UACM

Universidad Autónoma de la Ciudad de México

Nada humano me es ajeno

UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO

POSGRADO EN CIENCIAS GENÓMICAS

DESARROLLO DE META-ANÁLISIS DE DATOS COMPLEJOS EN BIOLOGÍA COMPUTACIONAL

T E S I S

QUE PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS GENÓMICAS

P R E S E N T A

BIOL. EXP. DANIEL ORTEGA BERNAL

DIRECTORA DE TESIS

DRA. CLAUDIA SELENE ZÁRATE GUERRA

MÉXICO, D.F. **ENERO, 2012**

UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

CA Coordinación Académica

SISTEMA BIBLIOTECARIO DE INFORMACIÓN Y DOCUMENTACIÓN

UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO
COORDINACIÓN ACADÉMICA

RESTRICCIONES DE USO PARA LAS TESIS DIGITALES

DERECHOS RESERVADOS©

COMITE TUTORIAL

DIRECTORA

Dra. Claudia Selene Zárate Guerra
Profesor Titular. Posgrado en Ciencias Genómicas
Universidad Autónoma de la Ciudad de México

ASESORES

Dra. Claudia Rangel Escareño
Jefe del Departamento de Genómica Computacional
Instituto Nacional de Medicina Genómica

Dr. Oscar Alberto Pérez González
Jefe del Laboratorio de Oncología Experimental
Instituto Nacional de Pediatría

Dr. Mauricio Castañón Arreola
Profesor Titular. Posgrado en Ciencias Genómicas
Universidad Autónoma de la Ciudad de México

Dr. Leon P. Martínez Castilla
Profesor Titular. Facultad de Química
Universidad Nacional Autónoma de México

El presente proyecto de investigación se realizo gracias al apoyo académico que me otorgó la Universidad Autónoma de la Ciudad de México, a la cual agradezco.

El presente trabajo se realizó en el laboratorio de Bioinformatica del Posgrado en Ciencias Genómicas de la Universidad Autónoma de la Ciudad de México bajo la dirección de la Dra. Claudia Selene Zárate Guerra.

EL presente proyecto se llevo a cabo gracias al apoyo económico del Instituto de Ciencia y Tecnología (ICyT DF) mediante el proyecto PICDS08-17.

Para María Teresa, Alfredo y toda mi familia.
Gracias por su apoyo

Gracias a la Dra. Claudia Selene Zárate Guerra
que me permitió implementar la creatividad
en la generación y desarrollo de este proyecto
y darme libertades para desarrollarme de forma
transdisciplinaria.
“En la investigación científica algo más importante que el conocimiento
es la imaginación.”
Albert Einstein

“Mandar sobre muchos es igual que mandar sobre pocos.
Sólo hay que organizarse”
Sun Tzu. El Arte de la Guerra

Las respuestas del gran cumulo de información genómica
está en la organización de ella misma.

INDICE GENERAL

LISTA DE FIGURAS

LISTA DE TABLAS

1 INTRODUCCIÓN

1.1 Antecedentes generales del tema

1.1.1 Genómica

La genética es a la vez la ciencia de cómo las células expresan y transmiten la información del DNA y una herramienta para estudiar procesos biológicos relacionados con las funciones celulares, así mismo la genómica es la ciencia de la estructura y evolución de los genomas y una herramienta para aprender sobre las funciones de los genes. Genética y genómica tienen escalas y enfoques diferentes; mientras que la genética emplea mutantes para identificar y estudiar los pocos genes que controlan un fenotipo en particular, la genómica recoge datos sobre todos los genes de un organismo. La genómica incluye los métodos para recopilar y analizar datos completos acerca de los genes, incluida su secuencia y la abundancia de mRNA, así como las propiedades de las proteínas para las que codifican estos genes (proteómica) (1).

El proyecto del genoma humano (PGH) fue uno de los grandes avances de la genómica, en los inicios del PGH (1991) los retos informáticos eran proveer de una conectividad funcional, flexible y barata para toda la comunidad científica, establecer la estandarización y compatibilidad, no sólo de los protocolos de comunicación, sino de los formatos en que se presentan los datos, y permitir la conectividad entre las diversas bases de datos, lo que requirió de la interacción de biólogos con conocimientos en informática, matemáticos y programadores (2). En el periodo 1993 – 1998 eran construir un mapa genético de alta resolución del genoma humano, desarrollar las capacidades de colectar, almacenar, distribuir y analizar los datos producidos y crear la tecnología apropiada para alcanzar estos objetivos (3). Para el periodo comprendido de 1998 – 2003 dentro de los retos bioinformáticos para el PGH destacaba el desarrollo de herramientas analíticas específicas para la gran cantidad de información generada (4).

1.1.2 Acumulación de información y análisis de datos

Las investigaciones realizadas con herramientas genómicas dan resultados que a menudo contienen mucha más información que la anticipada por los investigadores que la han generado. Este enorme volumen de información que se colecta en las bases de datos permite que se realicen nuevos tipos de análisis, bajo métodos distintos a los tradicionales (5), como el meta-análisis o mediante el análisis en conjunto de múltiples plataformas genómicas. Además de las secuencias de los genomas, otras herramientas como los microarreglos y la secuenciación masiva se han convertido en herramientas importantes para comprender los sistemas biológicos. Como resultado de esto, se han desarrollado nuevos métodos para visualizar la información presente en las matrices masivas de mediciones que estas tecnologías producen (6). Por ejemplo, para sistematizar la información producida por los estudios de microarreglos se han generado bases de datos que compilan los archivos generados en este tipo de experimentos, como Stanford Microarray Database (SMD) (7), ArrayExpress (8) o Gene Expression Omnibus (GEO) (9). Este tipo de recursos sirven para dos propósitos, como un archivo de datos, que permite a otros investigadores confirmar los resultados publicados, así como permitir un nuevo análisis de los datos, que va más allá de lo previsto en el estudio original. Un nuevo análisis podría consistir en volver a analizar los datos que se utilizaron para un estudio determinado pero con una pregunta diferente a contestar, o un meta-análisis que consiste en analizar de manera conjunta datos generados por separado (10). El uso en conjunto de los datos permite dar una mayor profundidad a la información con la que se cuenta, y la demanda para utilizar eficazmente estos conjuntos de datos ha aumentado, pues permite que se utilicen adicionalmente para el análisis y verificación de resultados de otros estudios (11). El meta-análisis no está limitado a un sólo tipo de datos, es posible y recomendable utilizar datos diversos, ya sean los provenientes de microarreglos, proteomas, interactomas, etc., en conjunto.

La continua acumulación de datos, desde secuencias de genomas completos, meta-genomas, ensayos de abundancia de mRNA o proteínas, etc., se traduce en la necesidad de desarrollar sistemas automáticos que permitan reunir la información

pertinente a una pregunta biológica dada. Hay, pues, una brecha creciente entre la generación de información y la capacidad de procesar y analizar los datos resultantes (12). En este sentido la biología computacional, puede definirse como el desarrollo y aplicación de métodos teóricos y de análisis, este enfoque es una herramienta que cada vez es más necesaria en la investigación biológica, donde uno de sus principales objetivos es el desarrollo de aplicaciones útiles que permitan comprender la información biológica de mejor manera (13).

1.1.3 Microarreglos

Entre las tecnologías de alto rendimiento que generan gran cantidad de datos se encuentran los microarreglos de RNAm, que es una metodología que hace miles de mediciones a nivel del transcriptoma. Entre los tipos de microarreglos se encuentran los que se basan en cDNA, que fueron empleados inicialmente por Schena *et al* (1995), dicha tecnología se basa en colocar de manera ordenada en una superficie sólida miles de marcadores de secuencias expresadas (ESTs). El RNAm se cuantifica de forma relativa por una hibridación competitiva a partir de cDNA proveniente de la retrotranscripción de RNAm a partir de 2 muestras diferentes (por ejemplo, células normales y células neoplásicas), cada una de estas muestras marcadas con dos fluoroforos (por ejemplo, Cy3 para el color verde y Cy5 para el color rojo) (14).

Otra tecnología para generar microarreglos utiliza la síntesis de oligonucleótidos por fotolitografía (Affymetrix, Santa Clara, CA). Esta metodología permite la síntesis *in situ de* cientos de miles de oligonucleotidos/cm^2 (con una longitud aproximada de 25 meros). Mientras los microarreglos de cDNA se caracterizan por el uso de una secuencia relativamente larga (>300 bases) para cada gen, los microarreglos de Affymetrix tienen hasta 20 oligonucleótidos cortos para cada gen o EST, esta tecnología emplea conjuntos de sondas (*probeset*) que representan al mismo gen (15). A diferencia de las sondas impresas en los microarreglos de cDNA, los microarreglos de Affymetrix miden la cantidad de mensajeros por grupos de sondas, que permiten la generación de valores absolutos de intensidad, donde para contrastar dos condiciones es necesario utilizar un microarreglo para cada muestra. Para evaluar la especificidad de hibridación

del blanco de cada oligonucleótidos se asocian sondas (PM: *Perfect Match*) a un control negativo (MM: *MisMatch*), donde su secuencia difiere en el nucleótido central con respecto a la sonda PM (Figura 1). La evaluación de las señales de hibridación de las sondas PM y MM da indicio de la especificidad de los PM para identificar una sonda en específico dentro del *probeset*, donde una señal fuerte en la sonda MM es una advertencia de la presencia de hibridación inespecífica, sin embargo Neaf *et al* (2001) demostró que esta sonda par (PM-MM) genera lecturas ambiguas, ya que en más del 30 % de los MM reportan una señal mayor que la de los PM (16).

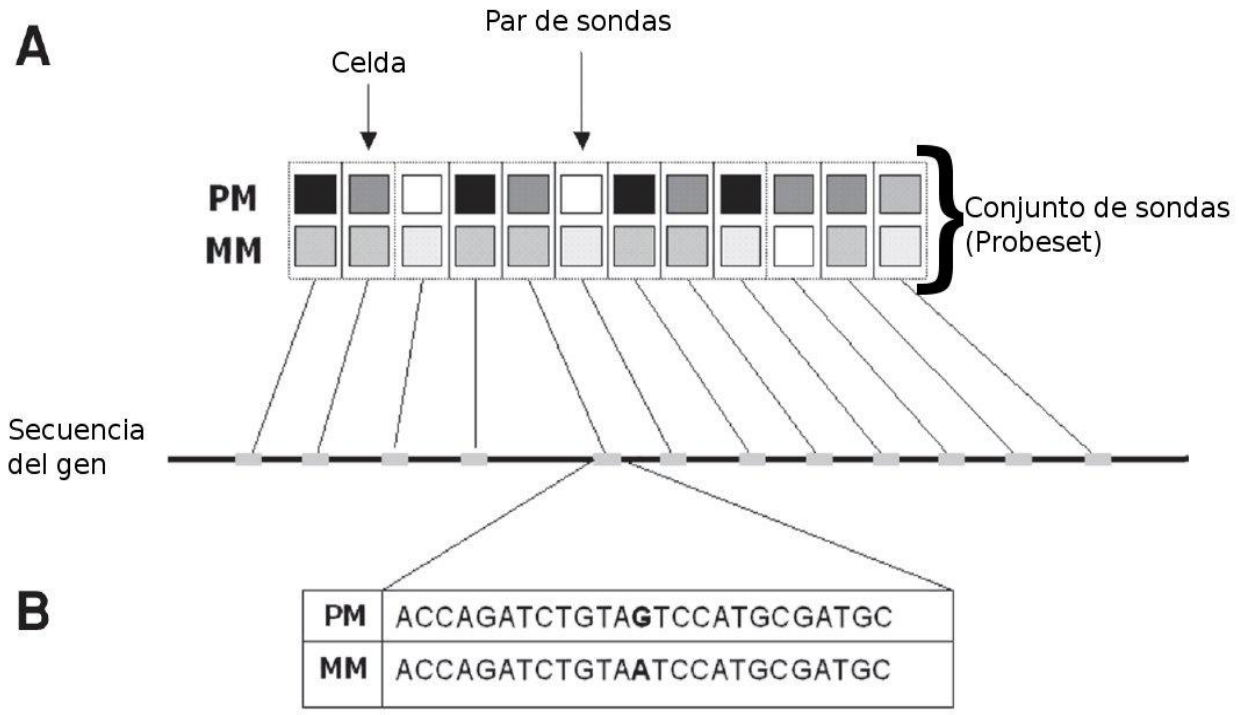


Figura 1. Esquema de un probeset. Cada gen o EST está representado por uno o varios conjuntos de sondas dispersos en el microarreglo. A) Cada conjunto de sondas está formado por un máximo de 20 pares de oligonucleótidos (sonda par (PM-MM)), y sus blancos están distribuidos a lo largo del gen de interés. B) Cada conjunto de sondas están formados por dos oligonucleótidos de la misma longitud, *perfect match* (PM) y *mismatch* (MM). La sonda PM tiene una secuencia correspondiente a la secuencia blanco; la sonda MM tiene la misma secuencia que la sonda PM, pero con un nucleótido central diferente (Tomada y modificada de 16).

1.1.3.1 Análisis de microarreglos

En general el proceso para la hibridación de microarreglos es largo e implica muchos pasos, además, las condiciones experimentales (degradación del RNA, presencia de inhibidores de la retrotranscriptasa, el rendimiento de la purificación del RNA, etc.)

pueden afectar fuertemente la hibridación en los microarreglos. Estos factores son fuente de error multiplicativo y afectan en gran medida la determinación de los niveles de expresión verdaderos (17), sobretodo en los genes que se expresan moderadamente (18). Por lo anterior, pre-procesar los datos crudos es esencial para comparar la matriz de datos de un microarreglo con otros, en los microarreglos de Affymetrix este proceso consiste en tres pasos, corrección de fondo, normalización e integración. La corrección de fondo consiste en eliminar el ruido de fondo inespecífico y las fluctuaciones locales del nivel de señal de los microarreglos individualmente (19). El objetivo de la normalización es eliminar la mayor cantidad de la variación sistemática (técnica y biológica) como sea posible, dejando intacta la variación biológica que se planea medir. El último paso en el pre-procesamiento es la integración, que consiste en la combinación de intensidades de las múltiples sondas de cada *probeset* para producir un valor de expresión (20).

El verdadero poder de análisis de microarreglos no es a partir del análisis de los archivos obtenidos de manera aislada, sino más bien, por el análisis de muchos archivos para identificar patrones de co-expresión (21) o expresión diferencial bajo dos condiciones diferentes (22). Bajo el enfoque de búsqueda de patrones de co-expresión el objetivo es identificar los genes que muestran patrones similares de expresión, por ejemplo, se puede usar el análisis de agrupamiento (*cluster analysis*) (6), para obtener un agrupamiento jerárquico, éste tiene la ventaja de ser simple y que el resultado se puede visualizar fácilmente mediante mapas de calor. Otro algoritmo es k-means (23) que requiere de conocimiento previo acerca del número de grupos en los que se dividen los datos y no produce grupos jerárquicos. Una tercer opción son los mapas auto-organizables (*Self Organizing Maps*, SOM) (24) que utiliza una red neuronal que lleva a cabo la agrupación por división, SOM asigna genes a una serie de grupos con base en la similitud de los vectores de expresión para los vectores de referencia que se definen para cada grupo. En cuanto al enfoque de expresión diferencial, se han generado numerosos métodos, ejemplos de dichos métodos son los análisis de varianza (25), SAM (*Significance Analysis of Microarray*) (26), LIMMA (Limma: *Linear Models for Microarray*) (27), entre otros.

1.1.4 Bases de Datos

Galperin *et al* (2011) reporta 1330 bases de datos de biología molecular, que van desde organismos específicos e información puntual hasta grandes bases de datos que colectan información de múltiples organismos e información diversa (28). Las bases de datos, los métodos de análisis computacional y la minería de datos son cada vez más necesarios en la investigación biológica, ya que proporcionan un acceso fácil a una gran cantidad de información biológica. Mientras que bases de datos individuales cubren información importante en ciertas áreas del conocimiento biológico, el uso integral de los datos públicos puede ser obstaculizado por el gran número y la fragmentación de dichas bases (29).

En el contexto de las bases de datos de rutas de señalización o redes, de acuerdo con Pathguide (12) estas bases se pueden agrupar en ocho grandes categorías en función del tipo de datos que contienen, el formato en el que se encuentran y el enfoque biológico (12) (Tabla 1).

1) Las bases de datos de interacción proteína-proteína reportan las interacciones de manera pareada, o bien, las interacciones entre una proteína y un complejo.
2) Las bases de datos de rutas metabólicas generalmente almacenan una serie de reacciones bioquímicas de las rutas involucradas en la conversión de metabolitos.
3) Las bases de datos de rutas de señalización reúnen grupos de interacciones moleculares y modificaciones químicas (como modificaciones post-traduccionales) implicados en rutas regulatorias.
4) Las redes de regulación génica almacenan los factores de transcripción y los genes que regulan.
5) Las bases de datos de interacción proteína-compuesto almacenan información de la interacción de proteínas con diversos ligandos.
6) Las bases de datos de rutas genéticas están compuestas de interacciones genéticas, como epístasis o letalidad sintética.
7) Las bases de datos de diagramas de rutas generalmente almacenan hipervínculos a imágenes de rutas.

8) El último tipo de bases de datos son las de secuencias de proteínas que almacenan la información secundaria.

Tabla 1. Estadísticas de Pathguide sobre bases de datos de información genómica.

Categoría	**Numero de bases de datos**
Interacción proteína-proteína	132
Rutas metabólicas	73
Rutas de señalización	62
Redes de regulación génica	51
Interacción proteína-compuesto	43
Redes de interacción génica	9
Diagramas de rutas	38
Secuencias de proteínas	20
Otras	14

Las categorías son aproximadas, una base de datos puede estar en varias categorías si contiene diversos tipos de datos.

1.1.5 Meta-análisis y minería de datos

1.1.5.1 Meta-análisis

El meta-análisis es un método de análisis integrativo que se define como una síntesis o revisión de los resultados a partir de datos que se obtienen de forma independiente pero que tienen una relación (30). El poder del meta-análisis puede ser mayor que el de cada uno de los análisis individuales de los que parten, esto amplifica la capacidad del análisis para descubrir información, por ejemplo, un gen que en un meta-análisis es identificado con alteraciones significativas en sus niveles de expresión, pudo no haber sido significativo en ninguno de los estudios individuales. A esto se le denomina "descubrimiento por integración-dirigida" (IDR del inglés) (31). El meta-análisis también puede ser importante cuando los estudios tienen un conflicto en sus conclusiones, ya que permite estimar un efecto promedio o destacar una variación sutil pero importante (30, 32).

A nivel de microarreglos, los estudios de Choi *et al* (2003) fueron de los primeros en aplicar meta-análisis. Para probar su enfoque utilizaron los datos provenientes de dos tipos de cáncer; por un lado utilizaron microarreglos de cDNA un hepatocarcinoma generados independientemente, pero bajo las mismas condiciones experimentales (meta-análisis con una plataforma). Por otro lado, emplearon microarreglos con muestras de cáncer de próstata, usando dos conjuntos de datos basados en microarreglos de cDNA y dos basados en microarreglos de oligonucleotidos (meta-análisis con dos plataformas diferentes). El método propuesto de meta-análisis se sustenta en que los cambios en la expresión génica se aprecian como "efecto del tamaño de la muestra", esto es que la lectura de la expresión de un determinado gen estará en función de la cantidad de mediciones de microarreglos del que se haya partido. Este trabajo demostró que la integración de datos permite descubrimiento de cambios pequeños en la expresión pero consistentes, con una mayor sensibilidad y fiabilidad, mientras que en análisis individuales estos cambios pequeños se apreciarían como falsos negativos. Los investigadores, por último, plantean el futuro del meta-análisis el cual debería incluir múltiples variables, y tomar en cuenta las interacciones entre los genes, así como sus co-relaciones (31). El enfoque propuesto por Choi *et al* (2003) fue retomado en GenMeta, un paquete de Bioconductor que puede aplicarse a datos de diferentes plataformas, asumiendo que estos datos tratan de responder la misma pregunta o que los aspectos de los experimentos son lo suficientemente similares que se pueda esperar realizar una mejor inferencia del conjunto de datos que de los datos de forma individual (33).

Otro enfoque para realizar meta-análisis con conjuntos de datos de microarreglos lo desarrollaron Parmigiani *et al* (2004) basado en el uso de correlaciones de integración (CI), esto es que se calculan todos los pares de correlaciones posibles a través de las muestras dentro de los proyectos individuales (cálculos dentro de la misma plataforma de microarreglos), de tal forma que al examinar si estas correlaciones son comparables entre los diferentes proyectos (cálculos en diferentes plataformas de microarreglos), se puede cuantificar la reproducibilidad de los resultados sin depender de la comparación directa de expresión entre las plataformas. Los autores compararon las asociaciones de los niveles de expresión de genes para el diagnostico diferencial de carcinoma de

células escamosas frente a un adenocarcinoma. El método propuesto mostró una correlación de 0.847 para todos los genes, tras solo elegir aquellos genes que tenían un alto grado de CI (50 % de los genes) se llegaba a una correlación del 0.921, donde solo pocos genes mostraron una reproducibilidad baja. Los problemas que mencionan los autores, con respecto a la discordancia de sus resultados, son la variación de la calidad técnica de las sondas y la impresión que podrían no reflejar las diferencias reales de expresión génica, sino mas bien las diferencias tecnológicas entre las plataformas; las diferencias biológicas entre las muestras y poblaciones consideradas en los tres estudios empleados por los autores; y por último una tercera fuente de variabilidad puede ser la variación aleatoria de las técnicas (34).

El método Rango de los Productos (RP) (35) es un método no paramétrico que se basa en las veces de cambio en la expresión como un método de selección para comparar y clasificar los genes dentro de cada conjunto de datos. Estos rangos (veces de cambio en los diversos conjuntos de datos) se combinan a continuación para producir una puntuación global de los genes a través de conjuntos de datos, obteniendo una lista de genes clasificados. Este meta-análisis se basa en la identificación de genes expresados diferencialmente y no procede de un modelo estadístico sofisticado, sino de un análisis por razonamiento biológico. Está sustentado en el cálculo de los rangos de veces de cambio entre los conjuntos de datos y por lo tanto por las replicas de los experimentos, por lo que es rápido y sencillo. Además, proporciona una sencilla y rigurosa determinación estadística de la importancia de cada gen, y permite un control flexible de la tasa de falsos positivos. Los resultados de los autores demuestran que este enfoque simple supera a herramientas como SAM, debido a que RP hace hipótesis relativamente débiles sobre los datos, esperando variaciones más o menos iguales para todos los genes. Por el contrario incluso los supuestos débiles de otras estadísticas no paramétricas pueden ser demasiado fuertes para los datos de microarreglos. Por otro lado, sus resultados confirmaron su enfoque basado en “intuición biológica” de que la regulación genética significativa es un proceso de tipo interruptor, de modo que los cambios son siempre grandes, donde cambios pequeños pueden tener significancia estadística, pero rara vez biológica. Este enfoque trabaja con datos muy pequeños, una ventaja más en comparación con pruebas que estiman la varianza, pues RP no se basa

en estimaciones de varianza para cada gen. Dichos resultados demostraron superioridad en pruebas con un número alto de replicas (11 ensayos de leucemias) en comparación con un análisis SAM (26). Este método además se puede usar con datos proteómicos y de metabolómica (35). En la figura 2 se muestra el comparativo de los diversos métodos de meta-análisis mencionados tanto bajo un mismo tipo de microarreglo (solo oligonucleótidos), como un meta-análisis con diferentes tipos de microarreglos (cDNA y oligonucleótidos), donde se aprecia que el meta-análisis basado en un solo tipo de microarreglo (de oligonucleótidos) tiende a tener tasas de error menor (con excepción de GenMeta) en comparación con un meta-análisis que incluya diversos tipos de microarreglos (en el estudio comparativo de Campain y Yang (2010) mezclando sólo microarreglos de oligonucleótidos y de cDNA).

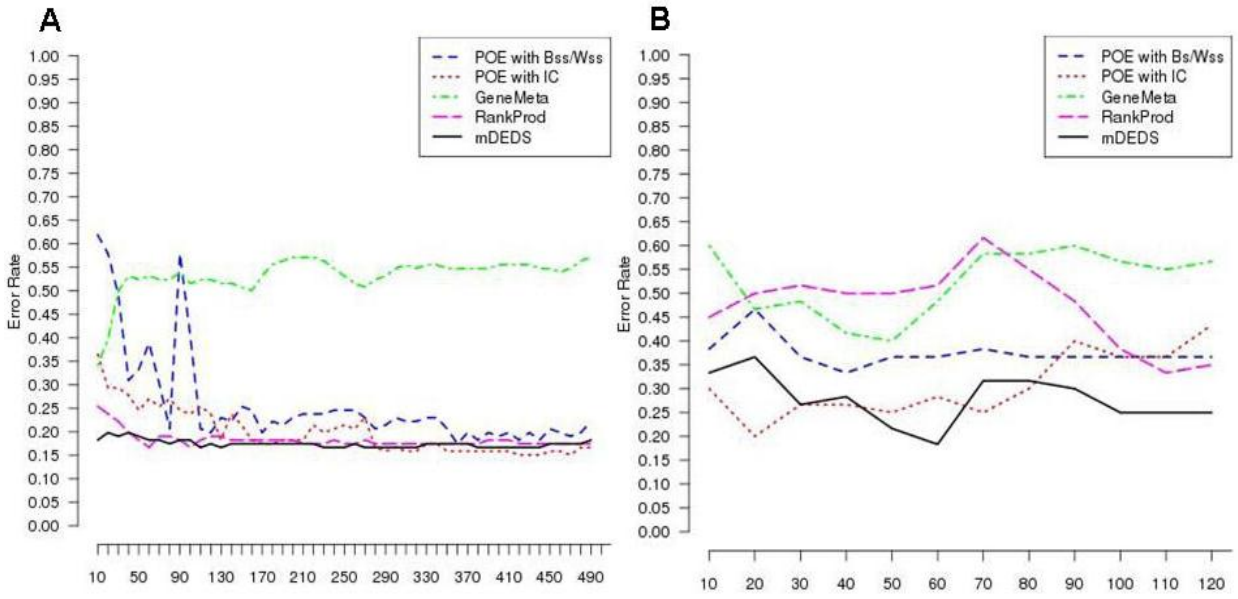


Figura 2. Gráficas comparativas de las metodologías de meta-análisis. A) Comparativo de las metodologías basado en microarreglos de la misma plataforma en un modelo en cáncer de mama. B) Comparativo de las metodologías basado en microarreglos de diferentes plataformas en linfoma. GenMeta representa al método de Choi *et al*, POE *with* IC equivale al meta-análisis IC y RankProd equivale al método RP (imagen tomada de 11).

1.1.5.2 Minería de datos

La minería de datos es utilizada a menudo para el descubrimiento de conocimiento y es uno de los subcampos más importantes en el manejo del información. La minería de datos tiene como objetivo analizar un conjunto de datos dado, con el fin de identificar

patrones potencialmente útiles (37) (Figura 3). Cuando se conjugan diversos tipos de datos se habla de minería a gran escala (Figura 4) (38). Siendo este enfoque relativamente nuevo dentro de las ciencias genómicas, aún se están concretando métodos para responder determinadas preguntas biológicas.

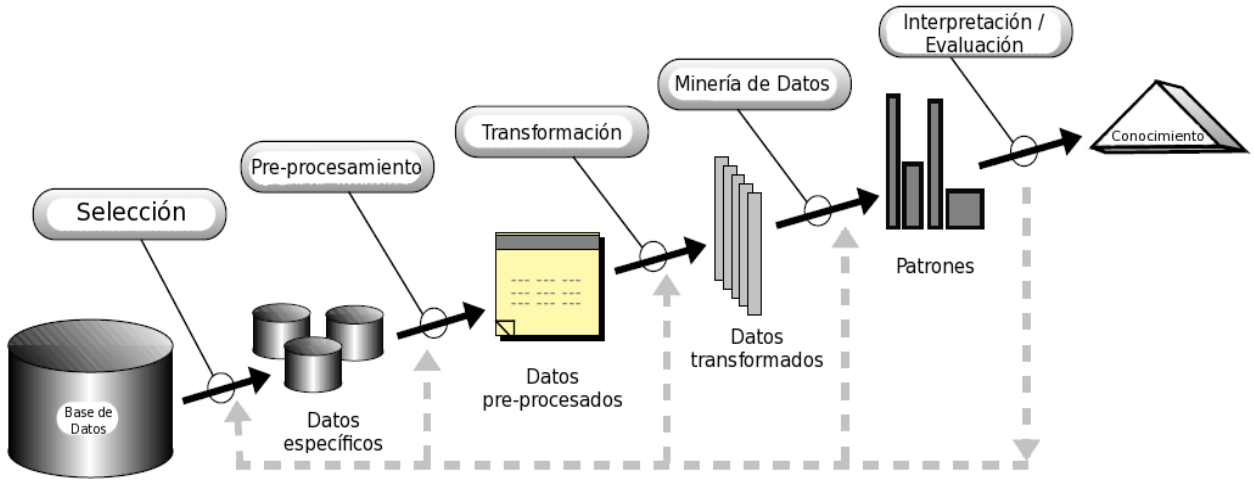


Figura 3. Pasos que componen el proceso de Descubrimiento de Conocimiento en Bases de Datos. (Tomada y modificada 37).

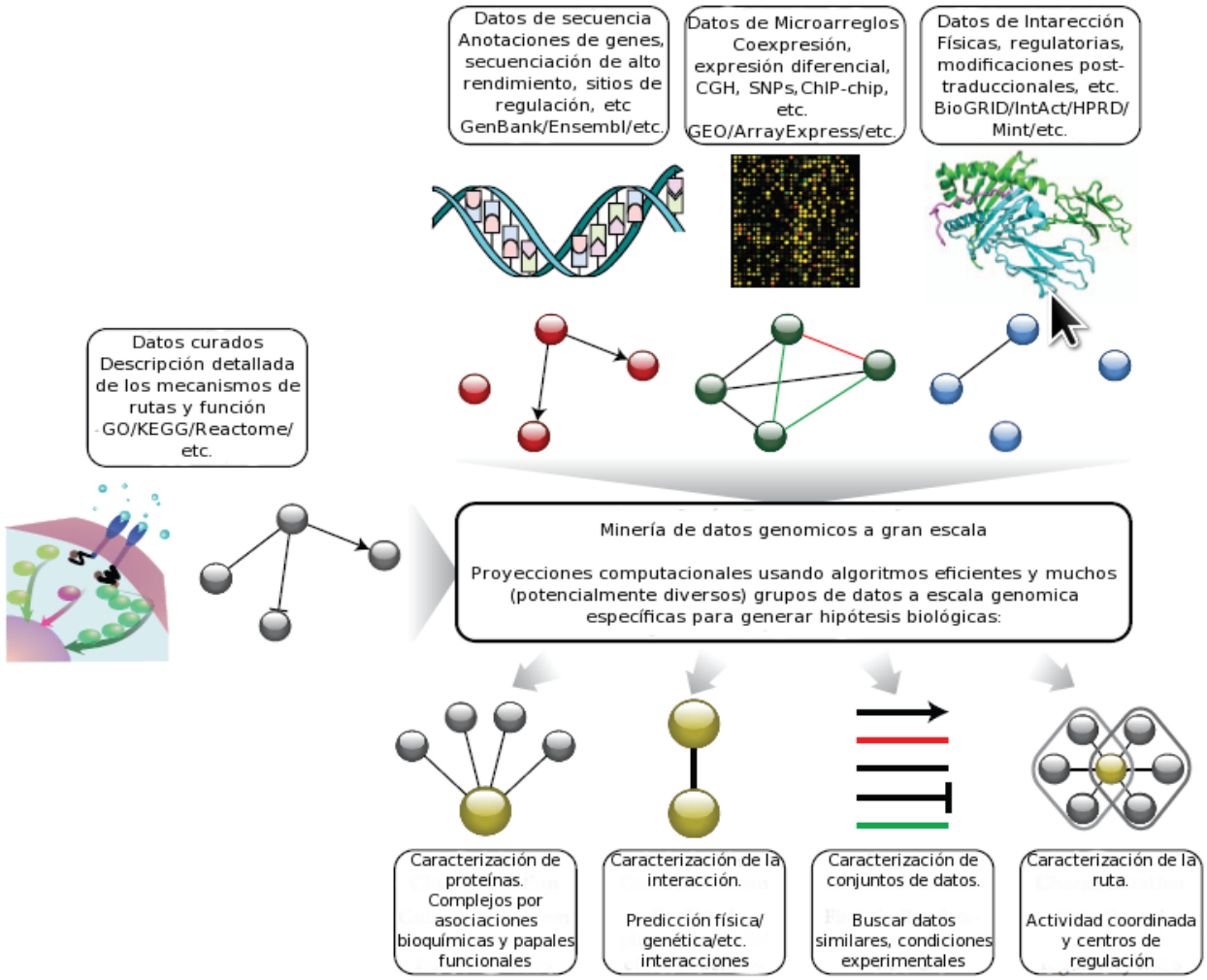


Figura 4. Estructura de la minería de datos genómicos a gran estala. En la figura se esquematizan las posibles entradas, el algoritmo que el investigador desarrolle o siga y por último las posibles salidas en función lo que busque el investigador (Tomada y modificada de 38).

La minería de datos aplicada a las ciencias genómica tiene múltiples retos:

1. Ya que la información (tipos de plataformas genómicas) que se desea comparar es muy diversa, así mismo el análisis de los datos genómicos de alto rendimiento tienen como resultado un gran número de resultados, donde muchos de ellos son falsos positivos (39).

2. El segundo reto son las grandes dimensiones de datos biológicos que se obtienen de los estudios genómicos (40).

3. El tercer desafío son los tamaños de muestra relativamente limitados,

comparados con las preguntas que se plantea resolver (39).

4. La limitación computacional es el cuarto reto, pues a menudo es prohibitivo resolver problemas de minería de datos genómicos utilizando una estrategia de búsqueda exhaustiva (41).

5. El quinto desafío es el ruido de los datos biológicos que se obtienen a partir de herramientas de alto rendimiento, ya que la información biológica y las señales de interés se observan a menudo junto con muchos otros factores al azar o de confusión.

6. Por último, esta el reto de la integración de datos genómicos con datos biológicos, como la función de los genes, los fenotipos observados y/o los parámetros clínicos del paciente (si es el caso) (5).

1.2 Antecedentes particulares

1.2.1 Minería de datos genómicos a gran escala

La minería de datos genómicos a gran escala cuenta con tres grandes categorías (38):

1. Las aplicaciones vía Internet que implican un sólo tipo de plataforma genómica.

2. Las aplicaciones de programación sofisticadas que permiten consultas de cómputo de diversas plataformas genómicas.

3. Las herramientas implementadas por el investigador que se basan en la obtención y el procesamiento manual de datos desde las bases públicas.

En la categoría de las aplicaciones vía Internet se incluyen interfaces en línea, los cuales sólo toman el análisis de plataformas genómicas individuales (por ejemplo KEGG (42) o Reactome (61)). Un ejemplo de este tipo de herramientas, pero que integra diversas fuentes de información es la aplicación STRING *(Search Tool for the*

Retrieval of Interacting Genes/Proteins), un recurso de base de datos dedicada a predecir interacciones proteína-proteína, incluyendo interacciones físicas o funcionales. STRING integra la información de numerosas fuentes, incluyendo bases de datos experimentales, métodos computacionales de predicción y publicaciones (43), otro ejemplo que integra diversos datos es BioMart la cual cuenta con un amplio repertorio de anotaciones de secuencia (44).

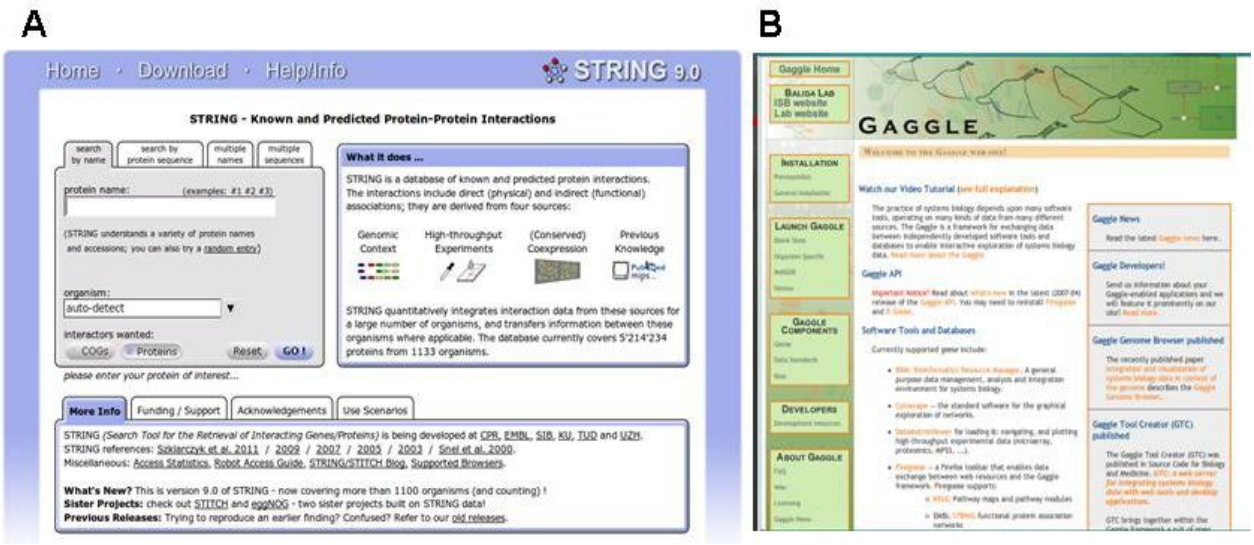


Figura 5. Programas de minería de datos vía Internet. A) Pantalla inicial del STRING. B) Pantalla inicial de Gaggle.

Las aplicaciones de programación sofisticadas de minería de datos a gran escala realiza consultas de los resultados de diversas bases de datos públicas y plataformas genómicas. Esto agrega un nivel de complejidad, ya que se debe decidir sobre el análisis posterior que se realizará, adecuado para los datos recuperados. La búsqueda se realiza de forma remota. Estos protocolos proporcionan una forma de plantear preguntas sofisticadas para un grupo de datos, dejando a examinar sólo los resultados finales. Un ejemplo de este enfoque es Gaggle, donde se integran diversas bases de datos (por ejemplo, KEGG, BioCyc), así como software (Cytoscape, String, DataMatrixViewer, DAVID, R y TIGR Microarray Expression Viewer) (45).

Tabla 2. Comparativo de diversas aplicaciones para minería de datos genómicos.

Características	**String (43)**	**BioMart (44)**	**DAVID (46, 47)**	**Gaggle (45)**
Bibliografía	X	X	X	X
Bases de datos	X	X	X	X
Rutas de señalización			X	X
Software				X
Uso de otras herramientas de minería de datos				X

El mayor nivel de flexibilidad en la minería de datos biológicos a gran escala lo da el tratamiento manual de los datos experimentales (herramientas implementadas por el investigador), y a su vez conlleva el mayor nivel de dedicación de tiempo y recursos computacionales. Este enfoque es actualmente una de las pocas maneras en que las que se pueden ejecutar las consultas multi-entradas. Ejemplo de ello es el trabajo realizado por Zhu *et al* (2008), donde realizaron un análisis integrador que construyó redes con más del 50% del genoma de *S. cerevisiae,* incorporando diferentes tipos de datos moleculares como genotipos, perfiles de expresión, sitios de unión a factores de transcripción, así como datos de interacción proteína-proteína (48).

1.3 Estado de arte

Como se menciono anteriormente la minería de datos genómicos a gran escala se divide en tres grandes categorías (las aplicaciones vía Internet, las aplicaciones de programación sofisticadas y las herramientas implementadas por el investigador). Una de las aplicaciones vía Internet más novedosas es BABELOMICS (49), una herramienta para integrar y analizar diferentes tipos de datos genómicos, dicha aplicación incluye métodos para el análisis de datos de expresión diferencial, ensayos de genotipificación, donde los resultados pueden integrarse a un análisis funcional mediante minería de datos con múltiples plataformas genomicas. BABELOMICS cuenta con diferentes métodos de enriquecimiento funcional o de enriquecimiento en conjunto, cuenta con fuentes de información de tipo biológica, funcional (GO, KEGG, Biocarta, Reactome, etc.), regulatoria (Transfac, Jaspar, ORegAnno, miRNAs, etc.), así como módulos de

minería de texto y de interacción proteína-proteína (Figura 6). A nivel de minería de datos a gran escala de tipo herramientas implementadas por el investigador se puede decir que hay una tendencia modesta a la alza (siendo un área relativamente nueva) a hacer investigación genómica a gran escala, como lo muestra la figura 7.

