Bioinformática aplicada a estudios del transcriptoma humano: análisis de expresión de genes, isoformas génicas y ncRNAs en muestras sanas y en cáncer.

Tesis Doctoral

Alberto Risueño Pérez

Diciembre 2012

AUTORIZACIÓN DEL DIRECTOR DE TESIS

El Dr. D. **Javier DE LAS RIVAS SANZ**, con D.N.I. nº 15949000H, Investigador Científico del Consejo Superior de Investigaciones Científicas (CSIC) director del grupo de Bioinformática y Genómica Funcional y profesor del Programa de Doctorado y Master del Centro de Investigación del Cáncer (CiC-IBMCC) de la Universidad de Salamanca (USAL), certifica que ha dirigido la Tesis Doctoral titulada "*Bioinformática aplicada a estudios del transcriptoma humano: análisis de expresión de genes, isoformas génicas y ncRNAs en muestras sanas y en cáncer*" presentada por D. **Alberto Risueño Pérez** alumno del programa de Doctorado del CiC-IBMCC de la Universidad de Salamanca; y autoriza la presentación de la misma, considerando completado todo el trabajo e investigaciones realizadas en los últimos años por el doctorando.

En Salamanca, a 10 de diciembre de 2012

El Director de la Tesis Doctoral,

Firma: Dr. Javier De Las Rivas Sanz
Investigador Científico del CSIC
Centro de Investigación del Cáncer (CiC-IBMCC, USAL/CSIC)

Índice

Introducción general

Bioinformática y transcriptómica

La **bioinformática**, también llamada **biología computacional**, es la aplicación de las tecnologías informáticas y computacionales al estudio de información y datos biológicos y biomoleculares. Se trata de una nueva área de conocimiento amplia y multidisciplinar que abarca dos grandes campos: el campo del cálculo, análisis, algorítmica y manejo de datos (que incluye disciplinas como estadística, ciencias de la computación, inteligencia artificial, etc) y el campo de la biología molecular moderna (que incluye bioquímica, biología celular, genómica y proteómica entre otros). La bioinformática es, por tanto, herramienta clave para permitir mejorar nuestros conocimientos en estudios experimentales de biología molecular actual que, por su gran complejidad y el gran volumen de datos e información que produce, deben ser manejados y tratados con herramientas computacionales.

En los últimos años, la bioinformática está contribuyendo enormemente al desarrollo y avance en nuevos conocimientos obtenidos en el estudio de genomas completos (**genómica**) y de proteomas (**proteómica**), en estudios sobre la expresión génica y sobre la transcripción de DNA a RNAs a nivel de genomas (**transcriptómica**), en estudios de estructura e interacción entre proteínas (**interactómica**), etc.

Los estudios de **expresión génica** analizan cómo y bajo qué circunstancias se "activan" los genes en una muestra biológica concreta, que pasan a ser "transcritos" copiándose de sus regiones codificantes del genoma (DNA) para producir cadenas de RNA. En este ámbito de la **transcripción**, una de las aportaciones tecnológicas más importantes en los últimos años ha sido la tecnología global ("ómica") de los microarrays de expresión; en concreto, **microarrays** de oligos de alta densidad que incluyen genomas completos. Esta tecnología permite obtener en un solo experimento, información sobre el nivel de expresión de todos o gran parte de los genes, transcritos y exones del organismo estudiado.

DNA, genes, RNAs y transcripción alternativa

Las moléculas de **DNA** que constituyen el **genoma** contienen la información genética hereditaria de cada organismo viviente y proporcionan todas las instrucciones necesarias para la construcción de un nuevo individuo. En los organismos metazoos superiores cada genoma está constituido por un conjunto determinado de moléculas de DNA de longitud específica, que cuando se pliegan son los cromosomas. La información biológica se codifica en la **secuencia** de nucleótidos específica que constituye cada molécula de DNA, que es un polímero de nucleótidos de doble cadena (polinucleótido). Cada nucleótido es un eslabón de ese

polímero que se empareja con su complementario en la otra cadena definiendo una secuencia lineal de eslabones. Cada nucleótido contiene a su vez una base nitrogenada, un azúcar y un grupo fosfato. Existen cuatro bases nitrogenadas diferentes: adenina (**A**), guanina (**G**), timina (**T**) y citosina (**C**), y su orden lineal es el que determina la secuencia genómica y a él están asociadas las funcionalidades específicas de cada región codificante del DNA. La complementariedad de las dos cadenas del DNA viene dada por la complementariedad de los nucleótidos que se unen por puentes de hidrógenos a través de las bases nitrogenadas (pares de bases) siempre de modo **A-T** y **G-C** (es decir, A con T y viceversa; G con C y viceversa) formando así una estructura de doble hebra de forma helicoidal (ver **figura 1**). Como se ha indicado, cada macromolécula de DNA existente en la célula se denomina **cromosoma**, y en el caso de la célula eucariota los cromosomas son estructuras formadas por la unión del DNA con distintos tipos de proteínas que le ayudan a plegarse en distintos niveles de complejidad. En la especie humana existen 24 moléculas distintas de DNA: 22 moléculas que constituyen los cromosomas homólogos o autosomas, más el cromosoma X y el cromosoma Y que constituyen los cromosomas sexuales. Todos ellos suman un total de más de 3 mil millones de pares de bases y forman así el genoma humano completo.

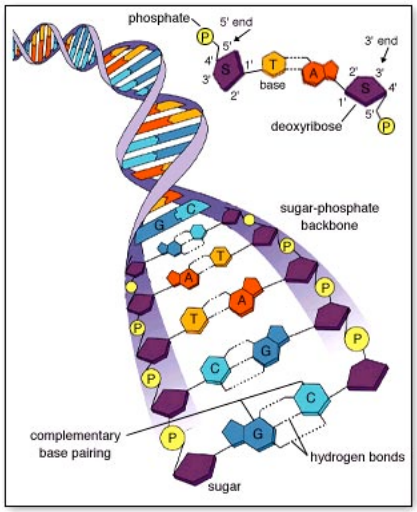


Figura 1. Esquema de doble hebra de DNA. (fuente: http://www.dna-sequencing-service.com)

A lo largo del **DNA** existen regiones codificantes llamadas **genes**. Estos genes contienen la información necesaria para generar moléculas funcionales de RNA. Estas moléculas de **RNA** se generan en un bioproceso bien regulado a partir de la copia específica del DNA, sustituyendo el nucleótido de base nitrogenada timina (**T**) por el nucleótido de base nitrogenada uracilo (**U**), de modo que las secuencias de RNA están codificadas por **A**, **U**, **G** y **C**. Además, habitualmente el RNA está conformado por una sola hebra, es decir es de cadena simple. El proceso de copia de DNA a RNA se llama **transcripción**. Existen numerosos tipos de RNA que tienen funciones específicas distintas: mRNA, tRNA, rRNA, miRNA, ncRNA, etc **(Lewin, 2004)**. Los **RNA mensajeros** (mRNAs) contienen la información que es utilizada en la síntesis específica de **proteínas** en el proceso denominado **traducción**, por eso se llaman a menudo *protein-coding RNAs* (pcRNAs). Los otros tipos de RNAs son muy variados y cada vez se van descubriendo más, tienen múltiples funciones celulares y suelen actuar de modo directo como macromoléculas en procesos catalíticos, reguladores, etc. Tras la transcripción de DNA en mRNA se producen los llamados mRNAs inmaduros ó **pre-mRNAs** que deben ser procesados en una serie de pasos post-transcripcionales para dar lugar a los mRNAs maduros. El principal de estos pasos post-transcripcionales es el llamado **corte y empalme alternativo** (*alternative splicing*), en el que se modifica el mRNA eliminando los fragmentos del transcrito inmaduro que no son codificantes para proteína. Los fragmentos eliminados se denominan intrones y los que permanecen hasta la traducción a proteína se denominan **exones** (ver **figura 2**). Los exones también pueden ser eliminados de forma selectiva, lo que significa que un mismo gen (o más propiamente un mismo *locus* génico) si se transcribe de distintos modos, es decir, si sufre varias lecturas alternativas de su secuencia de DNA para dar distintos mRNAs, puede codificar varias proteínas distintas que se denominan **isoformas**. Estos RNAs salidos de un mismo *locus* son **transcritos alternativos** y añaden una nueva capa de complejidad al proceso de expresión génica.

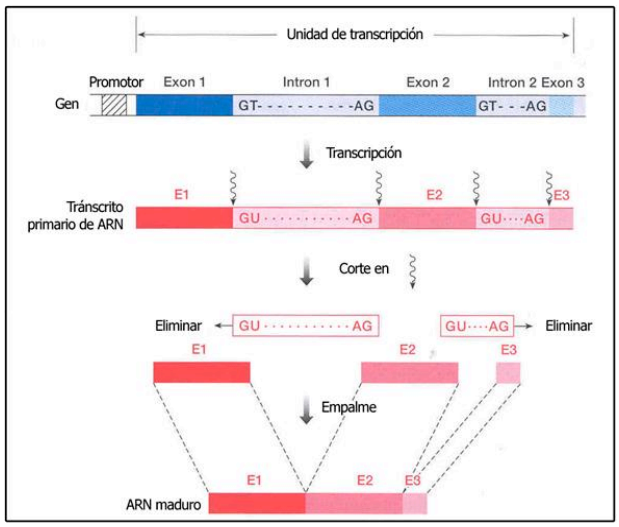


Figura 2. Proceso de transcripción a RNA y su posterior maduración a mRNA. (Fuente: http://www.down21.org)

Existen varios proyectos internacionales bioinformáticos y computacionales que han tratado de recoger y ordenar toda la información genética y biomolecular asociada a los cientos de **genomas** que se han secuenciado en los últimos años. El proyecto *Ensembl*, iniciado en 1999 **(Hubbard et al., 2009)** es uno de los más completos y ambiciosos ya que recoge datos de muchos genomas con mucho detalle a distintos niveles, incluyendo información sobre los genes, transcritos, proteínas, promotores, regiones reguladoras, etc, asociados a cada genoma, integrándolos con información procedente de otras bases de datos biológicas. Además, *Ensembl* dispone de uno de los navegadores genómicos (*genome browser*) de acceso *web* más avanzados y completos. Este proyecto es fruto de la colaboración entre el Instituto *European Bioinformatics Institute* (EBI), dependiente del Laboratorio Europeo de Biología Molecular (EMBL), y el *Wellcome Trust Sanger Institute* (WTSI), y se ha convertido en centro de referencia para investigadores de todo el mundo que trabajan con datos sobre genomas. En la presente Tesis Doctoral también servirá como referencia a la hora de anotar genes, transcritos y exones. Como dato meramente informativo se proporciona una tabla con la estadística del genoma humano (ver **tabla 1**) en la fecha de la presente escritura (noviembre de 2012).

Genes codificantes de proteína	20.476
Genes no codificantes de proteína	22.170
Pseudogenes	13.322
Transcritos	201.816
Exones	700.944

Tabla 1. Recuento de genes transcritos y exones identificados en el genoma humano (*Homo sapiens*) (datos de noviembre 2012).

Técnicas genómicas de alto rendimiento y microarrays de expresión

Las técnicas genómicas de alto rendimiento desarrolladas en los últimos años han automatizado y miniaturizado técnicas experimentales de biología molecular convencional para poder realizar análisis y obtener datos a gran escala, es decir, datos a nivel **ómico** o global que son derivados de análisis de genomas completos. Estas técnicas proporcionan a los investigadores un gran volumen de información normalmente en poco tiempo. Además, desde el origen de las técnicas genómicas hace casi dos décadas, que puede ubicarse en el primer secuenciador de DNA masivo creado a principio de los años noventa (*Massively Parallel Signature Sequencing* de *Lynx Therapeutic*), estas tecnologías han ido incrementando su potencial y reduciendo costes, llegando a convertirse en herramientas indispensables en el ámbito de la investigación biológica y biomédica actuales.

Las técnicas de alto rendimiento tratan de desarrollarse para abordar de modo global, completo, los distintos niveles biológicos y biomoleculares que se dan en un sistema biológico, no sólo el genoma, sino también el transcriptoma, el proteoma, etc. Es por esto por lo que también se las conoce como **técnicas ómicas**, que buscan obtener una visión global de los sistemas biológicos estudiados (organismos, células, etc).

