya expresión estuviese alterada en los tres tipos tumorales, ni al hacer comparaciones dos a dos. Esto puede deberse a que el método de corrección para comparaciones múltiples utilizado, Bonferroni, es muy exigente y aunque permite reducir considerablemente el número de falsos positivos obtenido, limita la sensibilidad para detectar positivos reales, generando muchos falsos

negativos. En consecuencia, se procedió a realizar de nuevo la misma comparación pero sin corregir por *test* múltiples. En este caso sí se identificaron varias decenas de genes alterados en ambos tipos de tumores. De hecho, se pudo comprobar que los cambios de expresión en estos genes observados en tumores con mutaciones en *SF3B1*, se producían siempre en el mismo sentido, y en la totalidad de los casos provocaban una disminución de la expresión. Estos datos podrían sugerir que mutaciones en *SF3B1* tienen un efecto sobre estos genes, independientemente del tipo de tumor estudiado, aunque será necesario disponer de más datos de mutaciones y de expresión para poder confirmarlo en una serie de validación independiente.

Por otra parte, el segundo objetivo de este trabajo consistió en explorar la posible existencia de fenómenos de *splicing* aberrante provocados por las mutaciones en *SF3B1*. Para ello se utilizaron los datos de RNA-Seq generados en el proyecto LLC. En este sentido, es necesario destacar que los procedimientos más habituales para el análisis de datos de RNA-Seq de muestras humanas se basan en la utilización de rutinas de trabajo en las que primero se alinean las secuencias al genoma de referencia, utilizando información de sitios de *splicing* canónicos, y posteriormente se procede a realizar filtros para identificar posibles sitios de *splicing* no canónicos o alternativos. Sin embargo, datos previos del laboratorio habían demostrado que la utilización de un paso de alineamiento al genoma de referencia provoca un sesgo en la identificación de sitios de *splicing* no anotados en las bases de datos, lo que puede limitar la utilidad de estas estrategias para identificar nuevos sitios de *splicing* alterados en cáncer. Por esta razón, en este trabajo se ha empleado una estrategia distinta que trata de identificar posibles cambios de *splicing* antes de proceder al alineamiento al genoma de referencia. Dicha estrategia se basa en el empleo de k-mers, en el ensamble *de novo* de los k-mers diferenciales y en su posterior alineamiento al genoma de referencia para identificar a qué gen pertenecen. Los resultados obtenidos han permitido identificar k-mers correspondientes a genes cuya expresión es distinta entre casos control y casos con mutaciones en *SF3B1*. El hecho de que estos genes no hubieran aparecido en el análisis de expresión diferencial llevado a cabo en el primer objetivo se debe a que en el anterior análisis se había establecido un punto de corte de expresión mínima, mientras que en el análisis de k-mers no se

estableció este tipo de corte (tan solo que hubiera al menos 4 copias de un k-mer en los casos con mutaciones en *SF3B1*). Dado el bajo nivel de expresión de estos genes, el posible papel de los mismos como mediadores de SF3B1 es limitado, aunque será necesario llevar a cabo análisis en muestras adicionales para confirmar si es un evento generalizado, y poder diseñar experimentos para verificar su posible papel en la transformación tumoral. En cuanto a la identificación de nuevos fenómenos de *splicing*, el único gen con un *splicing* nuevo identificado fue *KCTD17*, que codifica una subunidad de un canal de potasio recientemente implicado en distonia.