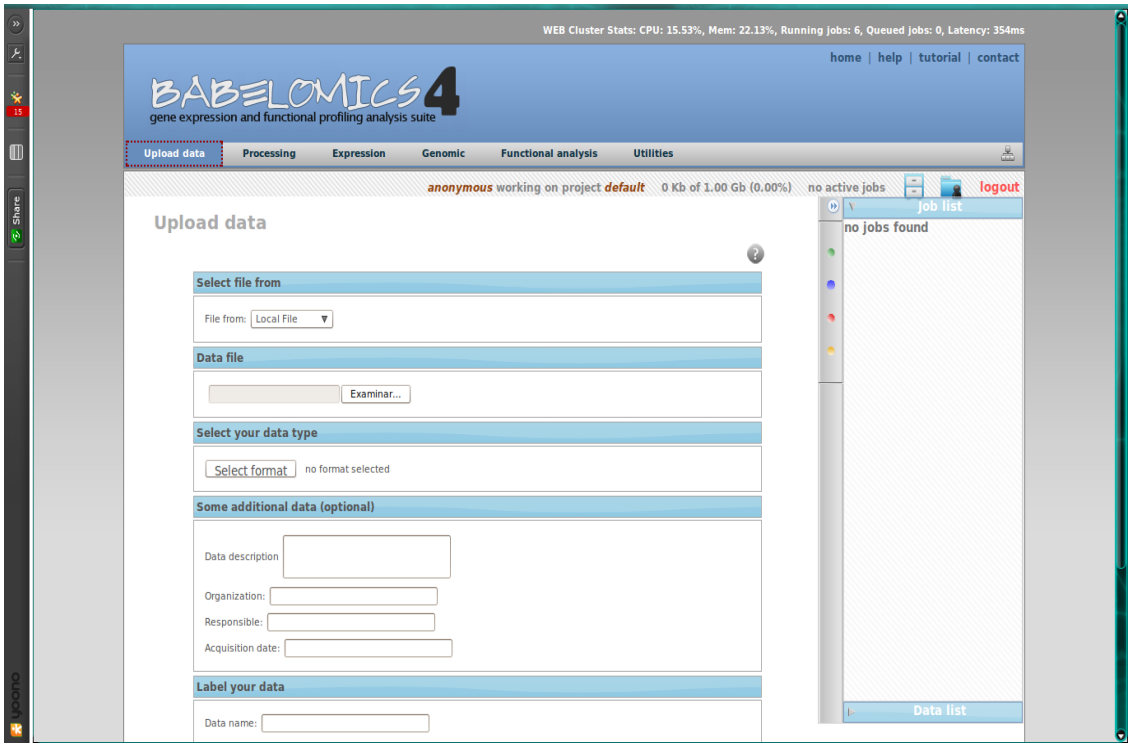


Figura 6. Pantalla inicial del sitio Babelomics. En ella se muestran las pestañas para acceder a cada uno de los análisis que se pueden realizar. A) Upload data: Microarreglos, listas de IDs, Anotación. B) Procesing: Criterios de pre-procesamiento. C) Expression: Tipo de análisis de expresión diferencial, agrupamientos, etc. D) Genomic: tipo de análisis (de asociación o estratificación). E) Funtional analysis: Análisis de enriquecimiento simple, por grupo, anotaciones funcionales. F) Utiilities: Diversos gráficos (Clustering, PCA, MA plot, etc).

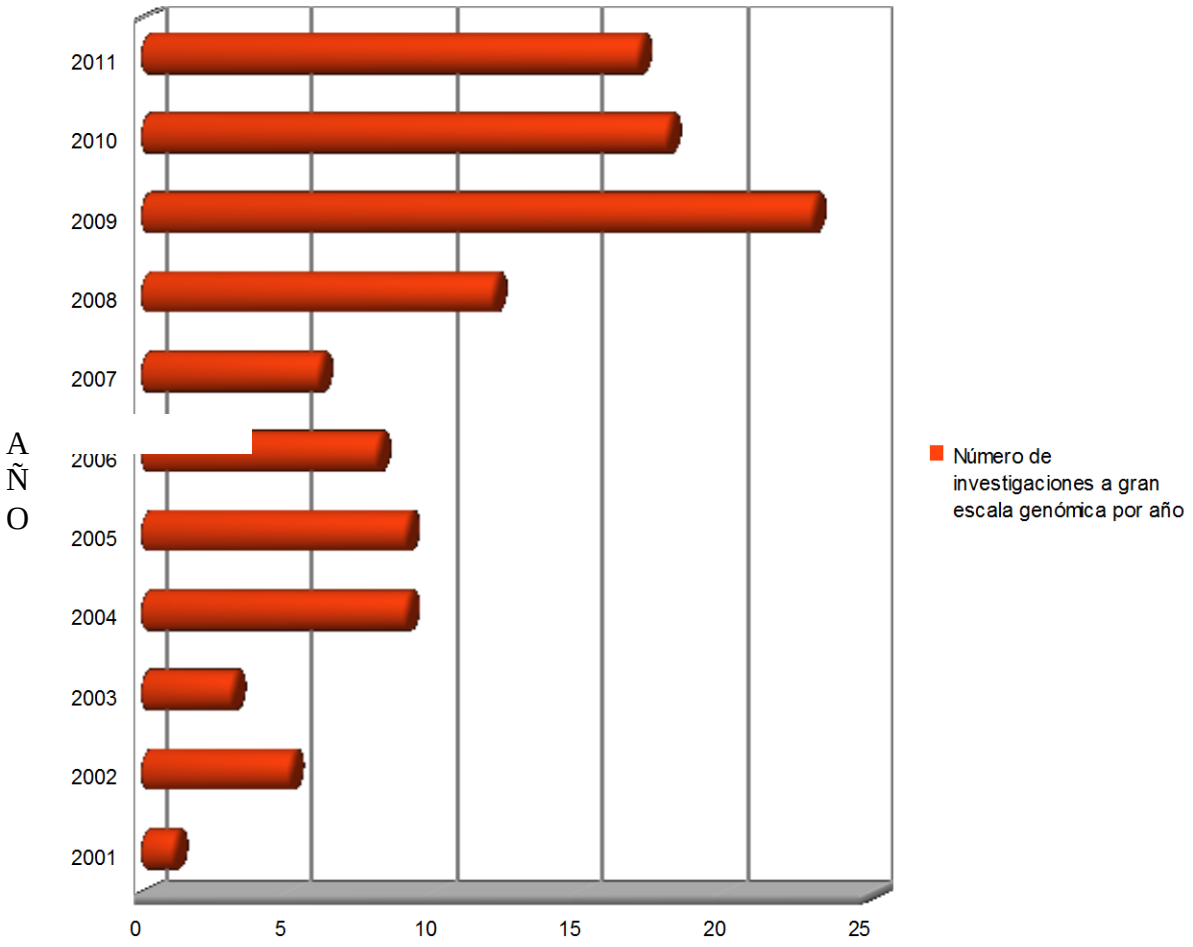


Figura 7. Investigación a gran escala genómica. Cantidad de investigación que se ha realizado a través del periodo 2001 – 2011, donde se aprecia una tendencia modesta a que investigaciones de este tipo continúen aumentando (resultados obtenidos de una búsqueda en PubMed con el termino de investigaciones similares a "*integrating large scale functional genomics data*")**.**

1.4 Modelo de estudio

Mediante la integración de meta-análisis en microarreglos con minería de datos genómica a gran escala se pretende generar resultados enriquecidos, para dicho fin es necesario contar con un modelo de estudio que este ampliamente investigado, de tal manera que se cuente con literatura que pueda respaldar los resultados encontrados, así como ayudar a la generación de hipótesis. En este sentido se tomó como modelo de estudio a las leucemias agudas linfoblasticas pediátricas (LALp), dicha neoplasia se encuentra ampliamente estudiada por la comunidad científica lo cual coadyuvará a darle un contexto adecuado a los resultados encontrados.

1.4.1 Leucemias, factores genéticos, biología molecular y clasificación

De forma normal la médula ósea produce células madre sanguíneas (células inmaduras) que con el tiempo se diferencian en células sanguíneas maduras. Las células madre sanguíneas se pueden convertir ya sea en una célula madre precursoras de las línea mieloide o linfoide (Figura 8). Las leucemias agudas son una neoplasia de tipo hematológico que se caracterizan por la proliferación y crecimiento incontrolado de células. Este tipo de cáncer pueden ser de tipo mielógeno o linfocítico dependiendo su linaje. Tanto las leucemias mieloides como las linfociticas tienen formas crónicas y agudas (50, 51).

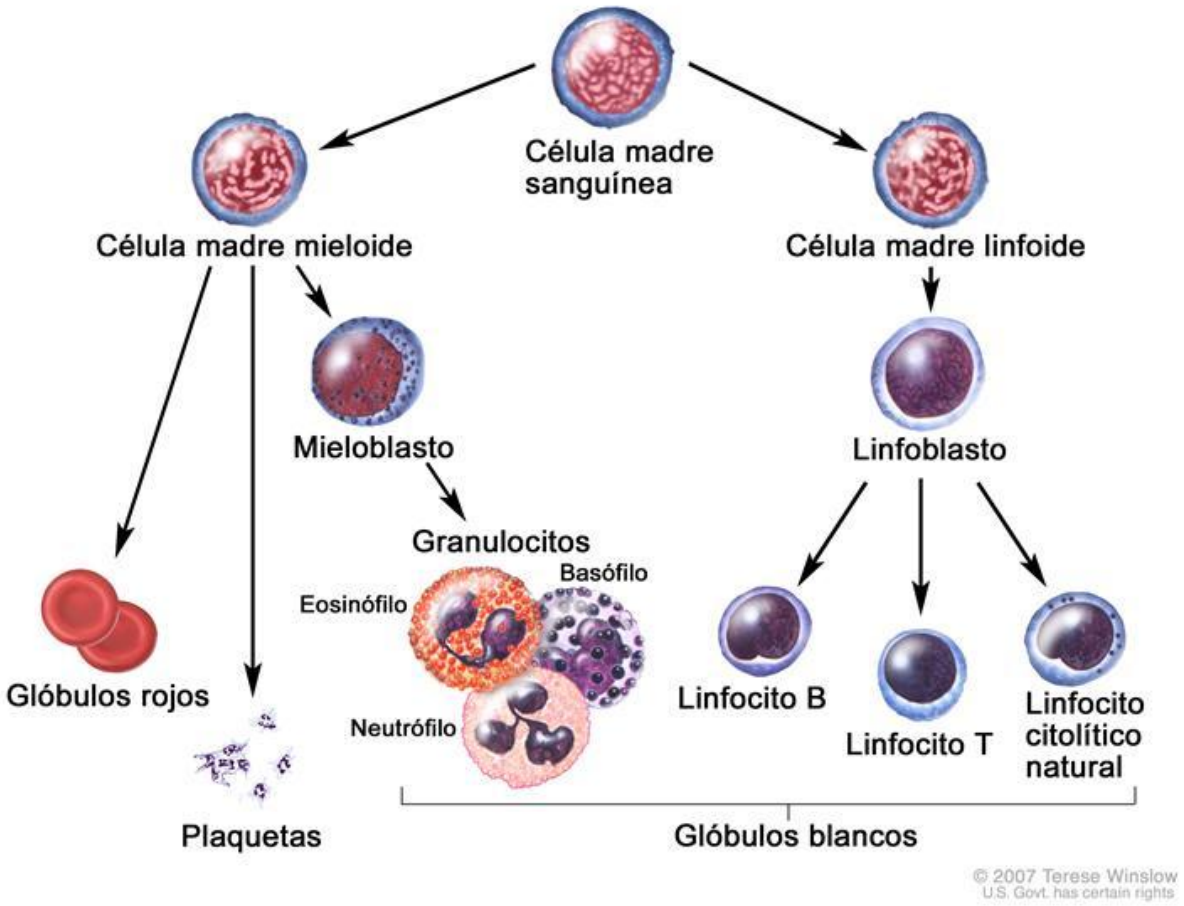


Figura 8. Desarrollo de las células sanguíneas. Las células madre mieloides se convierte en uno de tres tipos de células sanguíneas maduras: glóbulos rojos, plaquetas y granulocitos (glóbulos blancos). Las células madre linfoides se convierte en linfoblastos y de este se pueden generar linfocitos de 3 tipos: tipo B, tipo T y asesinas naturales. (Imagen tomada de http://www.cancer.gov/)

Durante el proceso de transformación de las células normales a células cancerosas ocurren varias alteraciones genéticas. En este proceso se presentan la pérdida del control de los mecanismos de replicación y reparación del ADN, así como la segregación del material genético. Se ha sugerido que las células cancerosas presentan mutaciones que inducen inestabilidad genómica y por lo tanto aceleran la tasa de mutación. Algunas de estas mutaciones afectan a genes que participan en el control del ciclo celular, así como la fidelidad e integridad de los sistemas de replicación o reparación de ADN. Una alta frecuencia de rearreglos génicos aberrantes puede condicionar una alta predisposición al desarrollo del cáncer (52).

Los rearreglos estructurales se pueden encontrar tanto en células diploides como en células con número cromosómico anormal (aneuploide). Con el refinamiento de los métodos citogenéticos para el estudio de las células leucémicas se han identificado translocaciones, inversiones y deleciones. Entre las alteraciones cromosómicas más frecuentes están:

Tabla 3. Características de los subtipos de LALp.

LAL	**Alteración cromosómica**	**Frecuencia (%)**	**Genes involucrados**
Pro-B	Ph+ o t(9;22)(q34q11)	3 – 5	BCR-ABL
Pro-B	t(12;21)(p13;q22)	25	TEL-AML1
Pro-B	rearreglos en11q23	6	MLL
Pre-B	t(1;19)(q23;p13)	5	E2A-PBX1
B	t(8;14), t(2;8), t(8;22)	2	MYC
B	Hiperdiploidia	7 – 8	-
T	rearreglos en 7q35 y 14q11	5	TCRb y TCRad respectivamente

(Tomado de 53)

1.4.2 Clasificación por microarreglos de las leucemias agudas linfoblásticas

La LAL es una enfermedad heterogénea, con subtipos de leucemia que difieren en la respuesta a la quimioterapia. La identificación de subtipos de leucemia en el pronóstico es un proceso impreciso y con un costo alto en horas hombre que requiere la experiencia combinada del hematólogo, el oncólogo, el patólogo y el citogenetista. Desde el 2002 se han usado los microarreglos para caracterizar a las LALp, investigadores como Yeoh *et al* (2002) (54), Ross *et al (2003)* (55) y recientemente Den Boer *et al* (2009) (56) han empleado esta tecnología para determinar un patrón diferencial de expresión entre los subtipos de LALp. Yeoh *et al (2002)* demostró mediante los microarreglos de la plataforma de Affymetrix (Santa Clara, CA) con el modelo HG_U95Av2 (~12,600 grupos de sondas (*probeset*)) que el perfil de expresión no sólo permite identificar con precisión los subtipos de leucemias de importancia pronóstica (en general alcanzaron una precisión diagnóstica del 96%), sino que mejora la capacidad para evaluar el comportamiento de la neoplasia bajo determinada terapia. Así mismo, los perfiles de expresión identificados incluyeron nuevos marcadores de diagnóstico y de subclasificación, así como genes candidatos para nuevas terapias. Por último, dicho análisis dio como resultado la identificación de un subtipo de leucemia nuevo (54) (Figura 9). Una de las observaciones más sorprendentes de dicho estudio es la notable diferencia en los perfiles de expresión de los distintos subtipos de leucemia, donde a pesar de que los subtipos tuvieran una morfología relativamente homogénea y una variabilidad limitada entre células T o B, cada subtipo de leucemia tenían un perfil de expresión diferencial que incluía un gran número de genes (54).

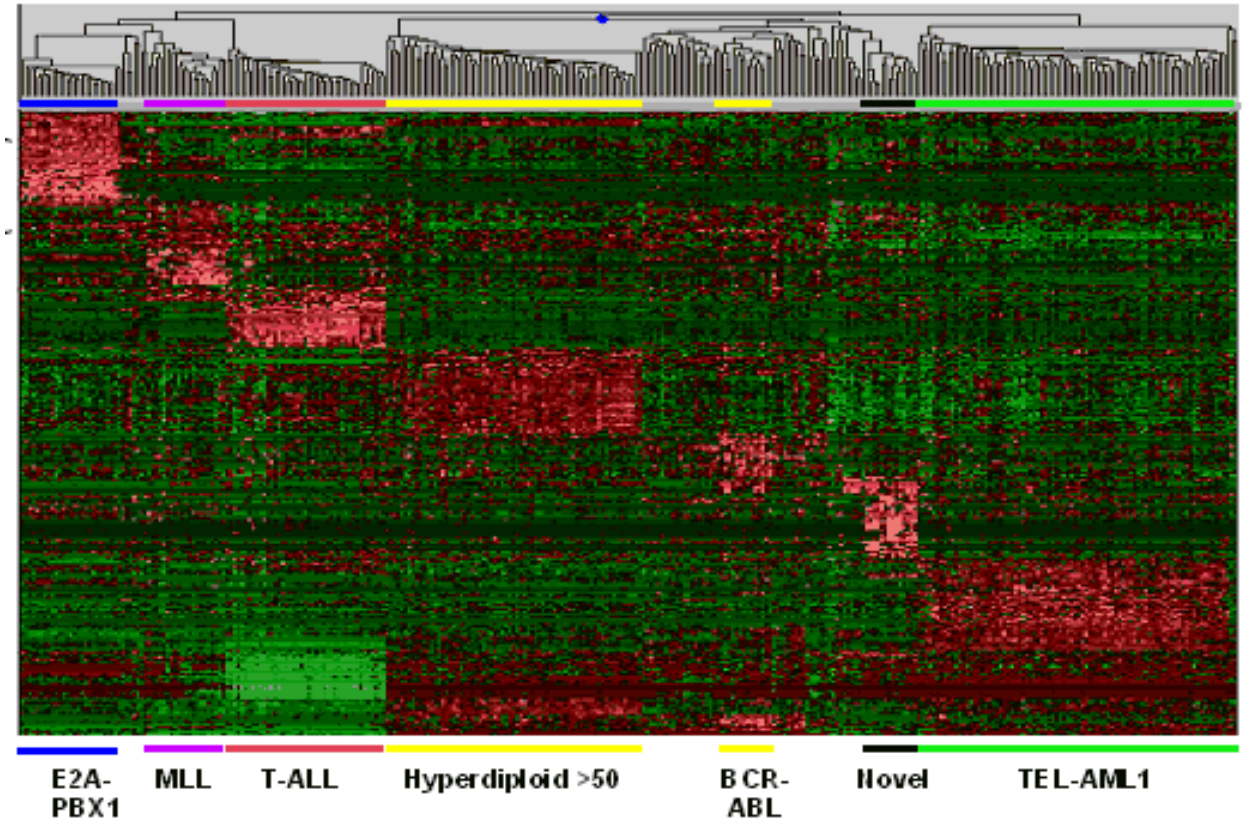


Figura 9. Mapa de calor con agrupamiento jerárquico obtenido por Yeoh *et al* (2002). De un total de 327 muestras de LALp mediante una reducción por árbol de decisión (filtro de varianza) y una prueba de Chi-cuadrada para la selección de los 40 mejores *probeset* expresados diferencialmente para cada uno de los siete subgrupos seleccionados. La selección representa 271 *probeset* donde determinados grupos de sondas fueron seleccionadas por más de un subgrupo de diagnóstico y por lo tanto removidas. Se muestran los 6 subtipos de leucemias en códigos de colores, así como un subtipo nuevo que no corresponde a alguna alteración cromosómica a la fecha de la investigación.

Bajo la misma metodología pero con dos modelos de microarreglos complementarios con mayor cobertura y complementarios (HGU133A y HGU133B, ~26,825 grupos de sondas) Ross *et al (2003)* obtuvieron como resultado 100 grupos de sondas por subtipo mejor calificadas para cada subtipo de LALp, donde 40% de sus resultados representaban a grupos de sondas ya reportadas por Yeoh *et al (2002)* (55).

Tabla 4. Correlación de genes encontrados entre los modelos de microarreglos U95Av2 y U133 por subgrupos de diagnóstico (Yeoh *et al* y Ross *et al* respectivamente). Existen probeset sinónimos por lo cual el número de genes es menor.

Subtipo	**Genes totales reportados por Ross *et al***	**Genes identificados tanto en Yeoh *et al* como en Ross *et al***	**Genes identificados en Ross *et al* no representados en el modelo HG_U95Av2 usado por Yeoh *et al***	**Genes identificados en Ross *et al* pero no identificados en Yeoh *et al***
BCR-ABL	89	19	41	29
E2A-PBX1	82	36	34	12
Hiperdiploidia	92	43	24	25
Rearreglos de MLL	83	26	41	16
LAL-T	75	53	16	6
TEL-AML	81	26	36	19
Totales	502	203	192	107
Porcentaje	100	40	39	21

El modelo U95Av2 (empleado por Yeoh et al) tiene menor cobertura de genes que U133 (empleado por Ross *et al*), de tal manera que en la columna 5 (Genes identificados en Ross *et al*. pero no identificados en Yeoh *et al*) muestra aquellos genes que están representados en ambos modelos, más que no se encontraron en el modelo U95Av2, esta discrepancia es debida a que las sondas empleadas en ambos modelos no son las mismas y la sensibilidad para detectar los niveles de RNAm es diferente.

2 JUSTIFICACIÓN

El ritmo con el que se acumulan datos biológicos continúa en aumento, frente a esto es necesario desarrollar herramientas de minería de datos que permitan agrupar, resumir y analizar la información obtenida de manera más práctica. Estas nuevas herramientas de biología computacional deben pensarse con miras a la escalabilidad y la eficiencia. De tal forma que la implementación de sistemas bioinformáticos semi-automatizados para la integración de diversos tipos de plataformas genómicas y la minería de datos es imprescindible para aprovechar la gran cantidad de información biológica depositada en las bases de datos, así como la que se puede generar en investigaciones con herramientas de alto rendimiento.

3 OBJETIVOS

3.1 Objetivo general

Implementar un sistema semi-automatizado para la integración y minería de datos genómicos humanos obtenidos a partir de distintas plataformas genómicas.

3.2 Objetivos particulares

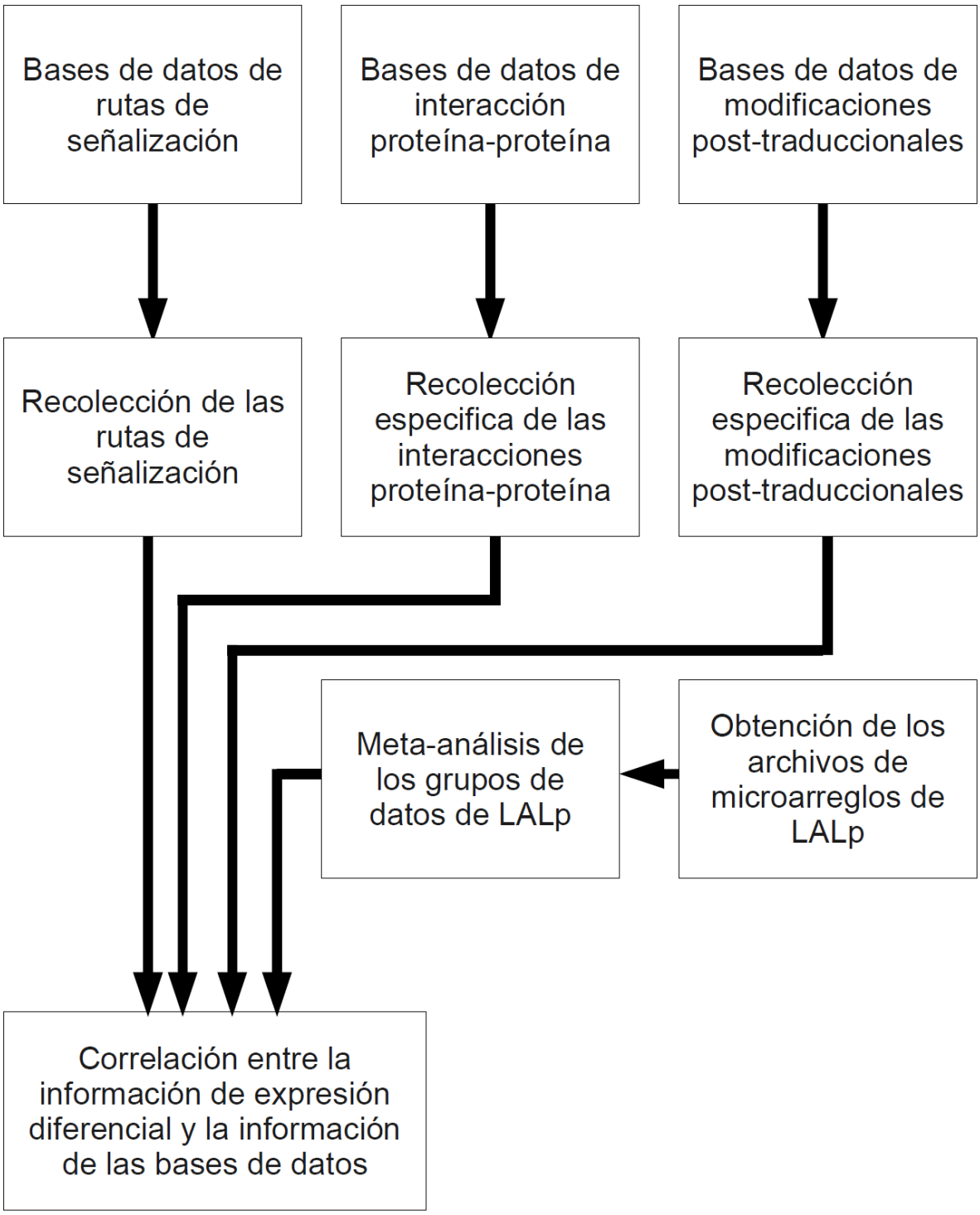
Obtener información de los genes, proteínas y rutas de señalización disponibles en diversas bases de datos y unificar los formatos en los que se presenta dicha información.

Realizar un meta-análisis de los grupos de datos disponibles de microarreglos de leucemias agudas linfoblasticas pediátricas para obtener resultados de expresión diferencial entre LAL-T y LAL-B (con alteraciones cromosómicas conocidas).

Correlacionar los datos de expresión diferencial obtenidos en el meta-análisis con los datos de las tres plataformas genómicas empleadas.

Analizar la importancia de los genes encontrados en el modelo de estudio planteado.

4 ESTRATEGIA EXPERIMENTAL



5 METODOLOGÍA

5.1 Obtención, unión y unificación de las plataformas genómicas empleadas a partir de diversas bases de datos

En primer lugar se llevó a cabo la identificación de las bases de datos que contienen las plataformas genómicas de interés (interacción proteína-proteína, modificaciones post-traduccionales y rutas de señalización); por otra parte se generó un programa con el nombre MADCBC.py (Meta-Análisis de Datos Complejos en Biología Computacional) bajo el lenguaje de programación Python, que contiene todos los procedimientos para llevar a cabo la minería de datos genómicos a gran escala, así como la correlación con resultados de expresión diferencial.

5.1.1 Selección de las bases de datos a emplear

Con la finalidad de obtener la información relativa a las interacciones proteína-proteína, las modificaciones post-traduccionales y las rutas de señalización, que permita hacer el análisis conjunto, se utilizo el sitio http://www.pathguide.org (12), con los criterios de búsqueda: "*Organism*: H. Sapiens" y "*Availability: Free for all users*". A partir del listado obtenido, se usó como criterio de selección que la base de datos sea primaria (*Data Source: Primary*). Para la selección de las bases de datos de interacción proteína-proteína (Tabla 5), así como para las bases de modificaciones post-traduccionales (Tabla 6), se buscó que la información se obtenga en formato de texto plano para sistematizar la extracción de datos, así como que se encontrara en archivos comprimidos o disponible desde un sitio FTP. Para la selección de las bases de datos de rutas de señalización se buscó que estuvieran en formato de texto plano o se pudieran extraer a partir del código html (Tabla 7).

Tabla 5. Bases de datos de interacción proteína-proteína.

Dase de datos	**Medio para la obtención de la información**	**Referencia**
Mint	Curación de literatura	Ceol *et al* (2010) (57)
MatrixDB	Curación de literatura	Chautard *et al* (2009) (58)
Domino	Curación de literatura	Ceol *et al* (2007) (59)
Interactome	Curación de literatura	Ferrari *et al* (2011) (60)
Reactome	Curación de literatura	Matthews *et al* (2008) (61)
Dip	Curación de literatura	Xenarios *et al* (2002) (62)
Biogrid	Curación de literatura	Breitkreutz *et al* (2008) (63)
IntAct	Curación de literatura	Aranda *et al* (2010) (64)
HPRD	Curación de literatura	Prasad *et al* (2008) (65)

Las bases de datos seleccionadas colectan las interacciones proteína-proteína a partir publicaciones que usan metodologías como co-cristalografía de rayos X, resonancia magnética nuclear, análisis bioquímicos, co-inmunoprecipitación, ensayos de doble híbrido, arreglos de proteínas, arreglos de anticuerpos, cromatografía por afinidad, co-localización por fluorescencia, estudio de complejos por espectrometría de masas, arreglos de péptidos, pull down etc. Se han omitido aquellas bases de datos que reportan interacciones por ortología o por predicción bioinformática.

Tabla 6. Bases de datos de modificaciones post-traduccionales.

Base de datos	**Tipo de modificación reportada**	**Referencia**
HPRD	Múltiples modificaciones +	Prasad *et al* (2008) (65)
Phosphosite	Múltiples modificaciones (archivos independientes por tipo de modificación) *	Hornbeck *et al* (2004) (66)
Phopho.ELM	Fosforilación	Diella *et al* (2008) (67)

+ Fosforilación, metilación, ubiquitinación, sumoilación, acetilación, glucosilación, corte proteolítico, puente disulfuro, s-nitrosilación, sulfación, palmitoilación, desfosforilación.
* Fosforilación, metilación, ubiquitinación, sumoilación, acetilación y glucosilación.

Tabla 7. Bases de datos de rutas de señalización.

Base de datos	**Referencia**
KEGG	Kanehisa *et al* (2005) (41)
REACTOME	Matthews *et al* (2008) (61)

5.1.2 Obtención automática de la información a emplear

5.1.2.1 Obtención del la terminología oficial de los genes

El diccionario que permite encontrar equivalencias entre la terminología oficial y la terminología Uniprot (por ser la que tiene mayor cobertura en todas las bases de datos empleadas con respecto a otros identificadores usados) se genero de la página Genenames (68) (http://www.genenames.org/) a través de las opciones a medida para bajar el archivo. Para la obtención automática de dicho diccionario (con nombre HUGO.txt) se generó el Script hugo.py (Anexo 1) que se integró al programa general MADCBC.py.

5.1.2.2 Interacción proteína-proteína y modificaciones post-traduccionales

Para generar el archivo que compilará toda la información de interés a partir de las bases de datos seleccionadas se generaron los *script*s ipp.py y modificaciones.py (Anexo 2 y 3 respectivamente) conformados por 4 partes y que están integrados en el programa general MADCBC.py.

- La primera parte obtiene el archivo en texto plano de cada una de las bases de datos (ver tablas 5 y 6).
- La segunda parte extrae información específica de *Homo sapiens*. Con respecto a las bases de interacción proteína-proteína se genera como salida un archivo de dos columnas, que corresponden a las proteínas implicadas en la interacción binaria de las proteínas. En cuanto a las bases de modificaciones post-traduccionales, se genera un archivo con el identificador de la proteína, el tipo de modificación y el residuo (incluida su posición) susceptible a la modificación

reportada.

- *La tercera parte* hace la conversión de la terminología usada por la base de datos a la terminología oficial usando el archivo HUGO.txt que se generó previamente.
- La cuarta y última parte de los *scripts* unen los archivos generados eliminando la información repetida en las bases de datos para generar los archivos IPP.txt y MODIFICACIONES.txt para las interacciones proteína-proteína y modificaciones post-traduccionales respectivamente.

5.1.2.3 Rutas de señalización

Para cada una de las bases de rutas de señalización se implementaron soluciones diferentes en función de las características de cada base de datos.

5.1.2.3.1 Reactome

Del sitio de Reactome (http://www.reactome.org/ReactomeGWT/entrypoint.html) (61) se obtuvo el archivo ReactomePathways.gmt.zip, donde se encuentra la totalidad de las rutas reportadas en dicho sitio. Para la extracción de datos se generó el *script* reactome.py (Anexo 4), el cual obtiene el archivo del sitio Reactome, descomprime dicho archivo y extrae la información necesaria (nombre de la ruta y los genes implicados en esa ruta, todo esto en una sola línea), para generar un archivo final de nombre REACTOME.txt.

5.1.2.3.2 Kegg

Para generar el archivo de las rutas de señalización de Kegg (http://www.genome.jp/kegg/) (42) se generó el *script* kegg.py (Anexo 5) con 3 procedimientos principales:

- La primera parte obtiene un listado de las rutas de señalización de humano reportadas en Kegg (http://www.genome.jp/dbget-bin/get_linkdb?-t+pathway+gn:T01001).

- A partir del listado obtenido la segunda parte del script obtiene la totalidad de las rutas reportadas en Kegg depurando la información no necesaria.

El tercer procedimiento une las rutas en un solo archivo bajo el nombre de KEGG.txt.

5.2 Meta-análisis de datos de microarreglos

5.2.1 Obtención de los archivos de microarreglos

En las base de datos de Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) (4) se realizó una búsqueda de microarreglos de leucemias agudas linfoblásticas pediátricas realizados con la plataforma Affymetrix modelo HU 133A Plus 2.0 (69). El criterio de inclusión fue que reportaran el tipo de aberración cromosómica que presentaban (fusión BCR-ABL, fusión TEL-AML1, fusión E2A-PBX, rearreglos que implican al gen MLL, hiperdiploidia o LAL-T). Con dichos parámetros se obtuvieron un total de 396 archivos de microarreglos, en la tabla 8 se muestran las características de estos.

Tabla 8. Fuente de los grupos de datos de microarreglos.

No. de acceso de GEO	**Referencia**	**Archivos totales**	**Archivos tomados para el meta-análisis**
GSE11877	Kang *et al* (2010) (70)	207	32
GSE7440	Bhojwani *et al* (2008) (71)	99	27
GSE7757	Campo *et al* (2007) (72)	99	23
GSE10609	Vlierberghe *et al* (2008) (73)	92	91
GSE10792	Bungaro *et al* (2009) (74)	160	52
GSE14062	Zangrado *et al* (2009) (75)	81	16
GSE14615	Winter *et al* (2007) (76)	50	50
GSE14834	Fulci *et al* (2009) (77)	19	13
GSE17459	Herzberg *et al* (2010) (78)	119	9
GSE18497	Staal *et al* (2010) (79)	82	14
GSE19475	Stam *et al* (2010) (80)	73	58
INP	Laboratorio Oncología Experimental	57	12

Tabla 9. Distribución de los microarreglos con base en el subtipo.

Rearreglo en MLL	TEL-AML t(12;21)	BCR-ABL t(9:22)	E2A-PBX t(1;19)	LAL-T	Hiper-diploidia	Totales
123	44	30	28	163	8	396

5.2.2 Pre-procesamiento

Se realizó el pre-procesamiento de 396 archivos de microarreglos utilizando las librerías de análisis de microarreglos de BioConductor (81) que utilizan la plataforma de análisis estadístico R. Los archivos se renombraron con base en un número que señala el subtipo de leucemia al que pertenece (1: Rearreglos que implican al gen MLL, 2: t(12;21) (TEL-AML), 3: t(1;19) (E2A-PBX), 4: Hiperdiploidia, 5: LAL-T y 6: t(9;22) (BCR-ABL)), seguida de un punto y un número consecutivo de dicho subtipo, por ejemplo: 1.031 corresponde a un archivo que tiene los datos de una leucemia que implica el rearreglo del gen MLL y que es la muestra 31 de ese subtipo. El pre-procesamiento se llevó a cabo utilizando los parámetros Background correction: RMA, Normalization: Quantiles, Summarization: medianpolish y pmcorrect: pmonly (Anexo 6, secuencia de comandos empleada para el pre-procesamiento de los datos de microarreglos). Dicho procedimiento generó la matriz de expresión (las 396 muestras con sus respectivas 54,765 lecturas de los probeset en escala logarítmica).

5.2.3 Análisis de expresión diferencial

Los resultados de expresión diferencial se obtuvieron con el método LIMMA (27), agrupando los datos con base en los 6 diferentes subtipos de LALp. Se construyó una matriz de contrastes para agrupar todos los subtipos, y se empleo junto con la matriz de expresión para la obtención de los resultados de expresión diferencial. Además se empleó el método removeBatchEffect para evitar el efecto de lote (donde los microarreglos se agrupan en función del estudio y no del tipo de muestra de la que se trate) (Anexo 7, secuencia de comandos empleada).

Para la obtención de los genes característicos del subtipo LAL-T se realizaron 5 contrastes (contra cada uno de los 5 subtipos restantes) para generar 5 tablas de resultados de expresión diferencial. Dichos resultados se generaron con 2 criterios de corte, que el valor de veces de cambio fuera igual o mayor a 1 logFCh (2 veces de cambio) y que el valor estadístico de B fuera igual o mayor a 3 (95.3 % de nivel de confianza de que el dato es verdadero).

Para obtener el listado de genes específicos y exclusivos del subtipo LAL-T los 5 resultados de expresión diferencial obtenidos por Bioconductor se separaron en 2 tablas (procedimiento realizado en hoja de cálculo) esquematizado en la tabla 10, una correspondiente a los *probeset* con expresión aumentada y otra correspondiente a *probeset* con expresión disminuida. Para obtener la lista de genes que se comportaban de la misma forma (sobre-expresados ó sub-expresados) en los resultados se generó el *script* probeset_iguales.py (Anexo 8) que selecciona aquellos probeset que se comparten en las 5 tablas de resultados (esquematizado en la tabla 11) y así generar un listado de sólo estos probesets compartidos.

Tabla 10. Relaciones para efectuar los contrastes para obtener la expresión diferencial del subtipo LAL-T con respecto al resto de subtipos.

Subtipo	**Subtipo contra el que se contrasto**	**Sub-tablas (Sobre-expresados / Sub-expresados)**
LAL-T	TEL/AML	TEL/AML Sobre-expresados
		TEL/AML Sub-expresados
	BCR/ABL	BCR/ABL Sobre-expresados
		BCR/ABL Sub-expresados
	MLL	MLL Sobre-expresados
		MLL Sub-expresados
	E2A/PBX	E2A/PBX Sobre-expresados
		E2A/PBX Sub-expresados
	Hiperdiploidia	Hiperdiploidia Sobre-expresados
		Hiperdiploidia Sub-expresados

Tabla 11. Comparación realizados mediante el *script* probeset_iguales.py para conseguir los *probeset* que se compartían el los 5 resultados de expresión diferencial de LAL-T.