En el área de la transcriptómica, los **microarrays** de oligos de alta densidad para medir la expresión han protagonizado una particular revolución tecnológica en el campo biomédico en los últimos años. Aún hoy en día, en el que la secuenciación de RNA (*RNA sequencing*, **RNA-seq**, utilizando técnicas de ultrasecuenciación *Next Generation Sequencing,* **NGS**) parece una realidad firme, los microarrays siguen siendo utilizados en multitud de estudios. Su coste asequible y su gran reproducibilidad han hecho de las plataformas de microarrays de oligos de alta densidad la manera más utilizada de hacer estudios de expresión génica. Existen multitud de publicaciones y multitud series de muestras accesibles en distintos repositorios públicos, por lo que el interés por estos dispositivos aún no ha cesado, aún cuando parece que la nueva revolución tecnológica en el campo de la secuenciación ha llegado definitivamente.

La práctica totalidad de los estudios presentados en esta Tesis Doctoral están basados en la utilización, mejora y análisis de datos de *GeneChips*, que son **microarrays de oligos de alta densidad** para RNA que miden expresión (manufacturados por la compañía norteamericana *Affymetrix*). Los microarrays de expresión de *Affymetrix* son dispositivos físicos que basan su funcionamiento en la hibridación del RNA procedente de la muestra estudiada con oligonucleótidos de DNA de longitud 25 llamados **sondas** (*probes*) montados en el array. Como se puede ver en la **figura 3**, el microarray es un pequeño cartucho con una ventana de aproximadamente 1 cm^2 donde se encuentra el array dividido en miles de micro celdas que contienen fijadas copias de oligonucleótidos (de secuencia específica) correspondientes a fragmentos de los genes humanos conocidos en el momento de su diseño. Cada **conjunto de sondas** oligos correspondientes a la secuencia representativa de un gen se denomina *probeset*, y están diseñados para ser capaces de detectar la expresión de dicho gen a partir de la hibridación con una muestra experimental de RNA que contenga dicho gen (ver **figura 4**).

Figura 3. Cartucho de microarray de oligos de alta densidad. (Fuente: www.affymetrix.com)

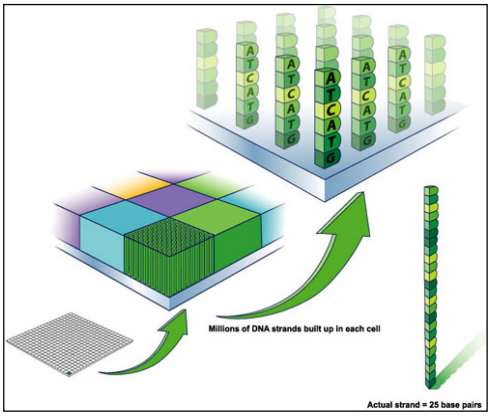


Figura 4. Conjunto de sondas presente en cada celda del array. (Fuente: www.affymetrix.com)

Según el modo de hibridación los microarrays de expresión de *Affymetrix* se dividen en dos tipos: los modelos 3' IVT y los modelos *Whole Transcript* (WT). En los **modelos 3' IVT**, para evitar el sesgo que se produce por los pasos de amplificación y PCR, las sondas se sitúan en las cercanías del extremo 3' del gen, muchas veces en las zonas no traducidas (*untranslated regions*, UTR) del gen. En los **modelos WT** más modernos, que siguen protocolos que evitan el sesgo de las amplifi-

caciones, las sondas están ubicadas a lo largo de toda la secuencia de cada *locus* génico, siendo la señal de expresión más fiable y siendo además posible la detección de los distintos exones del gen (ver **figura 5**).

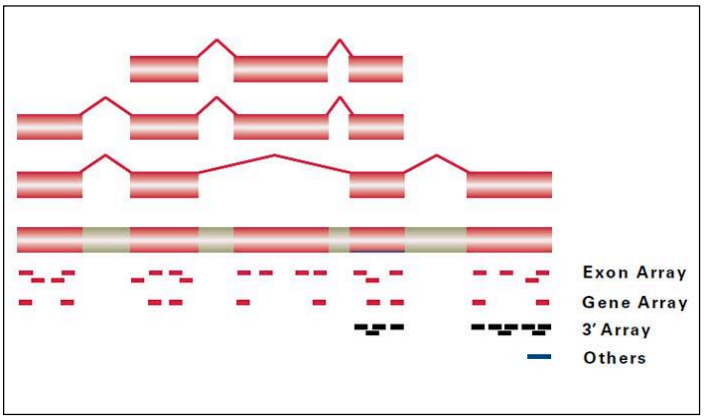


Figura 5. Diferencias en la localización de sondas en un locus génico para los distintos modelos de arrays. (Fuente: www.affymetrix.com)

El proceso de hibridación para los modelos 3' IVT comienza con el paso de transformación a cDNA de la muestra de RNA extraído del tejido o células sometidas a estudio. Tras este paso las muestras pueden ser almacenadas largo tiempo ya que la molécula de DNA es mucho más estable que la de RNA. Cuando llega la hora de hibridar el microarray el cDNA se convierte de nuevo a RNA mediante transcripción *in vitro* en el que el RNA es amplificado y etiquetado con biotina fluorescente. Este segundo RNA será la cadena complementaria del RNA original de la muestra (cRNA). Posteriormente este RNA es fragmentado en cadenas más cortas similares a los oligonucleótidos del array y es introducido en el dispositivo para lograr la hibridación. Tras un tiempo de incubación, se procede al lavado de la muestra permaneciendo exclusivamente el RNA de la muestra marcado con fluorescencia que ha hibridado, es decir, se ha unido con su sonda complementaria (ver **figura 6**). Tras un escáner, un láser detecta la cantidad de fluorescencia para cada celda generando una imagen. Un posterior análisis de esa imagen cuantifica el nivel de expresión de cada sonda identificándola por su posición concreta en los ejes X e Y de la matriz y volcando toda la información a un fichero. Estos ficheros son los denominados "datos crudos". En el caso de los modelos WT, existe un paso intermedio entre la transcripción *in vitro* del cRNA y su introducción en el microarray. Este paso consiste en volver a obtener la hebra del RNA inicial en pequeños fragmentos al azar siguiendo la técnica *random priming.* Al existir un paso más, en este caso los oligos no tendrán la secuencia del gen original sino que deben tener la secuencia complementaria para que se produzca la hibridación (ver **figura 7**).

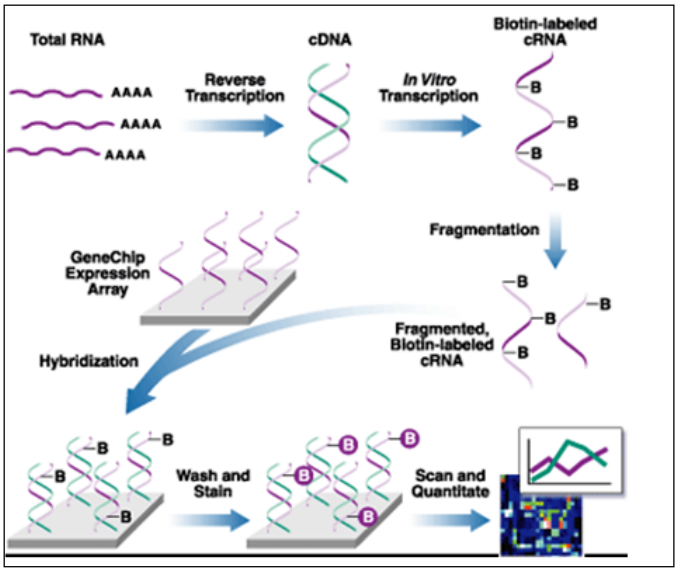


Figura 6. Protocolo de preparación mRNA extraído (cadenas poliA) para medir la expresión génica global usando microarrays: **(i)** copia por transcripción reversa a cDNA, **(ii)** transcripción a cRNA y etiquetado con sistema biotina, **(iii)** fragmentación, **(iv)** hibridación en dispositivos microarrays de alta densidad de oligonucleótidos que miden la expresión global (*genome-wide*), manufacturados por *Affymetrix*. (Fuente: www.affymetrix.com)

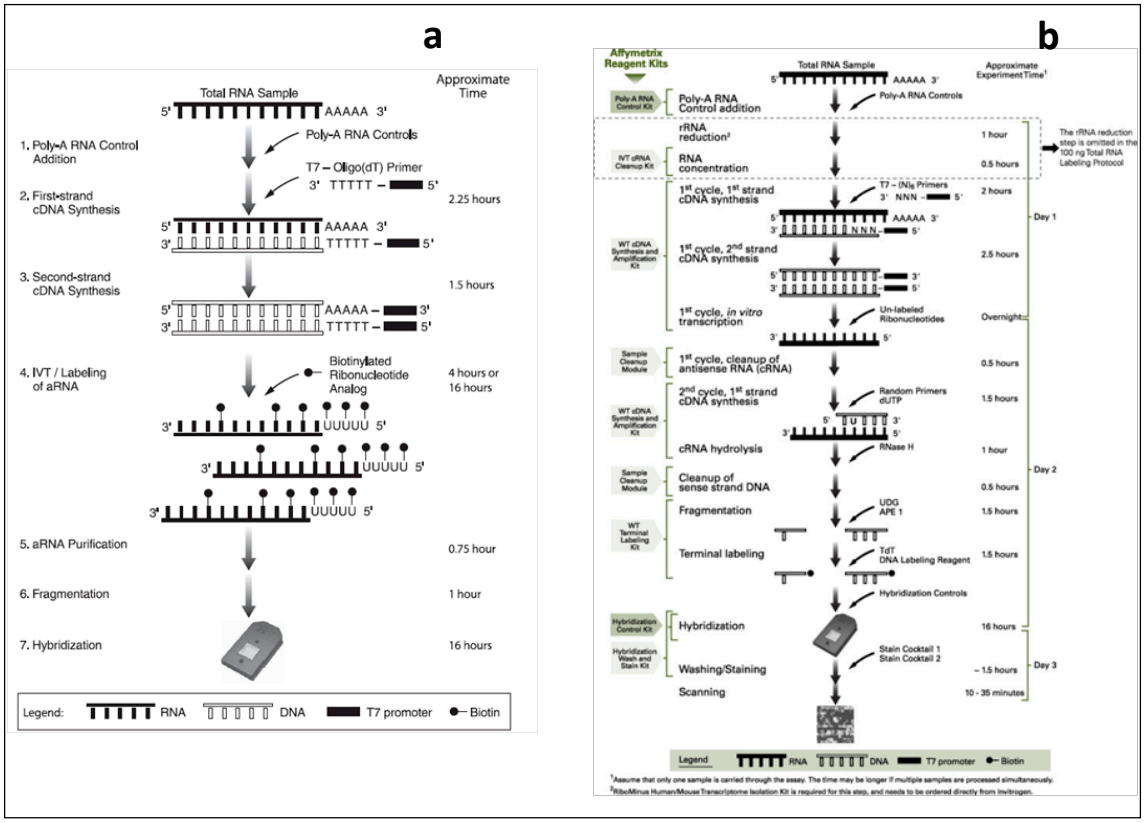


Figura 7. Diferencias en el protocolo de de etiquetado e hibridación entre los modelos 3' IVT (a) y los más modernos *Whole Transcripts* (b) basados en la técnica random priming.

La importancia de la tecnología de microarrays en los últimos años y su aportación a investigaciones de todo el mundo se puede ilustrar con el siguiente ejemplo: en el repositorio público *Gene Expression Omnibus* (**http://www.ncbi.nlm.nih.gov/geo/**) **(Barrett et al., 2005)** a fecha de Septiembre de 2012 el modelo de array "Human Genome 133 Plus 2.0" de *Affymetrix* (correspondiente a la plataforma identificada por el código GPL570 en GEO) cuenta con 70.831 muestras hibridadas, que proceden de 2.616 series de datos. También como ilustración del crecimiento indicar que en 8 meses de 2012 (de enero a agosto) el incremento de datos depositados en GEO correspondientes a este mismo microarray ha sido de un 12.8 %, es decir más de 8.000 nuevas muestras. La riqueza de datos disponible hace que las plataformas de *Affymetrix* sean merecedoras de estudios exhaustivos y para ello el estudio y análisis bioinformático computacional es esencial. Además, la profundización en el diseño y características biomoleculares de estas plataformas permite plantearse mejoras en la forma de analizar este tipo de datos.