Dado que esta aproximación no permitió identificar el sitio de *splicing* alternativo en 3' que se había descrito en *ATM* asociado a mutaciones puntuales en *SF3B1* (13), se estudió en detalle este gen, llevando a cabo un análisis puramente cuantitativo de este evento. Así, se pudo verificar con éxito que tanto el *splicing* canónico de *ATM* como el críptico ocurren con independencia del estado mutacional de *SF3B1* y que la alteración de este gen repercute significativamente en la frecuencia con la que se usa el sitio aceptor alternativo. En base a nuestros resultados, se puede añadir también que las diferencias en el proceso de *splicing* asociadas a mutaciones en *SF3B1* no son todo o nada sino que la alteración de este gen desencadena un efecto sutil que sólo se puede detectar mediante estudio cuantitativo. Este hallazgo está en sintonía con los resultados del *clustering* jerárquico y podría explicar porqué el análisis cualitativo de k-mers no fue decisivo a la hora de identificar nuevos eventos de *splicing* que dependan de la presencia de mutaciones en *SF3B1*. En conclusión, este estudio revela que no existen fenómenos de *splicing* aberrantes recurrentes provocados por mutaciones en *SF3B1*, y sugiere que estrategias basadas en análisis cuantitativos de k-mers pueden permitir la identificación de fenómenos de procesamiento del RNA favorecidos por la mutación en este factor de *splicing*.

7. CONCLUSIONES

Las conclusiones del estudio fueron las siguientes:

1. Las mutaciones en *SF3B1* provocan cambios significativos en los perfiles de expresión génica en LLC, cáncer de páncreas y cáncer de mama.
2. La mayoría de los genes con expresión alterada tanto en LLC como en cáncer de páncreas tienen un cambio de expresión asociado a mutaciones en *SF3B1* en el mismo sentido en ambos tipos tumorales, y en todos los casos hacia una expresión reprimida.
3. El análisis de datos de RNA-Seq mediante la utilización de k-mers permite la identificación de nuevos sitios de *splicing* alterados por mutaciones en *SF3B1*, incluyendo la activación de sitios crípticos 5' de *splicing* en el gen *KCTD17*.
4. El *splicing* críptico de *ATM* previamente descrito como asociado a mutaciones en *SF3B1*, no es específico de tumores con *SF3B1* mutado, sino que también se detecta en tumores sin mutaciones. Sin embargo, la alteración de este gen repercute significativamente en la frecuencia con la que se usa el sitio aceptor alternativo.
5. Los estudios futuros encaminados a identificar cambios en la maduración de RNAs causados por mutaciones en *SF3B1* deberían basarse en estrategias cuantitativas y no solo cualitativas.

8. BIBLIOGRAFÍA

1. Atlanta: American Cancer Society; 2015. American Cancer Society. Global Cancer Facts & Figures 3rd Edition. 2015.

2. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 9 de abril de 2009;458(7239):719-24.

3. Hanahan D, Weinberg RA. The Hallmarks of Cancer. Cell. 7 de enero de 2000;100(1):57-70.

4. (Chairperson) TJH, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. Nature. 15 de abril de 2010;464(7291):993-8.

5. Metzker ML. Sequencing technologies — the next generation. Nat Rev Genet. enero de 2010;11(1):31-46.

6. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 8 de marzo de 2007;446(7132):153-8.

7. McLendon R, Friedman A, Bigner D, Meir EGV, Brat DJ, Mastrogianakis GM, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 23 de octubre de 2008;455(7216):1061-8.

8. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 23 de octubre de 2008;455(7216):1069-75.

9. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science. 16 de noviembre de 2007;318(5853):1108-13.

10. Jones S, Zhang X, Parsons DW, Lin JC-H, Leary RJ, Angenendt P, et al. Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. Science. 26 de septiembre de 2008;321(5897):1801-6.

11. Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ, Angenendt P, et al. An Integrated Genomic Analysis of Human Glioblastoma Multiforme. Science. 26 de septiembre de 2008;321(5897):1807.

12. Puente XS, Pinyol M, Quesada V, Conde L, Ordóñez GR, Villamor N, et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature. 7 de julio de 2011;475(7354):101-5.

13. Ferreira PG, Jares P, Rico D, Gómez-López G, Martínez-Trillos A, Villamor N, et al. Transcriptome characterization by RNA sequencing identifies a major molecular and clinical subdivision in chronic lymphocytic leukemia. Genome Res. 2 de enero de 2014;24(2):212-26.

14. Quesada V, Conde L, Villamor N, Ordóñez GR, Jares P, Bassaganyas L, et al. Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nat Genet. enero de 2012;44(1):47-52.