	Probeset sobre-expresados	**Probeset sub-expresados**
Contraste 1	TEL/AML vs BCR/ABL	TEL/AML vs BCR/ABL
Contraste 2	Probeset compartidos TEL/AML-BCR/ABL vs MLL	Probeset compartidos TEL/AML-BCR/ABL vs MLL
Contraste 3	Probeset compartidos TEL/AML-BCR/ABL-MLL vs E2A/PBX	Probeset compartidos TEL/AML-BCR/ABL-MLL vs E2A/PBX
Contraste 4	Probeset compartidos TEL/AML-BCR/ABL-MLL-E2A/PBX vs Hiperdiploidia	Probeset compartidos TEL/AML-BCR/ABL-MLL-E2A/PBX vs Hiperdiploidia
Probeset compartidos entre los 5 resultados	TEL/AML-BCR/ABL-MLL-E2A/PBX-Hiperdiploidia	TEL/AML-BCR/ABL-MLL-E2A/PBX-Hiperdiploidia
Unión de resultados	Unión de los probeset sobre-expresados y sub-expresados	

El listado resultante de los contrastes realizados con el *script* probeset_iguales.py se relacionó con la matriz de expresión mediante el *script* selecmatriz.py (Anexo 9) para generar la matriz de expresión especifica de los resultados de expresión diferencial del meta-análisis. Dicha matriz de expresión se utilizó para construir en R el mapa de calor del agrupamiento jerárquico (Anexo 10 secuencia de comandos). Así mismo, con dicha matriz se realizó un análisis de componentes principales (PCA) con el software libre MeV (82), para observar la segregación de los datos de las muestras de LAL-T con relación al resto de los subtipos, usando los parámetros por default y agrupando por muestra.

Se realizó un segundo análisis de expresión diferencial contrastando los 163 archivos de LAL subtipo T y los 233 archivos correspondientes al resto de los subtipos, dicho resultado de expresión diferencial se consiguió mediante LIMMA (27). Con dicho procedimiento se obtuvieron las veces de cambio del listado de *probeset* con expresión aumentada y disminuida resultantes de los contrastes realizados con el *script* probeset_iguales.py.

Debido a que los resultados arrojados por Bioconductor tenían la terminología de los identificadores de Affymetrix, fue necesario generar el *script* MA-HUGOaffy.py (Anexo

11) para renombrarlos a la terminología oficial, usando como diccionario el archivo de anotaciones HG-U133_Plus_2.na31.annot.csv, proveído por Affymetrix.

5.3 Correlación de los diferentes datos obtenidos

Para esta última parte del *script* MADCBC.py se generaron dos procesos, el primero de ellos permite enriquecimiento de la información de los genes cuya expresión está regulada, y el segundo proceso vincula los genes con expresión diferencial a las rutas afectadas:

- El primer procedimiento genera el enriquecimiento de una lista de genes de interés específicos de *Homo sapiens* (con los datos de interacción proteína-proteína, modificaciones post-traduccionales y de rutas de señalización), alimentada manualmente o a través de un archivo con el listado de genes. A partir de este proceso se genera un archivo por cada gen introducido, donde la primera parte da un listado de las proteínas con las que interacciona (de la lista de genes proporcionada), la segunda parte muestra todas las modificaciones post-traduccionales que se reportan en las bases de datos, la tercer parte muestra todas las rutas en las que esta implicado y por último se da un listado con todas las proteínas con las que interaccionan (estén o no en el listado de genes introducida), esto en texto plano. Dichos archivos se guardan en una carpeta que se genera al concluir el proceso (Anexo 12 script generado (resultados.py)).

- El segundo proceso genera 6 archivos en formato .csv, 3 archivos correspondientes a Reactome y 3 a Kegg, dos de ellos (uno de Reactome y uno de Kegg) tienen información de las rutas que tienen genes con expresión disminuida, un segundo par tiene resultados de las rutas que tienen genes que aumentan su expresión y el tercer par contiene resultados combinados, donde hay genes con expresión aumentada, así como genes con expresión disminuida. Todos los archivos generados muestran resultados porcentuales de cuantos genes están implicados en la ruta y cuantos de estos tienen modificada su expresión, dichos archivos se guardan en una carpeta generada al final del procedimiento (Anexo 13 script generado (resultados2.py)).

6 RESULTADOS

6.1 Actualización de las bases de datos

Para el óptimo funcionamiento del programa desarrollado en este trabajo, es necesario que los datos se actualicen periódicamente, por lo que se implementó un sistema que lleve a cabo dicha actualización. No sólo a nivel de los datos de interacción proteína-proteína, modificaciones post-traduccionales o rutas de señalización, sino también con respecto a la asignación de nombres oficiales para los identificadores que emplean en las bases de datos, pues de todos estos datos depende que los resultados finales sean más robustos. En la tabla 12 se muestran las actualizaciones de los cuatro tipos de datos en un transcurso de 6 meses.

Tabla 12. Ganancia de datos en un periodo de 6 meses en los cuatro niveles de información empleada en *Homo sapiens*.

Tipo de información	**Marzo del 2011**	**Septiembre del 2011**	**Crecimiento (%)**
Interacción proteína-proteína	251187	285959	12.12%
Modificaciones post-traduccionales	16188	16308	0.73%
Rutas de señalización	1079 rutas (Reactome)	1139	5.20%
	218 rutas (Kegg)	245	11.00%
Terminología oficial	33523	35835	6.40%

6.2 Meta-análisis de los datos de microarreglos

6.2.1 Pre-procesamiento

La finalidad del pre-procesamiento es la de homogeneizar las mediciones de los diversos microarreglos antes de compararlos entre si, esto es los 54,765 *probesets* de cada uno de los 396 archivos que se seleccionaron. En las figuras 10 y 11 se muestran los resultados gráficos del pre-procesamiento de los datos, donde son apreciadas las mediciones a nivel del rango dinámico (ejes de las equis, entre 6 a 12 en los niveles de

intensidad de fluorescencia en escala logarítmica) que serán las mediciones empleadas en lo sucesivo para realizar el análisis de expresión diferencial (antes y después de este rango las mediciones son tomadas como ruido de fondo debido a hibridación inespecifica).

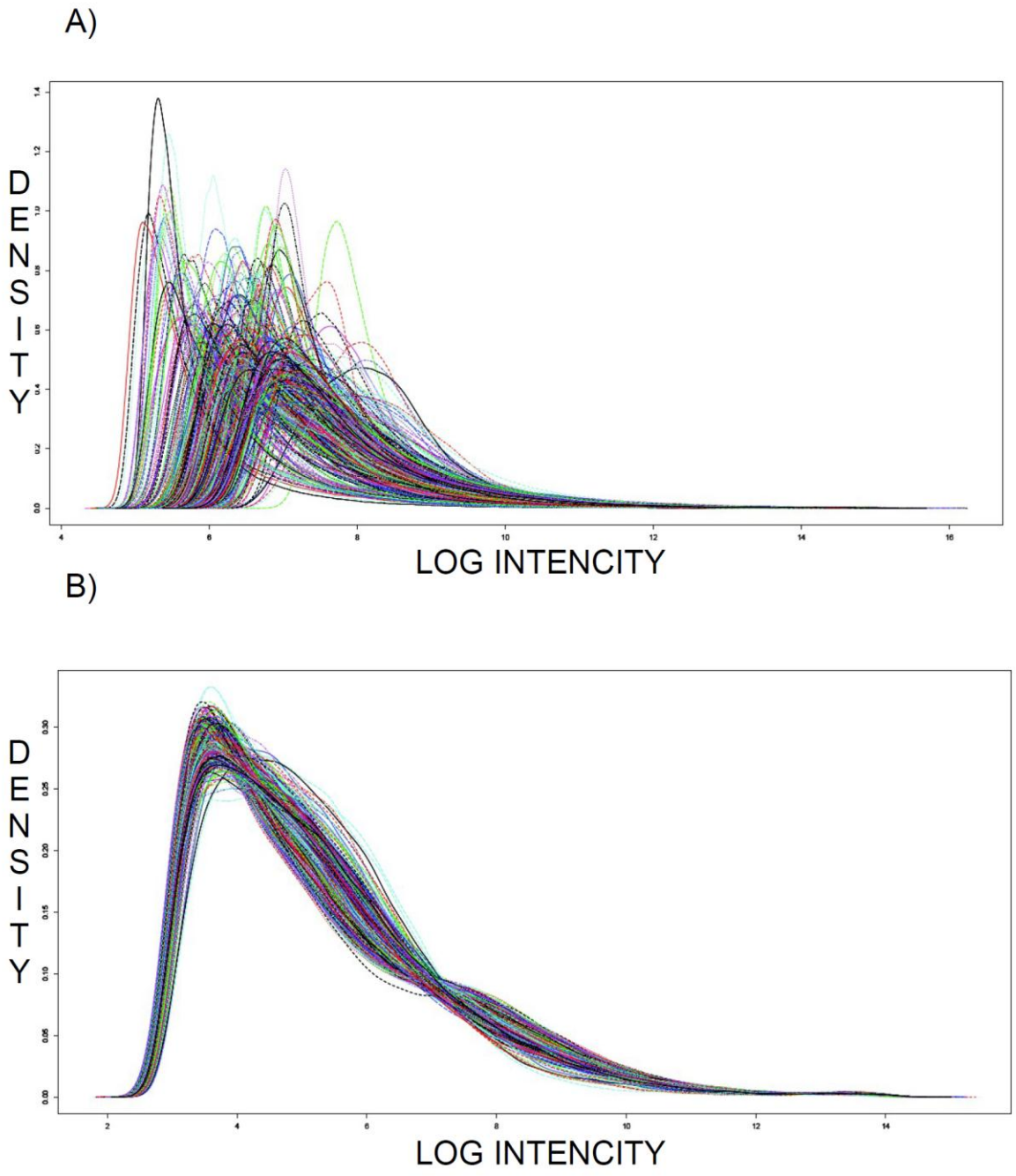


Figura 10. Histogramas de las distribuciones de intensidad de fluorescencia. Antes (A) y después (B) del pre-procesamiento. En el eje de las X se muestra la intensidad de fluorescencia en escala logarítmica y en el eje de las Y se muestra la densidad (cantidad de *probesets* en esa región).

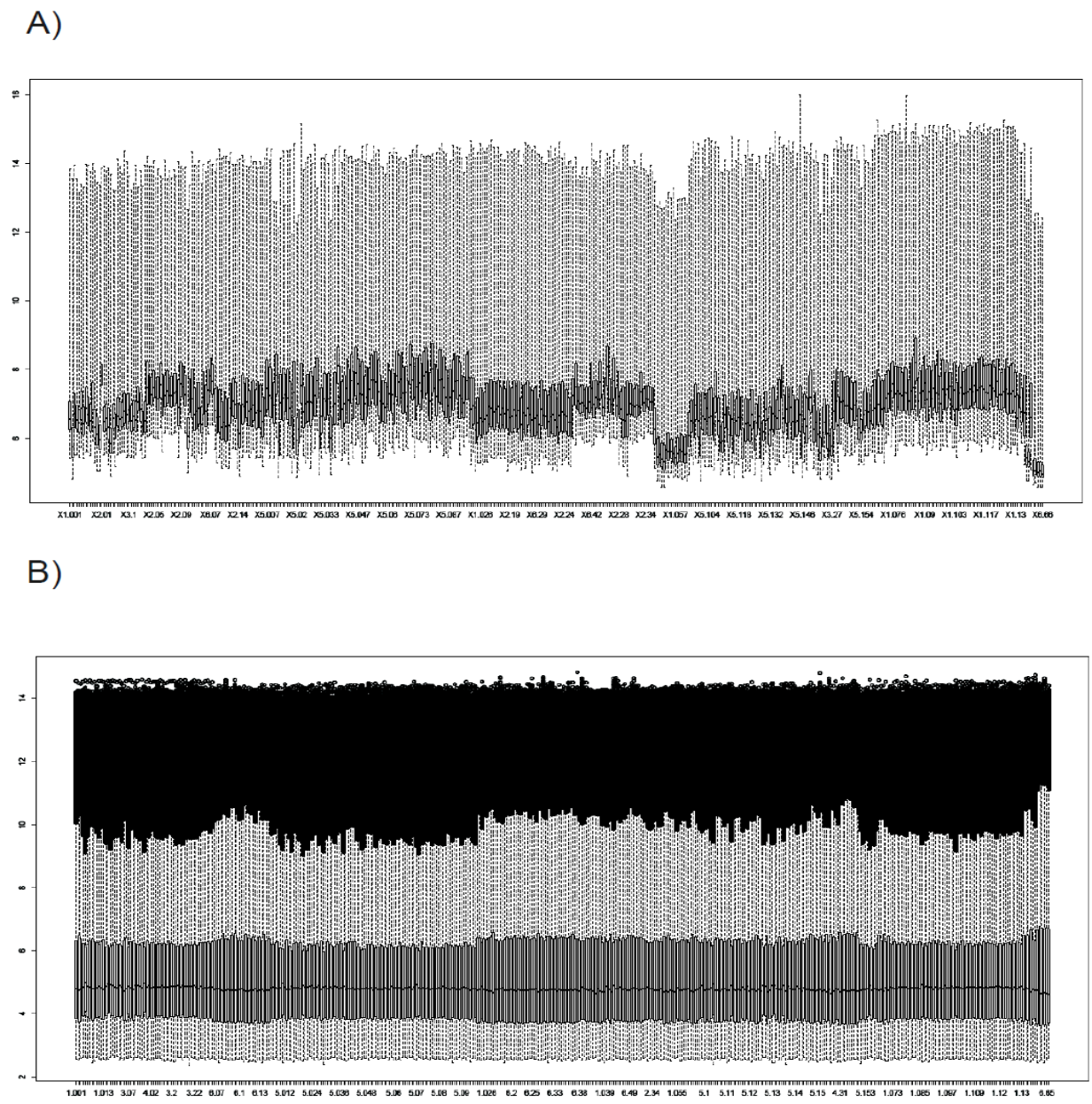


Figura 11. Diagramas de caja y bigote de las distribuciones de intensidad de fluorescencia. Antes (A) y después (B) del pre-procesamiento de las distribuciones de intensidad de fluorescencia. Tras el pre-procesamiento se aprecia como la distribución de las intensidades de fluorescencia son más homogéneas entre si.

Una vez hecho el análisis de los datos, mediante un mapa de calor de la correlación de Pearson de las intensidades de fluorescencia de los microarreloglos (Figura 12), se observó que se generaba un efecto de lote (donde las intensidades de fluorescencia se agrupan en función de la fuente de los datos y no del tipo de muestra de la que se trate), por ello se empleo el método removeBatchEffect() para eliminar dicho efecto (Figura 13).

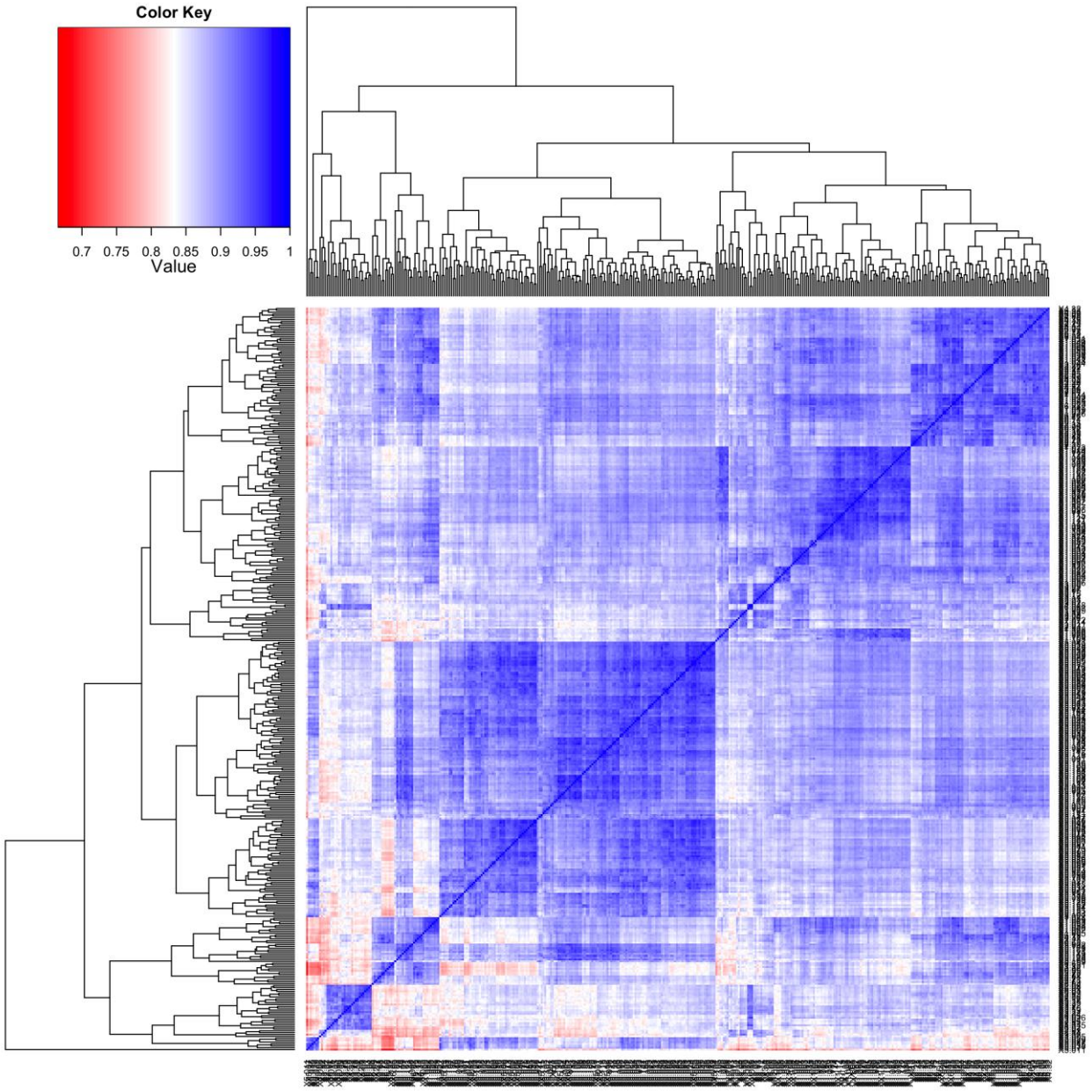


Figura 12. Mapa de calor de la correlación de Pearson. Los datos de intensidad de fluorescencia tienden a agruparse en función del grupo de trabajo de donde provienen. Cada línea del dendograma representa un archivo del microarreglos, el color azul representa mayor grado de cercanía entre las intensidades de los datos comparados.

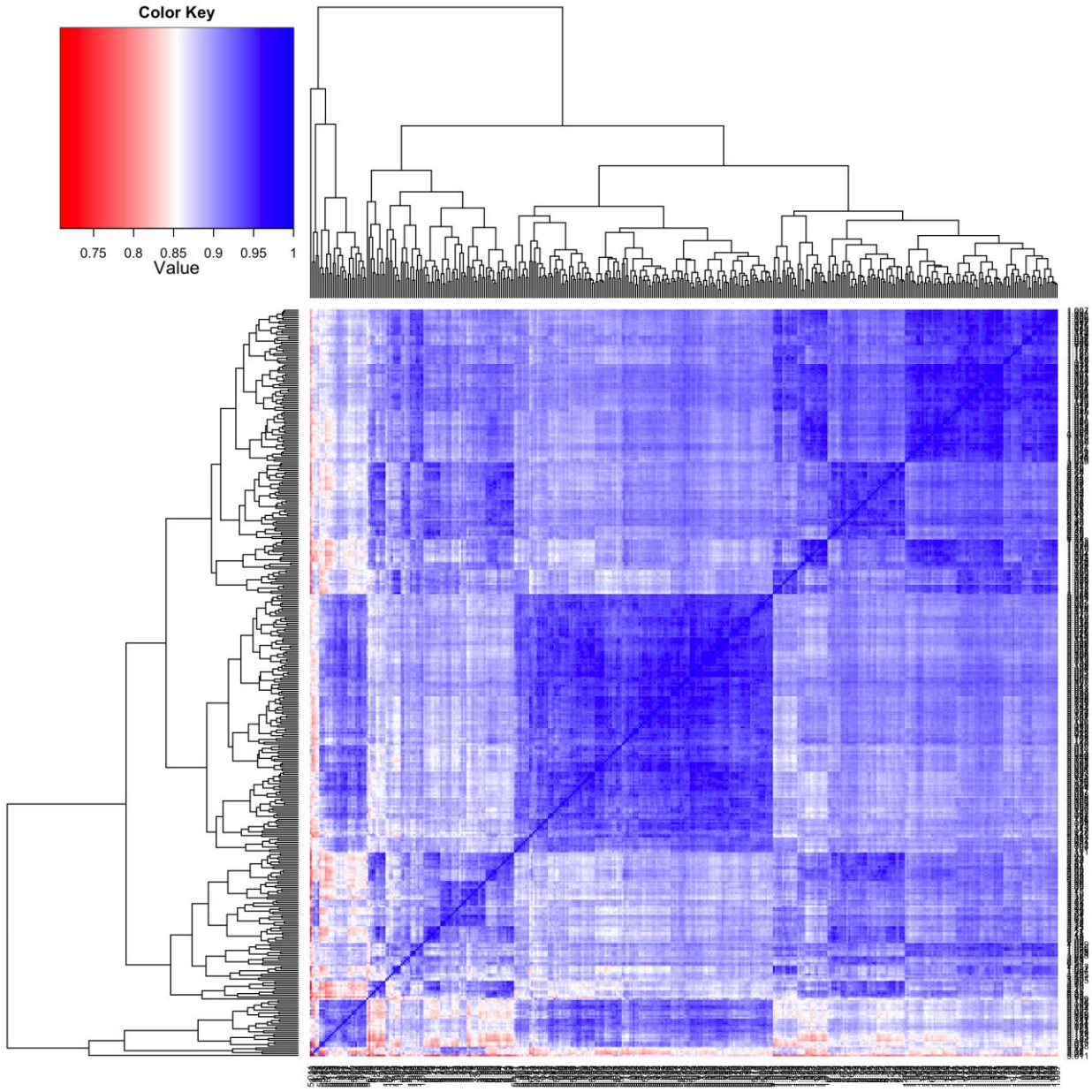


Figura 13. Mapa de calor de la correlación de Pearson tras aplicar el tratamiento removeBatchEffect. Tras remover el efecto de lote los microarreglos tienden a agruparse preferencialmente en función del subtipo de leucemia y no en función de la fuente de datos.

6.2.2 Expresión diferencial de LAL-T

Tras el meta-análisis realizado de los 12 grupos de datos (396 archivos de microarreglos), contrastando el subtipo LAL-T con cada uno de los subtipos restantes (MLL, BCR/ABL, E2A/PBX, TEL/AML e hiperdipliodia) se obtuvieron 775 *probeset* que se compartían en los cinco resultados de expresión diferencial (527 *probeset* con expresión disminuida y 248 con expresión aumentada), en la tabla 13 se presentan los resultados numéricos de los diferentes contrastes que se realizaron y en la figura 14 los resultados se representan en gráfica de volcán.

Tabla 13. Cantidad de *probeset* que se obtuvieron como resultados de expresión diferencial para cada uno de los contrastes de LAL-T con el resto de los subtipos.

Sub-tipo	**Subtipo con el que se contrasto**	**Sobre-expresados / Sub-expresados**	**Cantidad de *probeset***	***Probeset* totales**
LAL-T	TEL/AML	Sobre-expresados	1026	3782
		Sub-expresados	2756	
	BCR/ABL	Sobre-expresados	886	3325
		Sub-expresados	2439	
	MLL	Sobre-expresados	927	2963
		Sub-expresados	2036	
	E2A/PBX	Sobre-expresados	1758	4262
		Sub-expresados	2504	
	Hiperdiploidia	Sobre-expresados	727	4419
		Sub-expresados	3692	
	Compartidos	Sobre-expresados	248	775
		Sub-expresados	527	

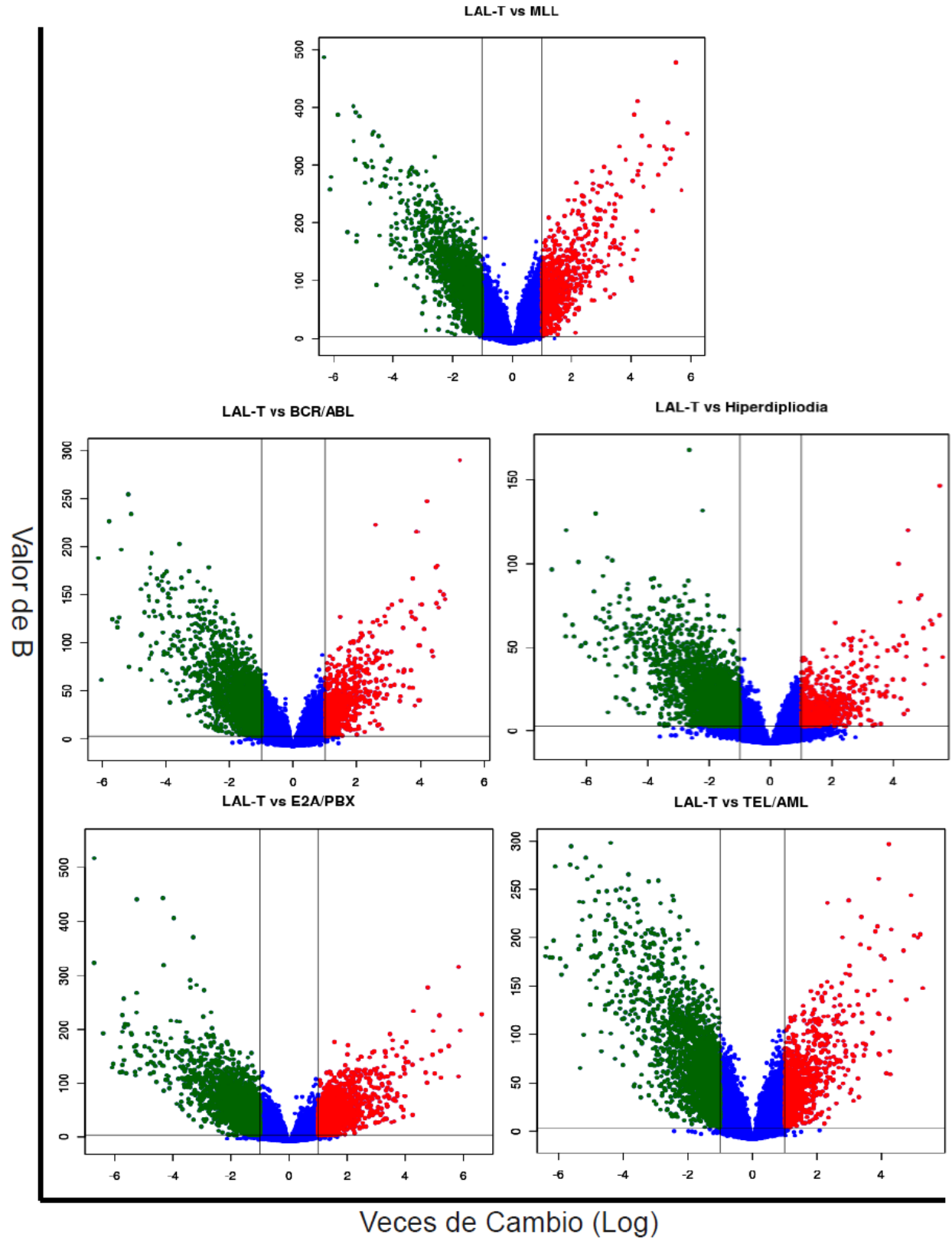


Figura 14. Gráficos de volcán de los resultados de expresión diferencial de cada uno de los contrastes que se realizaron. Las líneas representan los puntos de corte (eje de las X (Log Fold Change): 1 y -1 en escala logarítmica (equivalente a 2 veces de cambio), eje de las Y (Log Odds): Valor de B ≥ 3 (equivalente al 95 % de nivel de confianza que el resultado es verdadero).

Los 775 probeset que se identificaron son los que caracterizan específicamente al subtipo LAL-T sobre el resto de subtipos. Para identificar las veces de cambio se procedió a realizar un análisis de expresión diferencial de LAL-T contra el resto de los subtipos. Los resultados de expresión diferencial también se emplearon para generar el mapa de calor (figura 15) y así como el análisis de componentes principales (Figura 16) para visualizar el agrupamiento de las muestras LAL-T y LAL-B (MLL, BCR/ABL, E2A/PBX, TEL/AML e hiperdipliodia), en ambas gráficas es apreciable el agrupamiento, a excepción de una muestra correspondiente al subtipo BCR/ABL proveniente de Bungaro *et al* (2009), donde se emplearon diversas muestras con menos del 80 % de blastos totales (74), donde la alta cantidad de otros tipos celulares pudo influir en los niveles de mensajeros hibridados.

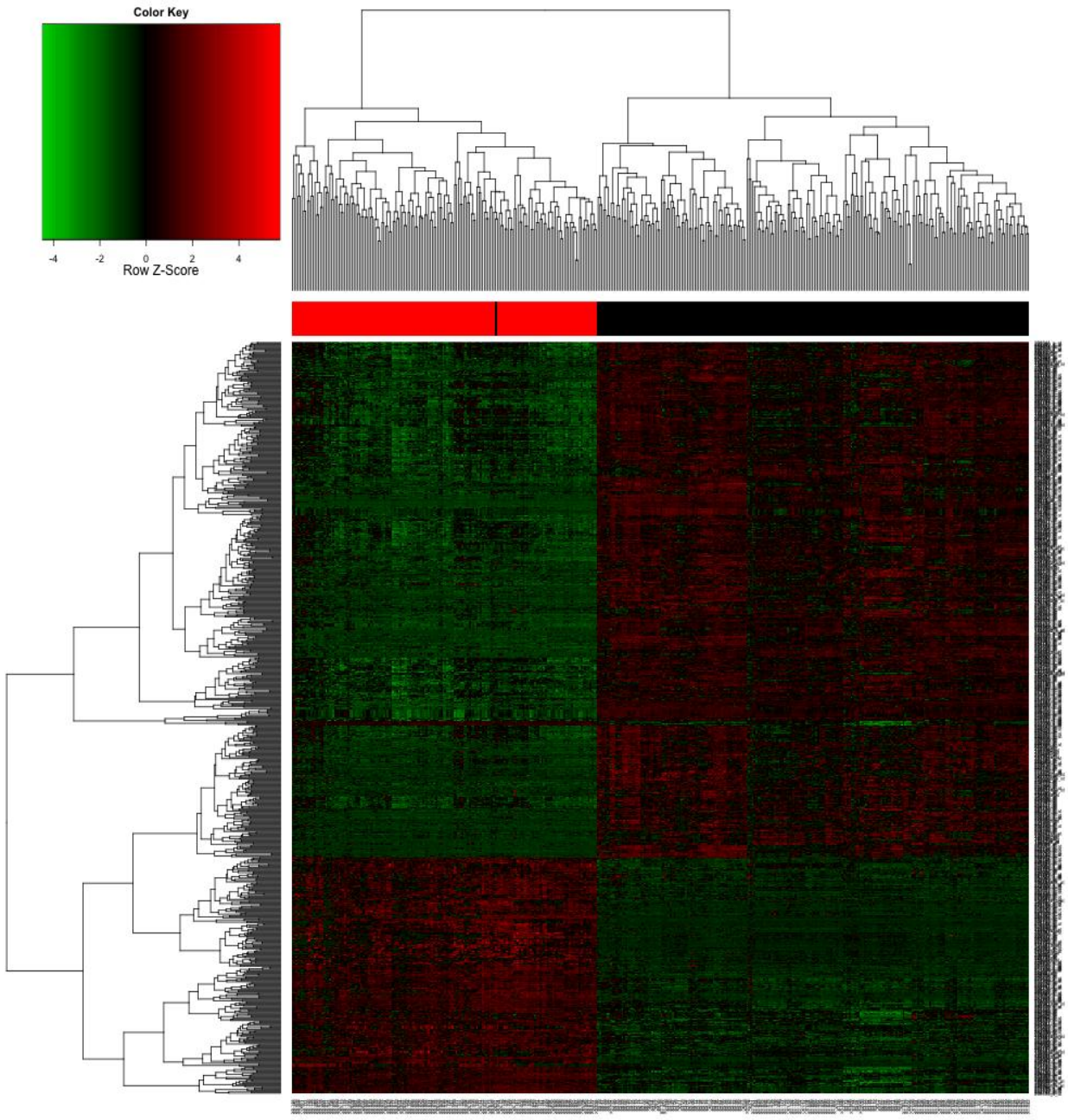


Figura 15. Mapa de calor basado en los resultados de expresión diferencial exclusivos de LAL-T. Grado de segregación alcanzado con los 775 probesets identificados, las barras rojas representa a las muestras LAL-T y la barra negra el resto de subtipos. Cada línea vertical (columna) del dendograma representa una muestra (396 muestras), cada línea horizontal (fila) del dendograma representa un probeset (775). La muestra BCR/ABL proveniente de Bungaro *et al* (2009) (73) es la línea que se aprecia en negro entre las muestras de LAL-T.

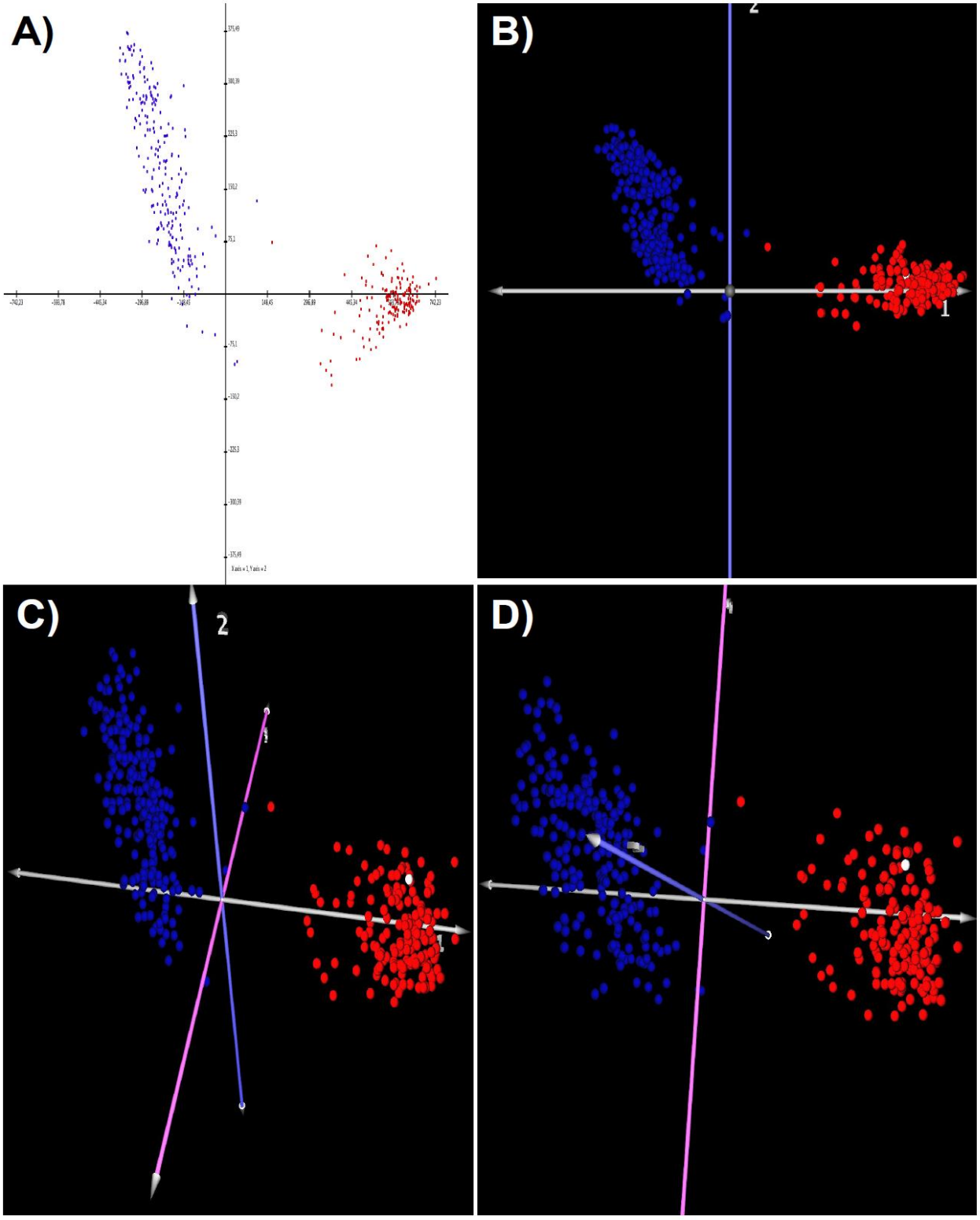


Figura 16. PCA LAL-T vs resto de subtipos. La imagen muestra diferentes perspectivas del PCA donde se aprecia la segregación obtenida por los 775 probesets que se obtuvieron del los resultados de expresión diferencial. En rojo LAL-T y en azul el resto de los subtipos. A) PCA en 2D, B), C) y D) PCA en 3D. La muestra BCR/ABL proveniente de Bungaro *et al* (2009) se muestra en blanco.

De los 775 *probeset* resultantes, 65 fueron removidos, pues representan ESTs sin datos ni nombres asignados quedando 710 *probeset*. Diversos probesets en los microarreglos de Affymetrix representan a un mismo gen, de tal forma que al transformar los identificadores empleados por Affymetrix a la terminología oficial, se procedió a remover manualmente los probeset sinónimos eligiendo aquellos con el número de B más alto. Teniéndose 299 probeset sinónimos, correspondientes a 121 probeset, se paso de un listado de 710 probeset a 532 (362 con expresión disminuida y 170 con expresión aumentada). Estos resultados fueron así mismo empleados para la ultima parte del programa MADCBC.py (correlación de los resultados de expresión diferencial con los múltiples datos adquiridos de las bases de datos), dicho listado de 532 *probeset* (genes) se muestra en el anexo 14.

6.3 Relación de las bases de datos generadas con los resultados de expresión diferencial

6.3.1 Enriquecimiento de los genes

Una de las finalidades de programa MADCBC.py es contar con información de interacción proteína-proteína, rutas de señalización y modificaciones post-traduccionales de los genes que se expresan diferencialmente; de los 532 genes introducidos al programa obtuvieron resultados en el proceso de enriquecimiento 472 genes. Las tablas generadas se agrupan con base a sus resultados en la tabla 14. En el Anexo 15 y 16 se muestran 2 ejemplos de las tablas resultantes del enriquecimiento de genes (TOP2B.txt y CD1B.txt).

Tabla 14. Resultados numéricos de los resultados de enriquecimiento de los genes de LAL-T.

Tipo de resultado	**Numero de genes**	**Porcentajes**
Sin resultados	60	11.3
Interacción proteína-proteína (IPP)	39	7.3
Modificaciones post-traduccionales	32	6
Rutas de señalización	9	1.6
IPP y Modificaciones	146	27.5
IPP y Rutas de señalización	41	7.7
Modificaciones y Rutas de señalización	8	1.5
Todos los resultados	197	36.8
Totales	532	100

6.3.2 Relación de genes con expresión diferencial y rutas de señalización

El segundo tipo de resultados que se obtienen del programa MADCBC.py es vincular el comportamiento de los genes con las rutas de señalización, llevando a cabo un análisis para determinar las rutas de señalización en las que participan los genes que cambian su expresión. Las rutas se pueden clasificar en 3 categorías, en las que los genes se encuentran con expresión disminuida, rutas que tienen genes con expresión aumentada y aquellas con genes con expresión tanto disminuida como aumentada. Como se muestra en la tabla 15 la mayoría de las rutas tienen genes con una expresión disminuida Los anexos 17, 18 y 19 muestran las tablas de resultados de KEGG (sub-expresadas, sobre-expresadas y equilibrio respectivamente) y los anexos 20, 21 y 22 los resultados de Reactome (sub-expresadas, sobre-expresadas y equilibrio respectivamente).

Tabla 15. Comportamiento de las rutas con base al comportamiento de los genes que se encontraron en los resultados de expresión diferencial.

Base de datos	**Comportamiento**		
	Sub-expresadas	**Sobre-expresadas**	**Equilibrio**
KEGG	55	19	91
Reactome	247	55	131

Nota: Cabe destacar que varias de estas rutas tuvieron menos de 5 % genes cuya expresión se vio alterada, o la ruta se señalización aparentemente sale del contexto del modelo de estudio, por lo que no se consideraron para el análisis posterior. (Ver Anexos 17 – 22).

Uno de los motivos por los que se eligió a las LALp es porque existen un gran número de estudios a nivel molecular, de tal forma que los resultados arrojados por el programa MADCBC.py serían relativamente fáciles de comprobar con la literatura disponible. Al ser contrastado el subtipo LAL-T contra el resto de subtipos, lo que parcialmente se esta haciendo es (pues para el modelo de estudio sólo se tomaron los archivos de microarreglos de muestras que tienen alteraciones cromosómicas definidas) contrastar las leucemias de células T contra las leucemias de células B.