Objetivos

La presente **Tesis Doctoral** tiene como objetivo fundamental la aplicación de técnicas y metodologías de **bioinformática y biología computacional** al estudio global de datos del **transcriptoma humano** obtenidos por plataformas genómicas. De modo concreto se centra en el análisis y cuantificación de la **expresión de genes** (mRNAs codificantes), de las **isoformas génicas** derivadas de procesos de *splicing* alternativo y de genes no codificantes para proteína llamados en general **ncRNAs**, (entre los que se incluyen los **microRNAs**). Todo ello se ha realizado utilizando datos derivados tanto de **muestras humanas** de individuos **sanos** como de distintas series de muestras de **pacientes** con **cáncer**.

El trabajo consta de cuatro partes, cada una de las cuales tiene unos objetivos concretos que a continuación se describen brevemente:

Objetivo 1.- Mejora del método análisis de datos de plataformas experimentales de **expresión génica global** –particularmente de datos de expresión producidos por **microarrays** de *Affymetrix* ampliamente usados en investigación biomédica– sustituyendo la anotación original proporcionada por el fabricante basada en sus sondas (*probe-oriented annotation*) por un **remapeo** y **anotación** alternativo, actualizado y centrado en las entidades biológicas (*gene-oriented annotation*) que toma como referencia los **genes**, **transcritos** y **exones** definidos en bases de datos biomoleculares actuales. Integración de los datos generados en una **plataforma *web* interactiva** con un **navegador genómico-transcriptómico** que permita explorar y visualizar de modo simple tanto la estructura de los *loci* génicos, como el mapeo de sondas de todos los microarrays de *Affymetrix* y ciertas series de datos experimentales de expresión. Todo ello implementado para el genoma de humano (*Homo sapiens*), ratón (*Mus musculus*) y rata (*Rattus norvegicus*).

Objetivo 2.- Desarrollo y aplicación de un análisis de **expresión diferencial** para identificar **genes** marcadores en varios conjuntos de datos de muestras de cáncer (i.e. distintos subtipos de leucemias y de mieloma múltiple) y para reconocimiento y asignación de microRNAs (**miRNAs**) que marquen las categorías o clases en los datos de mieloma múltiple.

Objetivo 3.- Diseño y desarrollo de nuevo **algoritmo** que permita la identificación robusta de eventos de ***splicing* alternativo** en genes a partir de datos de expresión obtenidos con microarrays de exones (Exon *1.0 Affymetrix*). Validación del algoritmo sobre un conjunto de datos conocidos sobre genes humanos que sufren *splicing* y aplicación de dicho algoritmo a un conjunto de muestras de cáncer.

Objetivo 4.- Desarrollo de un estudio transcriptómico global de **coexpresión de genes** humanos basado en datos de microarrays obtenidos para varias series de muestras de tejidos sanos. Identificación de conjuntos de genes que coexpresan, así como reconocimiento de **genes específicos de tejido** (*tissue-specific genes*) y genes generales de mantenimiento (*house-keeping genes*). Estudio evolutivo de ambos tipos de genes analizando su conservación en distintas especies. Estudio evolutivo aplicado a genes desregulados en cáncer.

Capítulo 1
Diseño y construcción de un explorador genómico y transcriptómico con mapeo de sondas de expresión a genes, transcritos, exones y ncRNAs: *GATExplorer*

1.1. Introducción

Desde que se presentó el primer borrador del genoma humano en el año 2001 **(Lander et al., 2001; Venter et al., 2001)**, las revisiones y actualizaciones de su secuencia consenso han sido continuas hasta hoy día. La versión GRCh37 de septiembre de 2009 incluye 33.868.498 pares de bases más que su predecesora NCBI36 lanzada en octubre de 2005 (3.286.906.305 y 3.253.037.807 pares de bases respectivamente). Si en lugar de observar cambios en la secuencia genómica observamos la evolución en el conocimiento de los genes, es decir, la parte funcional más conocida del genoma, vemos que los cambios son notables. El número de secuencias codificantes de RNA expresadas en las distintas células del organismo humano se ha incrementado enormemente en los últimos años **(Carninci et al., 2005; Kapranov et al., 2007)**. Este descubrimiento de nuevas secuencias de RNA y su posterior alineamiento sobre el genoma no sólo identifica nuevos genes, sino que puede fusionar varios genes considerados anteriormente como distintos. Esto supone que el número de genes conocidos no aumenta necesariamente con el tiempo pudiendo incluso mostrar, de forma paradójica, una reducción de su número. De esta manera, el número de genes codificantes de proteína catalogados en 2005 era aproximadamente 26.000 mientras que a inicios de 2012 el número es de poco más de 20.000 (**www.ensembl.org**). Sin embargo, el número de transcritos distintos para los genes humanos genes ha aumentado considerablemente lo que indica que los genes son considerablemente más complejos de lo estimado inicialmente. Todo esto significa que el conocimiento del transcriptoma humano, y del transcriptoma de metazoos en general, está aumentando dramáticamente en la última década. Las bases de datos como la del proyecto *Ensembl* **(Hubbard et al., 2009)** recogen esta información actualizada y puede ser utilizada para mejorar la precisión de los numerosos estudios transcriptómicos realizados con plataformas genómicas (*genome-wide platforms*), como los microarrays de expresión. La interpretación de los datos de estas plataformas en base a versiones más actuales y completas del genoma humano permite análisis más cercanos a la realidad biológica y mejor uso de los datos derivados de estos estudios.

1.1.1. Bases de datos de ncRNAs como complemento de la información de *Ensembl*

En estos últimos años también se han identificado multitud de secuencias nuevas de RNA no codificante de proteína (ncRNA) suscitando un creciente interés en este tipo de transcritos. Numerosas investigaciones se han llevado a cabo tratando de catalogar estas secuencias y de descubrir qué función cumplen. Estos estudios aún están comenzando, pero ya presentan a la célula como una máquina transcripcional de increíble complejidad **(Amaral et al., 2008)**. Varios trabajos han recopilado la información sobre ncRNA en bases de datos especializadas como ***RNAdb*** **(Pang et al., 2007)**. Este tipo de bases de datos pueden complementar a *Ensembl* en los estudios transcriptómicos para proporcionar una visión más global sobre los diferentes mecanismos moleculares de las distintas células que conforman un organismo.

Es evidente que este aumento de conocimiento a nivel transcripcional ha terminado por cambiar el concepto clásico de "gen" originalmente asociado a un RNA mensajero (mRNA) y a proteína concreta, ya que cada "locus génico" de un genoma como el humano puede dar lugar a muchos mensajeros diferentes y a su vez estos originar proteínas con pequeñas o grandes variaciones (llamadas isoformas) derivadas de procesos de transcripción y maduración diferencial que suelen suceder en distintos tipos celulares. De este modo, cada "locus génico" del genoma puede incluir una gran complejidad y sufrir regulación a distintos niveles, de modo que la definición de los genes humanos no es tan clara y debe ser estudiada y revisada constantemente.

1.1.2. Microarrays de oligos de alta densidad para medir la expresión génica a escala genómica global

A medida que el conocimiento del transcriptoma avanza va ampliando el catálogo de secuencias conocidas de RNA. Estudios posteriores se centran en comprender la regulación y función de las distintas secuencias identificadas. Una de las tecnologías más populares para hacer esto es la de microarrays de oligonucleótidos de alta densidad diseñados para medir la expresión de todos los genes de un genoma. Estos nano dispositivos de oligos –es decir, de secuencias cortas de DNA de cadena simple– son plataformas de escala genómica (*genome-wide scale*) que permiten medir la cantidad de miles de fragmentos de RNA a la vez. De este modo, en el microarray se incluyen cientos de miles de secuencias cortas de DNA sobre las que se pueden testar miles de genes a la vez, es decir, todos los transcritos presentes en un extracto celular concreto de estudio. La compañía americana *Affymetrix*, fundada en 1992, fue pionera en el diseño de estos dispositivos **(Lipshutz et al., 1999; Lockhart et al., 1996; Wodicka et al., 1997)** y es probablemente la empresa que con más éxito ha comercializado distintos modelos de microarrays de expresión de escala genómica. El modelo de microarrays humanos que más éxito ha tenido es el llamado *GeneChip Human Genome U133*, que fue lanzado en 2001 y su diseño estaba basado en la información presente en librerías de cDNA humano de aquel momento (en concreto, librerías de la base de datos *UniGene –build 133–* de abril de 2001). Estas librerías contienen colecciones de secuencias de RNA expresadas, denominadas *Expressed Sequence Tags* (ESTs), que han sido identificadas en humano a través de numerosos estudios experimentales sobre distintos tipos celulares y tejidos. Además, estos primeros modelos de microarrays de expresión se basaban en tecnología de secuenciación y copia por el extremo (modelos de tipo IVT 3'), y los modelos actualmente comercializados por *Affymetrix* siguen la tecnología *Whole Transcript* (WT) en la que el diseño está hecho en base a

la secuencia genómica del organismo estudiado, en lugar de su transcriptoma, utilizando para ello estudios combinados sobre varias bases de datos. Estos modelos WT fueron diseñados en 2006.

Todo esto hace ver que la evidencia biológica en la que ha sido basado el diseño de las sondas de estos microarrays genómicos no refleja el conocimiento actual, dada la celeridad de los avances en biología molecular, y por lo tanto existe una necesidad de reconstruir y reinterpretar los análisis llevados a cabo con esta tecnología en base a una información biológica más actualizada.

1.1.3. Mapeo global de las sondas (*probes*) presentes en microarrays de alta densidad

Como se ha indicado los microarrays genómicos de alta densidad están construidos incluyendo cientos de miles de oligonucleótidos de secuencias específicas (i.e. oligos de DNA de cadena simple). Los microarrays de *Affymetrix* agrupan varias sondas de oligos (*probes*) en conjuntos, denominados *probesets*, etiquetados con un identificador específico de tipo numérico (XXXXX_at). Estos conjuntos de sondas son asignados a un gen o entidad génica concreta del genoma correspondiente, siguiendo los mapeos realizados por *Affymetrix* durante el diseño y construcción del modelo concreto de microarray. En los modelos IVT 3' la mayoría de los *probesets* están formados por conjuntos de 11 sondas. En el caso de los modelos WT *Gene Array* la agrupación de sondas se realiza a dos niveles: *exon probeset* y *transcript cluster*. Para los modelos WT *Exon Array* existen los niveles *probeset*, *exon cluster* y *transcript cluster* **(Affymetrix, 2005c)**. Estos *probesets* de los modelos WT agrupan sondas próximas entre sí (*probesets*), que están ubicadas en el mismo exón (*exon probeset* y *exon cluster*) y sondas ubicadas en el mismo locus (*transcript cluster*) (ver **tabla 1.1**).

Nivel de agrupación	Modelo de microarray	Número de sondas	Descripción
Probeset	*Exon Array*	4	Grupo de 4 sondas ubicadas en regiones próximas
Exon Probeset	*Gene Array*	Variable	Sondas pertenecientes al mismo exón
Exon Cluster	*Exon Array*	Variable	Sondas pertenecientes al mismo exón
Transcrip Cluster	*Gene Array y Exon Array*	Variable	Sondas pertenecientes al mismo locus

Tabla 1.1. Niveles de agrupación de las sondas de los modelos de microarray *Whole Transcript* de *Affymetrix*.

De este modo, el diseño de la sondas de *Affymetrix* pretende asociar un *probeset* de los modelos IVT 3' con un único gen, de manera que existan varias sondas para detectar cada gen. Sin embargo, la asociación *probeset-to-gene* no es univoca ya que un mismo gen puede tener asociados varios *probesets*. En los modelos WT las sondas agrupadas bajo un mismo *transcript cluster* deben ir ubicadas en un solo gen, y un gen debería tener un único *transcript cluster* asociado.

Los p*robesets* del modelo WT *Exon Array* también se categorizan a su vez en tres conjuntos llamados *core*, *extended* y *full* que representan distintos niveles de evidencia biológica **(Affymetrix, 2005b)**: **(i)** los probesets *core* son los que está basados en la evidencia más fiable cuya información procede de las bases de datos *RefSeq* y las secuencias completas de *GenBank*. **(ii)** Los *probesets extended* suponen un nivel inferior de evidencia respecto a las

sondas *core* y su información procede de secuencias de *GenBank* no anotadas como completas, secuencias ESTs, genes de *Ensembl*, micro RNAs, etc. Finalmente, **(iii)** las sondas *full* sólo están basadas en información de genes predichos computacionalmente sin evidencia biológica real y cuya información procede de algoritmos predictivos como *GeneID*, *GenScan*, *RNAGene*, etc (ver **figura 1.1** y **tabla 1.2**).