15. Papaemmanuil E, Cazzola M, Boultwood J, Malcovati L, Vyas P, Bowen D, et al. Somatic SF3B1 Mutation in Myelodysplasia with Ring Sideroblasts. N Engl J Med. 13 de octubre de 2011;365(15):1384-95.

16. Network TCGA. Comprehensive molecular portraits of human breast tumours. Nature. 4 de octubre de 2012;490(7418):61-70.

17. Biankin AV, Waddell N, Kassahn KS, Gingras M-C, Muthuswamy LB, Johns AL, et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature. 15 de noviembre de 2012;491(7424):399-405.

18. Cazzola M, Rossi M, Malcovati L. Biologic and clinical significance of somatic mutations of SF3B1 in myeloid and lymphoid neoplasms. Blood. 10 de enero de 2013;121(2):260-9.

19. Dolatshad H, Pellagatti A, Fernandez-Mercado M, Yip BH, Malcovati L, Attwood M, et al. Disruption of SF3B1 results in deregulated expression and splicing of key genes and pathways in myelodysplastic syndrome hematopoietic stem and progenitor cells. Leukemia. mayo de 2015;29(5):1092-103.

20. Corrionero A, Miñana B, Valcárcel J. Reduced fidelity of branch point recognition and alternative splicing induced by the anti-tumor drug spliceostatin A. Genes Dev. 1 de marzo de 2011;25(5):445-59.

21. Nicolae M, Mangul S, Măndoiu II, Zelikovsky A. Estimation of alternative splicing isoform frequencies from RNA-Seq data. Algorithms Mol Biol AMB. 2011;6(1):9.

22. Jamison. Perl Programming for Biologists. John Wiley & Sons, Inc.; 2003.

23. Melsted P, Pritchard JK. Efficient counting of k -mers in DNA sequences using a bloom filter. BMC Bioinformatics. 10 de agosto de 2011;12(1):333.

24. Chikhi R, Limasset A, Jackman S, Simpson J, Medvedev P. On the representation of de Bruijn graphs. ArXiv14015383 Cs Q-Bio [Internet]. 21 de enero de 2014 [citado 9 de julio de 2015]; Recuperado a partir de: http://arxiv.org/abs/1401.5383

25. Kent WJ. BLAT—The BLAST-Like Alignment Tool. Genome Res. 4 de enero de 2002;12(4):656-64.

26. Tan Y-D, Xu H. A General Method for Accurate Estimation of False Discovery Rates in Identification of Differentially Expressed Genes. Bioinformatics. 14 de marzo de 2014;btu124.

27. Floratos A, Smith K, Ji Z, Watkinson J, Califano A. geWorkbench: an open source platform for integrative genomics. Bioinformatics. 15 de julio de 2010;26(14):1779-80.

28. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.

29. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. enero de 2009;37(1):1-13.

30. Balding DJ. A tutorial on statistical methods for population association studies. Nat Rev Genet. octubre de 2006;7(10):781-91.

APÉNDICE

A continuación se muestra, a modo de ejemplo, uno de los programas escritos en formato Perl para filtrar las matrices iniciales con el objetivo de seleccionar los genes con valor medio de expresión entre pacientes mayor o igual a 50. En concreto, éste fue escrito para el *set* de datos de leucemia linfática crónica:

```
$print="/home/angel/Blood_LabData/Output_Filtering50Probes.gWB.txt";
open FILE_print, ">$print";

open FILE,"$ARGV[0]";
while(<FILE>){
	chomp;
	$linea2=$_;

	if ($linea2=~/\"Probe_Set_ID\"/){
		print FILE_print "$linea2\n";
	}

	$suma=0;
	@GEN=split(/\t/,$linea2);
	for $t(1..$#GEN){
		$suma=$suma+$GEN[$t];
	}
	$promedio=$suma/($#GEN-1);
	if ($promedio<50){
		next;
	}
	else{
		print FILE_print "$linea2\n";
	}
}
close FILE;
close FILE_print;
```