6.3.2.1 Análisis de resultados de las rutas de señalización y los genes con expresión diferencial

Un primer resultado al analizar los resultados de genes con expresión diferencial es el referente a la ruta co-estimulación por CD28 con 32 genes implicados (ruta reportada en Reactome), la cual tiene una afectación de 3 genes con expresión disminuida (LYN con 4.9 veces de cambio, PDPK1 con 3.4 veces de cambio y PIK3R1 con 2.6) cabe destacar que LYN está implicada en la ruta de señalización de células B, mientras que PIK3R1 esta implicada en la ruta de señalización de células T (rutas reportadas en KEGG). Los resultados reportan 4 genes con expresión aumentada (CD28 con 10.3 veces de cambio, GRAP2 con 2.6 veces de cambio, LCK con 22.2 veces de cambio y RICTOR con 2.4) (Figura 17). En las células T inmaduras, la co-estimulación con CD28 aumenta el progreso del ciclo celular, potentemente estimula la expresión tanto de la linfocina mitogénica interleucina-2 (IL-2) y su receptor, esta ruta también estimula la activación de un programa anti-apoptótico (83). Los resultados arrojados no aportan

evidencia sobre la estimulación de la expresión de IL-2 y su receptor en el contexto de las LAL-T, de hecho la ruta de interleucina-2 (43 genes) tiene 5 genes con expresión disminuida (CSF2RB con 6.4 veces de cambio, INPP5D con 4.4, PIK3CD con 3.2, PIK3R1 con 2.6, y SYK con 3.5 veces de cambio), reportando sólo un gen con expresión aumentada (LCK). Si bien los resultados no son contundentes, estos aportan información en el sentido esperado, es decir, ruta característica de células T, así como activación anti-apoptótica y progresión del ciclo celular (dos comportamientos correspondientes a procesos neoplásicos).

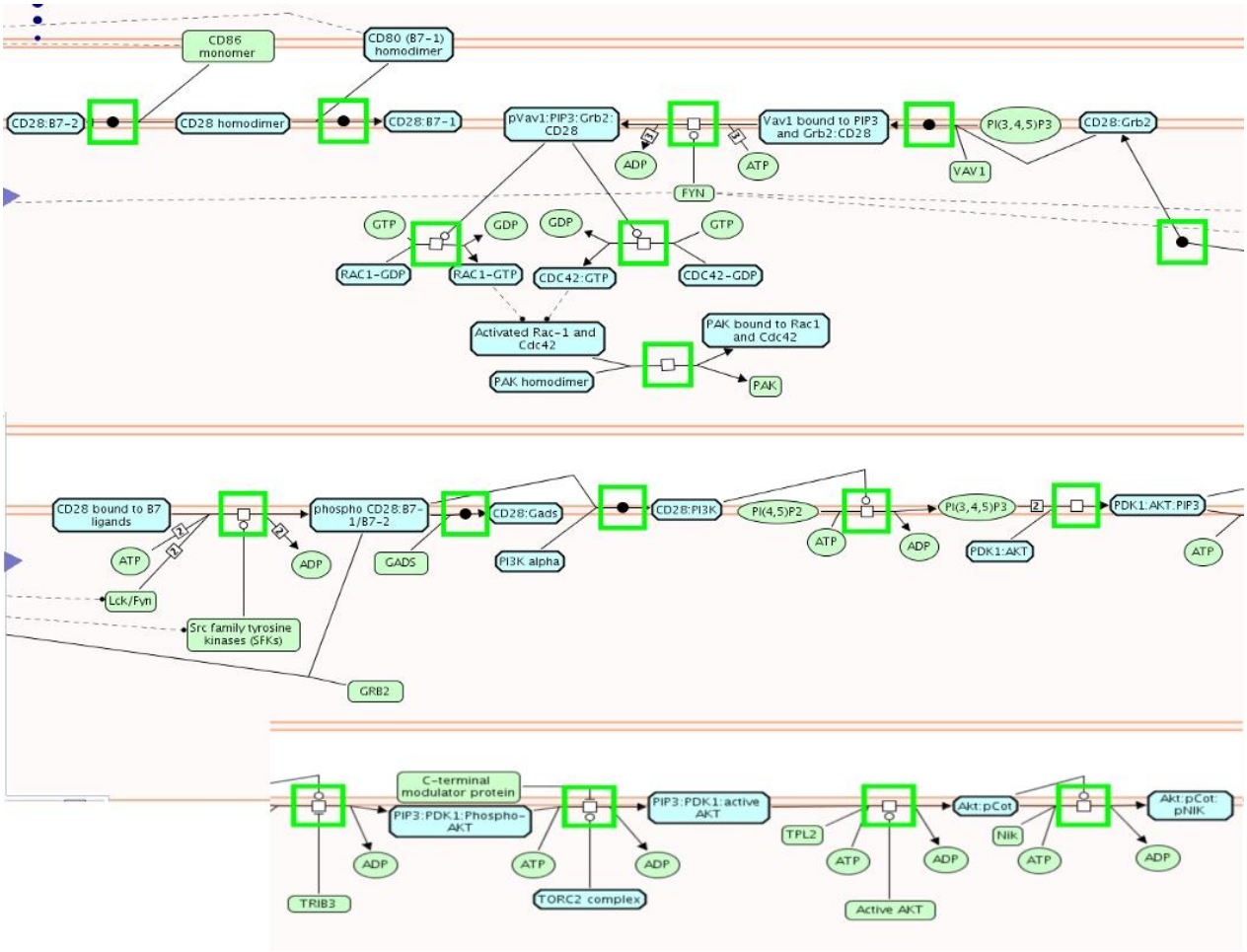


Figura 17. Ruta co-estimulación por CD28. En el esquema de la ruta se puede apreciar que CD28 lleva su función en estado dimérico, lo cual quiere decir que las expresión de esta debe de ser mayor que genes que actúen de forma monomérica, al tener valores de expresión altos (10.3 veces más) CD28 puede asegurar grandes cantidades de proteína en estado dimérico para llevar a cabo sus funciones dentro de la ruta, de la misma manera LCK (10.3 veces de cambio) que tiene función de cinasa al incrementar su expresión, tiene mayor probabilidad de llevar a cabo su función, y como lo indica la estereometría del diagrama, se necesitan dos de esta proteínas. (La imagen se encuentra seccionada en 3 niveles, debido a la extensión como se presenta por parte de Reactome).

Considerando que se contrastaron leucemias agudas linfoblasticas de células T con las de células B, era de esperarse que las rutas de señalización de los receptores de células B y T tengan grupos de genes con expresión disminuida y expresión aumentada respectivamente. En la ruta de señalización de los receptores de células B con 76 genes reportados a la fecha (reportada en KEGG) se reportan 18 genes (23.7 %) con expresión disminuida (MALT1, PIK3AP1, BLNK, IKBKB, INPP5D, LYN, PIK3CD, PIK3R1, PLCG2, PPP3CA, PPP3CC, SYK, BTK, CD19, CD72, CD79A y CD79B) (ver anexo 14 para observar las veces de cambio). Con respecto a la ruta de señalización de los receptores de células T con 109 genes reportados a la fecha (KEGG), se obtuvieron 13 genes (12 %) con expresión aumentada (ICOS, ITK, LCK, LCP2, PRKCQ, ZAP70, CD3D, CD3E, CD3G, CD247, CD8A, CD28 y GRAP2), sin embargo también se observaron 7 genes (6.4 %) con expresión disminuida (MALT1, IKBKB, PDK1, PIK3CD, PIK3R1, PPP3CA y PPP3CC), donde MALT1, IKBKB, PIK3CD, PIK3R1, PPP3CA y PPP3CC (6 de 7) también están implicados en la ruta de señalización de células B. Mostrando estos resultados coherencia con respecto al tipo de células de las que se trataba.

Fuera del contexto de rutas reportadas, encontramos que el gen BCL11B presenta un aumento en su expresión de 25 veces, la única información que se encontró en las bases de datos consultadas es que se reportan 15 interacciones proteína-proteína (CBX5, CHD4, HDAC1, HDAC2, HDAC3, MBD3, MTA1, MTA2, NR2F1, NR2F2, RBBP4, RBBP7, SP1, SUV39H1 y TAL1) y dos tipos de modificaciones post-traduccionales (fosforilación y acetilación). Huang *et al* (2009) reportaron que la supresión de la expresión de BCL11B por siRNA inhibe la proliferación e induce apoptosis en la línea celular Molt-4 de LAL-T, por lo que proponen que la expresión aumentada de BCL11B podría desempeñar un papel anti-apoptótico en LAL-T a través de la regulación de los genes corriente abajo (Figura 18) (84). En los resultado de expresión diferencial BCL2L1 (regulador de apoptosis), SP1 (SPP1) y CREBBP (ruta TGF-beta) (genes corriente abajo) no son reportados (3, 4 y 3 probeset presentes en el microarreglo HG-U133 Plus 2.0 respectivamente). Por otro lado los resultados de interacción proteína-proteína nos reportan que BCL11B (Anexo 23 (BCL11B.txt)) interacciona con SP1 (Anexo 24 (SP1.txt)) y ésta con CREBBP (Anexo 25

(CREBBP.txt)). Huang *et al* (2010) comentan que BCL11B y BCL2L1 tienen las mismas interacciones proteína-proteína e interactúan con BAK y BAX para regular la apoptosis de forma negativa (84), más los resultados de enriquecimiento del gen BCL11B en los resultados de interacción proteína-proteína no reportan a la fecha interacción con BAK o BAX (Anexo 23).

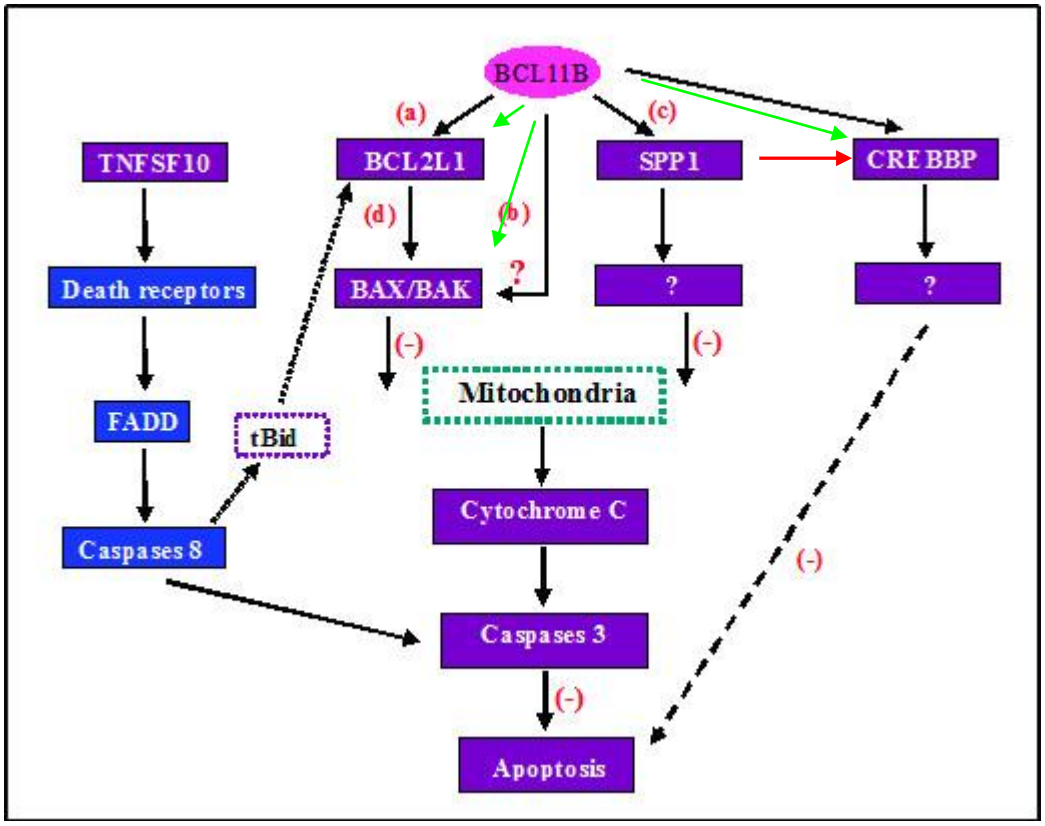


Figura 18. Representación esquemática de la red de regulación de la apoptosis por BCL11B y genes relacionados. (a) BCL2L1 se ve afectada por el gen BCL11B en la regulación transcripcional. (b, d) Según Huang *et al* (2010), BCL11B y BCL2L1 participan en las mismas interacciones proteína-proteína. BCL2L1 normalmente interfiere con la vía mitocondrial de apoptosis al secuestrar proteínas pro-apoptóticas como BAX y BAK1, (aunque los resultados generados por el programa no corroboran esto). (c) El gen SP1 (SPP1) puede ser un blanco en la regulación transcripcional BCL11B, el cual juega un papel consistente en los efectos anti-apoptóticos con el gen BCL11B por la disminución de la liberación del citocromo c mitocondrial. Los resultados que se obtuvieron por el programa MADCBC.py sugieren que BCL11B interactúa con SP1 (SPP1) (Anexo 23) y este con CREBBP (Anexo 24), el cual tiene interacciones con CHUK, IKBKB, RELA, TP53 o IKBKG que forman parte del la ruta de apoptosis (Anexo 25). La flecha roja hace referencia a la interacción proteína-proteína entre SPP1 y CREBBP según los resultados obtenidos por MADCBC.py, las flechas verdes señalan las interacciones que no se demostraron con los resultados de MADCBC.py. (Tomada y modificada de 84).

6.3.2.2 Comparativo con los resultados de la herramienta de minería de datos genómicos DAVID

Se busco comparar las herramientas DAVID y MADCBC.py desde la cantidad de pasos para llegar a los resultados (Figura 19), cobertura de plataformas genómicas y robustez de los resultados arrojados.

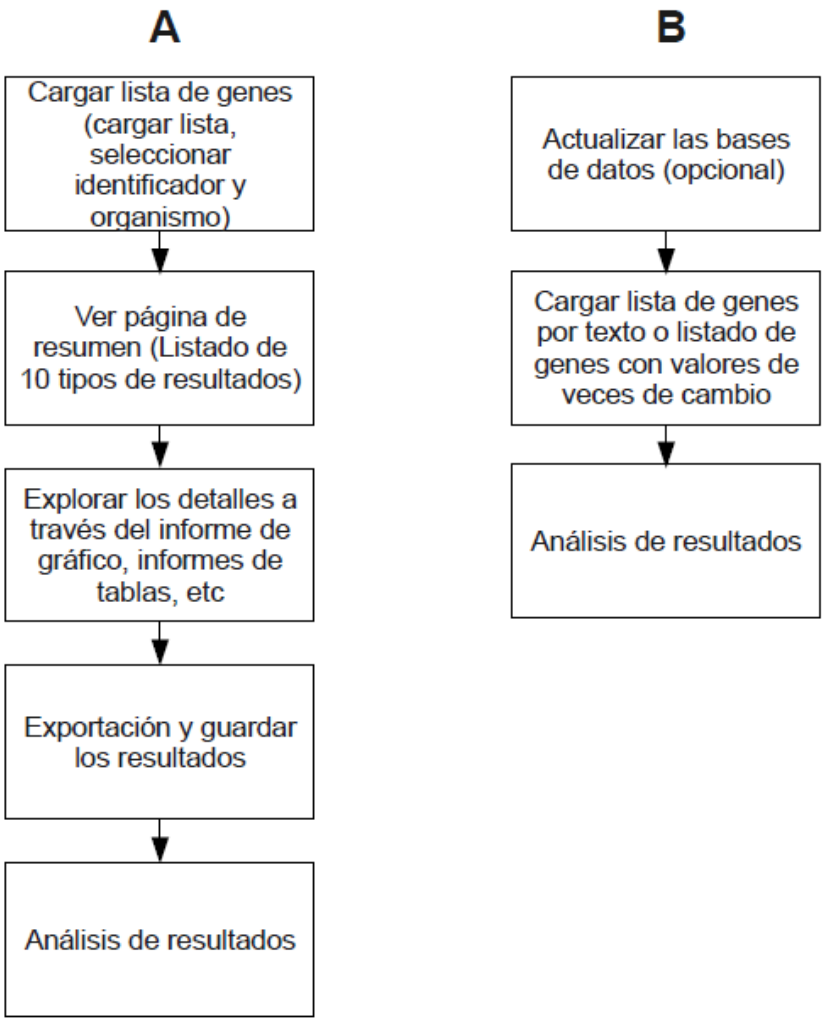


Figura 19. Comparativo de los pasos de MADCBC.py y DAVID. A) En DAVID se tienen 4 pasos previos al análisis de resultados. B. MADCBC.py sólo necesita cargar el listado de genes (a menos que se requiera actualizar las bases de datos) para obtener los resultados a ser analizados. Se puede observar que los pasos de DAVID implican un preanálisis de los resultados, pues en general este tipo de aplicaciones sólo requieren el listado de los genes y el resto es trabajo del usuario mediante análisis de los resultados arrojados.

A continuación se presenta una tabla con los resultados que se obtuvieron con los contrastes de MADCBC.py y DAVID (http://david.abcc.ncifcrf.gov/home.jsp).

Tabla 16. Cuadro comparativo de los resultados obtenidos con las dos herramientas.

	DAVID	**MADCBC.py**
597 probeset +	552 (92.4 %)	532 (89.1 %)
532 probeset *		
Genes con información	530 (99.6 %)	472 (88.7 %)
Rutas reportadas KEGG	31	165
Rutas reportadas Reactome	6	433
Bases de datos de rutas adicionales	4	-
Bases de interacción proteína-proteína	8	9
Información adicional	8 ▲	1 ●

+ Representando los *probeset* no sinónimos. En las columnas 2 y 3 se observan aquellos probeset que se les pudo asignar la terminología oficial y entre paréntesis su rendimiento porcentual.
* El resto de contrastes es a partir de los 532 *probeset* que tenían terminología asignada mediante el Script MA-HUGOaffy.py (Anexo 11) para renombrar a la terminología oficial a partir de los IDs de Affymetrix.
▲ Enfermedades, categorías funcionales, ontología, anotaciones generales, literatura, accesos principales, dominios proteicos y expresión en tejidos.
● Modificaciones post-traduccionales.

Un primer comparativo que se realizó fue el de analizar cuantos identificadores se podían encontrar mediante los dos métodos, esto con el fin de identificar si había perdida de información a este nivel. Con MADCBC.py de los 597 *probeset* hubo 10.9 % (65 *probeset*) que no tenían una terminología oficial asignada, por el contrario con DAVID, 7.6 % (45 *probeset*) no tenían una terminología oficial asignada.

Posteriormente se analizaron los resultados a partir de los 532 *probeset* que se les asigno un nombre mediante el *script* MA-HUGOaffy.py, donde se aprecia que MADCBC.py sólo reporta para 472 genes al menos un tipo de información y no teniendo reportes para 60 genes, mientras que DAVID sólo hay 2 genes que no reportan resultado alguno. En cuanto a los resultados de rutas de señalización se muestra una sensible baja en los obtenidos por parte de DAVID, esta herramienta reporta 31 rutas

para KEGG y 6 rutas para Reactome; mientras que en MADCBC.py las mismas bases tienen 165 y 433 rutas, esto se debe a que DAVID utiliza una prueba de Fisher modificada para mostrar solo aquellos resultados que tienen poca probabilidad de haber sido encontrados por azar. Es de resaltar que DAVID provee de 10 tipos de plataformas genómicas y tipos de bases de datos adicionales (enfermedades, categorías funcionales, ontología, anotaciones generales, literatura, bases de datos principales, dominios proteicos, expresión en tejidos, además de interacciones proteína-proteína y rutas de señalización), que sumado a los valores de p que ofrece en todos sus resultados, así como hipervínculos a otras bases de datos. A diferencia de DAVID, MADCBC.py ofrece todos los resultados encontrados, pues hasta la fecha no tiene estadísticos para restringir dichos resultados, así mismo esta herramienta muestra en cuanto a las plataformas genómicas de modificaciones post-traduccionales e interacción proteína-proteína resultados a partir de una base conjunta de todas las bases de datos colectadas, a diferencia de DAVID donde muestra los resultados con base a cada una de las bases de datos de las que se alimenta.

Tomando las conclusiones que proponen Huang *et al* (2010) referente a que la expresión aumentada de BCL11B podría desempeñar un papel anti-apoptótico en LAL-T, se procedió a analizar los resultados de los 532 genes introducidos a DAVID, así como de los genes CREBBP y SP1 (genes río abajo de BCL11B) de forma individual. Los resultados de DAVID concuerdan con lo encontrado por MADCBC.py, donde no se reportan las interacciones de BCL11B con BCL2L1, BAK1, BAX ni CREBBP que los autores sugieren, sin embargo si se reporta interacción con SP1. La única diferencia entre las herramientas DAVID y MADCBC.py con respecto a estos resultados es la interacción de SP1 con CREBBP, donde DAVID no la reporta. Dicha interacción esta reportada en la base de datos BioGrid (63), la cual no se encuentra entre las bases con las que se alimente DAVID, dicha interacción la reportan Goto *et al* (2006) y Leclerc *et al* (2010) mediante un ensayo de captura por afinidad mediante Western (85, 86).

7 DISCUCIÓN

El fundamento del método de minería de datos genómicos utilizado para desarrollar el proyecto MADCBC.py fue empleado anteriormente por Huttenhower C. y Hofmann O. (2010) quienes reunieron diversos tipos de plataformas genómicas con el fin de generar investigación a gran escala tomando como modelo biológico de estudio la levadura de cerveza (*Saccharomyces cerevisiae)*. Herramientas como DAVID (46, 47) han empleado el enfoque de reunir información a partir de diversas plataformas genómicas desde el 2003, este tipo de herramientas se sustentan en conjuntar información mediante el uso de entornos de creación de *scripts* (R, *python*, Perl , rubi, etc.), procurando ejecutarlos de forma eficiente para evitar tiempos excesivos de procesamiento, así como el empleo de métodos de almacenamiento sencillos (texto plano, XML, etc.) para evitar grandes cantidades de espacio tras la acumulación de datos (38). El proyecto MADCBC.py emplea para dichos fines el lenguaje de programación P*ython* para la generación de los *scripts*, y para el almacenamiento de la conjunción resultante de las bases de datos emplea archivos con formato en texto plano, que contienen la información mínima necesaria para obtener resultados donde se correlacionen los genes o proteínas con expresión diferencial a las rutas de señalización en las que participen, así como con los datos de interacción proteína-proteína y de modificaciones post-traduccionales.

Para el desarrollo de esta tesis, después de desarrollar el proyecto MADCBC.py se hizo una prueba analizando datos provenientes de algunos subtipos de leucemias; sin embargo, el programa se puede emplear para datos obtenidos con otro enfoque biológico, ya que este tipo de herramientas ayudan a sustentar y/o generar nuevas hipótesis a través de enriquecer la información acerca de los genes o proteínas consultadas (que pueden incluir resultados de expresión diferencial de microarreglos, geles en 2D, o simplemente un listado de genes de interés) provenientes de diversos tipos de investigaciones biológicas. Debido a la complejidad biológica, los programas de minería de datos en el estado actual, así como el análisis de grandes listas de genes con herramientas bioinformáticas son todavía un procedimiento exploratorio de los resultados y no una solución estadística pura, como sería deseable. Las mejores

conclusiones analíticas se realizan mediante el conocimiento biológico del investigador, bases de datos integradas, algoritmos de cálculo y el enriquecimiento de los valores de certidumbre estadísticos (47). Si bien los resultados aportados por MADCBC.py no poseen a la fecha un valor de certidumbre estadística que advierta de posibles falsos positivos, esta herramienta aporta, mediante la correlación de los diferentes datos obtenidos, información valiosa que el investigador debe de considerar en conjunto con lo reportado en la literatura o mediante la implementación de experimentos que corroboren los resultados o las nuevas hipótesis generadas.

Actualmente el programa generado para el desarrollo de esta tesis, MADCBC.py, recolecta 4 diferentes tipos de información depositada en las bases de datos (interacción proteína-proteína, modificaciones post-traduccionales, rutas de señalización y terminologías) y lo correlaciona con genes o proteínas resultantes de análisis de expresión diferencial (a partir de cualquier tecnología que pueda discernir entre dos condiciones) o simplemente de un listado de genes de interés, esto sólo para el organismo *Homo sapiens.* En la actualidad los diferentes programas con enfoques similares como GAGGLE (45), DAVID (46, 47) o BABELOMICS (56), son herramientas que tienen entre sus características el manejo de información de múltiples organismos, así como diversos tipos de información genómica. Tanto la inclusión de diversos organismos (mediante la sistematización de terminologías), así como la inclusión de otro tipo de información genómica (regulación, subunidades proteicas, etc.) son parte de las perspectivas de este proyecto.

Las investigaciones individuales a escala genómica por lo general resultan complejas y de difícil interpretación por sí mismas y propensas a errores (87), y con el uso cada vez más común de tecnologías que disminuyen costos, así como el desarrollo de diferentes tecnologías “ómicas”, se tiene como resultado un gran cúmulo de datos. En este sentido el análisis bioinformático es un paso limitante para el uso fructífero de los datos obtenidos (88), por lo que es necesario el desarrollo de programas que permitan la integración y minería de las grandes cantidades de datos “ómicos” heterogéneos (89), MADCBC.py tiene el potencial de contribuir junto con otras herramientas de este tipo a aumentar la exploración, administración y el análisis de los datos.

Uno de los pasos fundamentales de este tipo de herramientas para una apropiada administración y procesamiento de los datos es la homogeneización de las terminologías empleadas. Bajo el abordaje de este proyecto, la terminología de los datos de partida (que se encontraban bajo las terminologías de Uniprot, y los identificadores de Affymetrix) fue intercambiada por los símbolos oficiales propuestos por HGNC (68). La integración facilitada por la homogeneización de los datos es un requisito previo para mejorar la eficacia de la extracción y el análisis de la información biológica.

Los objetivos de la integración de datos son la formulación de consultas, informes y análisis complejos (análisis multidimensional y minería de datos). Sin embargo, la complejidad de los datos biológicos representan un desafío mayor en la integración de bases de datos de biología molecular, y que requiere una participación activa de los investigadores para su interpretación (88). Es por ello que se buscará que MADCBC.py integre más bases de datos para robustecer los datos biológicos con los que se cuenta a la fecha, así como de otros tipos de datos (ontología, regulación, literatura, etc.). Además se planea integrar valores estadísticos, lo que ayudará al usuario a interpretar los resultados obtenidos, de manera que tendrá el potencial de responder preguntas más complejas sobre los procesos biológicos que ocurren en las células bajo determinadas condiciones.

La investigación biológica futura, con la integración de los componentes computacionales y experimentales, permitirá la generación de nuevas hipótesis y propiciará el desarrollo de métodos experimentales para ponerlas a prueba. Los datos resultantes a su vez, serán utilizados para generar modelos más refinados que permitan mejorar la comprensión global de los mecanismos celulares (88). Uno de los grandes nichos de oportunidad de proyectos similares a MADCBC.py es la biología modular (procesos biológicos de interés, o módulos, que estudian la interacción compleja de macromoléculas que participan en un proceso dado), donde la integración de los datos actualmente disponibles de diversas fuentes nos ayuda a obtener una comprensión más precisa y completa de los procesos celulares (90, 91). Por lo tanto, la

integración de datos es deseable para poder modelar redes celulares con más precisión y para hacer inferencias funcionales (90).

Así mismo el presente proyecto contribuye con su enfoque integrador a la minería de datos genómicos a gran escala desde dos de las tres categorías en las que los agrupan Huttenhower C. y Hofmann O. (2010):

A nivel de aplicaciones de programación, MADCBC.py se presenta como una herramienta que tiene una gran cobertura en las bases de datos disponibles de las que se obtiene la información: 9 bases de datos primarias de interacción proteína-proteína y 3 bases de datos primarias de modificaciones post-traduccionales (todas las bases reportadas en http://www.pathguide.org (12) y que cumplieron los criterios de inclusión), teniendo con esto un gran respaldo en información, así mismo se sustenta en dos de las bases de datos más grandes en cuanto a rutas de señalización, Reactome (61) y Kegg (41).

Por otra parte bajo el enfoque de herramientas implementadas por el investigador, que ha tenido un desarrollo modesto en los últimos años (ver figura 7), MADCBC.py es una herramienta que fomenta este tipo de investigación, ya que una de las partes esenciales del programa es la de conjuntar la información desde diferentes bases de datos, homogeneizar terminologías y generar archivos de cada uno de los tipos de información. Mediante este procedimiento se generan 4 archivos, dos de ellos que compilan información de diversas bases de datos (interacción proteína-proteína y modificaciones post-traduccionales, IPP.txt y MODIFICACIONES.txt respectivamente), y dos más de rutas de señalización (REACTOME.txt y KEGG.txt) todos ellos almacenados en formato de texto plano. Cada uno de estos archivos puede ser empleados por los investigadores de forma individual o en conjunto (de manera independiente al programa MADCBC.py) para desarrollar investigación a gran escala sin tener que ocupar tiempo de investigación en recolectar dicha información. En la medida que se implementen otros tipos de datos, su potencial se verá maximizado. Así mismo se puede mencionar que el archivo IPP.txt puede ser empleado por el software

de código abierto para la visualización de redes e integración de datos Cytoscape (92), donde uno de los medios para robustecer los resultados obtenidos por este programa es mediante la introducción de un archivo de interacción proteína-proteína con el formato de salida del programa MADCBC.py.

Al momento de comparar MADCBC.py con una herramienta similar (DAVID) se evidenciaron grandes diferencias en cuanto a los alcances de estas, sobretodo en cuanto a los diversos tipos de plataformas genómicas y bases de datos que emplea DAVID (enfermedades, categorías funcionales, ontología, anotaciones generales, literatura, bases de datos principales, dominios proteicos, interacciones proteína-proteína y rutas de señalización, mientras que MADCBC.py sólo emplea 3 plataformas genómicas, interacciones proteína-proteína, modificaciones post-traduccionales y rutas de señalización). Algo que resaltó fue la perdida de *probeset* al utilizar el *script* MA-HUGOaffy.py, ya que algunos genes no tienen asignada un terminología oficial. Se debe tomar en cuenta que para renombrar a la terminología oficial los IDs de Affymetrix, el *script* MA-HUGOaffy.py empleó en el momento de desarrollo del programa MADCBC.py el archivo de anotaciones HG-U133_Plus_2.na31.annot.csv proveído por Affymetrix, sin embargo a la fecha de la escritura de este documento existe una versión más reciente de dicho archivo (la versión 32, HG-U133_Plus_2.na32.annot.csv).

Es de notar que gracias a la robustez en la cantidad de bases de datos de interacción proteína-proteína con las que se alimenta el programa MADCBC.py y al analizar los datos obtenidos de los diferentes microarreglos provenientes de pacientes pediátricos con LAL se pudo identificar como potencialmente relevante la interacción entre las proteínas PS1 y CREBBP. Esta interacción no se reportó en los resultados obtenidos con DAVID, a pesar de que posiblemente sea uno de los pasos requeridos en el efecto anti-apoptótico que se promueve debido al aumento en la expresión de BCL11B en las LALp subtipo T (84). De esta manera, el análisis de los datos con el programa MADCBC.py sugiere que sería importante determinar experimentalmente las posibles implicaciones en la vía de la apoptosis de la interacción entre PS1 y CREBBP y de su papel el proceso de carcinogénesis en este subtipo de leucemia.

Si bien el programa MADCBC.py todavía requiere de implementar diversas mejoras, como valores estadísticos, este proyecto se está desarrollando en un periodo que requiere de la generación de herramientas de minería de datos multientrada, las cuales buscan integrar la gran diversidad de plataformas genómicas. Esta integración permitirá correlacionar la información generada en los centros de investigación de biología molecular y aumentar los resultados obtenidos para generar nuevas hipótesis así como resultados sólidos, en la medida que este tipo de enfoques vayan tomando peso estadístico y mayor certidumbre biológica.

Aunque el programa MADCBC.py esta diseñado para colectar información referente a las modificaciones post-traduccionales, dicha información no tiene mucha relevancia en el estado actual del programa, sin embargo cobrará importancia en la medida que el programa se actualice para que integre información a nivel de secuencia de aminoácidos, provenientes de datos proteómicos, o bien de *splicing* alternativo que hacen referencia a las diferentes isoformas de las proteínas.

En general, podemos concluir que el programa MADCBC.py, así como el desarrollo de diversas herramientas de minería de datos a gran escala, son de gran ayuda para la comprensión y administración de la información genómica generada. Además, conforme este tipo de información se genere de forma más ágil por el abaratamiento de las tecnologías de alto rendimiento, así como el desarrollo de nuevas tecnologías de mayor capacidad de procesamiento de muestras, la demanda de herramientas de genómica computacional será cada vez mayor para el análisis de la información obtenida, así como aumentar las posibilidades de encontrar respuestas complejas a preguntas de orden biológico.

8 CONCLUSIONES

- El funcionamiento óptimo del programa requiere actualización continua de los datos, de tal forma que es importante que esta se lleve a cabo de manera automática, por lo que en el programa existe la opción de hacerlo.

- Una parte importante del proyecto es la integración que le dará el usuario a los resultados arrojados, pues es necesario que los resultados sean depurados para tener conclusiones apegadas a la realidad biológica del evento estudiado.

- Es necesario integrar diversos tipos datos genómicos al programa, para robustecer los resultados arrojados, para que se generen respuestas acordes a la complejidad celular.

- El programa desarrollado generó resultados que concuerdan con el modelo de estudio empleado (LAL-T). Se encontró evidencia que sugiere que el camino para inhibir la apoptosis por una expresión aumentada de BCL11B (modelo de Huang *et al* (2010)) es mediante la interacción de este con SP1 y de SP1 con CREBBP, y que concluye con la unión de CREBBP con una o varias de las proteínas de la ruta de apoptosis, inhibiendo su acción.

9 PERSPECTIVAS

9.1 A nivel del programa desarrollado

- Análisis estadísticos. Uno de los puntos pendientes en el programa es la certidumbre estadística que se le puede dar al investigador al interpretar los resultados obtenidos, de tal forma que es necesario implementar el cálculo de valores de certidumbre para los diversos resultados.

- Resumen de resultados. Los resultados se reportan para cada gen o para las rutas afectadas, se plantea que para actualizaciones futuras se genere una tabla de resumen de dichos resultado de tal forma que se identifiquen genes (que muestren expresión diferencial) que no cuenten con resultados (sin interacción proteína-proteína, modificaciones post-traduccionales ni rutas de señalización), de esta manera se puedan incluir para análisis específicos que el usuario pueda implementar (BLAST, análisis de co-expresión, etc.).

- Resultados gráficos. Se planea implementar un resultado gráfico que facilite la interpretación de los resultados obtenidos.

- Vínculos a los principales sitios Web sobre biología molecular. Uno de los objetivos del programa es ofrecer resultados resumidos, de tal forma que en la medida que el usuario requiera profundizar en ciertos resultados pueda acceder desde las tablas de resultados a las grandes bases de datos (NCBI, Ensambl, UCSC Genome Browser, etc.), mediante hipervínculos a dichas bases.

- Diversos organismos. La versión presente del programa ofrece sólo resultados para Homo Sapiens, esto en gran medida debido a que la información de este organismo esta más sistematizada. En la medida que se estandarice la información para otros organismos (tanto a nivel de bases de datos como en el

procesamiento de estas por parte del programa) se podrá ampliar la cobertura del programa a diversos organismos.

- Entorno gráfico. En la medida en que el programa se vuelva más complejo se requerirá que sea más interactivo, por ello sería deseable desarrollar un entorno gráfico a fin de que los usuarios menos familiarizados con la línea de comandos puedan emplear el programa con comodidad.

- La información genómica (que se complementa con la información proteómica) se encuentra en diversos formatos, en la medida que el proyecto siga desarrollándose se deberán agregar más bases de datos como fuente de información (ontologías, interacción con moléculas no proteicas, factores de transcripción (y su secuencia blanco), niveles de expresión de MicroRNAs, datos de splicing alternativo) para resultados integrales.

9.2 A nivel de aplicación al modelo de estudio

- Para este trabajo sólo se procesó la información del subtipo LAL-T (enriquecimiento de la información de los genes y grado de afectación de las rutas de señalización) con el programa MADCBC.py, sin embargo se cuenta con los resultados de expresión diferencial del resto de los subtipos, que serán procesados en el futuro.

- En los resultados de expresión diferencial de LAL-T se tienen 60 genes sin resultados en el enriquecimiento y 65 probeset que representan a EST que no tienen nombre ni información asignada y que siendo característicos del subtipo LAL-T son substrato de análisis a nivel de co-expresión y a nivel de secuencia (realizar BLAST, identificar dominios proteicos), si bien, dicho análisis está fuera de los alcances del proyecto, se puede llevar a cabo un análisis específico de estos probeset encontrados.

10 REFERENCIAS BIBLIOGRAFICAS

1. Murray A. Whither genomics?. *Genome Biology.* (2000) 1(1):003.1–003.6
2. Pearsons L, Soll D. The Human Genome Project: a paradigm for information management in the life sciences. FASEB J. (1991)5: 35-39
3. Understanding Our Genetic Inheritance. The U.S. Human Genome Project: The First Five Years FY 1991-1995. *National Technical Information Service, US. Dept. of Commerce. (1990) DOE/ER-0452P*
4. Collins F, Patrinos A, Jordan E, Chakravarti A, Gesteland R, Walters L y los miembros del DOE y el grupo de planificación del NIH. New Goals for the U.S. Human Genome Project: 1998 –2003. (1998) 282:682-689
5. Lee J, Williams P, Cheon S. Data Mining in Genomics. Clin Lab Med. (2008) 28(1): 145–viii
6. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA (1998) 95:14863-14868
7. Ball CA, Awad IA, Demeter J, et al. The Stanford Microarray Database accommodates additional microarray platforms and data formats. Nucleic Acids Res (2005) 33:D580-D582
8. Rocca-Serra P, Brazma A, Parkinson H, Sarkans U, Shojatalab M, Contrino S, Vilo J, Abeygunawardena N, Mukherjee G, Holloway E, et al. ArrayExpress: a public database of gene expression data at EBI. C R Biol (2003) 326(10-11):1075-1078
9. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res (2009) 37:D885–D890
10. Lee H, Hsu A, Sajdak J, Qin J, Pavlidis P. Coexpression Analysis of Human Genes Across Many Microarray Data Sets. *Genome Res.* (2004) 14:1085-1094
11. Campain A, Yang Y. Comparison study of microarray meta-analysis methods. Bioinformatics (2010) 11:408

12. Bader G, Cary M, Sander C. Pathguide: a Pathway Resource List. Nucleic Acids Research (2006) 34: D504–D506

13. Robles Porcada Víctor. Clasificación Supervisada basada en Redes Bayesianas. Aplicación en Biología Computacional. (2003). Tesis de Doctorado en Informática

14. Schena, M., Shalon, D., Davis, R.W. y Brown P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (1995) 270:467-470.

15. Silvia Saviozzi, Giovanni Iazzetti, Enrico Caserta, Alessandro Guffanti, and Raffaele A. Calogero. Microarray Data Analysis and Mining. Methods in Molecular Medicine, vol. 94: Molecular Diagnosis of Infectious Diseases, 2e Edited by: J. Decker and U. Reischl © Humana Press Inc., Totowa, NJ

16. Naef, F., Lim, D. A., Patil, N. y Magnasco, M. O. From features to expression: High density oligonucleotide array analysis revisited. Tech Report (2001) 1:1–9.

17. Hartemink, D. G., Jaakkola, I., and Young, R. Maximum likelihood estimation of optimal scaling factors for expression array normalization. Microarrays: Optical Technologies and Informatics (Proceedings of SPIE), (2001) p. 4266

18. Rocke, D. M. and Durbin, B. A model for measurement error for gene expression arrays. J. Comput. Biol. (2001) 8:557–569.

19. Freudenberg J.M. Comparison of background correction and normalization procedures for high-density oligonucleotide microarrays. Technical Report 3, Leipzig Bioinformatics Working Paper. 2005.