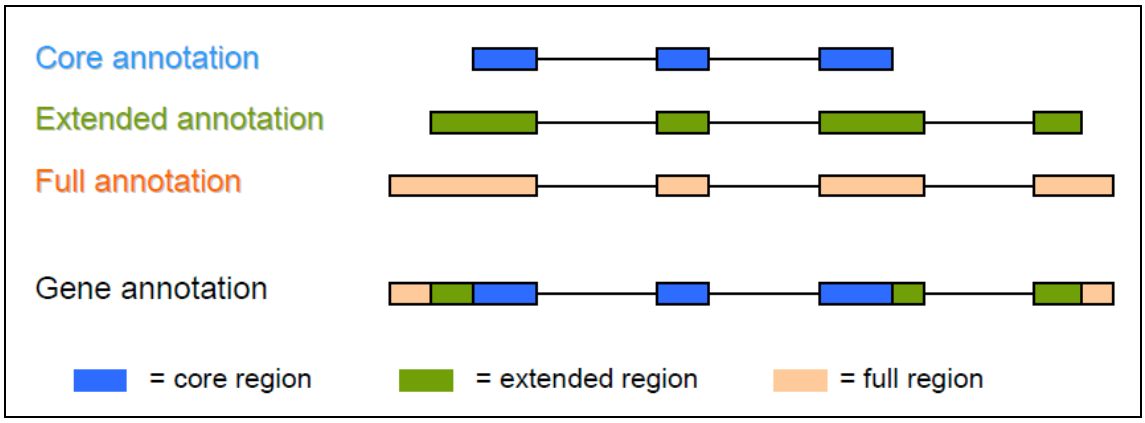


Figura 1.1. Sondas *core*, *extended* y *full* de los microarrays de exones de *Affymetrix*.

Nivel de evidencia biológica	Descripción	Fuentes de datos
Core	Evidencia biológica más fiable.	RefSeq y secuencias completas de mRNA de GenBank.
Extended	Evidencia biológica de cDNA que extiende el nivel Core.	Secuencias de mRNA de GenBank no anotadas como completas, secuencias ESTs, genes de Ensembl, genes mitocondriales de MitoMap, microRNAs y genes y pseudogenes de Vega.
Full	Basados en predicciones computacionales únicamente.	GeneID, GenScan, GenScanSubOptimal, exoniphy, RNAGene, sgpGene y Twinscan.

Tabla 1.2. Tipos de *probeset* en función de su evidencia biológica, pertenecientes al modelo de microarray *Exon Array* de *Affymetrix*.

La asociación entre *probeset* y gen es proporcionada por *Affymetrix* a través de unos ficheros de anotación. *Affymetrix* trata de mantener esta anotación actualizada en base al conocimiento biológico disponible en cada momento actualizando sus ficheros de anotación entre 1 y 3 veces por año. Con estas actualizaciones se tratan de resolver las inconsistencias que producen las actualizaciones de las bases de datos biológicas sobre el diseño original de los microarrays. Sin embargo, no cambian con el tiempo las sondas de secuencia concreta que se han incluido en el microarray. También cabe destacar que, en el caso de los modelos WT, incluso la primera versión de la anotación fue posterior al diseño de las sondas, y por lo tanto este tipo de inconsistencias puede existir en ciertos *probesets* desde el momento inicial del microarray. Estas inconsistencias se pueden resumir en tres tipos:

1. La **identificación de un gen nuevo** puede convertir a un *probeset* en ambiguo, es decir que detecte la señal procedente de dos genes distintos, lo que se conoce como hibridación múltiple o cruzada.
2. La **fusión de dos genes** contiguos puede hacer concurrir dos *probesets* hacia el mismo gen cuando anteriormente estaban anotados a genes distintos.
3. La **desaparición de un gen o de un exón** codificante que se creía existente en cierto

locus del genoma, o la modificación de su posición o tamaño (por ejemplo por modificación de las regiones UTR génicas). Esto pueda dejar fuera algún *probeset* descubriendo que realmente no está diseñado para detectar ninguna región codificante y por ello sería incapaz de detectar señal transcriptómica alguna.

Un ejemplo de ambigüedad está representado en el *transcript cluster* 7918813 del microarray WT *Human Gene*, el cual tiene como genes asociados NRAS y CSDE1 indicando que detecta la señal de esos dos genes siendo imposible distinguir la aportación de cada uno de ellos. Este *transcript cluster* tiene 33 sondas (ver **tabla 3**) de las cuales algunas pueden ser asignadas a NRAS y otras a CSD1 (ver **figura 1.2**).

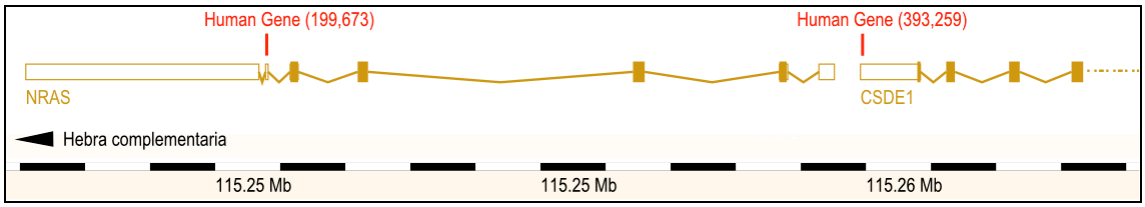


Figura 1.2. El *transcript cluster* 7918813 del microarray WT *Human Gene* de *Affymetrix* tiene sondas ubicadas en dos genes diferentes: NRAS y CSD1.

Esto muestra que la estructura rígida de agrupación de sondas de *Affymetrix*, implica utilizar en todos los casos el conjunto completo de sondas que constituye un *probeset*, anotando todas ellas de la misma manera. Como solución, en este trabajo, se opta por romper las entidades originales de *Affymetrix* tratando a cada sonda de forma independiente en un primer momento, realizando un remapeo completo a nivel de sonda (*probes level remapping*) agrupándolas después en base a entidades biológicas conocidas en el momento, en lugar de en base a los *probesets* originales.

Sonda	Sonda X	Sonda Y	Ubicación genómica
tcacgtttgcggtttggttctctgt	524	885	chr1:115250366-115250390 (-)
ctggggtggcagaggtgtgtttgtg	472	348	chr1:115250423-115250447 (-)
taacagggagtaacaagaggtgcat	495	510	chr1:115250474-115250498 (-)
tctggtcagacagccaagtgaggag	999	406	chr1:115250501-115250525 (-)
ttgtactaaactactgagagctggg	59	1044	chr1:115250560-115250584 (-)
cttgaaagtggctcttttctgacaa	912	613	chr1:115250775-115250799 (-)
ttgaaagtggctcttttctgacaaa	1037	747	chr1:115250776-115250800 (-)
aagtggctcttttctgacaaaactt	199	673	chr1:115250780-115250804 (-)
tgagtttttcattcggtactggcg	205	1030	chr1:115251212-115251236 (-)
ggtactggcgtatttctcttaccag	479	864	chr1:115251227-115251251 (-)
ggcgtatttctcttaccagtgtgta	824	96	chr1:115251233-115251257 (-)
gaatggaatcccgtaactcttggcc	729	291	chr1:115252217-115252241 (-)
tgtccttgttggcaaatcacacttg	570	938	chr1:115252268-115252292 (-)
tagcaccataggtacatcatccgag	336	779	chr1:115252301-115252325 (-)
tgtagaggttaatatccgcaaatga	150	869	chr1:115256422-115256446 (-)

caaatgacttgctattattgatggc	225	694	chr1:115256440-115256464 (-)
cttcgcctgtcctcatgtattggtc	663	324	chr1:115256482-115256506 (-)
gggatcatattcatctacaaagtgg	905	931	chr1:115258680-115258704 (-)
gctggattgtcagtgcgcttttccc	495	16	chr1:115258715-115258739 (-)
tgtcagtgcgcttttcccaacacca	162	668	chr1:115258722-115258746 (-)
tgcctctgctttggacagatttagg	576	902	chr1:115259291-115259315 (-)
cagatttaggaccacagccgggaaa	810	583	chr1:115259306-115259330 (-)
gaccacagccgggaaaaatgttgga	410	121	chr1:115259315-115259339 (-)
ggccccgcccgctacgtaatcagtc	700	152	chr1:115259387-115259411 (-)
gctacgtaatcagtcggcgccccag	935	274	chr1:115259397-115259421 (-)
ggcctccgaaccacgagtcatgcgg	79	480	chr1:115259450-115259474 (-)
ctcccacacgggacgtttcaataat	399	941	chr1:115259497-115259521 (-)
cacgggacgtttcaataatgaaagc	177	413	chr1:115259503-115259527 (-)
tttcaataatgaaagcgctaggtgc	599	150	chr1:115259512-115259536 (-)
gcttccattctttcgccattaacag	298	839	chr1:115259568-115259592 (-)
gagatcaaaacctcaaacgacaagg	838	946	chr1:115259651-115259675 (-)
tttacaggacacagtaaccaggcgg	833	583	chr1:115259862-115259886 (-)
aagaaaccgggtcctagaagctgca	393	259	chr1:115259963-115259987 (-)

Tabla 1.3. *Transcript cluster* 7918813 del microarray WT *Human Gene*.

1.1.4. Partes del trabajo desarrollado en este capítulo

Según lo descrito en la introducción presentamos a continuación las partes del trabajo que se ha realizado en este capítulo de la Tesis Doctoral:

1. **Re-mapeo** de todas las sondas de los distintos modelos de microarrays de expresión de *Affymetrix* en base al último conocimiento biológico existente en la base de datos de *Ensembl*. Los organismos elegidos son: **humano** (*Homo sapiens*), **ratón** (*Mus musculus*) y **rata** (*Rattus norvegicus*). El resultado es almacenado en base de datos y exportado posteriormente en forma de paquetes del programa estadístico R para su utilización.
2. Implementación de un **portal *web*** cuyo propósito es la interacción con el usuario de forma visual e interactiva en el que se integrará un **navegador genómico** (con datos de genomas, cromosomas, genes y demás entidades génicas) **y transcriptómico** (con datos de expresión y de mapeos) y en el que se incluirá un **repositorio** con distintos ficheros de anotación y con mapeos completos listos para ser descargados.

El resultado final es una aplicación bioinformática llamada ***GATExplorer*** (*Genomic and Transcriptomic Explorer*) accesible desde la dirección: **http://bioinfow.dep.usal.es/xgate**.

1.2. Materiales y métodos

1.2.1. Bases de datos utilizadas como fuente para los remapeos de sondas de microarrays

Para realizar el re-mapeo de las sondas de los microarrays de expresión de *Affymetrix* se utiliza como fuente de datos principal la información disponible en la base de datos de *Ensembl*. Esta base de datos biológica pública ofrece mediante su servidor ftp **(ftp://ftp.ensembl.org)**, ficheros que contienen todas las secuencias de cDNA conocidas para el transcriptoma de varias especies. Estos ficheros representan la información en formato FASTA, que asocia cada secuencia con su identificador en texto plano (ver **figura 1.3**) y cuyo fácil manejo lo ha convertido en un estándar a la hora de analizar y comparar secuencias de DNA. Los ficheros descargados proceden del directorio "cdna" correspondiente con la versión y el organismo seleccionados siguiendo la nomenclatura:

ftp://ftp.ensembl.org/pub/release-<nº_de_versión>/fasta/<organismo>/cdna.

```
>ENST00000397806 cdna:known chromosome:GRCh37:16:222889:223709:1 gene:ENSG00000188536
CACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGG
CCGCCTGGGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCT
GAGCCACGGCTCTGCCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAA
CGCCGTGGCGCACGTGGACGACATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGC
GCACAAGCTTCGGGTGGACCCGGTCAACTTCAAGCTCCTAAGCCACTGCCTGCTGGTGAC
CCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTGCGGTGCACGCCTCCCTGGACAAGTT
CCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCGTTAAGCTGGAGCCTCGGTAGC
CGTTCCTCCTGCCCGCTGGGCCTCCCAACGGGCCCTCCTCCCCTCCTTGCACCGGCCCTT
CCTGGTCTTTGAATAAAGTCTGAGTGGGCAGCA
>ENST00000251595 cdna:known chromosome:GRCh37:16:222846:223709:1 gene:ENSG00000188536
CATAAACCCTGGCGCGCTCGCGGGCCGGCACTCTTCTGGTCCCCACAGACTCAGAGAGAA
CCCACCATGGTGCTGTCTCCTGCCGACAAGACCAACGTCAAGGCCGCCTGGGGTAAGGTC
GGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAGAGGATGTTCCTGTCCTTCCCC
ACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTGCCCAGGTTAAGGGC
CACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGACATGCCC
AACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAAC
TTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTC
ACCCCTGCGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACC
TCCAAATACCGTTAAGCTGGAGCCTCGGTAGCCGTTCCTCCTGCCCGCTGGGCCTCCCAA
CGGGCCCTCCTCCCCTCCTTGCACCGGCCCTTCCTGGTCTTTGAATAAAGTCTGAGTGGG
CAGCA
```

Figura 1.3. Formato FASTA de secuencias cDNA procedentes de la base de datos *Ensembl*. Se muestra una pequeña parte del fichero en donde figura la secuencia de dos transcritos: ENST00000397806 y ENST00000251595.

Durante el desarrollo de la presente Tesis Doctoral se han utilizado las versiones **50** (julio de 2008), **53** (Marzo de 2009) y **57** (Marzo de 2010) de los organismos *Homo sapiens*, *Mus musculus* y *Rattus norvegicus* (ver **tabla 1.4**). Cada entrada de estos ficheros contiene un identificador de transcrito como el de la **figura 1.3** (ENST0000…) . La letra "T" indica que se trata de un transcrito. En *Ensembl* cada tipo de entidad concreta biomolecular concreta tiene su propia letra, de esta manera "G" indica gen, "T" transcrito, "E" exón y "P" proteína. Esta forma de codificar los distintos identificadores que tiene *Ensembl* es informativa para el investigador.

Junto a la base de datos de *Ensembl* para ubicar sobre los transcritos de las sondas de los microarrays de *Affymetrix* también es necesario contar con la secuencia de cada uno de los oligos de 25 nucleótidos que conforman dichos arrays. Esta información fue descargada de la página *web* corporativa de la empresa (**www.affymetrix.com**). Existe un fichero por cada modelo de microarray detallando su constitución completa y asociando una secuencia genómica de 25 nucleótidos a cada posición X e Y, que sirve como coordenada para ubicar cada sonda dentro de la topología del microarray (ver **tabla 1.4**). En el caso de los microarrays del modelo IVT 3', sólo las secuencias etiquetadas catalogadas como "*perfect match*" están

presentes en el fichero.

Organismo	Ensamblaje del genoma	Versión de Ensembl	Versión de RNAdb	Fecha de GATExplorer	Versión de GATExplorer
Human	GRCh37 (Sep.2009)	v 57 (Mar.2010)	2009	1.Sep.2010	v 3.0
Mouse	NCBIM37 (Apr.2007)	v 57 (Mar.2010)	2009	1.Sep.2010	v 3.0
Rat	RGSC3.4 (Dec.2006)	v 57 (Mar.2010)	-	1.Sep.2010	v 3.0
Human	NCBI36 (Oct.2005)	v 53 (Mar.2009)	2009	25.Jul.2009	v 2.0
Mouse	NCBIM37 (Apr.2007)	v 53 (Mar.2009)	2009	25.Jul.2009	v 2.0
Rat	RGSC3.4 (Dec.2006)	v 53 (Mar.2009)	-	25.Jul.2009	v 2.0
Human	NCBI36 (Oct.2005)	v 50 (Jul.2008)	-	Oct.2008	v 1.0
Mouse	NCBIM37 (Apr.2007)	v 50 (Jul.2008)	-	Oct.2008	v 1.0
Rat	RGSC3.4 (Dec.2006)	v 50 (Jul.2008)	-	Oct.2008	v 1.0

Tabla 1.4. Histórico de versiones de ensamblaje, *Ensembl* y *RNAdb* de ***GATExplorer***.

Para el mapeo sobre ncRNA son necesarios los archivos con las secuencias que se obtuvieron de otra base de datos específica para RNAs no codificantes: *RNAdb* (*A database of mammalian noncoding RNAs*); construida por el grupo australiano dirigido por el Profesor John Mattick **(Pang et al., 2007)**. Toda la información detallada y secuencias de ncRNAs fueron descargada de la *web* del grupo indicado (**http://research.imb.uq.edu.au**) en formato FASTA. Cada entrada de estos ficheros se corresponde con un identificador del transcrito. Al no disponer de datos para la especie *Rattus Norvegicus* solamente se descargaron los ficheros para humano y ratón.

Por último, ***GATExplorer*** también incorpora e integra datos de expresión obtenidos a partir de ciertos conjuntos de datos de microarrays de las tres especies que se incluyen en la *web*. Para el caso de humano que se utilizó el *set* de microarrays *GeneAtlas* (*G*EO ID GSE1133) **(Su et al., 2004)**. Estos microarrays fueron normalizados a nivel de sonda y almacenados en base de datos. La aplicación *web* recupera las sondas ubicadas en cada gen buscado por el usuario en tiempo real y presenta un perfil de expresión a lo largo de varios tejidos.

1.2.2. Utilización de la arquitectura LAMP (*Linux-Apache-MySQL-PHP*) para la construcción de una plataforma bioinformática

Durante todo el desarrollo de ***GATExplorer***, tanto en el entorno de desarrollo de los diferentes programas creados y utilizados, como en el entorno de pruebas y en el servidor de producción que aloja la versión final de la aplicación *web*, se ha utilizado la arquitectura LAMP. LAMP corresponde a la siglas de *Linux*, *Apache*, *MySQL* y PHP: herramientas informáticas de código abierto que por ser gratuitas, por su extendido uso y por permitir un alto nivel de programación son muy adecuadas para un proyecto de estas características.

Linux es el sistema operativo que controla las máquinas, *Apache* es el servidor de aplicaciones que proporcionará acceso remoto a la aplicación *web* (**http://httpd.apache.org/**), *MySQL* es el sistema gestor de base de datos (**http://www.mysql.com/**) y *PHP* es el lenguaje de programación en el que está implementada la aplicación *web* y que ejecuta las distintas instrucciones que la componen (**http://www.php.net/**).

Además de las herramientas indicadas, ***GATExplorer*** incluye un módulo llamado *Ming* que permite crear, mediante código *ActionScript*, herramientas *Flash* integradas en la aplicación *web*, suministrándole un considerable aumento de dinamismo e interactividad en su relación con el usuario final.

1.2.3. Algoritmo de alineamiento de secuencias: BLAST

Para ubicar las sondas de los microarrays de *Affymetrix* se utilizó el algoritmo *Basic Local Search Tool* (BLAST) **(Altschul et al., 1990)**. Este algoritmo de alineamiento compara de forma heurística una secuencia dada contra toda una librería de secuencias almacenadas en una base de datos, comprobando si existe alguna secuencia idéntica o similar en dicha librería, identificando la posición de la secuencia o secuencias homólogas encontradas y realizando un alineamiento entre lo encontrado y la secuencia problema.

Existen diversos programas que implementan el algoritmo BLAST, cada uno diseñado para manejar un tipo distinto de información (ver **tabla 1.5**). Estos programas han sido descargados de la página gubernamental de estadounidense *National Center for Biotechnology Information* (NCBI) accesible desde la dirección **http://blast.ncbi.nlm.nih.gov**.

Nombre del programa	Tipo de secuencia a buscar	Tipo de base de datos en donde busca
blastn	nucleótidos	nucleótidos
blastp	proteína	proteína
blastx	traducción de nucleótidos	proteína
tblastn	proteína	traducción de nucleótidos
tblastx	traducción de nucleótidos	traducción de nucleótidos

Tabla 1.5. Tipos de programas de alineamiento BLAST accesibles desde **http://blast.ncbi.nlm.nih.gov**.

Para el trabajo de alineamiento realizado en ***GATExplorer*** se lanzó el programa BLASTN, que es la versión diseñada para trabajar con nucleótidos. En concreto, se lanzaron miles de BLASTN correspondientes a cada una de las secuencias de cada sonda de 25 nucleótidos de cada uno de los microarrays utilizando como librerías de búsqueda los ficheros FASTA de *Ensembl* con todos los cDNAs correspondientes a cada genoma (humano, ratón y rata). El programa se configuró para admitir únicamente alineamientos perfectos (es decir, secuencias idénticas a la buscada) guardando el resultado en base de datos *MySQL* para su posterior proceso. El número de secuencias únicas para cada modelo de microarray se detalla en la **tabla 1.6**.

Organismo	Modelo de microarray	Número de secuencias distintas
Homo sapiens	HC_G110	30294
Homo sapiens	HG_Focus	97810
Homo sapiens	HG_U133A	241898
Homo sapiens	HG_U133A_2	241837
Homo sapiens	HG_U133B	248525
Homo sapiens	HG_U133_Plus_2	594532
Homo sapiens	HG_U95A	197599
Homo sapiens	HG_U95Av2	197582
Homo sapiens	HG_U95B	199191
Homo sapiens	HG_U95C	200491
Homo sapiens	HG_U95D	201274
Homo sapiens	HG_U95E	201012
Homo sapiens	Human_Exon_1.0	5270588
Homo sapiens	Human_Gene_1.0	804372
Homo sapiens	U133_X3P	631714
Mus musculus	MG_U74A	200843

Mus musculus	MG_U74Av2	197037
Mus musculus	MG_U74B	201514
Mus musculus	MG_U74Bv2	196971
Mus musculus	MG_U74C	200299
Mus musculus	MG_U74Cv2	182488
Mus musculus	MOE430A	245487
Mus musculus	MOE430B	247199
Mus musculus	Mouse430A_2	245487
Mus musculus	Mouse430_2	490490
Mus musculus	Mouse_Exon_1.0	4625878
Mus musculus	Mouse_Gene_1.0	833688
Mus musculus	Mu11KsubA	131205
Mus musculus	Mu11KsubB	118591
Rattus norvegicus	RAE230A	174975
Rattus norvegicus	RAE230B	168505
Rattus norvegicus	Rat230_2	341442
Rattus norvegicus	Rat_Exon_1.0	3997586
Rattus norvegicus	Rat_Gene_1.0	793624
Rattus norvegicus	RG_U34A	140057
Rattus norvegicus	RG_U34B	140293
Rattus norvegicus	RG_U34C	140252
Rattus norvegicus	RN_U34	21300
Rattus norvegicus	RT_U34	20407

Tabla 1.6. Número de secuencias distintas de los diferentes microarrays de expresión de *Affymetrix* para los organismos humano, ratón y rata.

Tras realizar todos los alineamientos contra el repositorio de cDNAs de *Ensembl* quedan bastantes sondas de oligonucleótidos huérfanas para las que no se encuentra una secuencia idéntica en ningún locus génico codificante y por ello se procedió a lanzar todas estas sondas contra la base de datos de ncRNAs citada **(Pang et al., 2007)**. Esto se realizó para las sondas de los microarrays de humano y de ratón, ya que en *RNAdb* no hay datos de rata.

El diseño original de las sondas de los microarrays de *Affymetrix* está basado en alineamientos consenso de ESTs (*Expressed Sequence Tags*), es decir en datos transcriptómicos globales sobre RNAs detectados en diversas muestras de la especie concreta para la cual se construyó el array. Esto supone que, como se ha comprobado luego, muchos de estos RNAs no corresponden a secuencias de un gen codificante para proteína. De este modo, no es de extrañar que haya multitud de sondas que no sean alineadas con transcritos codificantes cDNAs de *Ensembl*. En este trabajo se ha hecho un esfuerzo original para recuperar esas sondas huérfanas, tratando de asignarlas a nuevas entidades transcriptómicas mediante su alineación contra secuencias de la base de datos *RNAdb*.

1.2.4. Cambios de coordenadas: de cDNA a DNA genómico

Muchas de las secuencias de las sondas de los arrays que mapean sobre cDNAs quedan ubicadas cubriendo la unión de dos exones. El alineamiento de estas sondas sobre los ficheros FASTA de cDNA asegura poder incluirlas en el mapeo. Sin embargo, a la hora de representar la información a nivel genómico, las secuencias de cDNA no son las más adecuadas ya que los exones quedan separados por intrones en los locus génicos. En ***GATExplorer*** se ha buscado representar toda la información en su contexto genómico para permitir al usuario explorar

cada locus génico y cada región a lo largo de cada uno de los cromosomas. Es decir, se trata de un crear un navegador genómico que integre también datos transcriptómicos y permita visualizar ambos tipos de información de forma coherente. Para lograrlo cada locus presenta las sondas de oligonucleótidos en su lugar correspondiente, siendo por ello necesario un cambio de coordenadas de cDNA a coordenadas genómicas.