20. Gentleman R, Carey V, Huber W, Irizarry R y Dudoit S Bioinformatics and computational biology solutions using R and Bioconductor: New York: Springer. (2005) p 20

21. Quackenbush J. Computational analysis of microarray data. Genetics (2001) 2:418 – 427

22. Feten G., Aastveit AH., Snipen L. y Almoy T. A discussion concerning the inclusion of variety effect when analysis of variance is used to detect differentially expressed genes. Gene Regulation Systems Biol. (2007) 1:43-47.

23. Kohonen, T. Self Organizing Maps. Springer, Berlin. (1995)

24. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. y Church, G. M. Systematic determination of genetic network architecture. Nature Genet. (1999) 22:281–285.

25. Kerr MK, Martin M, Churchill GA: Analysis of variance for gene expression microarray data. J Comput Biol. (2000) 7:819-837

26. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response**.** Proc Natl Acad Sci. (2001) 98(9)**:**5116-5121

27. Smyth, G. K. Linear models and empirical Bayes methods for assessing dierential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology. (2004) Vol. 3, No. 1, Article 3.

28. Galperin M. Y. y Cochrane G. R. The 2011 *Nucleic Acids Research* Database Issue and the online Molecular Biology Database Collection. Nucl. Acids Res. *(2011) 39: D1-D6*

29. Cary M.P., Bader G.D. y Sander,C. Pathway information for systems biology. FEBS Lett. (2005) 579:1815–1820.

30. Normand SL. Meta-analysis: formulating, evaluating, combining, and reporting. Statistics in Medicine (1999) 18:321-359

31. Choi JK, Yu U, Kim S, Yoo OJ. Combining multiple microarray studies and modeling interstudy variation. Bioinformatics (2003) 19(Suppl 1):84-90

32. Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics (2008) 24:374-382

33. Gentleman R, Ruschhaupt M, Huber W, Lusa L. Meta-analysis for microarray experiments. Bioconductor (2008)

34. Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E. A cross-study comparison of gene expression studies for the molecular classification of lung cancer. Clinical Cancer Research (2004) 10:2922-2927

35. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Letters (2004) 573:83-92

36. Hong F, Breitling R, McEntee C, Wittner B, Jennifer L. Nemhauser J, Chory J. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. BMC Bioinformatics (2006) 22:2825–2827

37. Fayyad, U. M., Piatetsky-Shapiro, G., y Smyth, P. From Data Mining to Knowledge Discovery in Databases. AI Magazine (1996) 17(3):37-54.

38. Huttenhower C, Hofmann O. A Quick Guide to Large-Scale Genomic Data Mining. PLoS Comput Biol (2010) 6(5):e1000779

39. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A (2003) 100(16):9440–5

40. Hastie T, et al. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol (2000) 1(2):RESEARCH0003

41. Soukup M, Cho H, Lee JK. Robust classification modeling on microarray data using misclassification penalized posterior. Bioinformatics (2005) 21:i423–i430

42. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome Nucleic Acids Res (2004) 32:D277–D280

43. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. STRING 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res (2009) 37:D412–D416

44. Haider S, Ballester B, Smedley D, Zhang J, Rice P, et al. BioMart Central Portal–unified access to biological data. Nucleic Acids Res (2009) 37: W23–W27.

45. Shannon PT, Reiss DJ, Bonneau R, Baliga NS. The Gaggle: an open-source software system for integrating bioinformatics software and data sources. BMC Bioinformatics (2006) 7: 176

46. Huang DW, Sherman BT y Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protoc. (2009) 4(1):44–57*

47. Huang DW, Sherman BT y Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res. (2009) 37(1):1–13*

48. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE.. Integrating large scale functional genomic data to dissect the complexity of yeast regulatory networks. Nature Genetics (2008) 40: 854–861

49. Medina I, Carbonell J, Pulido L, Madeira S, Goetz S, Conesa A, Tárraga J, Pascual-Montano A, Nogales-Cadenas R, Santoyo J, García F, Marbá M, Montaner D y Dopazo J. Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Research. (2010) 38 : W210–W213

50. Acute lynphocytic leukemia. Publicación de The Leukemia and Lynphoma Society (2008) pp 6

51. Paredes R. Leucemias agudas. En: El niño con cáncer. (M. Uriel ed.) Editores de textos mexicanos. México D.F. (2007)

52. Chico P. Principios generales de la epidemiología del cáncer pediátrico. En: Oncología pediátrica. (L. Velásquez ed.) Ed. Intersistemas. México D.F. (2002)

53. Castillo V, Pérez P, Blanco B. Genética y cáncer en pediatría. En: Oncología pediátrica. (Velásquez L. ed.) Ed. Intersistemas. México D.F. (2002)

54. Yeoh E, Ross M, Shurtleff S, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer cell. (2002) 1:133-143

55. Ross ME, Zhou X, Song G, Shurtleff SA, Girtman K, Williams WK, Liu HC, Mahfouz R, Raimondi SC, Lenny N, Patel A, Downing JR. Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood. (2003) 102:2951-2959

56. Den Boer M, Slegtenhorst M, De Menezes M, Cheok M, Buijs-Gladdines J, Peters S, Van Zutven L, Beverloo B, Van der Spek P, Escherich G, Horstmann M, Janka-Schaub G, Kamps W, Evans W, y Pieters R. A subtype of childhood acute

lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. Lancet Oncol (2009) 10(2): 125–134

57. Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L y Cesareni G. MINT, the molecular interaction database: 2009 update. Nucleic Acids Res. (2010) 38:D532-9.

58. Chautard E, Ballut L, Thierry-Mieg N, Ricard-Blum S., MatrixDB, a database focused on extracellular protein protein and protein carbohydrate interactions, Bioinformatic.s (2009) 1;25(5):690-1.

59. Ceol A, Chatr-aryamontri A, Santonico E, Sacco R, Castagnoli L. y Cesareni G. DOMINO: a database of domain–peptide interactions *(2007) 35:D557-D560*

60. Ferrari E, Tinti M, Costa S, Corallino S, Pio Nardozza A, Chatraryamontri A, Ceol A, Cesareni A, y Castagnoli L. Identification of New Substrates of the Protein-tyrosine Phosphatase PTP1B by Bayesian Integration of Proteome Evidence. Genomics and Proteomics (*2011) 286:4173-4185.*

61. Matthews L, Gopinath G, Gillespie M, Caudy M, Croft D, de Bono B, Garapati P, Hemish J, Hermjakob H, Jassal B, Kanapin A, Lewis S, Mahajan S, May B, Schmidt E, Vastrik I, Wu G, Birney E, Stein L y D'Eustachio P. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. *(2008) 37:D619-D622.*

62. Xenarios I, Salwínski L, Duan XJ, Higney P, Kim SM, Eisenberg D. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res. (2002) 1;30(1):303-5

63. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bähler J, Wood V, Dolinski K, Tyers M. The BioGRID Interaction Database: 2008 update. Nucleic Acids Res. (2008) 36:D637-40

64. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian A, Kerrien S, Khadake J, Kerssemakers J, Leroy C, Menden M, Michaut M, Montecchi-Palazzi L, Neuhauser N, Orchard S, Perreau V, Roechert B, van Eijk K, y Hermjakob H. The IntAct molecular interaction database in 2010 *(2010) 38:D525-D531*

65. *Prasad TS, et al. Human Protein Reference Database—2009 update. Nucleic Acids Res. (2008) 37: D767-D772*

66. Hornbeck PV, Chabra I, Kornhauser JM, Skrzypek E, Zhang B. PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics. (2004) 4(6):1551-61

67. *Diella F, Gould CM, Chica C, Via A, Gibson TJ. Phospho.ELM: a database of phosphorylation sites--update 2008. Nucleic Acids Res. (2008) 36:D240-4.*

68. *Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA.* genenames.org: the HGNC resources in 2011. *Nucleic Acids Res. (2011) 39:D519-9*

69. *Data Sheet. GeneChip Human Genome U133 Arrays. Affymetrix.*

70. *Kang H, Chen IM, Wilson CS, Bedrick EJ et al. Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia. Blood. (2010) 18;115(7):1394-405*

71. Bhojwani D, Kang H, Menezes RX, Yang W et al. Gene expression signatures predictive of early response and outcome in high-risk childhood acute lymphoblastic leukemia: A Children's Oncology Group Study. J Clin Oncol (2008) 26(27):4376-84.

72. Campo Dell'Orto M, Zangrando A, Trentin L, Li R et al. New data on robustness of gene expression signatures in leukemia: comparison of three distinct total RNA preparation procedures. BMC Genomics (2007) 22;8:188.

73. Van Vlierberghe P, van Grotel M, Tchinda J, Lee C et al. The recurrent SET-NUP214 fusion as a new HOXA activation mechanism in pediatric T-cell acute lymphoblastic leukemia. Blood (2008) 1;111(9):4668-80.

74. Bungaro S, Dell'Orto MC, Zangrando A, Basso D et al. Integration of genomic and gene expression data of childhood ALL without known aberrations identifies subgroups with specific genetic hallmarks. Genes Chromosomes Cancer (2009) 48(1):22-38.

75. Zangrando A, Dell'orto MC, Te Kronnie G, Basso G. MLL rearrangements in pediatric acute lymphoblastic and myeloblastic leukemias: MLL specific and lineage specific signatures. BMC Med Genomics (2009) 23;2:36.

76. Winter SS, Jiang Z, Khawaja HM, Griffin T et al. Identification of genomic classifiers that distinguish induction failure in T-lineage acute lymphoblastic leukemia: a report from the Children's Oncology Group. Blood (2007) 1;110(5):1429-38.

77. Fulci V, Colombo T, Chiaretti S, Messina M et al. Characterization of B- and T-lineage acute lymphoblastic leukemia by integrated analysis of MicroRNA and mRNA expression profiles. Genes Chromosomes Cancer (2009) 48(12):1069-82.

78. Hertzberg L, Vendramini E, Ganmore I, Cazzaniga G et al. Down syndrome acute lymphoblastic leukemia, a highly heterogeneous disease in which aberrant expression of CRLF2 is associated with mutated JAK2: a report from the International BFM Study Group. Blood (2010) 4;115(5):1006-17.

79. Staal FJ, de Ridder D, Szczepanski T, Schonewille T et al. Genome-wide expression analysis of paired diagnosis-relapse samples in ALL indicates involvement of pathways related to DNA replication, cell cycle and DNA repair, independent of immune phenotype. Leukemia (2010) 24(3):491-9.

80. Stam RW, Schneider P, Hagelstein JA, van der Linden MH et al. Gene expression profiling-based dissection of MLL translocated and MLL germline acute lymphoblastic leukemia in infants. Blood (2010) 8;115(14):2835-44.

81. Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol (2004) 5:R80

82. Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, et al. TM4 microarray software suite. Methods in Enzymology. (2006) 411:134-93

83. Michel F y Acuto O. CD28 costimulation: a source of Vav-1 for TCR signaling with the help of SLP-76? Sci STKE. (2002) 144:pe35.

84. Huang X, Chen S, Shen Q, Yang L, Li B, Zhong L, Geng S, Du X y Li Y. Analysis of the expression pattern of the BCL11B gene and its relatives in patients with T-cell acute lymphoblastic leukemia. J Hematol Oncol. (2010) 16;3(1):44.

85. Goto T, Matsui Y, Fernandes RJ, Hanson DA, Kubo T, Yukata K, Michigami T, Komori T, Fujita T, Yang L, Eyre DR y Yasui N. Sp1 family of transcription factors regulates the human alpha2 (XI) collagen gene (COL11A2) in Saos-2 osteoblastic cells. J. Bone Miner. Res. (2006) 21:661-73

86. Leclerc GJ, Mou C, Leclerc GM, Mian AM y Barredo JC. Histone deacetylase inhibitors induce FPGS mRNA expression and intracellular accumulation of long-chain methotrexate polyglutamates in childhood acute lymphoblastic leukemia: implications for combination therapy. Leukemia (2010) 24(3):552-62

87. Yu, U., Lee, S.H., Kim, Y.J. and Kim, S. Bioinformatics in the post-genome era. J. Biochem. Mol. Biol. (2004) 37:75-82.

88. Gir Won Lee & Sangsoo Kim. Genome data mining for everyone. BMB Rep. (2008) 41(11):757-764

89. Qi Y, Ge H. Modularity and dynamics of cellular networks. PloS Comput Biol. (2006) 2(12): e174

90. Ge H, Walhout AJ, Vidal M. Integrating ‘‘omic’’ information: A bridge between genomics and systems biology. Trends Genet. (2003) 19: 551–560.

91. Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han J-DJ, et al. Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis. (2005) 436: 861–865.

92. Smoot M, Ono K, Ruscheinski J, Wang PL y Ideker T. Cytoscape 2.8: new features for data integration and network visualization Bioinformatics. (2011) 1; 27(3): 431–432

11 ANEXOS

11.1 **Anexo 1 (hugo.py).** Script para obtener el diccionario de la terminología oficial.

```
print 'ACTUALIZANDO EL DICCIONARIO DE TERMINOLOGIA'
def estadodescarga(blocknum, bs, size):
        estadodescargado = blocknum * bs
        print "Cantidad descargado: %s Bytes de %s Bytes totales" % (estadodescargado, size)
web = 'http://www.genenames.org/cgi-bin/hgnc_downloads.cgi?title=HGNC+output+data&submit=submit&hgnc_dbtag=on&col=gd_app_sym&col=md_prot_id&col=md_ensembl_id&status=Approved&status=Entry+Withdrawn&status_opt=2&=on&where=&order_by=gd_app_sym_sort&format=text&limit=&.cgifields=&.cgifields=chr&.cgifields=status&.cgifields=hgnc_dbtag'
arch = 'HUGO'
archivo = urllib.urlretrieve(web, arch)
```

11.2 **Anexo 2 (ipp.py).** Script para la obtención de los datos de interacción proteína-proteína.

```
## Esta parte pide la dirección y el nombre a guardar de la BD de interaccion proteina-proteina (ipp) y la baja de la red al directorio en donde se encuentra este Script
print 'DESCARGANDO ARCHIVOS DE INTERACCION PROTEINA-PROTEINA'
def estadodescarga(blocknum, bs, size):
   estadodescargado = blocknum * bs
   print "Cantidad descargado: %s Bytes de %s Bytes totales" % (estadodescargado, size)
bases='http://www.reactome.org/download/current/homo_sapiens.mitab.interactions.txt.gz,ftp://mint.bio.uniroma2.it/pub/humaninteractome/Human_interactome.tar.gz'
for i in bases.split(','):
   web = i
   arch = i[-24:]
   print i
   archivo = urllib.urlretrieve(web, arch, reporthook = estadodescarga)
## Parte para descomprimir los archivos que se estan bajando de los sitios (formatos .tgz y .gz)
   if arch[-7:] == '.tar.gz':
        unzip= tarfile.open(arch, 'r:gz')
        unzip.extractall()
        unzip.close()
   elif arch[-3:] == '.gz':
        unzip= gzip.open(arch, 'rb')
        outfile = open(arch[0:-3], 'w')
        outfile.write(unzip.read())
        unzip.close()
        outfile.close()
   os.remove(arch)
web ='http://www.ebi.ac.uk/intact/export?query=q%3Dtaxid_expanded_id%253A9605%26sort%3Drigid%2Basc%26rows%3D30%26start%3D0%26facet%3Dtrue%26facet.missing%3Dtrue%26facet.field%3Dexpansion%26facet.field%3DinteractorType_id&format=mitab&conversationContext=1'
arch = 'intact.txt'
print 'http://www.ebi.ac.uk/intact/'
```

```
        archivo = urllib.urlretrieve(web, arch, reporthook = estadodescarga)
        print 'PROCESANDO LA INFORMACION DE LAS BASES DE DATOS'
        out=''
        base=open('Human_interactome')
        for j in base:
            if j.split('|')[3] == j.split('|')[4][0:-1] == "Homo sapiens":
                out=out+j.split('|')[0]+ '\t' + j.split('|')[1]+'\n'
        base.close()
        base=open('Human_interactome','w')
        base.write(out)
        base.close()
        archi='dip.txt,matriz.txt,mint.txt,intact.txt,itab.interactions.txt,domino.txt'
        for i in archi.split(','):
            print 'PROCESANDO: ' + i
            out=''
            base=open(i)
            for j in base:
                if j.split('\t')[9] == "taxid:9606(Homo sapiens)" or 'taxid:9606(Human)' or 'taxid:9606':
                    if j.split('\t')[10] == "taxid:9606(Homo sapiens)" or 'taxid:9606(Human)' or 'taxid:9606':
                        out=out+j.split('\t')[0]+ '\t' + j.split('\t')[1]+'\n'
            base.close()
            base=open(i+'2','w')
            base.write(out)
            base.close()
        out=''
        print 'PROCESANDO: biogrid.txt'
        base=open('biogrid.txt')
        for j in base:
            if j.split('\t')[15] == j.split('\t')[16] == '9606':
                out=out+j.split('\t')[7]+ '\t' + j.split('\t')[8]+'\n'
        base.close()
        base=open('biogrid.txt2','w')
        base.write(out)
        base.close()
        out=''
        print 'PROCESANDO: hprd.txt'
        base=open('hprd.txt')
        for j in base:
            out=out+j.split('\t')[0]+ '\t' + j.split('\t')[3]+'\n'
        base.close()
        base=open('hprd.txt2','w')
        base.write(out)
        base.close()
        ## Parte para quitar regiones de la informacion que es estorbosa para el procesamiento de la
informacion
        print 'QUITANDO INFORMACION REPETITIVA'
        bases='dip.txt2,matriz.txt2,mint.txt2,intact.txt2,itab.interactions.txt2,domino.txt2'
        for j in bases.split(','):
            print 'PROCESANDO: ' + j
            out2=''
            ipps = open(j)
            for i in ipps:
                if 'uniprotkb:' in i.split('\t')[1]:
                    if 'uniprotkb:' in i.split('\t')[0]:
                        out2=                                    out2                                    +
i.split('\t')[0].replace(i.split('\t')[0],i.split('\t')[0][i.split('\t')[0].find('uniprotkb:')+len('uniprotkb:'):])    +    '\t'    +
i.split('\t')[1].replace(i.split('\t')[1],i.split('\t')[1][i.split('\t')[1].find('uniprotkb:')+len('uniprotkb:'):][0:-1])+'\n'
            ipps.close()
```

```
            ipps=open(j,'w')
            ipps.write(out2)
            ipps.close()
        bases='dip.txt2,matriz.txt2,mint.txt2,intact.txt2,itab.interactions.txt2,domino.txt2'
        for j in bases.split(','):
            print 'PROCESANDO: ' + j
            out=''
            ipps = open(j)
            for i in ipps:
                out=out + i.split('\t')[0][:6] + '\t' + i.split('\t')[1][:6] + '\n'
            ipps.close()
            ipps=open(j,'w')
            ipps.write(out)
            ipps.close()
        ## Parte para intercambiar el ID de Uniprot por la terminologia oficial
        print 'RENOMBRANDO A LA TERMINOLOGIA OFICIAL'
        bases='Human_interactome,dip.txt2,matriz.txt2,mint.txt2,intact.txt2,itab.interactions.txt2,domino.tx
t2'
        hugo = open('HUGO', 'r')
        diccionario = dict([(i.split("\t")[1],i.split("\t")[0]) for i in hugo])
        hugo.close()
        for m in bases.split(','):
            print 'PROCESANDO: ' + m
            bd = open(m)
            out3=""
            out4=""
            for h in bd:
                for i, j in diccionario.iteritems():
                    if i == h.split('\t')[0]:
                        out3 = out3 + h.split('\t')[0].replace(h.split('\t')[0],j) + '\t' + h.split('\t')[1]
            bd.close()
            bd = open(m,'w')
            bd.write(out3)
            bd.close()
            bd = open(m)
            for k in bd:
                for i, j in diccionario.iteritems():
                    if i == k.split('\t')[1][0:-1]:
                        out4 = out4 + k.split('\t')[0] + '\t' + k.split('\t')[1][0:-1].replace(k.split('\t')[1][0:-1],j) + '\n'
            bd.close()
            bd = open(m,'w')
            bd.write(out4)
            bd.close()
        ## Parte para integrar las diversas bases de datos editadas, tanto para su union como para quitar
aquellos datos repetitivos
        print 'UNIENDO LA INFORMACION EN UNA SOLA BASE DE DATOS'
        bases='Human_interactome,dip.txt2,matriz.txt2,mint.txt2,intact.txt2,itab.interactions.txt2,domino.tx
t2,biogrid.txt2,hprd.txt2'
        out6=[]
        out7=''
        out8=[]
        out9=''
        for m in bases.split(','):
            print 'UNIENDO ' + m +' A LA BASE FINAL'
            una=open(m)
            for i in una:
                out6.extend(i)
            una.close()
```

```
out7=out7+''.join(out6)
for j in out7.split('\n'):
   out8= out8 + sorted(j.split('\t'))
M=[]
for i in range(len(out8)/2):
   M.append([0]*2)
a=0
while a<(len(out8)-1):
   for i in range(len(out8)/2):
         for j in range(2):
            a=a+1
            M[i][j]=out8[a-1]
M.sort()
M2=[]
for i in M:
   if i not in M2:
         M2.append(i)
for i in M2:
   out9 = out9 + '\t'.join(i) + '\n'
bd=open('IPP.txt',"w")
bd.write(out9)
bd.close()
bases='Human_interactome,dip.txt2,matriz.txt2,mint.txt2,intact.txt2,itab.interactions.txt2,domino.tx
t2,biogrid.txt2,hprd.txt2,intact.txt,itab.interactions.txt'
for m in bases.split(','):
   os.remove(m)
```

11.3 **Anexo 3 (modificaciones.py).** Script para obtener los datos de modificaciones post-traduccionales.

```
def estadodescarga(blocknum, bs, size):
   estadodescargado = blocknum * bs
   print "Cantidad descargado: %s Bytes de %s Bytes totales" % (estadodescargado, size)
site='Phosphorylation_site_dataset.gz,Acetylation_site_dataset.gz,Methylation_site_dataset.gz,O-
GlcNAc_site_dataset.gz,Sumoylation_site_dataset.gz,Ubiquitination_site_dataset.gz'
for i in site.split(','):
## Esta parte pide la dirección y el nombre a guardar de la BD de modificaciones post-
traduccionles
## y la baja de la red al directorio en donde se encuentra este Script
   web = 'http://www.phosphosite.org/downloads/'+i
   arch = i
   print i
   archivo = urllib.urlretrieve(web, arch, reporthook = estadodescarga)
## Parte para descomprimir los archivos que se estan bajando de los sitios (formatos .tgz y .gz)
   if arch[-3:] == '.gz':
         unzip= gzip.open(arch, 'rb')
         outfile = open(arch[0:-3]+'.txt', 'w')
         outfile.write(unzip.read())
         unzip.close()
         outfile.close()
   os.remove(i)
web                                                                                                =
'http://phospho.elm.eu.org/dumps/phosphoELM_all_latest.dump.tgz'##http://phospho.elm.eu.org/dumps/p
hosphoELM_vertebrate_latest.dump.tgz
arch = 'phosphoELM.tgz'
```

```
        print arch
        archivo = urllib.urlretrieve(web, arch, reporthook = estadodescarga)
        unzip= tarfile.open(arch, 'r:gz')
        unzip.extractall()
        unzip.close()
        os.remove(arch)
        os.remove('Phospho.Elm_AcademicLicense.pdf')
        print 'EXTRAYENDO INFORMACION ESPECIFICA DE LAS BASES DE MODIFICACIONES POST-TRADUCCIONALES'
        bd='Phosphorylation_site_dataset.txt,Acetylation_site_dataset.txt,Sumoylation_site_dataset.txt,Ubiquitination_site_dataset.txt'
        for j in bd.split(','):
            base2=open(j, 'r')
            out=''
            for i in base2:
                if i.split('\t')[5] == 'human':
                    out = out + i.split('\t')[1] + '\t' + i.split('\t')[2] + '\t' + i.split('\t')[3][1:] + i.split('\t')[3][:1] + '\n'
            base2.close()
            rebase2=open(j, 'w')
            rebase2.write(out)
            rebase2.close()
        bd2='Methylation_site_dataset.txt,O-GlcNAc_site_dataset.txt'
        for j in bd2.split(','):
            base2=open(j, 'r')
            out=''
            for i in base2:
                if i.split('\t')[5] == 'human':
                    if j=='Methylation_site_dataset.txt':
                        out = out + i.split('\t')[1] + '\t' + 'METHYLATION' + '\t' + i.split('\t')[3][1:] + i.split('\t')[3][:1] + '\n'
                    if j=='O-GlcNAc_site_dataset.txt':
                        out = out + i.split('\t')[1] + '\t' + 'GLYCOSYLATION' + '\t' + i.split('\t')[3][1:] + i.split('\t')[3][:1] + '\n'
            base2.close()
            rebase2=open(j, 'w')
            rebase2.write(out)
            rebase2.close()
            out=''
        base2=open('phosphoELM_all_2010-09.dump', 'r')
        out=''
        for i in base2:
            if  i.split('\t')[8]=='Homo sapiens':
                out = out + i.split('\t')[1] + '\t' + 'PHOSPHORYLATION' + '\t' + i.split('\t')[3] + i.split('\t')[4] + '\n'
        base2.close()
        rebase2=open('phosphoELM_all_2010-09.dump', 'w')
        rebase2.write(out)
        rebase2.close()
        base2=open('hprdmodif.txt', 'r')
        out=''
        for i in base2:
            out = out + i.split('\t')[1] + '\t' + i.split('\t')[8].upper() + '\t' + i.split('\t')[4] + i.split('\t')[5] + '\n'
        base2.close()
        rebase2=open('hprd_modif.txt', 'w')
        rebase2.write(out.replace(';-',''))
        rebase2.close()
        ## Parte `para intercamibar los identificadores de Uniprot por la terminologia oficial
        print 'CAMBIANDO A LA TERMINOLOGIA OFICIAL'
```

```
	hugo = open('HUGO', 'r')
	bases='Phosphorylation_site_dataset.txt,Acetylation_site_dataset.txt,Methylation_site_dataset.txt,
Sumoylation_site_dataset.txt,Ubiquitination_site_dataset.txt,O-
GlcNAc_site_dataset.txt,phosphoELM_all_2010-09.dump'
	diccionario = dict([(i.split("\t")[1],i.split("\t")[0]) for i in hugo])
	hugo.close()
	for m in bases.split(','):
	   out=''
	   bd = open(m, 'r')
	   for h in bd:
	         for i, j in diccionario.iteritems():
	            if i == h.split('\t')[0]:
	               out = out + h.split('\t')[0].replace(h.split('\t')[0],j) + '\t' + h.split('\t')[1] + '\t' + h.split('\t')[2]
	   bd.close()
	   bde = open(m, 'w')
	   bde.write(out)
	   bde.close()
	## Parte para integrar las diversas bases de datos editadas de modificaciones post-
traduccionales, tanto para su union como para quitar aquellos datos repetitivos
	print 'UNIENDO LOS RESULTADOS DE LAS MODIFICACIONES POST-TRADUCCIONALES EN
UNA SOLA LISTA'
	bases='Phosphorylation_site_dataset.txt,Acetylation_site_dataset.txt,Methylation_site_dataset.txt,
Sumoylation_site_dataset.txt,Ubiquitination_site_dataset.txt,O-
GlcNAc_site_dataset.txt,phosphoELM_all_2010-09.dump,hprd_modif.txt'
	out6=[]
	out7=''
	for j in bases.split(','):
	   demas=open(j,"r")
	   for i in demas:
	         out6.extend(i)
	   demas.close()
	   os.remove(j)
	bd=open('MODIFICACIONES',"w")
	bd.write(''.join(out6))
	bd.close()
	## Aqui se tomara a la proteina y a su modificacion en especifico y se concatenaran todos los
sitios modificados en una sola linea
	bd=open('MODIFICACIONES')
	dic={}
	for i in bd:
	   clave=i.split('\t')[0]+'\t'+i.split('\t')[1]+'\t'
	   todas_claves = dic.keys()
	   if clave in todas_claves:
	         if i.split('\t')[2][0:-1] not in dic[clave]:
	            dic[clave].append(i.split('\t')[2][0:-1])
	   else:
	         dic[clave]=[i.split('\t')[2][0:-1]]
	for i,j in dic.iteritems():
	   out7=out7+''.join(i)+'\t'.join(sorted(j))+'\n'
	bd.close()
	todas=open('MODIFICACIONES','w')
	todas.write(out7)
	todas.close()
```

11.4 **Anexo 4 (reactome.py).** Script para obtener los datos de Reactome

```
def estadodescarga(blocknum, bs, size):
    estadodescargado = blocknum * bs
    print "Cantidad descargado: %s Bytes de %s Bytes totales" % (estadodescargado, size)
out=''
web = 'http://www.reactome.org/download/current/ReactomePathways.gmt.zip'
arch = 'ReactomePathways.gmt.zip'
archivo = urllib.urlretrieve(web, arch, reporthook = estadodescarga)
## Parte que descomprime el archivo optenido
unzip= open('ReactomePathways.gmt.zip', 'rb')
z = zipfile.ZipFile(unzip)
for name in z.namelist():
    outfile = open(name, 'wb')
    outfile.write(z.read(name))
    outfile.close()
unzip.close()
## Parte que quitara la cadena 'Reactome Pathway       ' presente en todas las rutas
reactome=open("ReactomePathways.gmt")
out=""
out2=''
for i in reactome:
    out=out + i
out2=out2+out.replace('Reactome Pathway'+'\t','')
ruta=open('REACTOME',"w")
ruta.write(out2)
ruta.close()
reactome.close()
# Para borrar los archivos
os.remove('ReactomePathways.gmt.zip')
os.remove('ReactomePathways.gmt')
```

11.5 **Anexo 5 (kegg.py).** Script para obtener los datos de KEGG

```
out=''
out2=''
out3=''
out4=''
web = 'http://www.genome.jp/dbget-bin/get_linkdb?-t+pathway+gn:T01001'
arch = 'hsa.txt'
archivo = urllib.urlretrieve(web, arch)
arch=open('hsa.txt')
for i in arch:
    out=out+i
out2=out2+out.split('---'+'\n')[1]
out3=out3+out2.split('\n'+'</pre><hr><a href')[0]
arch.close()
arch=open('hsa.txt','w')
arch.write(out3)
arch.close()
arch=open('hsa.txt')
for i in arch:
    out4=out4+i.split('">')[1].replace(i.split('">')[1],i.split('">')[1][:i.split('">')[1].find(' - Ho')])+'\n'
arch=open('hsa.txt','w')
arch.write(out4.replace('</a>            ','\t'))
arch.close()
## Seleccionara las rutas especificas del organismo de interes con base a la lista de rutas totales reportadas en kegg y la lista de las rutas especificas del organismo
```

```
## Parte para seleccionar las rutas especificas del organismo de interes
lista2=''
lista2_0=''
lista2_0_1=''
lista2_1=''
lista3=''
lista1=open('hsa.txt')
for i in lista1:
    lista2=lista2+i
for m in lista2.split('\n'):
    lista2_0_1=lista2_0_1+m.split('\n')[0]+'\n'
lista2_1=lista2_1+lista2_0_1.replace('/','-')
lista3=lista3+lista2_1.replace('              ','\t')
lista1.close()
## Parte para bajar las rutas de internet
for i in lista3.split('\n')[:-2]:
    clave=i[0:8]
    web="http://www.genome.jp/dbget-bin/get_linkdb?-t+genes+path:"+clave
    nombre=i[9:]+'.txt'
    archivo=urllib.urlretrieve(web, nombre)
    print nombre[:-4]
    ruta=i[9:]
    rutas=open(ruta+".txt", "r")
    out=[]
    for j in rutas:
            out=out + j.split("")
    redit=open(ruta+".txt", "w")
    redit.write(''.join(out))
    redit.close()
    redit=open(ruta+".txt", "r")
    out2=''
    for k in redit:
            out2=out2 + k
    redit2=open(ruta+'.txt', 'w')
    redit2.write(out2.split('-'*100+'\n')[1])
    redit2.close()
    redit.close()
    rutas.close()
    redit2=open(ruta+'.txt', 'r')
    out3=[]
    for l in redit2:
            out3=out3 + l.split(' ')
    redit=open(ruta+".txt", "w")
    redit.write(''.join(out3))
    redit2.close()
    redit.close()
    redit=open(ruta+".txt", "r")
    out4=''
    for x in redit:
            out4=out4 + x
    redit2=open(ruta+'.txt', 'w')
    redit2.write(out4.split('\n'+'</pre>')[0])
    redit.close()
    rutas.close()
    redit2.close()
    redit2=open(ruta+'.txt', 'r')
    out5=''
    out6=''
```

```
    out7=''
    for y in redit2:
        for c in y:
            out6 = out6 + c.replace(';',',')
    for d in out6.split('\n'):
        out5=out5 + d.split(',')[0] + '\n'
    for z in out5.split('\n'):
        out7 = out7 + z.split('</a>')[-1] + '\t'
    redit=open(ruta+".txt", "w")
    redit.write(out7)
    redit2.close()
    redit.close()
```

Este programa sera para unir las rutas de Kegg en un solo archivo, primero pedira el nombre del archivo, despues la columna donde se encuetre el nombre de las rutas.

por un proceso iterativo se pedira para cada archivo: que en un archivo nuevo pegue en la primer columna el nombre de la ruta, en las columnas siguientes todas las proteinas implicadas en esa ruta

```
rutall=open('KEGG','w')
out=''
genes=''
for i in lista3.split('\n')[:-2]:
    o=open(ruta+'.txt')
    ruta=i[9:]
    for j in o:
        out = out+ruta+'\t'+j+'\n'
    o.close()
rutall.write(out.replace('\t\t',''))
rutall.close()
# Para borrar los archivos
for i in lista3.split('\n')[:-2]:
    ruta=i[9:]
    os.remove(ruta+'.txt')
os.remove('hsa.txt')
```

11.6 **Anexo 6.** Secuencia de comandos de R para el pre-procesamiento de los datos de microarreglos.

```
> library(affy)
> library(limma)
> meta397=ReadAffy()
> meta397
AffyBatch object
size of arrays=1164x1164 features (148 kb)
cdf=HG-U133_Plus_2 (54675 affyids)
number of samples=396
number of genes=54675
annotation=hgu133plus2
notes=
> hist(meta396)
> boxplot(meta396)
>
sampleNames(meta396)=c(1.001,1.002,1.003,1.004,1.005,1.006,1.007,1.008,1.009,1.010,1.011,1.012,1.
013,1.014,2.010,2.020,2.030,3.010,3.020,3.030,3.040,3.050,3.060,3.070,3.080,3.090,3.100,3.110,3.120,3
.130,3.140,3.150,3.160,1.015,2.040,2.050,6.010,3.170,3.180,3.190,3.200,3.210,6.020,6.030,2.060,2.070,
2.080,2.090,6.040,1.016,3.220,2.100,2.110,6.050,1.017,6.060,3.230,3.240,1.018,1.019,6.070,5.001,2.120
```

```
,3.250,4.030,6.080,6.090,5.002,2.130,4.040,6.100,6.110,2.140,5.003,5.004,4.050,6.120,5.005,6.130,6.14
0,6.150,6.160,5.006,5.007,5.008,5.009,5.010,5.011,5.012,5.013,5.014,5.015,5.016,5.017,5.018,5.019,5.0
20,5.021,5.022,5.023,5.024,5.025,5.026,5.027,5.028,5.029,5.030,5.031,5.032,5.033,5.034,5.035,5.036,5.
037,5.038,5.039,5.040,5.041,5.042,5.043,5.044,5.045,5.046,5.047,5.048,5.049,5.050,5.051,5.052,5.053,5
.054,5.055,5.056,5.057,5.058,5.059,5.060,5.061,5.062,5.063,5.064,5.065,5.066,5.067,5.068,5.069,5.070,
5.071,5.072,5.073,5.074,5.075,5.076,5.077,5.078,5.079,5.080,5.081,5.082,5.083,5.084,5.085,5.086,5.087
,5.088,5.089,5.090,5.091,5.092,5.093,5.094,5.095,5.096,1.022,1.023,1.024,1.025,1.026,1.027,1.028,1.02
9,2.150,2.160,2.170,2.180,2.190,2.200,1.030,6.230,2.210,1.031,1.032,1.033,2.220,1.034,2.230,2.240,1.0
35,2.250,6.390,2.260,6.400,1.036,1.037,1.038,1.039,1.040,1.041,1.042,2.270,6.460,6.470,2.280,2.290,2.
300,2.310,1.043,2.320,1.044,2.330,1.045,2.340,6.510,1.046,1.047,1.048,1.049,1.050,1.051,1.052,1.053,1
.054,1.055,1.056,1.057,1.058,1.059,1.060,1.061,1.062,1.063,5.097,5.098,5.099,5.100,5.101,5.102,5.103,
5.104,5.105,5.106,5.107,5.108,5.109,5.110,5.111,5.112,5.113,5.114,5.115,5.116,5.117,5.118,5.119,5.120,5
.121,5.122,5.123,5.124,5.125,5.126,5.127,5.128,5.129,5.130,5.131,5.132,5.133,5.134,5.135,5.136,5.137,
5.138,5.139,5.140,5.141,5.142,5.143,5.144,5.145,5.146,5.147,5.148,5.149,5.150,6.520,6.530,6.540,6.550
,1.064,1.065,1.066,3.260,3.270,4.310,4.320,4.330,4.340,4.350,2.640,2.650,2.660,2.670,5.151,5.152,5.15
4,5.155,5.156,5.157,5.158,5.159,5.160,5.161,5.162,5.163,5.164,1.073,1.074,1.075,1.076,1.077,1.078,1.0
79,1.080,1.081,1.082,1.083,1.084,1.085,1.086,1.087,1.088,1.089,1.090,1.091,1.092,1.093,1.094,1.095,1.
096,1.097,1.098,1.099,1.100,1.101,1.102,1.103,1.104,1.105,1.106,1.107,1.108,1.109,1.110,1.111,1.112,1.
113,1.114,1.115,1.116,1.117,1.118,1.119,1.120,1.121,1.122,1.123,1.124,1.125,1.126,1.127,1.128,1.129,1.
130,2.680,2.690,2.700,6.630,2.710,6.640,2.720,3.320,2.730,1.131,6.650,6.660)
>
meta396=meta396[,c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,34,50,55,59,60,174,175,176,177,178,179,180,181,
188,191,192,193,195,198,203,204,205,206,207,208,209,217,219,221,224,225,226,227,228,229,230,231,
232,233,234,235,236,237,238,239,240,241,300,301,302,327,328,329,330,331,332,333,334,335,336,337,
338,339,340,341,342,343,344,345,346,347,348,349,350,351,352,353,354,355,356,357,358,359,360,361,
362,363,364,365,366,367,368,369,370,371,372,373,374,375,376,377,378,379,380,381,382,383,384,394,
15,16,17,35,36,45,46,47,48,52,53,63,69,73,182,183,184,185,186,187,190,194,196,197,199,201,210,213,
214,215,216,218,220,222,310,311,312,313,385,386,387,389,391,393,18,19,20,21,22,23,24,25,26,27,28,2
9,30,31,32,33,38,39,40,41,42,51,57,58,64,303,304,392,65,70,76,305,306,307,308,309,62,68,74,75,78,83,
84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100,101,102,103,104,105,106,107,108,109,110,111,112
,113,114,115,116,117,118,119,120,121,122,123,124,125,126,127,128,129,130,131,132,133,134,135,136,
137,138,139,140,141,142,143,144,145,146,147,148,149,150,151,152,153,154,155,156,157,158,159,160,
161,162,163,164,165,166,167,168,169,170,171,172,173,242,243,244,245,246,247,248,249,250,251,252,
253,254,255,256,257,258,259,260,261,262,263,264,265,266,267,268,269,270,271,272,273,274,275,276,
277,278,279,280,281,282,283,284,285,286,287,288,289,290,291,292,293,294,295,314,315,316,317,318,
319,320,321,322,323,324,325,326,37,43,44,49,54,56,61,66,67,71,72,77,79,80,81,82,189,200,202,211,21
2,223,296,297,298,299,388,390,395,396)]
>
meta396norm=expresso(meta396,bgcorrect.method='rma',normalize.method='quantiles',pmcorrect.metho
d='pmonly',summary.method='medianpolish')
> write.table(exprs(meta396norm),file='meta396.txt')
```

11.7 **Anexo 7.** Secuencia de comandos en R para obtener los resultados de expresión diferencial.

```
> r=read.table('meta396.txt')
> heatmap.2(cor(r),col=bluered(75),key=TRUE,symkey=FALSE,trace='none')
> heatmap.2(cor(r),col=redblue(75),key=TRUE,symkey=FALSE,trace='none')
>
matriz=model.matrix(~0+factor(c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1
,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,
2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,4,4,4,4,4,4,4,4,5,5,5,5,
5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,
5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,
```

```
5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,5,6,6,6,6,6,6,6,6,6,
6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6)))
> colnames(matriz)=c('mll','telaml','e2apbx','hipcrom','lalt','bcrabl')
> meta396lote=removeBatchEffect(meta396norm,lote396,matriz)
> heatmap.2(cor(meta396lote),col=redblue(75),key=TRUE,symkey=FALSE,trace='none')
> cont=makeContrasts('mll - telaml','mll - e2apbx','mll - hipcrom','mll - lalt','mll - bcrabl','telaml - e2apbx','telaml - hipcrom','telaml - lalt','telaml- bcrabl','e2apbx - hipcrom','e2apbx - lalt','e2apbx - bcrabl','hipcrom - lalt','hipcrom - bcrabl','lalt – bcrabl',levels=matriz)
> fit=lmFit(meta396lote,matriz)
> fit2=contrasts.fit(fit,cont)
> fit3=eBayes(fit2)
> B=3
> M=1
> TTmll=topTable(fit3,coef=1,adjust='fdr',n=2000)
> selectedmll = TTmll[(TTmll$B>B & abs(TTmll$logFC)>M),]
> write.csv(selectedmll,file='mlltelaml.csv')
> TTmll2=topTable(fit3,coef=2,adjust='fdr',n=2000)
> selectedmll2 = TTmll2[(TTmll2$B>B & abs(TTmll2$logFC)>M),]
> write.csv(selectedmll2,file='mlle2apbx.csv')
> TTmll3=topTable(fit3,coef=3,adjust='fdr',n=2000)
> selectedmll3 = TTmll3[(TTmll3$B>B & abs(TTmll3$logFC)>M),]
> write.csv(selectedmll3,file='mllhipcrom.csv')
> TTmll4=topTable(fit3,coef=4,adjust='fdr',n=2000)
> selectedmll4 = TTmll4[(TTmll4$B>B & abs(TTmll4$logFC)>M),]
> write.csv(selectedmll4,file='mlllalt.csv')
> TTmll5=topTable(fit3,coef=5,adjust='fdr',n=2000)
> selectedmll5 = TTmll5[(TTmll5$B>B & abs(TTmll5$logFC)>M),]
> write.csv(selectedmll5,file='mllbcrabl.csv')
> TTtel=topTable(fit3,coef=6,adjust='fdr',n=2000)
> selectedtel = TTtel[(TTtel$B>B & abs(TTtel$logFC)>M),]
> write.csv(selectedtel,file='telamle2apbx.csv')
> TTtel3=topTable(fit3,coef=8,adjust='fdr',n=2000)
> selectedtel3 = TTtel3[(TTtel3$B>B & abs(TTtel3$logFC)>M),]
> write.csv(selectedtel3,file='telamllalt.csv')
> TTtel4=topTable(fit3,coef=9,adjust='fdr',n=2000)
> selectedtel4 = TTtel4[(TTtel4$B>B & abs(TTtel4$logFC)>M),]
> TTe2a=topTable(fit3,coef=10,adjust='fdr',n=2000)
> selectede2a = TTe2a[(TTe2a$B>B & abs(TTe2a$logFC)>M),]
> write.csv(selectede2a,file='e2apbxhipcrom.csv')
> e2apbxlalt=topTable(fit3,coef=11,adjust='fdr',n=2000)
> write.csv(selectede2a,file='e2apbxlalt.csv')
> e2apbxbcrabl=topTable(fit3,coef=12,adjust='fdr',n=2000)
> TThip=topTable(fit3,coef=13,adjust='fdr',n=2000)
> selectedhip = TThip[(TThip$B>B & abs(TThip$logFC)>M),]
> write.csv(selectedhip,file='hipcromlalt.csv')
> TThip2=topTable(fit3,coef=14,adjust='fdr',n=2000)
> selectedhip2 = TThip2[(TThip2$B>B & abs(TThip2$logFC)>M),]
> write.csv(selectedhip2,file='hipcrombcrabl.csv')
> TTlalt=topTable(fit3,coef=15,adjust='fdr',n=2000)
> selectedlalt = TTlalt[(TTlalt$B>B & abs(TTlalt$logFC)>M),]
> write.csv(selectedlalt,file='laltbcrabl.csv')
> TTmll=topTable(fit3,coef=1,adjust='fdr',n=2000)
> selectedmll = TTmll[(TTmll$B>B & abs(TTmll$logFC)>M),]
> write.csv(selectedmll,file='mlltelaml.csv')
> TTmll2=topTable(fit3,coef=2,adjust='fdr',n=2000)
> selectedmll2 = TTmll2[(TTmll2$B>B & abs(TTmll2$logFC)>M),]
> write.csv(selectedmll2,file='mlle2apbx.csv')
> TTmll3=topTable(fit3,coef=3,adjust='fdr',n=2000)
```

```
> selectedmll3 = TTmll3[(TTmll3$B>B & abs(TTmll3$logFC)>M),]
> write.csv(selectedmll3,file='mllhipcrom.csv')
> TTmll4=topTable(fit3,coef=4,adjust='fdr',n=2000)
> selectedmll4 = TTmll4[(TTmll4$B>B & abs(TTmll4$logFC)>M),]
> write.csv(selectedmll4,file='mlllalt.csv')
> TTmll5=topTable(fit3,coef=5,adjust='fdr',n=2000)
> selectedmll5 = TTmll5[(TTmll5$B>B & abs(TTmll5$logFC)>M),]
> write.csv(selectedmll5,file='mllbcrabl.csv')
> TTtel=topTable(fit3,coef=6,adjust='fdr',n=2000)
> selectedtel = TTtel[(TTtel$B>B & abs(TTtel$logFC)>M),]
> write.csv(selectedtel,file='telamle2apbx.csv')
> TTtel2=topTable(fit3,coef=7,adjust='fdr',n=2000)
> selectedtel2 = TTtel2[(TTtel2$B>B & abs(TTtel2$logFC)>M),]
> write.csv(selectedtel2,file='telamlhipcrom.csv')
> TTtel3=topTable(fit3,coef=8,adjust='fdr',n=2000)
> selectedtel3 = TTtel3[(TTtel3$B>B & abs(TTtel3$logFC)>M),]
> write.csv(selectedtel3,file='telamllalt.csv')
> TTtel4=topTable(fit3,coef=9,adjust='fdr',n=2000)
> selectedtel4 = TTtel4[(TTtel4$B>B & abs(TTtel4$logFC)>M),]
> write.csv(selectedtel4,file='telamlbcrabl.csv')
> TTe2a=topTable(fit3,coef=10,adjust='fdr',n=2000)
> selectede2a = TTe2a[(TTe2a$B>B & abs(TTe2a$logFC)>M),]
> write.csv(selectede2a,file='e2apbxhipcrom.csv')
> TTe2a2=topTable(fit3,coef=11,adjust='fdr',n=2000)
> selectede2a2 = TTe2a2[(TTe2a2$B>B & abs(TTe2a2$logFC)>M),]
> write.csv(selectede2a,file='e2apbxlalt.csv')
> TTe2a3=topTable(fit3,coef=12,adjust='fdr',n=2000)
> selectede2a3 = TTe2a3[(TTe2a3$B>B & abs(TTe2a3$logFC)>M),]
> write.csv(selectede2a3,file='e2apbxbcrabl.csv')
> TThip=topTable(fit3,coef=13,adjust='fdr',n=2000)
> selectedhip = TThip[(TThip$B>B & abs(TThip$logFC)>M),]
> write.csv(selectedhip,file='hipcromlalt.csv')
> TThip2=topTable(fit3,coef=14,adjust='fdr',n=2000)
> selectedhip2 = TThip2[(TThip2$B>B & abs(TThip2$logFC)>M),]
> write.csv(selectedhip2,file='hipcrombcrabl.csv')
> TTlalt=topTable(fit3,coef=15,adjust='fdr',n=2000)
> selectedlalt = TTlalt[(TTlalt$B>B & abs(TTlalt$logFC)>M),]
> write.csv(selectedlalt,file='laltbcrabl.csv')
> write.csv(selectede2a2,file='e2apbxlalt.csv')
```

11.8 **Anexo 8 (probeset_iguales.py).** Script para seleccionar los resultados de expresión diferencial que se comparten.

```
#!/usr/bin/env python
listapru=raw_input("Nombre del listado a contrastar: ")
listacontraste=raw_input("Nombre del listado de referencia: ")
eset=raw_input("Nombre del archivo a generar: ")
esetdif=open(eset,'w')
lista1=open(listapru)
lista2=open(listacontraste)
out=""
out2=''
for l in lista1:
        out2=out2+l
for j in lista2:
```

```
        lista1=open(listapru)
        for i in lista1:
                if i.split('\n')[0] == j.split('\n')[0]:
                                if j.split('\n')[0]+'\n' not in out:
                                        out=out+j.split('\n')[0]+'\n'
        lista1.close()
esetdif.write(out)
esetdif.close()
lista1.close()
```

11.9 **Anexo 9 (selecmatriz.py).** Script para seleccionar los datos específicos desde la matriz de expresión.

```
#!/usr/bin/env python
import string
matriz=raw_input("Nombre de la matriz de expresion: ")
listado=raw_input("Nombre del listado de referencia: ")
eset=raw_input("Nombre del archivo que compilara mas datos: ")
esetdif=open(eset,'a')
matrix=open(matriz)
lista=open(listado)
out=""
out2=''
for l in matrix:
        out2=out2+l
matrix.close()
for j in lista:
        for i in out2.split('\n'):
                if i.split(' ')[0] == j.split('\t')[1][0:-1]:
                        out=out+i+'\n'
esetdif.write(out)
esetdif.close()
lista.close()
```

11.10 **Anexo 10.** Comandos en R para realizar el mapa de calor de los datos extraídos.

```
r =as.matrix(read.table(‘meta673.txt’))
heatmap.2(r, col=greenred(75), key=TRUE, symkey=FALSE, trace="none", cexCol=0.5,scale='row',ColSideColors = c(rep("purple",123), rep("green",44),rep('blue',28),rep('yellow',8),rep('red',163),rep('orange',30)))
```

11.11 **Anexo 11 (MA-HUGOaffy.py).** Script para renombrar los ID de Affymetrix.

```
#!/usr/bin/env python
hugobd = raw_input("BD que tiene las nomenclaturas: ")
nombre = raw_input("Nombre del archivo con los ID de affymetrix: ")
renombre = raw_input("Nombre del archivo creada con la terminologia de HUGO: ")
hugo = open(hugobd, 'r')
bd = open(nombre, 'r')
clave = int(raw_input("Columna donde esta la terminologia de la BD a procesar: "))
significado = int(raw_input("Columna donde esta la nomenclatura propuesta por HUGO: "))
diccionario = dict([(i.split("\t")[clave-1],i.split("\t")[significado-1]) for i in hugo])
out3=""
for h in bd:
        for i, j in diccionario.iteritems():
                if i == h.split('\t')[0]:
                        out3 = out3 + h.split('\t')[0].replace(h.split('\t')[0],j) + '\t' + h.split('\t')[1] + '\t' +
h.split('\t')[2]
bd.close()
bde = open(renombre, 'w')
bde.write(out3)
bd.close()
bde.close()
hugo.close()
```

11.12 **Anexo 12 (resultados.py).** Script para el enriquecimiento de los genes.

```
## Parte para introducir los genes para su enriquecimiento, son dos opciones una para introducirla por el
techado y otra para introducir un archivo con la lista de ellos
pet1=raw_input('Se introducirá un conjunto de proteínas (genes) mediante el teclado [digite 1] o un
archivo con el conjunto de datos de expresión diferencial [digite 2] para enriquecer sus resultados: ')
dir1=os.getcwd()
gen=[]
genes=''
if pet1 == '1':
        gen=raw_input('Introduce el nombre de el(los) gen(es) separados por un espacio simple (debes
introducir mas de uno): ')
        otro=gen.split(' ')
        genes='\n'.join(otro)
        os.mkdir('DATOS_ENRIQUECIDOS')
if pet1 == '2':
        arch=raw_input('Introduce el nombre del archivo: ')
        archi=open(arch, 'r')
        colgen=int(raw_input('En que columna se encuentra el nombre de los genes: '))
        FCh=int(raw_input('Culumna donde se encuentran las veces de cambio: '))
        for i in archi:
                genes = genes+i.split('\t')[colgen-1]+'\n'
        archi.close()
        os.mkdir('DATOS_ENRIQUECIDOS_'+arch[:-4])
## Parte para reportar las modificaciones post-traduccionales
for i in genes.split('\n'):
```

```
	out=''
	rep=open(i+'.txt','w')
	mod=open('MODIFICACIONES')
	print 'PROCESANDO: ' +i
	for j in mod:
		if i == j.split('\t')[0]:
			out=out+j.split('\n')[0][len(i)+1:]+'\n'
	mod.close()
	rep.write('MODIFICACIONES POST-TRADUCCIONALES:' + '\n' + out + '\n'  + '\n' + '\t-'*50 + '\n' + '\t-'*50 + '\n' + '\n')
	rep.close()
## Parte para el resultado de las rutas en las que estan implicados los genes
bdruta='REACTOME,KEGG'
for i in genes.split('\n'):
		print 'PROCESANDO: ' +i
		out=''
		for l in bdruta.split(','):
			bd=open(l,'r')
			for j in bd:
				for k in j.split('\t'):
					if i == k:
						out=out+l+'\t'+j.split('\n')[0]+'\n'+'\n'
			bd.close()
		rep=open(i+'.txt','a')
		rep.write('RUTAS EN LAS QUE ESTA IMPLICADO:' + '\n' + out)
		rep.close()
## Parte para reportar las proteinas con las que interactura, asi como las que interactuan de la lista que se proporciono
for i in genes.split('\n'):
	out=''
	out2=''
	out4=''
	IPP=open('IPP.txt')
	print 'PROCESANDO: ' +i
	for j in IPP:
		if i == j.split('\t')[0]:
			if j.split('\t')[1] not in out:
				out = out + j.split('\t')[1]
		if i == j.split('\t')[1][0:-1]:
			if j.split('\t')[0] not in out:
				out = out + j.split('\t')[0] + '\n'
	for k in out.split('\n'):
		for l in genes.split('\n'):
			rep=open(i+'.txt')
			if l==k.split('\n')[0]:
				out4=out4+k.split('\n')[0]+'\t'
	rep=open(i+'.txt','r')
	for k in rep:
		out2=out2+k
	rep.close()
	if pet1 == '1':
		os.chdir('DATOS_ENRIQUECIDOS')
	if pet1 == '2':
		os.chdir('DATOS_ENRIQUECIDOS_'+arch[:-4])
	rep=open(i+'.txt','w')
	rep.write('Resultados'+'  '+i+':'+'\n'+'PROTEINAS CON LAS QUE INTERACTUA DE LA LISTA PROPORCIONADA:' + '\n' + out4 + '\n' + '\t-'*50 + '\n' + '\t-'*50 + '\n' + '\n' + out2 + '\n' + '\t-'*50 + '\n' + '\t-'*50 + '\n' + '\n' + 'PROTEINAS CON LAS QUE INTERACTUA:' +'\n' + out)
```

```
	rep.close()
	os.chdir(dir1)
	IPP.close()
for i in genes.split('\n'):
	os.remove(i+'.txt')
```

11.13 **Anexo 13 (resultados2.py).** Script para generar los resultados de expresión diferencial unidos a las rutas de señalización.

```
	print 'LOS RESULTADOS SE BASARAN EN LAS BASES DE REACTOME Y KEGG, EL NOMBRE DEL ARCHIVO EMPESARA POR EL NOMBRE DE LA BASE DE DATOS SEGUIDA DEL NOMBRE DEL ARCHIVO DE EXPRESION DIFERENCIAL'
	os.mkdir('RUTAS_'+arch[:-4])
	dir1=os.getcwd()
	bd='REACTOME,KEGG'
	for z in bd.split(','):
		expdif=open(arch)
		out=''
		for i in expdif:
			vias=open(z)
			for j in vias:
				for k in j.split('\t'):
					if i.split('\t')[1] == k:
						if j not in out:
							out=out+j
			vias.close()
		expdif.close()
		expdif=open(arch)
		for w in expdif:
			for q in out.split('\n'):
				for e in q.split('\t'):
					if w.split('\t')[colgen-1] == e:
						out=out.replace('\t'+e+'\t','\t'+'['+w.split('\t')[FCh-1]+']'+e+'\t')
		exprdifr=open(z+'_'+arch,'w')
		exprdifr.write(out)
		exprdifr.close()
		expdif.close()
	##Parte para generar los resultados estadisticos
		exprdifr=open(z+'_'+arch)
		out=''
		out2=''
		for i in exprdifr:
			out=out+i
		for j in out.split('\n'):
			cont1=0
			cont2=0
			for k in j.split('\t'):
				if k[:2]=='[-':
					cont1=cont1+1 #Cuenta los genes sub-expresados
				cont2=cont2+1 #Cuenta los genes totales en la ruta
			porcent= cont1*100.0/(cont2) #% de genes desregulados en la ruta
			cont3=0
			for k in j.split('\t'):
				if re.search('\\[[0-9]',k[:2]):
```

```
                                        cont3=cont3+1 #Cuenta los genes sobre-expresados
                        porcent2= cont3*100.0/(cont2) #% de genes desregulados en la ruta
                        out2=out2+j.split('\t')[0].upper().replace('\t',' ')+'\t'+str(cont2) + '\t'+ str(cont1) + '\t'+str(porcent) + '\t'+str(cont3) + '\t'+str(porcent2) + '\n'
                res=open('ESTADISTICA'+'_'+z+'_'+arch,'w')
                exprdifr.close()
                res.write(out2)#.replace('\n\t-2 Genes implicados en la ruta.\t0 Genes sub-expresados:\t-0.0 % de genes sub-expresados.\t0 Genes sobre-expresados:\t-0.0 % de genes sobre-expresados.\n',''))
                res.close()
        ##Parte para generar archivos de resultados de rutas diferenciales
                res=open('ESTADISTICA'+'_'+z+'_'+arch)
                out=''
                out2=''
                out3=''
                for m in res:
                        if int(m.split('\t')[2][:1])>0:
                                if int(m.split('\t')[4][:1])==0:
                                        out2=out2+m.split('\n')[0]+'\n'
                        if int(m.split('\t')[2][:1])==0:
                                if int(m.split('\t')[4][:1])>0:
                                        out3=out3+m.split('\n')[0]+'\n'
                        if int(m.split('\t')[2][:1])>0:
                                if int(m.split('\t')[4][:1])>0:
                                        out=out+m.split('\n')[0]+'\n'
                os.chdir('RUTAS_'+arch[:-4])
                up=open('UP'+'_'+z+'_'+arch[:-3]+'csv','w')
                down=open('DOWN'+'_'+z+'_'+arch[:-3]+'csv','w')
                up.write('RUTAS\tGenes implicados en la ruta\tGenes sub-expresados\tGenes sub-expresados (%)\tGenes sobre-expresados\tGenes sobre-expresados (%)\n'+out3)
                up.close()
                down.write('RUTAS\tGenes implicados en la ruta\tGenes sub-expresados\tGenes sub-expresados (%)\tGenes sobre-expresados\tGenes sobre-expresados (%)\n'+out2)
                down.close()
                intermedio=open('equilibrio'+'_'+z+'_'+arch[:-3]+'csv','w')
                intermedio.write('RUTAS\tGenes implicados en la ruta\tGenes sub-expresados\tGenes sub-expresados (%)\tGenes sobre-expresados\tGenes sobre-expresados (%)\n'+out)
                intermedio.close()
                os.chdir(dir1)
                os.remove('ESTADISTICA'+'_'+z+'_'+arch)
                os.remove(z+'_'+arch)
```

11.14 **Anexo 14.** Resultado del análisis de expresión de LA-T de los 532 genes.

Símbolo oficial	ID Affymetrix	Veces de cambio	Valor de p ajustado	B	Símbolo oficial	ID Affymetrix	Veces de cambio	Valor de p ajustado	B
ABCD3	1554878_a_at	3.63	3.26E-059	128.21	LRP12	219631_at	6.77	1.02E-102	229.82
ABHD15	226796_at	-4.72	1.28E-073	161.97	LRRC8C	223533_at	-5.66	1.75E-105	236.29
ACAP2	1552472_a_at	-2.69	2.37E-082	182.34	LRRFIP1	201861_s_at	-3.18	1.84E-081	180.24
ACOT9	221641_s_at	-2.62	6.82E-	115.72	LY86	205859_at	-5.9	1.65E-	189.69

			054					085	
ACPL2	226925_at	4.44	1.86E-062	135.81	LYN	202625_at	-4.89	1.90E-062	135.79
ACSL1	201963_at	-9.32	5.44E-078	172.18	LYST	203518_at	-3.12	1.76E-037	77.2
ADCY9	204497_at	-5.54	3.01E-074	163.45	MACROD2	235278_at	-12.73	5.96E-121	272.27
AKAP2	202759_s_at	4.41	1.60E-072	159.39	MAFG	204970_s_at	-2.73	9.05E-089	197.33
ALDH1A2	207016_s_at	16.34	5.93E-063	136.98	MAL	204777_s_at	41.64	1.76E-125	282.96
AMMECR1	204976_s_at	-2.57	8.65E-048	101.4	MALT1	210017_at	-3.94	4.98E-067	146.53
APAF1	204859_s_at	-2.71	4.68E-058	125.49	MANBA	203778_at	-2.53	9.86E-101	225.19
APP	200602_at	-4.96	1.16E-039	82.32	MAP1A	203151_at	2.39	4.46E-042	87.98
AQP3	39248_at	6.77	6.56E-073	160.3	MAP3K5	203837_at	-4.41	3.71E-074	163.23
ARF1	1565651_at	-2.51	5.16E-061	132.43	MARCH3	213256_at	-5.06	5.47E-112	251.51
ARHGAP25	204882_at	-2.57	2.00E-059	128.7	MAZ	207824_s_at	2.5	5.83E-053	113.54
ARHGEF3	218501_at	3.41	9.70E-042	87.19	MBNL2	203640_at	-3.14	1.69E-051	110.12
ARL4C	202206_at	10.2	5.07E-076	167.58	MBP	225407_at	-3.23	4.60E-067	146.61
ARSD	225286_at	-3.61	2.88E-086	191.47	MEF2A	214684_at	-5.7	4.07E-127	286.77
ASAP1	224791_at	-3.07	1.26E-058	126.83	MEF2C	209199_s_at	-46.21	2.41E-177	403.58
ASAP2	206414_s_at	-7.21	1.04E-089	199.52	MEX3B	223627_at	2.69	5.14E-046	97.24
ATHL1	219359_at	3.39	3.58E-059	128.11	MGC5566	220449_at	3.73	8.01E-071	155.43
ATP11A	230875_s_at	-3.07	3.07E-044	93.06	MICAL1	218376_s_at	-2.81	1.16E-064	140.97
ATP1B1	201242_s_at	6.28	7.26E-058	125.04	MICALL1	55081_at	-2.43	1.17E-054	117.51
ATP2A3	207521_s_at	2.3	6.64E-032	64.1	MKNK2	218205_s_at	-3.92	2.06E-102	229.11
ATP6V1C1	226463_at	-2.14	6.35E-034	68.84	MLLT11	211071_s_at	3.51	2.26E-066	144.99
ATXN7	209964_s_at	2.55	2.39E-066	144.93	MLXIP	202519_at	-3.94	1.33E-079	175.93
AUTS2	212599_at	-8.69	6.81E-078	171.95	MOBKL2B	226844_at	-3.05	9.12E-061	131.85
BACH2	227173_s_at	-3.01	1.38E-042	89.18	MSRA	219281_at	-4.08	1.22E-093	208.69
BAG3	217911_s_at	4.66	5.75E-048	101.82	MTMR10	225810_at	-2.58	4.82E-071	155.94
BANK1	219667_s_at	-21.41	9.44E-125	281.22	MYEF2	222771_s_at	3.01	6.77E-039	80.52

BCL11A	1559078_at	-4.59	1.69E-080	178.02	MYLIP	223130_s_at	-5.58	1.75E-083	184.98
BCL11B	222895_s_at	25.28	1.44E-140	317.99	MYO1G	227799_at	2.91	5.88E-054	115.87
BCL2L2	209311_at	-2.46	2.07E-054	116.93	MYO7B	235383_at	7.84	6.85E-090	199.94
BEX2	224367_at	6.15	3.11E-039	81.31	N4BP2L2	221899_at	-2.66	1.96E-051	109.97
BIN2	219191_s_at	4.5	9.95E-077	169.24	NADK	208918_s_at	-2.57	1.13E-049	105.84
BLNK	207655_s_at	-42.81	1.68E-149	338.96	NAT14	223284_at	2.31	4.97E-069	151.22
BTK	205504_at	-9.38	3.66E-145	328.76	NCF1	204961_s_at	-5.03	6.81E-086	190.59
C10orf18	227777_at	-3.2	5.35E-073	160.51	NCF1C	214084_x_at	-4.47	3.90E-087	193.5
C11orf80	204922_at	2.39	8.85E-087	192.67	NCF2	209949_at	-3.07	2.53E-043	90.91
C12orf57	224719_s_at	2.93	6.67E-067	146.23	NCF4	207677_s_at	-7.16	5.04E-077	169.93
C14orf139	219563_at	-5.43	2.79E-090	200.86	NDFIP1	222422_s_at	3.73	1.02E-091	204.21
C17orf63	222641_s_at	2.28	1.56E-074	164.11	NDST3	220429_at	3.94	1.23E-081	180.65
C1orf38	207571_x_at	-7.84	1.78E-119	268.85	NEK4	204634_at	-3.68	2.18E-065	142.66
C1QTNF4	223708_at	-4.38	1.06E-038	80.07	NEK6	223158_s_at	-2.81	1.26E-076	169
C21orf2	238965_at	-3.2	1.99E-073	161.52	NGFRAP1	217963_s_at	20.82	6.65E-108	241.94
C4orf32	227856_at	-6.5	1.03E-072	159.84	NGRN	224281_s_at	2.28	3.03E-063	137.66
C5orf62	223276_at	-7.41	4.89E-088	195.62	NKX3-1	209706_at	8	8.60E-069	150.66
C6orf204	228007_at	4.59	2.90E-078	172.82	NLRP1	210113_s_at	-2.33	9.95E-061	131.75
C7orf23	204215_at	-3.23	4.25E-071	156.07	NME7	227556_at	6.11	1.26E-086	192.31
C7orf40	225699_at	2.48	3.15E-054	116.51	NPTN	202228_s_at	-2.23	5.49E-037	76.04
CAMK4	229029_at	5.86	1.61E-079	175.74	NPY	206001_at	-11	1.20E-076	169.05
CAP1	213798_s_at	-2.81	7.46E-095	211.53	NR1D2	225768_at	-2.85	1.25E-040	84.59
CAPN3	210944_s_at	-4.76	5.70E-091	202.46	NR3C1	201865_x_at	-3.46	1.55E-071	157.09
CARD8	1554479_a_at	-3.66	1.80E-048	103.01	NRIP1	202599_s_at	-4.29	2.43E-039	81.56
CASP8	207686_s_at	3.05	1.70E-050	107.77	NT5E	203939_at	-3.16	1.01E-030	61.32
CCDC50	225331_at	-6.63	3.24E-110	247.39	NUCB2	229838_at	3.12	3.79E-050	106.95
CCNDBP1	223084_s_at	-2.17	1.80E-	131.15	NUP88	202900_s_at	-2.5	3.88E-	151.47

			060					069	
CCR9	207445_s_at	3.29	5.98E-042	87.68	OCIAD2	225314_at	8.17	8.92E-128	288.3
CD19	206398_s_at	-29.04	1.11E-184	420.72	OFD1	203569_s_at	-3.14	1.24E-096	215.69
CD1B	206749_at	13.45	3.96E-076	167.83	OGFRL1	226810_at	-12.3	5.78E-095	211.79
CD1E	215784_at	17.03	9.03E-061	131.86	OGN	222722_at	3.43	4.07E-059	127.98
CD2	205831_at	10.41	2.92E-086	191.45	OTUD1	226140_s_at	-4.56	2.96E-089	198.47
CD22	204581_at	-4.17	1.52E-083	185.12	P2RX1	210401_at	-2.81	7.73E-070	153.11
CD24	266_s_at	-28.05	2.51E-108	242.94	PABPC4L	238865_at	3.01	7.60E-049	103.89
CD247	210031_at	5.54	6.09E-102	228.01	PAG1	225626_at	-5.43	5.05E-067	146.51
CD28	206545_at	10.27	3.65E-091	202.91	PAOX	50400_at	-2.53	3.71E-074	163.24
CD3D	213539_at	43.41	5.03E-234	534.67	PAX5	221969_at	-71.01	5.05E-224	511.33
CD3E	205456_at	19.56	5.58E-208	474.5	PCDH10	228635_at	2.55	2.56E-034	69.77
CD3G	206804_at	16.68	2.53E-193	440.62	PDE4B	203708_at	-6.02	1.01E-052	112.99
CD5	230489_at	4.99	1.39E-126	285.52	PDK1	226452_at	-3.1	8.34E-050	106.15
CD58	211744_s_at	-4.06	5.04E-052	111.35	PDLIM1	208690_s_at	-14.62	1.46E-088	196.85
CD6	213958_at	4.56	8.20E-114	255.72	PDPK1	204524_at	-3.36	2.66E-101	226.52
CD7	214049_x_at	14.52	9.69E-162	367.47	PELO	218472_s_at	6.23	5.44E-097	216.54
CD72	215925_s_at	-18.25	1.67E-121	273.56	PEX5L	222910_s_at	2.53	1.85E-064	140.5
CD74	209619_at	-9.71	1.38E-130	294.78	PFKFB3	202464_s_at	-2.93	7.27E-035	71.06
CD79A	205049_s_at	-5.24	5.21E-085	188.54	PGAP1	220576_at	4.2	3.33E-064	139.9
CD79B	205297_s_at	-9.19	1.38E-116	262.17	PHACTR1	213638_at	-4.63	4.09E-085	188.78
CD8A	205758_at	4.86	5.51E-056	120.63	PHLPP1	212719_at	-4.35	6.02E-088	195.41
CD9	201005_at	-18.13	5.26E-090	200.22	PICALM	212511_at	-3.05	6.33E-068	148.64
CD99	201028_s_at	5.58	3.38E-076	168	PIK3AP1	226459_at	-22.78	2.66E-165	375.69
CDH2	203440_at	3.25	7.24E-057	122.71	PIK3CD	203879_at	-3.25	2.66E-079	175.23
CDK14	204604_at	-4.69	1.47E-092	206.18	PIK3R1	212239_at	-2.64	9.18E-041	84.9
CDKN1A	202284_s_at	-3.56	5.26E-057	123.03	PION	213142_x_at	-4.59	8.17E-054	115.53

CDYL	203098_at	-2.48	1.01E-061	134.1	PIP5K1B	205632_s_at	-3.07	6.82E-048	101.65
CELF2	242268_at	-4.11	5.65E-064	139.36	PLCG2	204613_at	-5.31	1.22E-123	278.62
CEP170	207719_x_at	-2.77	2.10E-047	100.5	PLCL1	205934_at	2.55	1.67E-050	107.79
CHD7	226123_at	-6.96	2.25E-115	259.32	PLDN	222826_at	2.6	1.54E-059	128.97
CHI3L2	213060_s_at	21.26	4.73E-107	239.96	PLXNB2	208890_s_at	-6.45	2.79E-101	226.46
CHRNA3	211772_x_at	2.23	1.97E-046	98.22	POU2AF1	205267_at	-7.11	1.96E-061	133.41
CHST15	203066_at	-9.99	2.35E-108	243.01	PPFIA1	202066_at	-2.58	2.54E-059	128.46
CIB2	205008_s_at	3.41	9.17E-118	264.89	PPIL4	1559263_s_at	-13.09	6.36E-124	279.29
CIITA	205101_at	-4.76	2.08E-103	231.45	PPP1R16B	233813_at	-3.61	1.18E-045	96.39
CLIP1	201975_at	-2.51	6.56E-067	146.24	PPP2R5C	201877_s_at	-3.27	2.11E-045	95.79
CLIP3	212358_at	3.56	8.21E-064	138.98	PPP3CA	202425_x_at	-6.32	8.20E-117	262.7
CLN8	229958_at	-4.69	1.52E-060	131.33	PPP3CC	32541_at	-4	3.36E-092	205.34
CPEB4	224831_at	-3.25	2.86E-055	118.95	PRDM2	203056_s_at	-3.39	2.53E-067	147.23
CPNE2	225129_at	-3.58	5.10E-053	113.68	PRDX5	1560587_s_at	2.68	4.86E-062	134.84
CRIM1	202552_s_at	-4.23	8.63E-056	120.17	PRKAR2A	225011_at	-3.43	3.17E-067	147
CRNDE	238022_at	4.23	1.20E-049	105.77	PRKCQ	210039_s_at	4.08	1.86E-116	261.86
CSF2RB	205159_at	-6.41	3.65E-067	146.85	PRKX	204061_at	-3.34	4.73E-059	127.83
CSRP2	207030_s_at	-28.05	6.08E-133	300.28	PRTFDC1	222803_at	7.67	1.52E-109	245.8
CTBP2	210554_s_at	-6.5	1.24E-096	215.68	PSD3	203355_s_at	-11.55	1.44E-111	250.53
CTGF	209101_at	-41.07	2.86E-110	247.52	PTCD3	228590_at	-2.62	9.70E-061	131.78
CTNNA1	200765_x_at	-6.19	4.19E-096	214.45	PTCRA	211837_s_at	2.81	1.53E-055	119.58
CUL1	207614_s_at	-3.86	8.90E-094	209.02	PTK2	208820_at	-6.19	1.59E-088	196.76
CXADR	203917_at	3.53	2.10E-048	102.85	PTP4A2	208615_s_at	2.85	1.19E-088	197.05
CYFIP1	208923_at	-4	3.36E-046	97.67	PTPN12	202006_at	-3.97	4.11E-058	125.63
CYTH3	225147_at	-4.44	1.43E-089	199.19	PTPN22	206060_s_at	2.68	8.98E-046	96.66
DARS	201623_s_at	-2.69	4.43E-064	139.61	PTPN7	204852_s_at	3.36	1.02E-071	157.51
DCAF12	224789_at	-2.75	8.02E-	92.08	PTPRE	221840_at	-6.11	1.10E-	183.12

			044					082	
DCLRE1C	222233_s_at	-3.76	4.03E-077	170.16	PXK	225796_at	-7.57	1.29E-150	341.6
DDX21	208152_s_at	-2.69	2.55E-072	158.91	QKI	212636_at	-3.34	1.87E-057	124.08
DENND3	212975_at	-3.81	3.19E-082	182.03	QSER1	229982_at	2.99	3.02E-075	165.78
DGKD	208072_s_at	-3.32	2.72E-085	189.19	RAB27B	228708_at	2.35	8.35E-029	56.81
DHCR24	200862_at	3.12	3.32E-056	121.16	RAB32	204214_s_at	5.28	9.53E-095	211.28
DHTKD1	209916_at	-3.58	4.44E-082	181.69	RAB33A	206039_at	3.53	2.41E-056	121.48
DKFZp451A211	1556114_a_at	-3.07	2.05E-057	123.99	RAB34	224710_at	-5.5	3.02E-070	154.08
DOCK11	226875_at	-3.39	5.89E-083	183.75	RABEP2	74694_s_at	-3.46	5.70E-071	155.77
DOK3	223553_s_at	-11.39	2.69E-136	308.06	RAPGEF6	1555247_a_at	-2.99	6.50E-087	192.98
DPP8	220939_s_at	-2.58	1.14E-060	131.62	RARRES3	204070_at	2.55	1.16E-049	105.81
DUSP22	226440_at	-2.68	1.36E-073	161.9	RASA1	202677_at	-3.39	2.65E-069	151.86
DUSP6	208892_s_at	-4.23	3.22E-032	64.84	RASSF2	203185_at	-5.58	3.94E-094	209.84
DYNLT1	201999_s_at	-3.78	3.23E-057	123.53	RCAN1	208370_s_at	-5.66	2.91E-057	123.63
EBF1	227646_at	-77.17	1.81E-195	445.73	RCBTB1	218352_at	-3.61	1.72E-066	145.27
EEA1	225885_at	-2.89	6.82E-049	104	RDX	212397_at	-2.89	5.34E-052	111.29
EGFR	1565483_at	2.64	1.46E-053	114.95	REEP3	225785_at	-5.62	2.84E-079	175.16
EGLN1	223046_at	-5.62	3.66E-056	121.05	REEP5	208873_s_at	-2.79	3.65E-046	97.59
EIF2C2	225569_at	-2.93	2.61E-047	100.28	REV3L	208070_s_at	-2.43	2.37E-044	93.32
EIF2C4	227930_at	-2.97	7.47E-065	141.42	RFTN1	212646_at	-3.46	6.61E-057	122.8
EIF4G3	201935_s_at	-2.71	5.63E-048	101.84	RGPD1	235597_s_at	3.63	1.04E-054	117.63
ELK3	221773_at	-6.77	8.37E-057	122.56	RGPD5	210676_x_at	6.41	3.63E-073	160.91
ELL2	226099_at	-4.92	2.75E-051	109.63	RGS10	204319_s_at	9.13	9.49E-100	222.92
ELL3	219518_s_at	-2.77	8.22E-050	106.16	RHBDF2	219202_at	-3.03	1.15E-079	176.08
ELOVL4	219532_at	12.91	4.53E-095	212.05	RHOB	212099_at	-4.92	1.28E-035	72.83
ELOVL5	215082_at	2.69	1.49E-059	129.01	RHOH	204951_at	2.93	1.15E-055	119.88
ENG	201809_s_at	-2.97	2.96E-064	140.03	RHOQ	212120_at	-2.81	7.34E-073	160.18

ENO2	201313_at	5.74	3.05E-103	231.06	RHOU	223168_at	7.62	6.35E-074	162.68
ENTPD4	204076_at	-3.76	1.05E-068	150.46	RICTOR	228248_at	2.46	6.15E-045	94.7
EPHB6	204718_at	6.63	4.08E-105	235.42	RIPK2	209545_s_at	-4.59	5.44E-076	167.51
ERN1	235745_at	3.07	1.15E-052	112.86	ROCK2	202762_at	-2.51	4.55E-048	102.06
ERO1LB	231944_at	-4.06	1.30E-091	203.97	ROGDI	218394_at	-2.51	2.72E-080	177.54
ETNK1	219017_at	-2.69	1.62E-039	81.98	RP2	205191_at	-3.36	2.20E-067	147.37
ETS1	224833_at	4.47	3.76E-063	137.44	RPL32P3	226877_at	-2.71	3.66E-081	179.55
F13A1	203305_at	-6.54	1.27E-054	117.42	RRM2B	223342_at	-2.93	3.33E-064	139.9
FAM107B	223058_at	-5.03	1.88E-067	147.53	RUFY3	229334_at	4.11	1.61E-070	154.72
FAM129C	230983_at	-9.78	1.68E-098	220.03	S100A13	202598_at	-3.32	8.45E-079	174.07
FAM135A	223497_at	-2.35	1.61E-053	114.85	SAMD13	229402_at	2.71	2.84E-061	133.04
FAM3C	201889_at	-5.24	2.07E-063	138.05	SCARB2	224983_at	-3.53	2.18E-060	130.96
FAM43A	227410_at	-3.58	1.17E-052	112.84	SCD	200832_s_at	16.45	5.52E-166	377.41
FAM69B	229002_at	-5.31	4.30E-074	163.08	SCGB3A1	230378_at	7.26	2.65E-063	137.8
FBXO33	226970_at	-2.85	2.54E-052	112.04	SCPEP1	218217_at	-6.5	2.03E-104	233.81
FGFR1	210973_s_at	2.73	4.08E-061	132.67	SELO	233167_at	-3.89	3.13E-066	144.65
FLI1	204236_at	-3.68	2.22E-061	133.29	SEPT1	227552_at	3.97	6.46E-070	153.3
FLJ39653	1553145_at	-4.5	1.23E-087	194.68	SEPW1	1555851_s_at	5.7	1.08E-095	213.49
FMNL2	226184_at	-6.54	2.01E-071	156.82	SETDB2	235339_at	-3.36	9.99E-057	122.38
FOXO1	202723_s_at	-11.96	5.05E-107	239.89	SGK223	235085_at	-3.76	1.35E-067	147.87
FRMD6	225464_at	3.23	7.90E-049	103.85	SH2D1A	210116_at	32.9	3.43E-182	414.76
FXYD2	207434_s_at	15.78	1.01E-074	164.55	SH2D2A	207351_s_at	2.66	8.87E-047	99.03
GAB1	229114_at	-9.51	3.27E-087	193.69	SH3KBP1	1554168_a_at	-2.97	6.53E-046	96.99
GALNT1	201724_s_at	-4.99	1.03E-069	152.82	SHOC2	202777_at	-2.41	5.44E-048	101.88
GALNT14	219271_at	-3.68	3.66E-066	144.49	SIDT2	56256_at	-3.07	1.73E-070	154.64
GALNT6	219956_at	8.46	2.51E-119	268.49	SIT1	205484_at	2.57	9.65E-050	106
GALNT7	218313_s_at	-2.95	2.26E-	102.78	SIVA1	203489_at	3.23	3.96E-	214.51