Para realizar el cambio de coordenadas de cDNA a DNA genómico, se utilizó la información de *Ensembl* en donde se almacena la estructura y posición de cada uno de los locus, transcritos y exones que conforman un organismo. Localizando en esta base de datos cada uno de los transcritos (ENST…), es posible ubicar cada sonda en su posición absoluta dentro del cromosoma. La **figura 1.4** muestra el alineamiento de 2 sondas pertenecientes a 2 *probesets* del array HG-U133A de *Affymetrix* sobre el transcrito "ENST00000345358". Los distintos exones de este transcrito se representan alternando los colores **gris** y **rojo**. La sonda ubicada en las coordenadas (x,y: 640,75) pertenece al *probeset* "21833_s_at" y se encuentra completamente dentro de un exón, siendo posible su alineamiento tanto en cDNA como en DNA genómico. La sonda (132,51) perteneciente al probeset "208478_s_at" se encuentra entre 2 exones por lo que es posible su alineamiento completo sobre cDNA pero no sobre DNA genómico. La transformación de cambio de coordenadas de cDNA a DNA genómico permite tener en cuenta la posición de cada exón en el cromosoma. De esta manera una determinada sonda queda asignada a un exón (o dos), un transcrito y un gen.

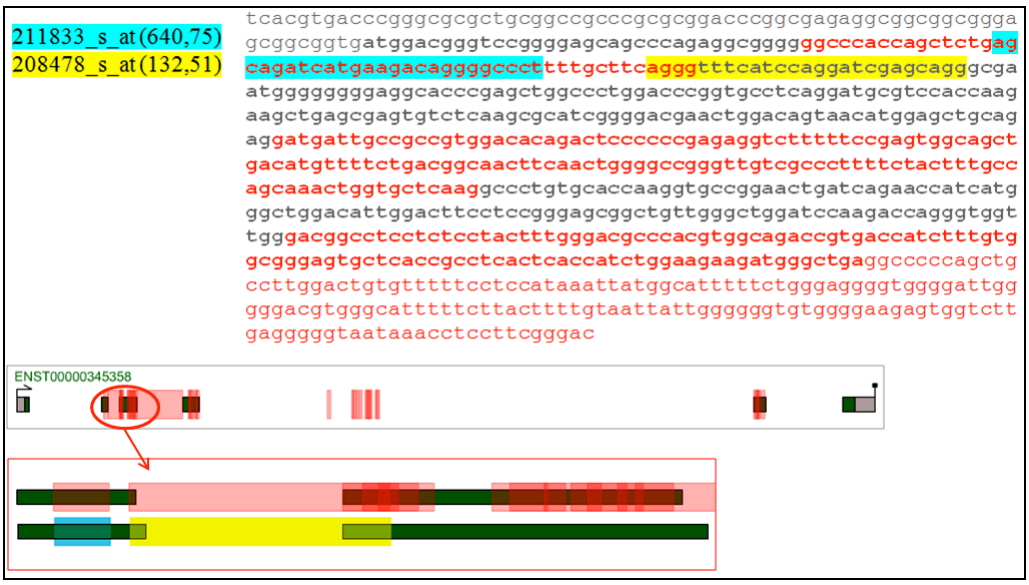


Figura 1.4. Alineamiento de dos sondas pertenecientes a dos probesets del modelo de microarray HG-U133A de *Affymetrix* sobre el transcrito ENST00000345358. Se muestra las diferencias entre el alineamiento sobre cDNA y el alineamiento sobre DNA genómico.

En la **figura 1.5** se presenta de modo gráfico un diagrama de flujo (*pipeline*) de los procesos y pasos que se han dado para la construcción de la plataforma bioinformática ***GATExplorer***, indicando las principales fuentes y bases de datos que se integran (cDNAs, mRNAs y ncRNAs; microarrays; expresión; genomas) así como los principales métodos necesarios para el re-mapeo de todas las sondas de arrays de *Affymetrix* a RNAs (mRNAs y ncRNAs), la ubicación de genes y expresión a escala global en el genoma, y el cálculo de expresión para una batería de datos de distintos tejidos. También se indican los archivos *outputs* que la plataforma produce.

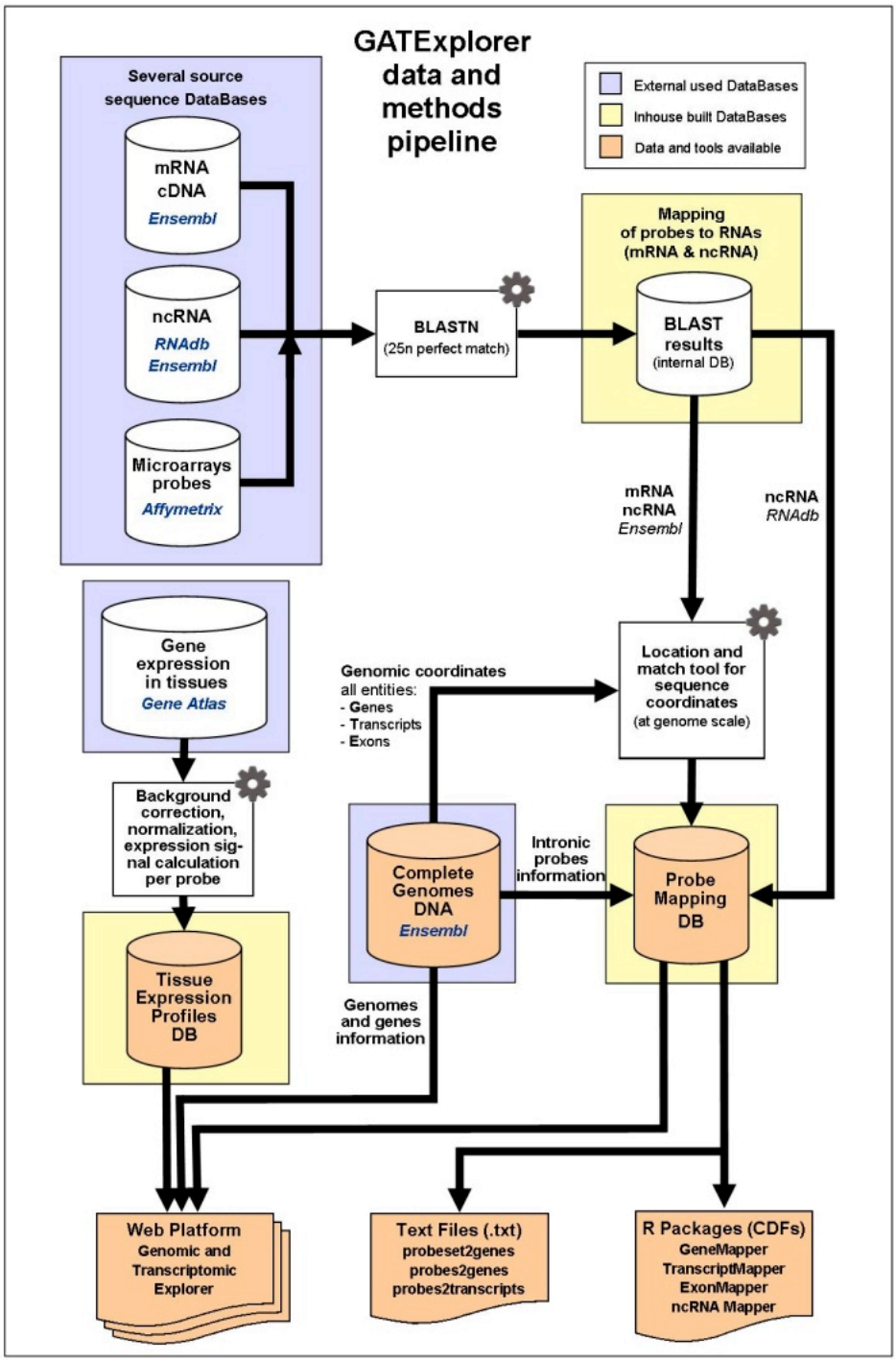


Figura 1.5. Representación esquemática del flujo de trabajo y procesos que se integran en ***GATExplorer***, indicando las bases de datos y métodos utilizados en su construcción, así como los principales tipos de datos que se producen (*output files*).

En resumen, como se ha indicado, las bases de datos externas utilizadas son: *Ensembl*, *RNAdb*, *Affymetrix* y *GeneAtlas.* El algoritmo BLASTN se utilizó para alinear las secuencias de 25 nucleótidos de cada sonda de los principales microarrays de expresión de *Affymetrix* de humano, ratón y rata, permitiendo únicamente alineamientos perfectos de longitud 25. Las librerías de búsqueda para RNAs codificantes de proteína (mRNA) proceden de *Ensembl*, mientras que para las secuencias de RNA no codificantes de proteína (ncRNA) se utilizó *RNAdb*. Tras el mapeo inicial, las sondas se ubicaron en contexto genómico realizando un cambio de coordenadas de cDNA a DNA genómico utilizando *Ensembl*. Los resultados son proporcionados

en un portal *web* interactivo, así como también en ficheros de texto plano (.txt) y paquetes para utilizar con el programa estadístico R **(Ihaka and Gentleman, 1996)**.

En la **figura 1.6** se presenta el flujo de las diferentes tareas a realizar en el desarrollo de un re-mapeo completo y re-construcción de los datos contenidos en ***GATExplorer*** de principio a fin, indicando su dependencia secuencial temporal, así como su grado de paralelización. Este esquema sirve de flujo de trabajo para la actualización de ***GATExplorer***.

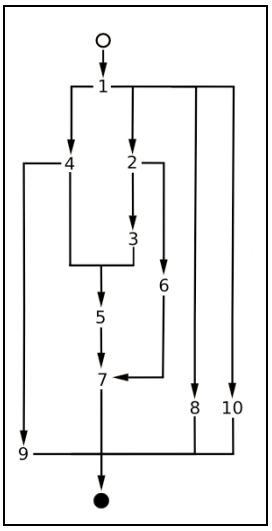


Figura 1.6. Esquema de las diferentes tareas en el proceso de re-mapeo. Las flechas indican la dependencia que tienen en el tiempo y determinan el grado de paralelización.

A continuación se detallan los distintos procesos o tareas específicas correspondientes a cada paso numerado en el flujo de trabajo.

Paso previo

Descripción: Descarga de ficheros de microarrays de la web de *Affymetrix* con información de las sondas. Por ejemplo: 15 ficheros de 15 microarrays en humano.

Procesos

Descripción	Duración aprox.	Recursos
Descarga de ficheros de texto tabulados. (Human_Exon_1.0 525MB) (Human_Gene_1.0 82MB)		Conexión a internet

Paso 1

Descripción: Descarga de datos de *Ensembl* e instalación en la BBDD.

Procesos

Descripción	Duración aprox.	Recursos
Descarga de datos del FTP de *Ensembl* (ftp.ensembl.org) ej. homo_sapiens_core_53_36o: - Script de creación de BBDD MySQL: • 2.4GB comprimida, 11GB sin comprimir - Fichero fasta cDNA (134MB) - Fichero fasta DNA (3.5GB) - Fichero fasta ncRNA (1.9MB)	6 horas	Acceso a internet
Descompresión de los datos.	30 minutos	Proceso + acceso a disco HD
Instalación de la BBDD en servidor.	2 horas	Proceso + acceso a disco HD

Paso 2

Descripción:
Asignación de sondas mapeadas por *Ensembl* a sus locus en función de sus coordenadas genómicas.

Procesos

Descripción	Duración aprox.	Recursos
Sentencias SQL.	25 horas	Proceso + acceso a disco HD

Paso 3

Descripción:
Obtención de secuencias de las sondas en función de su posición en el genoma y comparación.