			048					096	
GAS2L3	238756_at	-4.47	4.15E-056	120.92	SKAP2	241074_at	7.31	7.53E-096	213.86
GATA3	209604_s_at	12.47	1.60E-094	210.76	SLA2	232234_at	5.39	2.71E-123	277.8
GBE1	203282_at	-3.16	1.14E-055	119.89	SLC22A16	232232_s_at	-2.79	4.65E-038	78.56
GGA2	208913_at	-5.66	5.13E-126	284.21	SLC25A13	203775_at	-2.58	2.06E-069	152.11
GNG7	206896_s_at	-4.76	6.62E-096	213.99	SLC4A4	203908_at	2.13	1.03E-039	82.43
GNS	212334_at	-2.25	5.33E-036	73.72	SLIT1	213601_at	3.48	4.11E-067	146.73
GPBP1	224648_at	-2.43	9.63E-039	80.16	SMAD5	225223_at	-2.41	6.30E-049	104.08
GPD1L	212510_at	-2.87	2.56E-070	154.24	SMARCA2	217707_x_at	-4.35	6.15E-075	165.05
GPR18	210279_at	-3.27	2.34E-039	81.6	SMCHD1	212569_at	-2.23	4.73E-044	92.62
GPSM1	226043_at	-3.29	7.91E-061	131.99	SMEK2	233759_s_at	-2.23	1.11E-035	72.97
GRAP2	208406_s_at	2.6	2.12E-095	212.81	SNX18	226683_at	-3.32	6.01E-057	122.89
GRB10	210999_s_at	-3.12	2.14E-073	161.45	SNX2	202114_at	-6.36	1.36E-165	376.43
GXYLT2	235371_at	33.59	1.35E-153	348.52	SNX30	226249_at	-4.44	5.96E-073	160.4
HBA1	214414_x_at	-5.24	1.34E-026	51.62	SNX9	223027_at	-2.77	9.49E-025	47.27
HBEGF	203821_at	-7.89	2.87E-080	177.49	SOD2	215078_at	-3.56	1.87E-044	93.56
HBG1	213515_x_at	-7.67	2.56E-040	83.86	SP3	232529_at	-4.82	1.25E-113	255.28
HES4	227347_x_at	11.71	1.70E-090	201.36	SPNS3	235900_at	4.29	1.21E-093	208.7
HEXB	201944_at	-3.46	2.54E-082	182.27	SPOPL	225658_at	-4.99	9.10E-096	213.66
HHEX	204689_at	-6.59	8.76E-080	176.35	SQLE	213577_at	2.41	1.13E-072	159.74
HHIP	237466_s_at	4.53	4.98E-070	153.57	SRGAP2P1	228628_at	-2.46	3.13E-058	125.91
HIST3H2A	221582_at	3.36	8.02E-049	103.83	SSBP3	217991_x_at	2.75	4.75E-082	181.62
HIVEP3	233884_at	2.71	2.79E-045	95.51	STARD4	226390_at	3.39	1.02E-058	127.04
HLA-DMA	217478_s_at	-17.03	4.06E-147	333.35	STK38	202951_at	-4	5.18E-082	181.53
HLA-DMB	203932_at	-8.28	6.09E-133	300.26	STK4	223746_at	-2.69	1.95E-055	119.34
HLA-DOA	226878_at	-8.06	1.57E-149	339.07	STRBP	229513_at	-5.21	4.07E-086	191.11
HLA-DPA1	211990_at	-19.7	3.72E-135	305.42	STRN3	204496_at	-2.33	6.19E-049	104.1

HLA-DPB1	201137_s_at	-12.64	4.18E-128	289.07	STX3	209238_at	-3.36	1.23E-084	187.67
HLA-DQA1	212671_s_at	-9	5.98E-054	115.85	STX7	212632_at	-3.43	3.60E-069	151.55
HLA-DQB1	209823_x_at	-6.54	1.02E-101	227.48	SUN1	212074_at	-2.35	1.35E-036	75.12
HLA-DRA	210982_s_at	-13.09	6.81E-125	281.56	SWAP70	209306_s_at	-5.43	1.26E-097	218.01
HLA-DRB1	208306_x_at	-7.73	3.03E-125	282.39	SYK	207540_s_at	-3.53	3.04E-070	154.07
HLA-DRB6	217362_x_at	-3.14	3.59E-088	195.93	SYVN1	225090_at	-2.62	3.14E-062	135.28
HLA-F	221978_at	-2.95	2.86E-067	147.1	TACC1	217437_s_at	-2.85	9.94E-056	120.03
HLX	214438_at	-6.68	6.04E-102	228.03	TAF1D	222728_s_at	-2.89	2.92E-067	147.08
HMGCR	202539_s_at	2.73	1.77E-049	105.37	TAP2	225973_at	-2.66	1.08E-048	103.53
HMGCS1	205822_s_at	2.71	6.81E-061	132.15	TAPT1	227407_at	-5.66	4.93E-108	242.26
HN1	217755_at	4.72	1.46E-103	231.82	TBK1	218520_at	-2.46	1.44E-062	136.07
HNRPLL	225386_s_at	8.69	1.79E-106	238.6	TCF4	203753_at	-9.45	4.87E-074	162.95
HOOK3	236192_at	-2.36	5.11E-055	118.35	TCF7	205254_x_at	5.54	4.14E-121	272.64
HSPA8	210338_s_at	3.16	1.84E-068	149.89	TCL1A	209995_s_at	-24.93	4.61E-094	209.68
ICOS	210439_at	5.74	2.40E-065	142.56	TEP1	228670_at	-2.55	3.20E-047	100.07
ICOSLG	228976_at	-3.89	1.30E-081	180.59	TERF2	229790_at	-3.36	1.50E-071	157.12
ID1	208937_s_at	6.77	4.16E-043	90.41	TFDP2	244043_at	4.38	4.66E-086	190.97
ID2	213931_at	-6.23	1.16E-065	143.32	TFEB	50221_at	-6.19	4.17E-133	300.68
IGF2BP3	203820_s_at	-12.73	4.49E-097	216.74	TGFBR1	224793_s_at	-3.76	7.47E-044	92.15
IGHD	213674_x_at	-4.56	3.47E-056	121.11	TLR1	210176_at	-5.39	6.56E-104	232.63
IGHM	209374_s_at	-16	5.59E-110	246.8	TMED4	224676_at	2.57	4.15E-067	146.72
IGSF10	1556579_s_at	3.32	1.19E-039	82.29	TMX4	201581_at	-4.86	3.04E-083	184.42
IKBKB	209341_s_at	-2.27	8.00E-040	82.7	TNFRSF10D	227345_at	-3.14	4.00E-053	113.93
IL32	203828_s_at	7.78	5.24E-061	132.41	TNS1	221748_s_at	-5.13	7.26E-054	115.65
INPP5D	203332_s_at	-4.44	6.92E-121	272.11	TNS3	217853_at	-7.11	7.14E-073	160.21
INSR	213792_s_at	-13.18	4.06E-147	333.34	TOP2B	211987_at	-2.95	1.19E-100	225
IRF8	204057_at	-4.23	3.62E-	137.48	TOX	204529_s_at	5.62	1.70E-	145.28

			063					066	
IRS2	209184_s_at	-6.68	1.36E-079	175.9	TRA@	217143_s_at	9.78	1.34E-044	93.9
IRX5	210239_at	3.32	3.28E-042	88.29	TRAF3IP2	215411_s_at	-2.71	5.94E-064	139.31
ITK	211339_s_at	4.53	1.48E-056	121.98	TRAK1	214924_s_at	-2.43	8.06E-077	169.46
ITM2A	202747_s_at	13.45	3.10E-092	205.42	TRAT1	217147_s_at	14.62	4.00E-135	305.32
JAM3	212813_at	5.17	2.06E-092	205.84	TRBC1	211796_s_at	49.52	5.03E-169	384.46
JMJD1C	228793_at	-7.26	3.15E-078	172.74	TRD@	213830_at	7.16	7.72E-044	92.12
JUP	201015_s_at	-4.66	4.45E-075	165.38	TRIB1	202241_at	-4.76	1.26E-047	101.02
KANK2	218418_s_at	-5.78	1.86E-072	159.23	TRIM44	217759_at	-2.83	7.10E-049	103.96
KCNK5	219615_s_at	2.58	2.82E-072	158.81	TRIO	209012_at	-2.95	1.59E-065	143
KCTD1	226245_at	5.13	1.32E-088	196.95	TSHR	215442_s_at	3.76	8.92E-065	141.24
KCTD15	222668_at	-5.7	4.55E-082	181.67	TSNAX	203983_at	-3.66	3.37E-062	135.21
KCTD6	238077_at	-2.99	4.88E-073	160.6	TSPAN13	217979_at	-14.32	8.77E-116	260.28
KDM1B	227021_at	-2.81	7.04E-052	111.01	TSPAN14	223314_at	-3.89	1.79E-106	238.6
KIAA0368	214356_s_at	-2.93	8.71E-039	80.26	TST	209605_at	-2.79	1.30E-044	93.93
KIAA0748	222920_s_at	4.08	4.50E-103	230.66	UBA3	209115_at	-2.35	5.71E-055	118.24
KIAA1033	212795_at	-5.5	1.77E-109	245.64	UBASH3A	220418_at	4.63	3.77E-107	240.2
KIAA1432	226222_at	-2.55	1.54E-058	126.63	UBASH3B	228359_at	-3.68	9.82E-069	150.53
KIF16B	219570_at	-2.97	1.24E-069	152.63	UBE2I	208760_at	-2.31	3.99E-044	92.79
KIF26A	234307_s_at	-2.48	2.86E-067	147.1	UBR1	226921_at	-2.85	1.75E-050	107.74
KIN	205664_at	-2.36	2.04E-048	102.88	UBR5	1555888_at	-4.11	2.04E-074	163.84
KLF2	219371_s_at	-4.14	4.27E-062	134.97	USP4	202682_s_at	-2.43	2.92E-086	191.45
KLHL2	219157_at	-8.75	5.39E-089	197.86	USP44	224048_at	2.35	4.66E-057	123.15
KLRAQ1	1553955_at	-2.75	6.61E-039	80.54	USP48	225925_s_at	-2.48	3.03E-066	144.68
KPNA1	202056_at	-2.33	7.52E-047	99.2	UXS1	225583_at	-5.7	1.39E-122	276.16
KRT2	207908_at	2.28	5.09E-029	57.31	VANGL1	229997_at	7.78	4.25E-139	314.55
LAPTM5	201721_s_at	-2.99	8.76E-084	185.68	VAT1	208626_s_at	3.68	6.76E-097	216.31

LAT	211005_at	16.22	1.12E-144	327.61	VCL	200931_s_at	-3.81	2.56E-067	147.22
LAT2	221581_s_at	-7.52	7.50E-074	162.5	VPREB1	221349_at	-12.73	4.87E-074	162.94
LCK	204890_s_at	22.16	5.00E-161	365.77	VPREB3	220068_at	-16.8	5.08E-103	230.53
LCOR	228454_at	-2.81	6.01E-063	136.96	VRK2	205126_at	-2.58	7.67E-051	108.58
LCORL	240592_at	-2.64	1.31E-058	126.79	VSIG10	226485_at	-5.78	5.80E-078	172.11
LCP2	244556_at	2.19	9.24E-062	134.19	WASF1	204165_at	-3.27	7.93E-034	68.62
LDOC1L	223228_at	5.7	1.03E-077	171.53	WDFY1	224800_at	-5.39	1.05E-103	232.16
LILRA2	207857_at	-14.93	2.55E-141	319.77	WDR37	211383_s_at	-2.53	4.35E-053	113.84
LILRB1	229937_x_at	-6.45	1.37E-072	159.55	WDSUB1	226668_at	-3.89	1.14E-083	185.42
LILRB2	207697_x_at	-4.2	1.64E-084	187.37	YIPF5	221423_s_at	2.3	1.45E-065	143.09
LIME1	219541_at	3.68	9.89E-109	243.9	YPEL2	227020_at	-5.31	9.19E-088	194.97
LOC100130458	239214_at	-11.31	7.62E-116	260.43	YPEL3	223179_at	-3.18	1.86E-072	159.23
LOC202181	220609_at	-2.43	3.83E-051	109.29	ZAK	225662_at	-3.03	7.37E-038	78.09
LOC220729	230077_at	-5.03	1.05E-077	171.51	ZAP70	1555613_a_at	7.21	7.63E-106	237.14
LOC284757	236846_at	2.68	8.50E-053	113.16	ZBTB8A	235142_at	8	2.40E-072	158.97
LOC285957	243601_at	3.01	3.36E-051	109.42	ZCCHC14	212655_at	-2.85	1.81E-059	128.81
LOC389043	240546_at	4.38	8.20E-062	134.31	ZCCHC7	226496_at	-31.12	1.59E-182	415.65
LOC441453	217551_at	2.36	1.21E-044	94.01	ZEB2	203603_s_at	-17.39	2.35E-151	343.34
LOC541471	225799_at	9.38	5.67E-102	228.1	ZHX2	203556_at	-3.76	1.76E-057	124.15
LOC728606	1559276_at	5.1	6.67E-099	220.96	ZMYND8	209048_s_at	-6.11	4.05E-122	275.05
LOC729082	225332_at	-2.93	1.61E-045	96.07	ZNF70	243816_at	-3.53	2.69E-082	182.21
LOC90925	210450_at	-2.71	1.13E-047	101.13	ZNRF1	223383_at	2.55	1.40E-044	93.86
LPGAT1	202651_at	-3.12	4.68E-059	127.84	ZSWIM6	226208_at	-4.47	1.27E-071	157.29
LPIN1	212274_at	-3.1	2.52E-087	193.95	ZYG11B	225350_s_at	-2.53	6.29E-056	120.5

11.15 **Anexo 15 (TOP2B.txt).** Ejemplo de los resultados de enriquecimiento de genes (1)

Resultados TOP2B:

PROTEINAS CON LAS QUE INTERACTUA DE LA LISTA PROPORCIONADA:

CD3E TOP2B

- - - - - - - - - - - - -
- - - - - - - - - - - - -
- - - - - - - - - - - - -
- - - - - - - - - - -

MODIFICACIONES POST-TRADUCCIONALES:

PHOSPHORYLATION 1126S 1127S 1128S 1130S 1132T 1134S 1207S 1231S 1287T 1331S
1335S 1337S 1339S 1365Y 1366T 1370S 1395S 1398T 1408S 1416Y 1417T 1419S
1436S 1447S 1448Y 1449S 1452S 1456S 1461S 1466S 1468S 1471S 1496S 1517S
1519S 1521S 152Y 1545S 1547S 1553Y 1570T 1571S 1576S 1587T 1591S 1604Y
1608S 32S 40S 41S 634T 635S 971S 973Y 98Y

ACETYLATION 360K 362K 368K 987K

- - - - - - - - - - - - -
- - - - - - - - - - - - -
- - - - - - - - - - - - -
- - - - - - - - - - -

RUTAS EN LAS QUE ESTA IMPLICADO:

- - - - - - - - - - - - -
- - - - - - - - - - - - -
- - - - - - - - - - - - -
- - - - - - - - - - -

PROTEINAS CON LAS QUE INTERACTUA:

BAZ1B
CD3E
CSNK2B
ESR1
HDAC1
HDAC2
MAPK3
MTA2
NOP56
OLA1
PARP1
PML
RARA
RFC3
SRRM1
SUMO1
SUMO2
SUMO3

SUMO4
TOP2A
TOP2B
TOPBP1
TP53
USF1
UVRAG
XPO1
XRCC5

11.16 **Anexo 16 (CD1B.txt).** Ejemplo de los resultados de enriquecimiento de genes (2)

Resultados CD1B:

PROTEINAS CON LAS QUE INTERACTUA DE LA LISTA PROPORCIONADA:

- - - - - - - - - - - - -
- - - - - - - - - - - - -
- - - - - - - - - - - - -
- - - - - - - - - - -

MODIFICACIONES POST-TRADUCCIONALES:

GLYCOSYLATION 258N 75N

- - - - - - - - - - - - -
- - - - - - - - - - - - -
- - - - - - - - - - - - -
- - - - - - - - - - -

RUTAS EN LAS QUE ESTA IMPLICADO:

KEGG Natural killer cell mediated cytotoxicity CD24 CR1 CR2 CSF1 CSF1R CSF2 CSF2RA
CSF3 CSF3R CD55 DNTT EPO EPOR FCER2 FCGR1A FLT3 FLT3LG
GP1BA GP1BB GP5 GP9 ANPEP GYPA HLA-DRA HLA-DRB1 HLA-DRB3
HLA-DRB4 HLA-DRB5 IL1A IL1B IL1R1 IL2RA IL3 IL3RA IL4 IL4R
IL5 IL5RA IL6 IL6R IL7 IL7R IL9R IL11 IL11RA ITGA6 ITGA1 ITGA2
ITGA2B ITGA3 ITGA4 ITGA5 ITGAM ITGB3 KIT KITLG MME TFRC THPO
TNF TPO IL1R2 CD1A CD1B CD1C CD1D CD1E CD2 CD3D CD3E CD3G
CD4 CD5 CD7 CD8A CD8B CD9 CD14 CD19 MS4A1 CD22 CD33 CD34
CD36 CD37 CD38 CD44 CD59

- - - - - - - - - - - - -
- - - - - - - - - - - - -
- - - - - - - - - - - - -
- - - - - - - - - - -

PROTEINAS CON LAS QUE INTERACTUA:

B2M
PSAP

11.17 **Anexo 17.** Resultado de las rutas KEGG con genes con expresión disminuida.

RUTAS	Genes		
	Implicados en la ruta	Sub-expresados	%
ASTHMA	32	8	25.000
B CELL RECEPTOR SIGNALING PATHWAY	76	18	23.684
STAPHYLOCOCCUS AUREUS INFECTION	57	8	14.035
LEISHMANIASIS	74	10	13.514
OTHER GLYCAN DEGRADATION	18	2	11.111
FC GAMMA R-MEDIATED PHAGOCYTOSIS	96	10	10.417
TYPE II DIABETES MELLITUS	49	5	10.204
GLYCOSAMINOGLYCAN DEGRADATION	20	2	10.000
PHAGOSOME	156	15	9.615
VIBRIO CHOLERAE INFECTION	55	5	9.091
SULFUR RELAY SYSTEM	11	1	9.091
INSULIN SIGNALING PATHWAY	139	12	8.633
CHRONIC MYELOID LEUKEMIA	74	6	8.108
SHIGELLOSIS	63	5	7.937
PHOSPHATIDYLINOSITOL SIGNALING SYSTEM	79	6	7.595
AMYOTROPHIC LATERAL SCLEROSIS (ALS)	54	4	7.407
NICOTINATE AND NICOTINAMIDE METABOLISM	27	2	7.407
NON-HOMOLOGOUS END-JOINING	15	1	6.667
GLYCOSPHINGOLIPID BIOSYNTHESIS	16	1	6.250
SMALL CELL LUNG CANCER	86	5	5.814
LYSOSOME	122	7	5.738
SNARE INTERACTIONS IN VESICULAR TRANSPORT	37	2	5.405
OOCYTE MEIOSIS	115	6	5.217
INOSITOL PHOSPHATE METABOLISM	58	3	5.172
AMOEBIASIS	107	5	4.673
PROGESTERONE-MEDIATED OOCYTE MATURATION	88	4	4.545
CIRCADIAN RHYTHM	24	1	4.167
AMINO SUGAR AND NUCLEOTIDE SUGAR METABOLISM	49	2	4.082
GLUTAMATERGIC SYNAPSE	127	5	3.937
MATURITY ONSET DIABETES OF THE YOUNG	26	1	3.846
GLYCEROPHOSPHOLIPID METABOLISM	83	3	3.614
COLLECTING DUCT ACID SECRETION	28	1	3.571
UBIQUITIN MEDIATED PROTEOLYSIS	140	5	3.571
STARCH AND SUCROSE METABOLISM	56	2	3.571
CYTOSOLIC DNA-SENSING PATHWAY	63	2	3.175
SYNAPTIC VESICLE CYCLE	65	2	3.077
PRION DISEASES	36	1	2.778
CYSTEINE AND METHIONINE METABOLISM	36	1	2.778
FRUCTOSE AND MANNOSE METABOLISM	37	1	2.703

SALMONELLA INFECTION	88	2	2.273
FATTY ACID METABOLISM	45	1	2.222
DILATED CARDIOMYOPATHY	91	2	2.198
LYSINE DEGRADATION	50	1	2.000
GLYCEROLIPID METABOLISM	51	1	1.961
GLUTATHIONE METABOLISM	51	1	1.961
JAK-STAT SIGNALING PATHWAY	156	3	1.923
TASTE TRANSDUCTION	53	1	1.887
PATHOGENIC ESCHERICHIA COLI INFECTION	58	1	1.724
AMINOACYL-TRNA BIOSYNTHESIS	64	1	1.563
COMPLEMENT AND COAGULATION CASCADES	70	1	1.429
LONG-TERM DEPRESSION	71	1	1.408
PERTUSSIS	75	1	1.333
PARKINSON'S DISEASE	131	1	0.763
OXIDATIVE PHOSPHORYLATION	133	1	0.752
OLFACTORY TRANSDUCTION	389	1	0.257

11.18 **Anexo 18.** Resultado de las rutas KEGG con genes con expresión aumentada.

RUTAS	Genes		
	Implicados en la ruta	Sobre-expresados	%
TERPENOID BACKBONE BIOSYNTHESIS	16	2	12.500
STEROID BIOSYNTHESIS	20	2	10.000
SYNTHESIS AND DEGRADATION OF KETONE BODIES	10	1	10.000
BIOSYNTHESIS OF UNSATURATED FATTY ACIDS	22	2	9.091
PROXIMAL TUBULE BICARBONATE RECLAMATION	24	2	8.333
DORSO-VENTRAL AXIS FORMATION	26	2	7.692
MINERAL ABSORPTION	52	2	3.846
GLYCOSYLPHOSPHATIDYLINOSITOL(GPI)-ANCHOR BIOSYNTHE	26	1	3.846
GLYCOSAMINOGLYCAN BIOSYNTHESIS	27	1	3.704
PROTEIN DIGESTION AND ABSORPTION	82	3	3.659
BASAL CELL CARCINOMA	56	2	3.571
THYROID CANCER	30	1	3.333
BUTANOATE METABOLISM	31	1	3.226
CARDIAC MUSCLE CONTRACTION	78	2	2.564
VALINE, LEUCINE AND ISOLEUCINE DEGRADATION	45	1	2.222
OTHER TYPES OF O-GLYCAN BIOSYNTHESIS	47	1	2.128
RETINOL METABOLISM	65	1	1.538
GLYCOLYSIS - GLUCONEOGENESIS	67	1	1.493
SPLICEOSOME	129	1	0.775

11.19 **Anexo 19.** Resultado de las rutas KEGG con genes con expresión mixta.

RUTAS	Genes				
	Implicados en la ruta	Sub-expresados	%	Sobre-expresados	%
PRIMARY IMMUNODEFICIENCY	36	6	16.667	6	16.667
T CELL RECEPTOR SIGNALING PATHWAY	109	7	6.422	13	11.927
HEMATOPOIETIC CELL LINEAGE	89	6	6.742	9	10.112
INTESTINAL IMMUNE NETWORK FOR IGA PRODUCTION	50	9	18.000	3	6.000
CELL ADHESION MOLECULES (CAMS)	136	12	8.824	8	5.882
BILE SECRETION	72	2	2.778	4	5.556
MEASLES	135	4	2.963	7	5.185
PANCREATIC SECRETION	104	1	0.962	5	4.808
CHAGAS DISEASE (AMERICAN TRYPANOSOMIASIS)	105	4	3.810	5	4.762
ALDOSTERONE-REGULATED SODIUM REABSORPTION	43	5	11.628	2	4.651
CARBOHYDRATE DIGESTION AND ABSORPTION	45	2	4.444	2	4.444
PROSTATE CANCER	90	6	6.667	4	4.444
VIRAL MYOCARDITIS	73	10	13.699	3	4.110
ADHERENS JUNCTION	74	5	6.757	3	4.054
ENDOCRINE AND OTHER FACTOR-REGULATED CALCIUM REABS	50	2	4.000	2	4.000
ENDOMETRIAL CANCER	53	4	7.547	2	3.774
AUTOIMMUNE THYROID DISEASE	55	9	16.364	2	3.636
NATURAL KILLER CELL MEDIATED CYTOTOXICITY	142	8	5.634	5	3.521
LEUKOCYTE TRANSENDOTHELIAL MIGRATION	117	9	7.692	4	3.419
MUCIN TYPE O-GLYCAN BIOSYNTHESIS	31	3	9.677	1	3.226
EPITHELIAL CELL SIGNALING IN HELICOBACTER PYLORI I	69	5	7.246	2	2.899
MELANOMA	72	3	4.167	2	2.778
RNA DEGRADATION	72	1	1.389	2	2.778
ARRHYTHMOGENIC RIGHT VENTRICULAR CARDIOMYOPATHY (A	75	2	2.667	2	2.667
PEROXISOME	79	3	3.797	2	2.532
ANTIGEN PROCESSING AND PRESENTATION	79	11	13.924	2	2.532
ALLOGRAFT REJECTION	40	9	22.500	1	2.500
MRNA SURVEILLANCE PATHWAY	85	1	1.176	2	2.353
OSTEOCLAST DIFFERENTIATION	129	15	11.628	3	2.326

BLADDER CANCER	43	1	2.326	1	2.326
GRAFT-VERSUS-HOST DISEASE	44	9	20.455	1	2.273
ABC TRANSPORTERS	45	1	2.222	1	2.222
VASOPRESSIN-REGULATED WATER REABSORPTION	45	2	4.444	1	2.222
SALIVARY SECRETION	90	2	2.222	2	2.222
TYPE I DIABETES MELLITUS	46	9	19.565	1	2.174
PATHWAYS IN CANCER	328	13	3.963	7	2.134
NOTCH SIGNALING PATHWAY	48	1	2.083	1	2.083
MTOR SIGNALING PATHWAY	53	3	5.660	1	1.887
HTLV-I INFECTION	268	19	7.090	5	1.866
NON-SMALL CELL LUNG CANCER	55	5	9.091	1	1.818
ALZHEIMER'S DISEASE	168	4	2.381	3	1.786
CHOLINERGIC SYNAPSE	113	5	4.425	2	1.770
HEDGEHOG SIGNALING PATHWAY	57	1	1.754	1	1.754
ACUTE MYELOID LEUKEMIA	59	4	6.780	1	1.695
CALCIUM SIGNALING PATHWAY	178	6	3.371	3	1.685
NOD-LIKE RECEPTOR SIGNALING PATHWAY	60	4	6.667	1	1.667
COLORECTAL CANCER	63	3	4.762	1	1.587
NEUROTROPHIN SIGNALING PATHWAY	128	9	7.031	2	1.563
AXON GUIDANCE	131	6	4.580	2	1.527
GLIOMA	66	4	6.061	1	1.515
TIGHT JUNCTION	133	2	1.504	2	1.504
TOXOPLASMOSIS	134	13	9.701	2	1.493
MAPK SIGNALING PATHWAY	273	13	4.762	4	1.465
SYSTEMIC LUPUS ERYTHEMATOSUS	139	8	5.755	2	1.439
ADIPOCYTOKINE SIGNALING PATHWAY	70	4	5.714	1	1.429
P53 SIGNALING PATHWAY	70	3	4.286	1	1.429
LONG-TERM POTENTIATION	71	3	4.225	1	1.408
PANCREATIC CANCER	71	4	5.634	1	1.408
RENAL CELL CARCINOMA	71	4	5.634	1	1.408
PPAR SIGNALING PATHWAY	72	2	2.778	1	1.389
RIG-I-LIKE RECEPTOR SIGNALING PATHWAY	72	2	2.778	1	1.389
BACTERIAL INVASION OF EPITHELIAL CELLS	72	7	9.722	1	1.389
GASTRIC ACID SECRETION	75	2	2.667	1	1.333
WNT SIGNALING PATHWAY	152	7	4.605	2	1.316
VEGF SIGNALING PATHWAY	77	6	7.792	1	1.299
FC EPSILON RI SIGNALING PATHWAY	80	8	10.000	1	1.250
PROTEIN PROCESSING IN ENDOPLASMIC RETICULUM	168	4	2.381	2	1.190
TGF-BETA SIGNALING PATHWAY	86	4	4.651	1	1.163
APOPTOSIS	87	10	11.494	1	1.149
ERBB SIGNALING PATHWAY	88	7	7.955	1	1.136
GAP JUNCTION	91	2	2.198	1	1.099
RHEUMATOID ARTHRITIS	93	9	9.677	1	1.075
CHEMOKINE SIGNALING PATHWAY	190	9	4.737	2	1.053
GNRH SIGNALING PATHWAY	102	3	2.941	1	0.980

MELANOGENESIS	102	2	1.961	1	0.980
PYRIMIDINE METABOLISM	102	3	2.941	1	0.980
ENDOCYTOSIS	205	9	4.390	2	0.976
TOLL-LIKE RECEPTOR SIGNALING PATHWAY	103	5	4.854	1	0.971
REGULATION OF ACTIN CYTOSKELETON	215	9	4.186	2	0.930
METABOLIC PATHWAYS	1139	18	1.580	10	0.878
VASCULAR SMOOTH MUSCLE CONTRACTION	127	3	2.362	1	0.787
CELL CYCLE	129	2	1.550	1	0.775
HEPATITIS C	135	6	4.444	1	0.741
CYTOKINE-CYTOKINE RECEPTOR INTERACTION	276	3	1.087	2	0.725
RNA TRANSPORT	154	3	1.948	1	0.649
NEUROACTIVE LIGAND-RECEPTOR INTERACTION	319	3	0.940	2	0.627
PURINE METABOLISM	165	5	3.030	1	0.606
INFLUENZA A	177	15	8.475	1	0.565
TUBERCULOSIS	181	18	9.945	1	0.552
HUNTINGTON'S DISEASE	184	2	1.087	1	0.543
FOCAL ADHESION	201	6	2.985	1	0.498

11.20 **Anexo 20.** Resultado de las rutas Reactome con genes con expresión disminuida.

RUTAS	Genes		
	Implicados en la ruta	Sub-expresados	%
TRANSCRIPTIONAL ACTIVATION OF CELL CYCLE INHIBITOR P21	3	1	33.333
TRANSCRIPTIONAL ACTIVATION OF P53 RESPONSIVE GENES	3	1	33.333
POLYAMINES ARE OXIDIZED TO AMINES, ALDEHYDES AND H2O2 BY PAOS	3	1	33.333
REGULATION OF SIGNALING BY CBL	19	5	26.316
FORMATION OF APOPTOSOME	4	1	25.000
SHC ACTIVATION	4	1	25.000
INTERCONVERSION OF POLYAMINES	4	1	25.000
SIGNAL ATTENUATION	14	3	21.429
AMINE OXIDASE REACTIONS	5	1	20.000
ACTIVATION OF PKB	5	1	20.000
ACTIVATION OF CASPASES THROUGH APOPTOSOME-MEDIATED CLEAVAGE	6	1	16.667
CYTOCHROME C-MEDIATED APOPTOTIC RESPONSE	6	1	16.667
IRS ACTIVATION	6	1	16.667
DARPP-32 EVENTS	25	4	16.000
INTERLEUKIN-3, 5 AND GM-CSF SIGNALING	46	7	15.217
INTERFERON GAMMA SIGNALING	74	11	14.865
P75NTR RECRUITS SIGNALLING COMPLEXES	14	2	14.286
RSK ACTIVATION	7	1	14.286
VIRAL DSRNA:TLR3:TRIF COMPLEX ACTIVATES TBK1/IKK EPSILON	7	1	14.286

INTERLEUKIN RECEPTOR SHC SIGNALING	29	4	13.793
ERK/MAPK TARGETS	22	3	13.636
DCC MEDIATED ATTRACTIVE SIGNALING	15	2	13.333
APOPTOTIC FACTOR-MEDIATED RESPONSE	8	1	12.500
NEGATIVE REGULATION OF THE PI3K/AKT NETWORK	8	1	12.500
ACTIVATION OF BAD AND TRANSLOCATION TO MITOCHONDRIA	8	1	12.500
TAK1 ACTIVATES NFKB BY PHOSPHORYLATION AND ACTIVATION OF IKKS COMPLEX	25	3	12.000
NUCLEAR EVENTS (KINASE AND TRANSCRIPTION FACTOR ACTIVATION)	25	3	12.000
PKA ACTIVATION	17	2	11.765
P75NTR SIGNALS VIA NF-KB	17	2	11.765
PKA ACTIVATION IN GLUCAGON SIGNALLING	18	2	11.111
PKA-MEDIATED PHOSPHORYLATION OF CREB	18	2	11.111
GLYCOGEN SYNTHESIS	9	1	11.111
BINDING OF RNA BY INSULIN-LIKE GROWTH FACTOR-2 MRNA BINDING PROTEINS (IGF2BPS/IMPS/VICKZS)	9	1	11.111
IKK COMPLEX RECRUITMENT MEDIATED BY RIP1	9	1	11.111
NICOTINATE METABOLISM	9	1	11.111
ORGANIC CATION TRANSPORT	9	1	11.111
ACTIVATION OF IRF3/IRF7 MEDIATED BY TBK1/IKK EPSILON	9	1	11.111
INTEGRIN ALPHAIIB BETA3 SIGNALING	28	3	10.714
VIRAL DSRNA:TLR3:TRIF COMPLEX ACTIVATES RIP1	19	2	10.526
FORMYL PEPTIDE RECEPTORS BIND FORMYL PEPTIDES AND MANY OTHER LIGANDS	10	1	10.000
ROLE OF ABL IN ROBO-SLIT SIGNALING	10	1	10.000
SEMA4D IN SEMAPHORIN SIGNALING	30	3	10.000
AKT PHOSPHORYLATES TARGETS IN THE NUCLEUS	10	1	10.000
ENDOSOMAL/VACUOLAR PATHWAY	10	1	10.000
ELEVATION OF CYTOSOLIC CA2 LEVELS	10	1	10.000
INTERFERON SIGNALING	111	11	9.910
MAPK TARGETS/ NUCLEAR EVENTS MEDIATED BY MAP KINASES	31	3	9.677
TRIF MEDIATED TLR3 SIGNALING	75	7	9.333
TOLL LIKE RECEPTOR 3 (TLR3) CASCADE	75	7	9.333
TOLL LIKE RECEPTOR 4 (TLR4) CASCADE	97	9	9.278
MYD88-INDEPENDENT CASCADE INITIATED ON PLASMA MEMBRANE	76	7	9.211
TOLL RECEPTOR CASCADES	109	10	9.174
ADENYLATE CYCLASE ACTIVATING PATHWAY	11	1	9.091
INFLAMMASOMES	22	2	9.091
COPI MEDIATED TRANSPORT	11	1	9.091
GOLGI TO ER RETROGRADE TRANSPORT	11	1	9.091
IRAK1 RECRUITS IKK COMPLEX	11	1	9.091
IRAK1 RECRUITS IKK COMPLEX UPON TLR7/8 OR 9 STIMULATION	11	1	9.091
PURINE CATABOLISM	11	1	9.091
GP1B-IX-V ACTIVATION SIGNALLING	11	1	9.091
NETRIN MEDIATED REPULSION SIGNALS	11	1	9.091
PROTEOLYTIC CLEAVAGE OF SNARE COMPLEX PROTEINS	22	2	9.091

GLUCAGON SIGNALING IN METABOLIC REGULATION	34	3	8.824
TRAF6 MEDIATED INDUCTION OF PROINFLAMMATORY CYTOKINES	69	6	8.696
TRAF6 MEDIATED NF-KB ACTIVATION	23	2	8.696
MICRORNA (MIRNA) BIOGENESIS	23	2	8.696
REGULATORY RNA PATHWAYS	23	2	8.696
ACTIVATED TLR4 SIGNALLING	93	8	8.602
NFKB AND MAP KINASES ACTIVATION MEDIATED BY TLR4 SIGNALING REPERTOIRE	72	6	8.333
GROWTH HORMONE RECEPTOR SIGNALING	24	2	8.333
TOLL LIKE RECEPTOR 9 (TLR9) CASCADE	86	7	8.140
MAP KINASE ACTIVATION IN TLR CASCADE	50	4	8.000
ANTIGEN PRESENTATION: FOLDING, ASSEMBLY AND PEPTIDE LOADING OF CLASS I MHC	25	2	8.000
SEMA4D INDUCED CELL MIGRATION AND GROWTH-CONE COLLAPSE	25	2	8.000
BOTULINUM NEUROTOXICITY	25	2	8.000
MYD88:MAL CASCADE INITIATED ON PLASMA MEMBRANE	88	7	7.955
TOLL LIKE RECEPTOR 2 (TLR2) CASCADE	88	7	7.955
TOLL LIKE RECEPTOR TLR1:TLR2 CASCADE	88	7	7.955
TOLL LIKE RECEPTOR TLR6:TLR2 CASCADE	88	7	7.955
PLATELET AGGREGATION (PLUG FORMATION)	38	3	7.895
INSULIN RECEPTOR RECYCLING	26	2	7.692
ERKS ARE INACTIVATED	13	1	7.692
G BETA:GAMMA SIGNALLING THROUGH PI3KGAMMA	26	2	7.692
PYRIMIDINE CATABOLISM	13	1	7.692
REGULATION OF PYRUVATE DEHYDROGENASE (PDH) COMPLEX	13	1	7.692
TRAF6 MEDIATED INDUCTION OF NFKB AND MAP KINASES UPON TLR7/8 OR 9 ACTIVATION	81	6	7.407
MYD88 DEPENDENT CASCADE INITIATED ON ENDOSOME	82	6	7.317
TOLL LIKE RECEPTOR 7/8 (TLR7/8) CASCADE	82	6	7.317
INTERLEUKIN-1 SIGNALING	41	3	7.317
MYD88 CASCADE INITIATED ON PLASMA MEMBRANE	83	6	7.229
TOLL LIKE RECEPTOR 10 (TLR10) CASCADE	83	6	7.229
TOLL LIKE RECEPTOR 5 (TLR5) CASCADE	83	6	7.229
ADENYLATE CYCLASE INHIBITORY PATHWAY	14	1	7.143
INHIBITION OF ADENYLATE CYCLASE PATHWAY	14	1	7.143
ADVANCED GLYCOSYLATION ENDPRODUCT RECEPTOR SIGNALING	14	1	7.143
THE NLRP3 INFLAMMASOME	14	1	7.143
AKT PHOSPHORYLATES TARGETS IN THE CYTOSOL	14	1	7.143
NF-KB IS ACTIVATED AND SIGNALS SURVIVAL	14	1	7.143
ACTIVATION OF DNA FRAGMENTATION FACTOR	14	1	7.143
APOPTOSIS INDUCED DNA FRAGMENTATION	14	1	7.143
ORGANIC CATION/ANION/ZWITTERION TRANSPORT	14	1	7.143
G-PROTEIN BETA:GAMMA SIGNALLING	29	2	6.897
COMMON PATHWAY	15	1	6.667
SOS-MEDIATED SIGNALLING	15	1	6.667
PLATELET ADHESION TO EXPOSED COLLAGEN	15	1	6.667

TRAF3-DEPENDENT IRF ACTIVATION PATHWAY	15	1	6.667
PLATELET DEGRANULATION	79	5	6.329
METABOLISM OF POLYAMINES	16	1	6.250
BETA-CATENIN PHOSPHORYLATION CASCADE	16	1	6.250
GRB2:SOS PROVIDES LINKAGE TO MAPK SIGNALING FOR INTERGRINS	16	1	6.250
P130CAS LINKAGE TO MAPK SIGNALING FOR INTEGRINS	16	1	6.250
SIGNALING BY TGF BETA	16	1	6.250
RESPONSE TO ELEVATED PLATELET CYTOSOLIC CA2	84	5	5.952
CIRCADIAN CLOCK	34	2	5.882
TRANSPORT OF RIBONUCLEOPROTEINS INTO THE HOST NUCLEUS	34	2	5.882
PLATELET SENSITIZATION BY LDL	17	1	5.882
JNK (C-JUN KINASES) PHOSPHORYLATION AND ACTIVATION MEDIATED BY ACTIVATED HUMAN TAK1	17	1	5.882
SHC-RELATED EVENTS	18	1	5.556
ACTIVATION OF BH3-ONLY PROTEINS	18	1	5.556
ACTIVATED TAK1 MEDIATES P38 MAPK ACTIVATION	18	1	5.556
ACTIVATED AMPK STIMULATES FATTY-ACID OXIDATION IN MUSCLE	19	1	5.263
VPR-MEDIATED NUCLEAR IMPORT OF PICS	38	2	5.263
PYRUVATE METABOLISM	19	1	5.263
TIE2 SIGNALING	19	1	5.263
SYNTHESIS AND INTERCONVERSION OF NUCLEOTIDE DI- AND TRIPHOSPHATES	19	1	5.263
ACTIVATION OF GABAB RECEPTORS	39	2	5.128
GABA B RECEPTOR ACTIVATION	39	2	5.128
PROSTACYCLIN SIGNALLING THROUGH PROSTACYCLIN RECEPTOR	20	1	5.000
INTERACTIONS OF VPR WITH HOST CELLULAR PROTEINS	41	2	4.878
G BETA:GAMMA SIGNALLING THROUGH PLC BETA	21	1	4.762
ADP SIGNALLING THROUGH P2Y PURINOCEPTOR 12	22	1	4.545
PRESYNAPTIC FUNCTION OF KAINATE RECEPTORS	22	1	4.545
REGULATION OF INSULIN SECRETION BY GLUCAGON-LIKE PEPTIDE-1	44	2	4.545
SEMAPHORIN INTERACTIONS	67	3	4.478
NEPHRIN INTERACTIONS	23	1	4.348
THROMBOXANE SIGNALLING THROUGH TP RECEPTOR	24	1	4.167
BMAL1:CLOCK/NPAS2 ACTIVATES GENE EXPRESSION	24	1	4.167
SIGNALING BY BMP	24	1	4.167
CYTOSOLIC TRNA AMINOACYLATION	25	1	4.000
PYRIMIDINE METABOLISM	25	1	4.000
ADP SIGNALLING THROUGH P2Y PURINOCEPTOR 1	26	1	3.846
ACTIVATION OF G PROTEIN GATED POTASSIUM CHANNELS	26	1	3.846
INHIBITION OF VOLTAGE GATED CA2 CHANNELS VIA GBETA/GAMMA SUBUNITS	26	1	3.846
NUCLEAR RECEPTOR TRANSCRIPTION PATHWAY	52	2	3.846
GABA RECEPTOR ACTIVATION	54	2	3.704
REGULATION OF LIPID METABOLISM BY PEROXISOME PROLIFERATOR-ACTIVATED RECEPTOR ALPHA (PPARALPHA)	56	2	3.571
TRANSFERRIN ENDOCYTOSIS AND RECYCLING	28	1	3.571

KINESINS	28	1	3.571
CREB PHOSPHORYLATION THROUGH THE ACTIVATION OF RAS	28	1	3.571
SCF(SKP2)-MEDIATED DEGRADATION OF P27/P21	57	2	3.509
G-PROTEIN ACTIVATION	29	1	3.448
RECYCLING PATHWAY OF L1	29	1	3.448
MEIOTIC SYNAPSIS	58	2	3.448
INHIBITION OF INSULIN SECRETION BY ADRENALINE/NORADRENALINE	30	1	3.333
REGULATION OF GLUCOKINASE BY GLUCOKINASE REGULATORY PROTEIN	30	1	3.333
ACTIVATION OF KAINATE RECEPTORS UPON GLUTAMATE BINDING	31	1	3.226
TRAF6 MEDIATED IRF7 ACTIVATION	31	1	3.226
INWARDLY RECTIFYING K CHANNELS	32	1	3.125
SIGNAL AMPLIFICATION	32	1	3.125
NUCLEAR IMPORT OF REV PROTEIN	32	1	3.125
PACKAGING OF TELOMERE ENDS	32	1	3.125
ER-PHAGOSOME PATHWAY	65	2	3.077
INTERFERON ALPHA/BETA SIGNALING	65	2	3.077
SIGNALING BY ROBO RECEPTOR	33	1	3.030
CYCLIN E ASSOCIATED EVENTS DURING G1/S TRANSITION	66	2	3.030
FORMATION OF FIBRIN CLOT (CLOTTING CASCADE)	33	1	3.030
THROMBIN SIGNALLING THROUGH PROTEINASE ACTIVATED RECEPTORS (PARS)	33	1	3.030
PURINE METABOLISM	33	1	3.030
TRANSPORT OF THE SLBP INDEPENDENT MATURE MRNA	33	1	3.030
CYCLIN A:CDK2-ASSOCIATED EVENTS AT S PHASE ENTRY	67	2	2.985
DEGRADATION OF BETA-CATENIN BY THE DESTRUCTION COMPLEX	68	2	2.941
SIGNALING BY WNT	68	2	2.941
GLUCAGON-TYPE LIGAND RECEPTORS	34	1	2.941
REV-MEDIATED NUCLEAR EXPORT OF HIV-1 RNA	34	1	2.941
TRANSPORT OF THE SLBP DEPENDANT MATURE MRNA	34	1	2.941
NEGATIVE REGULATORS OF RIG-I/MDA5 SIGNALING	34	1	2.941
METABOLISM OF NUCLEOTIDES	71	2	2.817
INTERACTIONS OF REV WITH HOST CELLULAR PROTEINS	36	1	2.778
NEP/NS2 INTERACTS WITH THE CELLULAR EXPORT MACHINERY	36	1	2.778
TRANSPORT OF MATURE MRNA DERIVED FROM AN INTRONLESS TRANSCRIPT	36	1	2.778
EXPORT OF VIRAL RIBONUCLEOPROTEINS FROM NUCLEUS	37	1	2.703
TRANSPORT OF MATURE MRNAS DERIVED FROM INTRONLESS TRANSCRIPTS	37	1	2.703
IRON UPTAKE AND TRANSPORT	38	1	2.632
ANTIGEN PROCESSING-CROSS PRESENTATION	76	2	2.632
GLUCOSE TRANSPORT	41	1	2.439
PYRUVATE METABOLISM AND CITRIC ACID (TCA) CYCLE	41	1	2.439
INTEGRATION OF ENERGY METABOLISM	126	3	2.381
TRNA AMINOACYLATION	43	1	2.326
HEXOSE TRANSPORT	43	1	2.326

MEIOSIS	86	2	2.326
MITOTIC PROMETAPHASE	93	2	2.151
M PHASE	97	2	2.062
REGULATION OF INSULIN SECRETION	99	2	2.020
METABOLISM OF NON-CODING RNA	50	1	2.000
SNRNP ASSEMBLY	50	1	2.000
MUSCLE CONTRACTION	50	1	2.000
CLASS I MHC MEDIATED ANTIGEN PROCESSING & PRESENTATION	252	5	1.984
AMYLOIDS	52	1	1.923
METABOLISM OF VITAMINS AND COFACTORS	52	1	1.923
METABOLISM OF WATER-SOLUBLE VITAMINS AND COFACTORS	52	1	1.923
TRANSPORT OF MATURE MRNA DERIVED FROM AN INTRON-CONTAINING TRANSCRIPT	52	1	1.923
G1/S TRANSITION	110	2	1.818
SCF-BETA-TRCP MEDIATED DEGRADATION OF EMI1	55	1	1.818
TRANSPORT OF MATURE TRANSCRIPT TO CYTOPLASM	56	1	1.786
S PHASE	113	2	1.770
P53-DEPENDENT G1 DNA DAMAGE RESPONSE	58	1	1.724
P53-DEPENDENT G1/S DNA DAMAGE CHECKPOINT	58	1	1.724
TELOMERE MAINTENANCE	60	1	1.667
G1/S DNA DAMAGE CHECKPOINTS	61	1	1.639
LOSS OF NLP FROM MITOTIC CENTROSOMES	63	1	1.587
LOSS OF PROTEINS REQUIRED FOR INTERPHASE MICROTUBULE ORGANIZATION�Ã�, ÂFROM THE CENTROSOME	63	1	1.587
DNA REPLICATION	201	3	1.493
TRANSCRIPTIONAL REGULATION OF WHITE ADIPOCYTE DIFFERENTIATION	70	1	1.429
PHASE 1 - FUNCTIONALIZATION OF COMPOUNDS	70	1	1.429
ORC1 REMOVAL FROM CHROMATIN	71	1	1.408
SWITCHING OF ORIGINS TO A POST-REPLICATIVE STATE	71	1	1.408
ANTIGEN PROCESSING: UBIQUITINATION & PROTEASOME DEGRADATION	214	3	1.402
INFLUENZA LIFE CYCLE	145	2	1.379
REMOVAL OF LICENSING FACTORS FROM ORIGINS	73	1	1.370
CENTROSOME MATURATION	73	1	1.370
RECRUITMENT OF MITOTIC CENTROSOME PROTEINS AND COMPLEXES	73	1	1.370
INFLUENZA INFECTION	150	2	1.333
REGULATION OF DNA REPLICATION	76	1	1.316
REGULATION OF APC/C ACTIVATORS BETWEEN G1/S AND EARLY ANAPHASE	78	1	1.282
CHROMOSOME MAINTENANCE	80	1	1.250
APC/C-MEDIATED DEGRADATION OF CELL CYCLE PROTEINS	83	1	1.205
REGULATION OF MITOTIC CELL CYCLE	83	1	1.205
G2/M TRANSITION	85	1	1.176
MITOTIC G2-G2/M PHASES	88	1	1.136
MITOTIC M-M/G1 PHASES	179	2	1.117
CLASS B/2 (SECRETIN FAMILY RECEPTORS)	91	1	1.099

LATE PHASE OF HIV LIFE CYCLE	95	1	1.053
SYNTHESIS OF DNA	97	1	1.031
TRANSPORT OF GLUCOSE AND OTHER SUGARS, BILE SALTS AND ORGANIC ACIDS, METAL IONS AND AMINE COMPOUNDS	97	1	1.031
POTASSIUM CHANNELS	100	1	1.000
REGULATION OF GENE EXPRESSION IN BETA CELLS	103	1	0.971
HIV LIFE CYCLE	114	1	0.877
REGULATION OF BETA-CELL DEVELOPMENT	115	1	0.870
CELL CYCLE CHECKPOINTS	118	1	0.847
GENERIC TRANSCRIPTION PATHWAY	245	2	0.816
THE CITRIC ACID (TCA) CYCLE AND RESPIRATORY ELECTRON TRANSPORT	131	1	0.763
PROCESSING OF CAPPED INTRON-CONTAINING PRE-MRNA	139	1	0.719
BIOLOGICAL OXIDATIONS	140	1	0.714
GENE EXPRESSION	589	4	0.679
MRNA PROCESSING	158	1	0.633
METABOLISM OF AMINO ACIDS AND DERIVATIVES	175	1	0.571
FORMATION AND MATURATION OF MRNA TRANSCRIPT	186	1	0.538

11.21 **Anexo 21.** Resultado de las rutas Reactome con genes con expresión aumentada.

RUTAS	Genes		
	Implicados en la ruta	Sobre-expresados	%
CAMK IV-MEDIATED PHOSPHORYLATION OF CREB	4	1	25.000
NEF MEDIATED DOWNREGULATION OF CD28 CELL SURFACE EXPRESSION	4	1	25.000
CREB PHOSPHORYLATION THROUGH THE ACTIVATION OF CAMKK	5	1	20.000
ACTIVATION, MYRISTOLYATION OF BID AND TRANSLOCATION TO MITOCHONDRIA	5	1	20.000
NEF AND SIGNAL TRANSDUCTION	10	2	20.000
FGFR1C AND KLOTHO LIGAND BINDING AND ACTIVATION	5	1	20.000
CHOLESTEROL BIOSYNTHESIS	23	4	17.391
FASL/ CD95L SIGNALING	6	1	16.667
CD28 DEPENDENT VAV1 PATHWAY	12	2	16.667
TRAIL SIGNALING	7	1	14.286
NADE MODULATES DEATH SIGNALLING	7	1	14.286
PREGNENOLONE BIOSYNTHESIS	7	1	14.286
TNF SIGNALING	8	1	12.500
HIGHLY SODIUM PERMEABLE ACETYLCHOLINE NICOTINIC RECEPTORS	8	1	12.500
KLOTHO-MEDIATED LIGAND BINDING	8	1	12.500
ATTACHMENT OF GPI ANCHOR TO UPAR	8	1	12.500
ACTIVATION OF PRO-CASPASE 8	10	1	10.000

CASPASE-8 IS FORMED FROM PROCASPASE-8	10	1	10.000
HIGHLY CALCIUM PERMEABLE NICOTINIC ACETYLCHOLINE RECEPTORS	10	1	10.000
BICARBONATE TRANSPORTERS	10	1	10.000
REDUCTION OF CYTOSOLIC CA²⁺ LEVELS	11	1	9.091
PASSIVE TRANSPORT BY AQUAPORINS	12	1	8.333
HIGHLY CALCIUM PERMEABLE POSTSYNAPTIC NICOTINIC ACETYLCHOLINE RECEPTORS	12	1	8.333
CASPASE-MEDIATED CLEAVAGE OF CYTOSKELETAL PROTEINS	13	1	7.692
PRESYNAPTIC NICOTINIC ACETYLCHOLINE RECEPTORS	13	1	7.692
FGFR1C LIGAND BINDING AND ACTIVATION	13	1	7.692
DEATH RECEPTOR SIGNALLING	14	1	7.143
EXTRINSIC PATHWAY FOR APOPTOSIS	14	1	7.143
ACETYLCHOLINE BINDING AND DOWNSTREAM EVENTS	15	1	6.667
ACTIVATION OF NICOTINIC ACETYLCHOLINE RECEPTORS	15	1	6.667
POSTSYNAPTIC NICOTINIC ACETYLCHOLINE RECEPTORS	15	1	6.667
GRB2 EVENTS IN EGFR SIGNALING	15	1	6.667
SHC1 EVENTS IN EGFR SIGNALING	16	1	6.250
FGFR1 LIGAND BINDING AND ACTIVATION	16	1	6.250
ABCA TRANSPORTERS IN LIPID HOMEOSTASIS	18	1	5.556
SIGNAL TRANSDUCTION BY L1	36	2	5.556
GABA SYNTHESIS, RELEASE, REUPTAKE AND DEGRADATION	20	1	5.000
FGFR LIGAND BINDING AND ACTIVATION	25	1	4.000
BASIGIN INTERACTIONS	26	1	3.846
G0 AND EARLY G1	26	1	3.846
POST-TRANSLATIONAL MODIFICATION: SYNTHESIS OF GPI-ANCHORED PROTEINS	27	1	3.704
SHC-MEDIATED CASCADE	30	1	3.333
STEROID HORMONES	30	1	3.333
NEUROTRANSMITTER RELEASE CYCLE	37	1	2.703
METABOLISM OF STEROID HORMONES AND VITAMINS A AND D	37	1	2.703
ABC-FAMILY PROTEINS MEDIATED TRANSPORT	39	1	2.564
FRS2-MEDIATED CASCADE	39	1	2.564
NEGATIVE REGULATION OF FGFR SIGNALING	41	1	2.439
CHEMOKINE RECEPTORS BIND CHEMOKINES	55	1	1.818
DESTABILIZATION OF MRNA BY AUF1 (HNRNP D0)	55	1	1.818
REGULATION OF MRNA STABILITY BY PROTEINS THAT BIND AU-RICH ELEMENTS	87	1	1.149
TRANSPORT OF INORGANIC CATIONS/ANIONS AND AMINO ACIDS/OLIGOPEPTIDES	95	1	1.053
POST-TRANSLATIONAL PROTEIN MODIFICATION	124	1	0.806
METABOLISM OF MRNA	219	1	0.457
METABOLISM OF PROTEINS	297	1	0.337

11.22 **Anexo 22.** Resultado de las rutas Reactome con genes con expresión mixta.

RUTAS	Genes				
	Implicados en la ruta	Sub-expresados	%	Sobre-expresados	%
GENERATION OF SECOND MESSENGER MOLECULES	43	8	18.605	9	20.930
TRANSLOCATION OF ZAP-70 TO IMMUNOLOGICAL SYNAPSE	30	8	26.667	6	20.000
PHOSPHORYLATION OF CD3 AND TCR ZETA CHAINS	32	9	28.125	6	18.750
PD-1 SIGNALING	35	8	22.857	6	17.143
TCR SIGNALING	70	15	21.429	11	15.714
DOWNSTREAM TCR SIGNALING	53	14	26.415	8	15.094
CD28 DEPENDENT PI3K/AKT SIGNALING	22	2	9.091	3	13.636
SYNTHESIS OF VERY LONG-CHAIN FATTY ACYL-COAS	15	1	6.667	2	13.333
COSTIMULATION BY THE CD28 FAMILY	78	13	16.667	10	12.821
CD28 CO-STIMULATION	32	3	9.375	4	12.500
FATTY ACYL-COA BIOSYNTHESIS	19	1	5.263	2	10.526
THE ROLE OF NEF IN HIV-1 REPLICATION AND DISEASE PATHOGENESIS	30	1	3.333	3	10.000
NEF MEDIATED CD4 DOWN-REGULATION	11	1	9.091	1	9.091
NEF-MEDIATES DOWN MODULATION OF CELL SURFACE RECEPTORS BY RECRUITING THEM TO CLATHRIN ADAPTERS	23	1	4.348	2	8.696
NF-KB ACTIVATION THROUGH FADD/RIP-1 PATHWAY MEDIATED BY CASPASE-8 AND -10	13	1	7.692	1	7.692
PECAM1 INTERACTIONS	13	2	15.385	1	7.692
SIGNAL REGULATORY PROTEIN (SIRP) FAMILY INTERACTIONS	14	1	7.143	1	7.143
EGFR INTERACTS WITH PHOSPHOLIPASE C-GAMMA	34	2	5.882	2	5.882
GPVI-MEDIATED ACTIVATION CASCADE	34	6	17.647	2	5.882
REGULATION OF KIT SIGNALING	17	1	5.882	1	5.882
TRIGLYCERIDE BIOSYNTHESIS	35	2	5.714	2	5.714
IMMUNOREGULATORY INTERACTIONS BETWEEN A LYMPHOID AND A NON-LYMPHOID CELL	126	4	3.175	7	5.556
ION TRANSPORT BY P-TYPE ATPASES	37	1	2.703	2	5.405
CELL SURFACE INTERACTIONS AT THE VASCULAR WALL	95	4	4.211	5	5.263
APOPTOTIC CLEAVAGE OF CELLULAR PROTEINS	39	1	2.564	2	5.128
PLATELET CALCIUM HOMEOSTASIS	20	1	5.000	1	5.000
GAB1 SIGNALOSOME	40	7	17.500	2	5.000
ION CHANNEL TRANSPORT	62	1	1.613	3	4.839
NETRIN-1 SIGNALING	43	2	4.651	2	4.651
CTLA4 INHIBITORY SIGNALING	22	3	13.636	1	4.545
LYSOSOME VESICLE BIOGENESIS	25	2	8.000	1	4.000
EFFECTS OF PIP2 HYDROLYSIS	26	1	3.846	1	3.846
SIGNALING BY SCF-KIT	79	7	8.861	3	3.797

APOPTOTIC EXECUTION PHASE	53	2	3.774	2	3.774
CAM PATHWAY	27	2	7.407	1	3.704
CALMODULIN INDUCED EVENTS	27	2	7.407	1	3.704
PHOSPHOLIPASE C-MEDIATED CASCADE	55	2	3.636	2	3.636
GOLGI ASSOCIATED VESICLE BIOGENESIS	55	4	7.273	2	3.636
EGFR DOWNREGULATION	28	1	3.571	1	3.571
GLYCOLYSIS	28	1	3.571	1	3.571
ADAPTIVE IMMUNE SYSTEM	483	26	5.383	17	3.520
CA-DEPENDENT EVENTS	29	2	6.897	1	3.448
PI-3K CASCADE	58	6	10.345	2	3.448
PIP3 ACTIVATES AKT SIGNALING	29	4	13.793	1	3.448
ADHERENS JUNCTIONS INTERACTIONS	30	2	6.667	1	3.333
CDO IN MYOGENESIS	30	3	10.000	1	3.333
MYOGENESIS	30	3	10.000	1	3.333
INTRINSIC PATHWAY FOR APOPTOSIS	31	2	6.452	1	3.226
CLATHRIN DERIVED VESICLE BUDDING	62	5	8.065	2	3.226
TRANS-GOLGI NETWORK VESICLE BUDDING	62	5	8.065	2	3.226
DAG AND IP3 SIGNALING	32	2	6.250	1	3.125
NOD1/2 SIGNALING PATHWAY	32	2	6.250	1	3.125
GLUCONEOGENESIS	32	1	3.125	1	3.125
CELL DEATH SIGNALLING VIA NRAGE, NRIF AND NADE	65	1	1.538	2	3.077
DOWNSTREAM SIGNALING OF ACTIVATED FGFR	101	8	7.921	3	2.970
POST NMDA RECEPTOR ACTIVATION EVENTS	34	1	2.941	1	2.941
PLC-GAMMA1 SIGNALLING	35	2	5.714	1	2.857
SIGNALING BY EGFR	110	10	9.091	3	2.727
ACTIVATION OF NMDA RECEPTOR UPON GLUTAMATE BINDING AND POSTSYNAPTIC EVENTS	38	1	2.632	1	2.632
PI3K/AKT ACTIVATION	38	6	15.789	1	2.632
SIGNALING BY FGFR	115	8	6.957	3	2.609
CYCLIN D ASSOCIATED EVENTS IN G1	39	2	5.128	1	2.564
G1 PHASE	39	2	5.128	1	2.564
REGULATION OF WATER BALANCE BY RENAL AQUAPORINS	41	3	7.317	1	2.439
METABOLISM OF LIPIDS AND LIPOPROTEINS	293	3	1.024	7	2.389
RHO GTPASE CYCLE	126	4	3.175	3	2.381
SIGNALING BY RHO GTPASES	126	4	3.175	3	2.381
INTERLEUKIN-2 SIGNALING	43	5	11.628	1	2.326
P75 NTR RECEPTOR-MEDIATED SIGNALLING	87	3	3.448	2	2.299
IMMUNE SYSTEM	788	42	5.330	18	2.284
PLC BETA MEDIATED EVENTS	44	2	4.545	1	2.273
G-PROTEIN MEDIATED EVENTS	45	2	4.444	1	2.222
HOST INTERACTIONS OF HIV FACTORS	136	3	2.206	3	2.206
G ALPHA (Z) SIGNALLING EVENTS	46	2	4.348	1	2.174
DOWNSTREAM SIGNAL TRANSDUCTION	94	8	8.511	2	2.128
L1CAM INTERACTIONS	95	2	2.105	2	2.105
AQUAPORIN-MEDIATED TRANSPORT	48	3	6.250	1	2.083

NRAGE SIGNALS DEATH THROUGH JNK	48	1	2.083	1	2.083
ACTIVATION OF CHAPERONES BY IRE1ALPHA	49	1	2.041	1	2.041
HEMOSTASIS	468	23	4.915	9	1.923
NUCLEOTIDE-BINDING DOMAIN, LEUCINE RICH REPEAT CONTAINING RECEPTOR (NLR) SIGNALING PATHWAYS	52	4	7.692	1	1.923
SIGNALLING BY NGF	222	14	6.306	4	1.802
FATTY ACID, TRIACYLGLYCEROL, AND KETONE BODY METABOLISM	113	3	2.655	2	1.770
CELL-CELL JUNCTION ORGANIZATION	60	2	3.333	1	1.667
SIGNALING BY PDGF	123	8	6.504	2	1.626
GLUCOSE METABOLISM	63	3	4.762	1	1.587
TRANSMISSION ACROSS CHEMICAL SYNAPSES	191	3	1.571	3	1.571
CELL-CELL COMMUNICATION	130	4	3.077	2	1.538
AXON GUIDANCE	267	8	2.996	4	1.498
HIV INFECTION	201	3	1.493	3	1.493
MEMBRANE TRAFFICKING	134	5	3.731	2	1.493
UNFOLDED PROTEIN RESPONSE	67	1	1.493	1	1.493
NGF SIGNALLING VIA TRKA FROM THE PLASMA MEMBRANE	137	11	8.029	2	1.460
NEUROTRANSMITTER RECEPTOR BINDING AND DOWNSTREAM TRANSMISSION IN THE POSTSYNAPTIC CELL	137	3	2.190	2	1.460
PLATELET ACTIVATION, SIGNALING AND AGGREGATION	206	14	6.796	3	1.456
NCAM SIGNALING FOR NEURITE OUT-GROWTH	71	1	1.408	1	1.408
PI3K CASCADE	71	5	7.042	1	1.408
TRANSMEMBRANE TRANSPORT OF SMALL MOLECULES	428	7	1.636	6	1.402
APOPTOSIS	149	4	2.685	2	1.342
RIG-I/MDA5 MEDIATED INDUCTION OF IFN-ALPHA/BETA PATHWAYS	77	3	3.896	1	1.299
G ALPHA (12/13) SIGNALLING EVENTS	78	5	6.410	1	1.282
OPIOID SIGNALLING	81	6	7.407	1	1.235
PLATELET HOMEOSTASIS	82	3	3.659	1	1.220
IRS-MEDIATED SIGNALLING	82	5	6.098	1	1.220
IRS-RELATED EVENTS	82	5	6.098	1	1.220
CELL JUNCTION ORGANIZATION	85	2	2.353	1	1.176
INTEGRIN CELL SURFACE INTERACTIONS	86	3	3.488	1	1.163
INSULIN RECEPTOR SIGNALLING CASCADE	87	6	6.897	1	1.149
NEURONAL SYSTEM	290	5	1.724	3	1.034
DEVELOPMENTAL BIOLOGY	495	14	2.828	5	1.010
G ALPHA (I) SIGNALLING EVENTS	201	5	2.488	2	0.995
SIGNAL TRANSDUCTION	1512	40	2.646	15	0.992
SIGNALING BY INTERLEUKINS	107	10	9.346	1	0.935
SIGNALING BY INSULIN RECEPTOR	110	7	6.364	1	0.909
G ALPHA (S) SIGNALLING EVENTS	126	3	2.381	1	0.794
FACTORS INVOLVED IN MEGAKARYOCYTE DEVELOPMENT AND PLATELET PRODUCTION	126	5	3.968	1	0.794
METABOLISM OF CARBOHYDRATES	127	4	3.150	1	0.787

MITOTIC G1-G1/S PHASES	136	2	1.471	1	0.735
CLASS A/1 (RHODOPSIN-LIKE RECEPTORS)	306	3	0.980	2	0.654
SIGNALING BY GPCR	962	15	1.559	6	0.624
GPCR DOWNSTREAM SIGNALING	867	12	1.384	5	0.577
G ALPHA (Q) SIGNALLING EVENTS	187	5	2.674	1	0.535
PEPTIDE LIGAND-BINDING RECEPTORS	187	2	1.070	1	0.535
GPCR LIGAND BINDING	411	4	0.973	2	0.487
CYTOKINE SIGNALING IN IMMUNE SYSTEM	221	22	9.955	1	0.452
DIABETES PATHWAYS	230	2	0.870	1	0.435
SLC-MEDIATED TRANSMEMBRANE TRANSPORT	251	2	0.797	1	0.398
INNATE IMMUNE SYSTEM	263	11	4.183	1	0.380
METABOLISM OF RNA	265	1	0.377	1	0.377
CELL CYCLE, MITOTIC	331	5	1.511	1	0.302
METABOLISM	922	9	0.976	1	0.108

11.23 Anexo 23 (BCL11B.txt).

PROTEINAS CON LAS QUE INTERACTUA DE LA LISTA PROPORCIONADA:

MODIFICACIONES POST-TRADUCCIONALES:

ACETYLATION 780K 851K
PHOSPHORYLATION 120T 129S 256S 260T 277S 313T 318S 358S 375S 376T
381S 398S 401S 406T 417T 483S 488S 493S 496S 497S 502T 678S
682S 753S 754T 765S 772S 776T 97S

RUTAS EN LAS QUE ESTA IMPLICADO:

PROTEINAS CON LAS QUE INTERACTUA:

CBX5
CHD4
HDAC1
HDAC2
HDAC3
MBD3
MTA1
MTA2
NR2F1
NR2F2
RBBP4
RBBP7
SP1
SUV39H1
TAL1

11.24 Anexo 24 (SP1.txt).

PROTEINAS CON LAS QUE INTERACTUA DE LA LISTA PROPORCIONADA:

CREBBP SP1

MODIFICACIONES POST-TRADUCCIONALES:

PROTEOLYTIC CLEAVAGE	590D										
SUMOYLATION	16K										
PHOSPHORYLATION	101S	278T	2S	375T	40S	41S	42S	43T	453T	46S	
	48S	56S	579T	59S	617S	641S	651T	668T	670S	681T	728S	732S
	739T	7S										
GLYCOSYLATION	491S	612S	640T	641S	698S	702S						
ACETYLATION	2S	703K										

RUTAS EN LAS QUE ESTA IMPLICADO:

KEGG TGF-beta signaling pathway (se omite el listado de genes de la ruta)
KEGG Huntington's disease (se omite el listado de genes de la ruta)

PROTEINAS CON LAS QUE INTERACTUA (se separan por comas y no por salto de línea):
AATF, ABL1, AHR, AR, ARHGAP21, ARNT, ASH2L, ATF7IP, ATF7IP2, BCL11B, BCL6, BCOR, BRCA1, BRCA2, C14orf106, C3orf17, CABP1, CASP3, CBX5, CCNA1, CCNA2, CCND1, CD2, CDK2, CDK4, CEBPB, CHD3, CREBBP, CRK, CRX, CSNK2A1, CTBP1, CTCFL, DCAF6, DDX5, DNMT1, DROSHA, E2F1, E2F2, E2F3, EGR1, ELF1, EP300, EPAS1, ESR1, ESR2, ESRRA, ESRRB, ESRRG, ETS1, EXOSC10, FYN, GABPA, GATA1, GATA3, GATA4, GATA6, GRB2, GRIN1, GTF2B, HBZ, HCFC1, HDAC1, HDAC2, HDAC3, HDAC4, HIF1A, HMGA1, HNF4A, HOXC11, HSP90AA1, HSPA8, HTT, JUN, KIF1A, KLF10, KLF4, KLF6, LDB1, LMO2, MAN1A2, MAPK1, MAPK11, MAPK14, MAPK3, MAX, MECP2, MEF2C, MEF2D, MGEA5, MIER1, MSX1, MT-ATP6, MT-ND4, MT-ND5, MYC, MYCN, MYOD1, MYOG, NAP1L1, NCK1, NCL, NCOR1, NCOR2, NFKB1, NFKB2, NFYA, NFYB, NFYC, NKX3-1, NOS3, NR2E1, NR2F1, NR5A1, OGT, OSBP, PARP1, PER3, PFDN5, PGR, PHC2, PHOX2A, PIAS2, PML, POGZ, POU2F1, PPARG, PPIG, PPP1R13L, PRKCZ, PRKDC, PSIP1, PSMC5, PURA, RAD51, RARA, RB1, RBBP4, RBL1, REL, RELA, REST, RNASEN, RNF19A, RNF4, RUNX1, RXRA, S1PR1, SENP6, SF3A1, SFPQ, SHC1, SIN3A, SKIV2L2, SKP2, SMAD2, SMAD3, SMAD4, SMARCA4, SMARCC1, SMARCC2, SOX10, SOX8, SP1, SP4, SRC, SREBF1, SREBF2, SRF, SRY, STAT3, SUB1, SUMO1, SYK, TAF4, TAF7, TAL1, TBP, TBX21, TLX3, TP53, TPI1, UBC, USP39, VHL, YY1, ZBTB16, ZBTB7A y ZBTB7B

11.25 Anexo 25 (CREBBP.txt).

PROTEINAS CON LAS QUE INTERACTUA DE LA LISTA PROPORCIONADA:

CREBBP SP1

MODIFICACIONES POST-TRADUCCIONALES:

METHYLATION	601R	702R	714R	742R	768R						
ACETYLATION	1014K	1165K	1178K	1203K	1216K	1376K	1535K	1545K	1548K	1549K	1553K
	1554K	1557K	1559K	1583K	1586K	1587K	1588K	1591K	1592K	1595K	1597K	1627K
	1703K	1706K	1711K	1741K	1744K	1797K	976K					
PHOSPHORYLATION	1001T	1030S	1034S	1038S	1072S	1076S	121S	124S	1382S	1386S		
	1391Y	1450Y	1457Y	1460Y	1717S	1725S	1733S	1755S	1763S	1771S	2025S	2038S
	2040S	2041S	2063S	2065S	2076S	2078S	2079S	2284S	2322S	2334S	2339S	2344S
	2352T	2356S	2368S	2372S	2377S	2382S	2390T	2394S	2406S	274S	302S	437S

659Y 963T 992S

RUTAS EN LAS QUE ESTA IMPLICADO:

(Se omiten 33 rutas en que participa (19 de Reactome y 14 de KEGG))

PROTEINAS CON LAS QUE INTERACTUA (se separan por comas y no por salto de línea):
ABCA1, ABCB4, ACADM, ACOX1, ACSL1, ACTA2, ACVR1, ADIPOQ, AGT, AIRE, ALX1, ANAPC2, ANAPC5, ANAPC7, ANGPTL4, ANKRD1, AP1B1, APOA1, APOA2, APOA5, AR, ATF1, ATF2, ATF3, ATF4, ATXN3, BCL3, BCL6, BRCA1, CALCOCO1, CAMK4, CARM1, CCNC, CD36, CDC16, CDC20, CDC25B, CDH1, CDH2, CDK19, CDK5RAP3, CDK8, CDKN1A, CDX2, CEBPA, CEBPB, CEBPD, CENPJ, CHUK, CITED1, CITED2, CNOT3, CPSF6, CPT1A, CPT2, CREB1, CREBBP, CREM, CRTC2, CRX, CSK, CSNK2A1, CSNK2A2, CTBP1, CTGF, CTNNB1, CUX1, CYP1A1, CYP4A11, CYP7A1, DACH1, DAXX, DDX17, DDX5, DEK, E2F1, E2F3, E2F5, EBF1, EGR1, EID3, EIF2B1, ELF3, ELK1, EP300, ESR1, ESR2, ETS1, ETS2, EWSR1, FABP4, FADS1, FASN, FGFR1, FHL2, FOS, FOSB, FOSL1, FOXM1, FOXO1, FOXO3, FOXO4, GABPA, GAK, GATA1, GATA2, GCM1, GLI3, GLIPR1, GMEB1, GMEB2, GRHL1, GRIP1, GTF2B, H3F3A, H3F3B, HDAC1, HDAC2, HDAC3, HES6, HIF1A, HIPK2, HIST1H3A, HIST1H4A, HIST3H3, HIST4H4, HLF, HMGA1, HMGB1, HMGB2, HNF1A, HNF1B, HNF4A, HOXA10, HOXA11, HOXA9, HOXB1, HOXB2, HOXB3, HOXB4, HOXB6, HOXB7, HOXB9, HOXD10, HOXD4, HSF1, HTT, IFNA10, IFNA13, IFNA14, IFNA16, IFNA17, IFNA2, IFNA21, IFNA4, IFNA5, IFNA6, IFNA7, IFNA8, IFNAR2, IFNB1, IKBKB, IKBKG, ING1, IRF1, IRF3, IRF5, IRF7, IRF9, JDP2, JUN, JUNB, KAT2A, KAT2B, KAT5, KHDRBS1, KLF1, KLF13, KLF4, KLF5, KPNA2, KPNA6, LDLR, LEP, LPL, MAF, MAFB, MAFG, MAFK, MAML1, MAML2, MAML3, MAP3K5, MAPK10, MBD2, MDC1, MDM2, ME1, MECOM, MED1, MED10, MED11, MED12, MED13, MED13L, MED14, MED15, MED16, MED17, MED18, MED19, MED20, MED21, MED22, MED23, MED24, MED25, MED26, MED27, MED29, MED30, MED31, MED4, MED6, MED7, MED8, MED9, MEIS1, MGMT, MIER1, MKNK1, MLL, MSH2, MSH6, MSX1, MTDH, MTF1, MYB, MYBL1, MYBL2, MYC, MYOD1, MYST3, N4BP2, NAP1L1, NCOA1, NCOA2, NCOA3, NCOA6, NCOR1, NCOR2, NEUROG1, NFATC2, NFATC4, NFE2, NFE2L2, NFIA, NFIC, NKX2-1, NLK, NMI, NOTCH1, NOTCH2, NOTCH3, NOTCH4, NPAS2, NPAT, NR3C1, NR5A1, NUP98, ONECUT1, PAPOLA, PARP1, PAX5, PCK1, PCMT1, PELP1, PEX11A, PHOX2A, PIAS1, PIAS3, PLAGL1, PLIN1, PLIN2, PLTP, PML, POLR2A, POU1F1, POU2F3, PPARA, PPARG, PPARGC1A, PRKCD, PROX1, PTMA, PTMS, RARA, RBBP4, RBBP7, RBCK1, RBPJ, REL, RELA, RGL1, RPA2, RPS6KA1, RPS6KA2, RPS6KA3, RPS6KA5, RUNX1, RUNX2, RUVBL1, RXRA, RXRG, SERTAD1, SERTAD2, SERTAD3, SETD1A, SH3GL1, SIN3A, SIN3B, SLC2A2, SMAD1, SMAD2, SMAD3, SMAD4, SMARCA4, SMARCB1, SMARCC1, SMARCC2, SND1, SNIP1, SNW1, SOX9, SP1, SP3, SPI1, SPIB, SRC, SRCAP, SREBF1, SREBF2, SRF, SS18L1, STAT1, STAT2, STAT3, STAT4, STAT5A, STAT5B, STAT6, SULT2A1, SUV39H1, TACC2, TAF6L, TBL1X, TBL1XR1, TBX21, TCF12, TCF3, TDG, TGS1, THRA, TIAM2, TNFRSF21, TP53, TP73, TRERF1, TRIB3, TRIM21, TRIP10, TRIP4, TXNRD1, UBC, UBTF, UCP1, UGT1A9, VDR, WT1, WWTR1, XRCC6, YWHAH, YY1, ZBTB17, ZBTB2 y ZNF639

El jurado designado por la Universidad Autónoma de la Ciudad de México aprobó esta tesis el día 27 de enero del 2012, en la Ciudad de México D.F. para optar al Grado de Maestro en Ciencias Genómicas, al Biol. Exp. Daniel Ortega Bernal

Dra. Claudia Selene Zárate Guerra _____

Dra. Eliza Irene Azuara Liceaga _____

Dra. Claudia Rangel Escareño _____

Dr. Oscar Alberto Pérez González _____