Procesos

Descripción	Duración aprox.	Recursos
Sentencias SQL.	10 horas	Proceso + acceso a disco HD
Proceso Java (*Probesequencer*): Recupera secuencias de *contigs* de la BBDD e identifica la zona de la sonda para obtener su secuencia.	8 horas	Proceso + acceso a disco HD
Proceso Java (*ComplementaryStrand*): Recupera secuencias de sondas de BBDD, realiza su inverso complementario y posteriormente inserta el resultado en BBDD. Es necesario para sondas de microarrays de exones (Human_Gene_1.0 y Human_Exon_1.0)	20 minutos	Proceso + acceso a disco HD

Paso 4

Descripción: BLAST sobre cDNA de todas las secuencias distintas de todos los microarrays. Una vez insertados en BBDD se seleccionan los que mapean sense y los antisense y se asignan a los arrays de expresión y de exones respectivamente. Posteriormente se identifican las sondas inter-exónicas ((PosicionFinalPosicionInicial+1)>25)

Procesos

Descripción	Duración aprox.	Recursos
Se concatenan los ficheros cDNA y ncRNA en uno solo fichero fasta (shell de Linux).	<1 minuto	Acceso a disco duro
Formateo del fichero fasta (*formatdb* de *blastall*).	<1 minuto	Proceso + acceso a disco HD
Proceso PHP: Recupera todas las secuencias distintas (7399361 en 15 microarrays de humano) y lanza un proceso BLAST (*blastn* de *blastall*) sobre el fichero fasta por cada secuencia. BLAST devuelve los resultados en formato XML que son parseados para insertarlos en BBDD.	48 horas	8 procesadores (el número de sondas a alinear se divide entre 8 procesos trabajando en paralelo) 6 GB de RAM
Proceso Java (*Complementary* Strand): Invierte las secuencias que mapean en la hebra complementaria para hacer la unión con la tabla de sondas de *Affymetrix* e identificar su id de sonda, id de probeset y posición en el microarray de exones.	20 minutos	Proceso + acceso a disco HD
Proceso Java (*Interexonic*): Recupera la información del mapeo de BBDD y realiza un cambio de coordenadas de cDNA a DNA de cada sonda sumando la longitud de los intrones de cada locus.	20 minutos	Acceso a disco duro (Se lanzan varios procesos concurrentes pero la principal limitación es el acceso a BBDD así que no hay gran beneficio con una CPU multiprocesador).
Sentencias SQL.	2-3 horas	Proceso + acceso a disco HD

Paso 5

Descripción:
Se crea una tabla que unifica la información del mapeo del paso 4 (gen, transcrito) con la información de *Affymetrix* (modelo de microarray, posición (x;y) en el microarray, probeset, sonda, hebra y secuencia)

Procesos

Descripción	Duración aprox.	Recursos
Sentencias SQL	15 minutos	Proceso + acceso a disco HD

Paso 6

Descripción: Se obtienen las sondas mapeadas por *Ensembl* sobre exones exclusivamente (actualmente no se usa en la aplicación)

Procesos

Descripción	Duración aprox.	Recursos
Sentencias SQL	1.25 horas	Proceso + acceso a disco HD

Paso 7

Descripción
Se localizan las sondas intrónicas y se insertan a la tabla del paso 5. Se realiza el recuento de sondas por gen, genes por sonda, transcritos por sonda y número de exones por sonda (para *Human_Exon_1.0*)

Procesos

Descripción	Duración aprox.	Recursos
Sentencias SQL.	3 horas	Proceso + acceso a disco HD

Paso 8

Descripción: Preparación del fichero fasta de DNA para poder utilizar el algoritmo BLAST en la aplicación online. Formateo de fichero fasta de DNA.

Procesos

Descripción	Duración aprox.	Recursos
Formateo de fichero fasta de dna (*formatdb* de *blastall*).	6 minutos	Proceso + acceso a disco HD

Paso 9

Descripción: Obtención de sondas que mapean en la hebra complementaria de cada gen a partir de las tablas generadas en el paso 4.

Procesos

Descripción	Duración aprox.	Recursos
Sentencias SQL.	30 minutos	Proceso + acceso a disco HD

Paso 10

Descripción: Creación de BBDD que hará corresponder la información de dominios de proteínas a su posición en el genoma.

Procesos

Descripción	Duración aprox.	Recursos
Sentencias SQL	30 minutos	Proceso + acceso a disco HD
Proceso Java (*Domains*): Calcula el inicio y el fin de cada dominio sobre el locus génico a partir de su posición en la proteína.	15 minutos	Proceso + acceso a disco HD

A la hora de escribir el código fuente de todos estos procesos hemos utilizado un entorno de desarrollo integrado IDE (*Integrated Development Environment*). Este tipo de herramientas ayuda al desarrollador a la edición del código mediante el resaltado de la sintaxis, generación automática de código, ayudas contextuales y facilidades para el trabajo en grupo. También ayuda en las tareas de compilación, depuración y empaquetamiento de aplicaciones. El entorno de desarrollo elegido para este proyecto fue *Eclipse*. Este programa es de código abierto y está accesible desde la dirección **http://www.eclipse.org/**. *Eclipse* es un programa muy versátil que cuenta, gracias a una comunidad grande de usuarios y desarrolladores, con multitud de *plugins* o extensiones que implementan distintas funcionalidades. Para este proyecto se utilizaron dos configuraciones distintas de *Eclipse*, una para los programas *PHP*, y otra para los programas *Java*.

1.3. Resultados

1.3.1. Mapeo completo de sondas de expresión a *loci* génicos

El resultado del re-mapeo ubica las sondas de oligonucleótidos de todos los microarrays de expresión en diferentes regiones de los *loci* génicos del genoma humano, de ratón y de rata. Estas regiones en los *loci* de los genes codificantes de proteína pueden ser: UTR-5' inicial, UTR-3' final, exones codificantes de proteína, regiones inter-exónicas y regiones intrónicas no codificantes. El mapeo sobre genes no codificantes de proteína es muy similar, con la única diferencia de que ninguno de sus exones será finalmente traducido a proteína.

El mapeo de secuencias es específico a cada una de las dos hebras del DNA (la hebra directa o la hebra complementaria), como lo es la localización de cada *locus* génico concreto. Sin embargo, nuestro mapeo también ha detectado un gran número de sondas ubicadas en la hebra complementaria de genes diana. Esto puede indicar errores o falta de conocimiento en la información genómica de referencia que da *Ensembl* tomada como fuente por *Affymetrix* a la hora de diseñar sus chips y asignarlos a genes concretos. También hay un gran número de sondas que no han podido alinearse sobre ninguna entidad transcripcional conocida hasta la fecha: ni mRNAs, ni ncRNAs. Estas sondas no asignadas pueden deberse a errores en el diseño original como sondas de expresión o, más probablemente, a cambios en las secuencias de referencia de los genes y regiones génicas, o a secuencias detectadas por expresión (ESTs) que se creían funcionales pero luego no han sido consideradas relevantes y no han sido incluídas y anotadas en las bases de datos de genomas completos.

Respecto al mapeo de sondas, la diferencia entre la ubicación de las sondas entre los dos tipos de microarrays de *Affymetrix* es muy grande. En la **figura 1.7** puede verse la diferencia entre el modelo *HG-U133 Plus 2.0* (array de tipo: IVT 3') y el modelo de exones *Human Exon 1.0* (array de tipo: *Whole Transcript*) sobre el gen PLCB3 (*phospholipase C, beta 3*). La tecnología de marcaje e hibridación molecular proporcionada por *Affymetrix*, hacía necesario que los arrays tipo IVT 3' –que son los más antiguos– concentrasen sus sondas en la región UTR 3' del *locus* génico diana, buscando el máximo aprovechamiento de las distintas amplificaciones del RNA. Sin embargo, los nuevos arrays tipo *Whole Transcript* distribuyen sus sondas a lo largo de todo el *locus* génico diana, con objeto de lograr mapear todos o casi todos los exones. Esto es posible gracias a que utilizan la tecnología de *Random Primer Labeling* en el procesado, marcaje e hibridación molecular proporcionado por *Affymetrix*. La incorporación de sondas a

lo largo de todo el *locus* hace posible ir más allá de la simple medida de expresión del gen y permite otro tipo de análisis nuevo, como el diseño de una estrategia de detección de *splicing* alternativo. Además, frente a los modelos antiguos como el *HG-U133 Plus 2.0* que asignan 11 sondas (*probes*) por cada conjunto de sondas (*probeset*), los nuevos *Human Exon* asignan en torno a 84 sondas por *locus*. Este considerable incremento en el número de sondas supone una mejora sustancial en las medidas de expresión permitiendo análisis estadísticos más robustos.

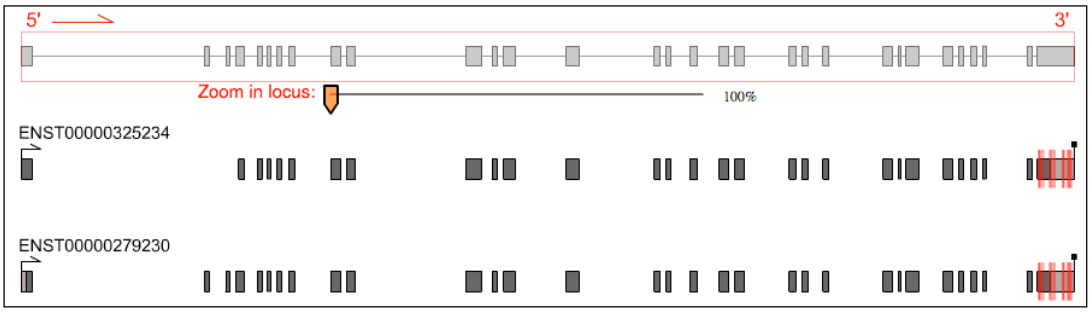


Figura 1.7a. Representación de tres transcritos del gen PLCB3 . En rojo se representan la ubicación de las 11 sondas (*probes*) del conjunto de sondas (*probeset*) para este gen incluido en el array HG-U133 (de tipo IVT 3'). Como puede verse, dichas sondas están concentradas en el extremo UTR 3'.

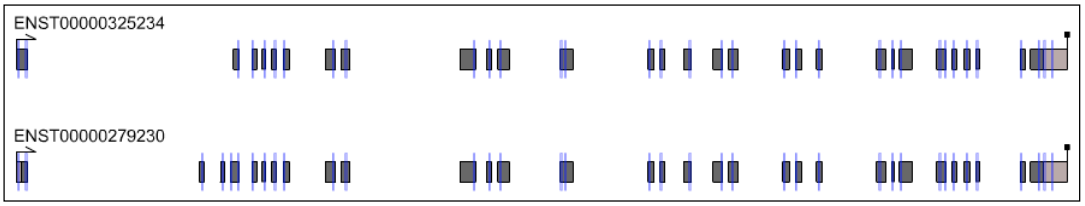


Figura 1.7b. Representación de los transcritos del gen PLCB3. En azul se representan la ubicación de las 84 sondas del array de exones *Human Exon*, ubicadas en todos los exones permitiendo la medición más precisa de la expresión del gen a lo largo de todo el *locus*.

1.3.2. Análisis y estadísticas de los resultados del mapeo de sondas

Los resultados del mapeo se pueden resumir en la **figura 1.8**. Esta figura presenta en datos estadísticos la ubicación de las sondas de los cuatro modelos más populares de microarrays de expresión humanos. Este re-mapeo se hizo utilizando datos de la versión 57 de *Ensembl* que corresponde al ensamblaje GRCh37 (hg19) del genoma humano y de la versión de 2009 de *RNAdb*. Se representa en **verde** la parte proporcional de sondas que mapean sobre mRNAs, en **azul** la parte mapeada sobre ncRNAs y en **rojo** el número de sondas que no fue posible asignar a ninguna entidad biomolecular conocida. Parte de estas sondas no asignadas puede explicarse por la ubicación de las mismas en los arrays tipo IVT 3', los cuales ubican muchas de sus sondas las regiones 3' de los genes cerca del extremo final del *locus* génico. Se ha observado que estas regiones UTR se encuentran no bien definidas para muchos genes humanos **(Muro et al., 2008)** y varían bastante entre las distintas versiones del genoma afectando al mapeo de ciertas sondas. Respecto a los transcritos ncRNA se ha hecho distinción entre 3 grupos: **(i)** mapeo a transcritos procedentes de *Ensembl* (que incluye algunos ncRNAs) **(ii)** mapeo a transcritos procedentes de *RNAdb* y **(iii)** sondas ubicadas en intrones, que son consideradas como parte posible de algún RNA ya que el diseño original de *Affymetrix* –sobre todo en los modelos IVT 3'– colocó sondas en algunos intrones basándose en alguna evidencia biológica quizás no recogida en todas las bases de datos. Se puede apreciar que la proporción de sondas ubicadas

en regiones no codificantes de proteína (en **azul**) y de sondas no asignadas (en **rojo**) es muy variable entre los diferentes modelos de arrays. Se puede observar cómo los modelos *HG-U133A* y *Human Gene 1.0* están más orientados a la detección de genes ya que la parte **verde**, indicando cobertura sobre mRNA, es mayor: 81,7% y 73,3% respectivamente. En el caso de *Human Gene 1.0*, al ser un modelo más moderno y con mayor número de sondas, también incrementa su cobertura sobre ncRNA respecto al array tipo *HG-U133A*. El array *HG-U133 Plus 2.0* es la unión de los modelos *HG-U133A* y *HG-U133B*. Este modelo B se diseñó de forma complementaria al A cubriendo regiones de menor evidencia biológica, lo que se ve reflejado en un gran porcentaje de sondas mapeando ncRNA (25,5%) y de sondas no asignadas (16,5%) en el array *HG-U133 Plus 2.0*. Finalmente, las cifras del modelo *Human Exon 1.0* evidencian que fue diseñado para mucho más que para la simple medición de la expresión de genes, contando con un gran número de sondas (más de 5 millones) y situando un gran porcentaje de ellas fuera de transcritos codificantes conocidos.

Mapping of microarray probes to transcribed entities:	HG U133A		HG U133 Plus 2.0		Human Gene 1.0		Human Exon 1.0	
mRNAs (EXONs encoding for proteins)	197742	81.7%	344936	58.0%	589695	73.3%	1408591	26.7%
ncRNAs	19949	8.2%	151370	25.5%	175348	21.8%	1686071	32.0%
INTRONs (as putative RNAs)	11654		85724		23476		1395112	
ncRNAs (*Ensembl*)	4301		30805		144014		228785	
ncRNAs (*RNAdb*)	9404		82380		25793		169983	
Not Assigned to RNAs (NA)	24207	10.0%	98226	16.5%	39329	4.9%	2175926	41.3%
TOTAL number of probes	241898		594532		804372		5270588	

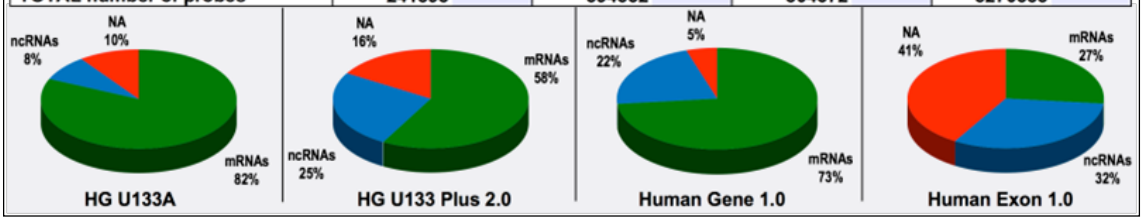


Figura 1.8. Estadística resultante del mapeo sobre los 4 modelos más populares de microarrays humanos de *Affymetrix* de expresión. Se comparan el número de sondas totales y la proporción de sondas ubicadas en mRNAs (**verde**), en ncRNAs (**azul**) de distinto tipo y las no asignadas (**rojo**). La figura se muestra en inglés por mostrar directamente los datos incluidos en la aplicación bioinformática ***GATExplorer*** (construída integramente en inglés).

1.3.3. Valoración de la cobertura y eficiencia de los microarrays para medir la expresión génica global

Con los resultados obtenidos del re-mapeo global de sondas se puede valorar la precisión de los arrays en términos de **cobertura** y **eficiencia** (*coverage and efficiency*). La información detallada sobre la cobertura y precisión del mapeo de cada microarray se presenta en la **tabla 1.7** y en la **tabla 1.8** .

Definimos la **cobertura** como la proporción (porcentaje, %) de genes que un modelo concreto de microarray puede identificar del total de genes conocidos presentes dentro del genoma para la especie correspondiente.

El conjunto de genes que constituyen un genoma se ha determinado como el número de genes codificante de proteína identificados y reportados por *Ensembl*. Se ha utilizado este criterio debido a que la mayoría de estos microarray se definió para la identificación de genes codificantes de proteína (es decir genes que se transcriben a mensajeros mRNAs que luego son traducidos), y porque el conocimiento de estos genes es mucho mayor que el de los transcritos

no codificantes.

Definimos la **eficiencia** como la proporción de sondas de un modelo concreto de microarray que mapea en algún gen de *Ensembl* de forma única no ambigua, esto es, sin hibridación cruzada con otros genes, y sin considerar las sondas únicamente localizadas en intrones.

	Transcripts				Gene Loci					
	Unique mapped		All mapped		Unique mapped		All mapped		TOTAL Number of transcripts	TOTAL Number of Gene Loci
Microarray	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**	**Nº**	**%**		
Human										
HG_U133A	5646	**5,63%**	47376	**47,23%**	12299	**57,79%**	13415	**63,04%**	100299	21281
HG_U133A_2	5646	**5,63%**	47376	**47,23%**	12299	**57,79%**	13415	**63,04%**	100299	21281
HG_U133B	4270	**4,26%**	24072	**24,00%**	7433	**34,93%**	8139	**38,25%**	100299	21281
HG_U133_Plus_2	10561	**10,53%**	68147	**67,94%**	17724	**83,29%**	18950	**89,05%**	100299	21281
HG_U95A	3626	**3,62%**	31634	**31,54%**	8545	**40,15%**	9561	**44,93%**	100299	21281
HG_U95Av2	3627	**3,62%**	31615	**31,52%**	8546	**40,16%**	9560	**44,92%**	100299	21281
HG_U95B	2719	**2,71%**	17303	**17,25%**	5483	**25,76%**	5955	**27,98%**	100299	21281
HG_U95C	2041	**2,03%**	14519	**14,48%**	3996	**18,78%**	5195	**24,41%**	100299	21281
HG_U95D	1448	**1,44%**	9211	**9,18%**	2572	**12,09%**	3433	**16,13%**	100299	21281
HG_U95E	2178	**2,17%**	13717	**13,68%**	4004	**18,81%**	4638	**21,79%**	100299	21281
HG_Focus	3198	**3,19%**	28725	**28,64%**	8156	**38,33%**	9017	**42,37%**	100299	21281
HC_G110	533	**0,53%**	5914	**5,90%**	1343	**6,31%**	1845	**8,67%**	100299	21281
U133_X3P	9953	**9,92%**	63392	**63,20%**	17583	**82,62%**	18787	**88,28%**	100299	21281
Human_Gene_1.0	17117	**17,07%**	97192	**96,90%**	19213	**90,28%**	20192	**94,88%**	100299	21281
Human_Exon_1.0	39350	**39,23%**	99816	**99,52%**	20238	**95,10%**	21012	**98,74%**	100299	21281
Mouse										
MG_U74A	4429	**6,29%**	19521	**27,73%**	7371	**32,32%**	8815	**38,65%**	70406	22806
MG_U74Av2	4870	**6,92%**	20934	**29,73%**	8123	**35,62%**	9330	**40,91%**	70406	22806
MG_U74B	3170	**4,50%**	12574	**17,86%**	5179	**22,71%**	5711	**25,04%**	70406	22806
MG_U74Bv2	3843	**5,46%**	15176	**21,55%**	6311	**27,67%**	6880	**30,17%**	70406	22806
MG_U74C	1160	**1,65%**	4502	**6,39%**	1757	**7,70%**	2498	**10,95%**	70406	22806
MG_U74Cv2	2165	**3,08%**	7561	**10,74%**	3399	**14,90%**	3928	**17,22%**	70406	22806
Mouse430_2	11967	**17,00%**	44667	**63,44%**	17037	**74,70%**	18402	**80,69%**	70406	22806
Mouse430A_2	7850	**11,15%**	33714	**47,89%**	12572	**55,13%**	13795	**60,49%**	70406	22806
MOE430A	7850	**11,15%**	33714	**47,89%**	12572	**55,13%**	13795	**60,49%**	70406	22806
MOE430B	4996	**7,10%**	15321	**21,76%**	6853	**30,05%**	7379	**32,36%**	70406	22806
Mu11KsubA	2621	**3,72%**	12435	**17,66%**	4530	**19,86%**	6026	**26,42%**	70406	22806
Mu11KsubB	1754	**2,49%**	9477	**13,46%**	3023	**13,26%**	3873	**16,98%**	70406	22806
Mouse_Gene_1.0	19692	**27,97%**	69162	**98,23%**	21390	**93,79%**	22354	**98,02%**	70406	22806
Mouse_Exon_1.0	39114	**55,55%**	69962	**99,37%**	21506	**94,30%**	22412	**98,27%**	70406	22806
Rat										
RG_U34A	3426	**10,39%**	8254	**25,03%**	4406	**19,21%**	5664	**24,69%**	32971	22938
RG_U34B	2348	**7,12%**	4871	**14,77%**	3117	**13,59%**	3453	**15,05%**	32971	22938
RG_U34C	2622	**7,95%**	5536	**16,79%**	3499	**15,25%**	3970	**17,31%**	32971	22938
Rat230_2	9253	**28,06%**	19554	**59,31%**	12065	**52,60%**	13428	**58,54%**	32971	22938
RAE230A	6828	**20,71%**	14859	**45,07%**	8986	**39,18%**	10209	**44,51%**	32971	22938
RAE230B	3114	**9,44%**	6513	**19,75%**	4201	**18,31%**	4498	**19,61%**	32971	22938
RN_U34	545	**1,65%**	1463	**4,44%**	723	**3,15%**	896	**3,91%**	32971	22938
RT_U34	434	**1,32%**	982	**2,98%**	536	**2,34%**	735	**3,20%**	32971	22938
Rat_Gene_1.0	21469	**65,11%**	32451	**98,42%**	21787	**94,98%**	22464	**97,93%**	32971	22938
Rat_Exon_1.0	22442	**68,07%**	32463	**98,46%**	21773	**94,92%**	22483	**98,02%**	32971	22938

Tabla 1.7. Número y porcentaje de genes codificantes de proteína que son mapeados por las sondas de cada modelo de microarray de expresión de *Affymetrix*. La tabla se muestra en inglés por mostrar directamente los datos incluidos en la aplicación bioinformática ***GATExplorer***.

Para optimizar la precisión de la tecnología de microarrays aplicada a para medir la expresión génica global –a escala ómica– es necesario minimizar la posible **hibridación cruzada** debida a sondas que mapean en varios *loci* génicos o en varias entidades transcritas (tanto mRNAs como ncRNAs).

	Transcripts				Gene Loci						
	1		>1 (ambiguous)		1		>1 (ambiguous)		TOTAL Number of probes mapping	TOTAL Number of probes in microarray	Mapping efficiency
Microarray	**Number**	**%**	**Number**	**%**	**Number**	**%**	**Number**	**%**			**%**
Human											
HG_U133A	65103	**32,22%**	136940	**67,78%**	192213	**95,13%**	9830	**4,87%**	202043	241898	**83,52%**
HG_U133A_2	65081	**32,22%**	136932	**67,78%**	192191	**95,14%**	9822	**4,86%**	202013	241837	**83,53%**
HG_U133B	56958	**45,56%**	68073	**54,44%**	120712	**96,55%**	4319	**3,45%**	125031	248525	**50,31%**
HG_U133_Plus_2	149924	**39,90%**	225817	**60,10%**	360264	**95,88%**	15477	**4,12%**	375741	594532	**63,20%**
HG_U95A	53670	**31,85%**	114860	**68,15%**	160229	**95,07%**	8301	**4,93%**	168530	197599	**85,29%**
HG_U95Av2	53685	**31,86%**	114839	**68,14%**	160219	**95,07%**	8305	**4,93%**	168524	197582	**85,29%**
HG_U95B	46773	**42,00%**	64599	**58,00%**	108609	**97,52%**	2763	**2,48%**	111372	199191	**55,91%**
HG_U95C	36330	**43,66%**	46878	**56,34%**	79842	**95,95%**	3366	**4,05%**	83208	200491	**41,50%**
HG_U95D	24633	**49,94%**	24694	**50,06%**	47522	**96,34%**	1805	**3,66%**	49327	201274	**24,51%**
HG_U95E	37240	**44,70%**	46072	**55,30%**	80104	**96,15%**	3208	**3,85%**	83312	201012	**41,45%**
HG_Focus	29521	**32,70%**	60753	**67,30%**	85854	**95,10%**	4420	**4,90%**	90274	97810	**92,30%**
HC_G110	7548	**28,70%**	18752	**71,30%**	24687	**93,87%**	1613	**6,13%**	26300	30294	**86,82%**
U133_X3P	159798	**40,73%**	232564	**59,27%**	374931	**95,56%**	17431	**4,44%**	392362	631714	**62,11%**
Human_Gene_1.0	294841	**40,19%**	438868	**59,81%**	673873	**91,84%**	59836	**8,16%**	733709	804372	**91,22%**
Human_Exon_1.0	619903	**37,86%**	1017473	**62,14%**	1543530	**94,27%**	93846	**5,73%**	1637376	5270588	**31,07%**
Mouse											
MG_U74A	63534	**48,92%**	66344	**51,08%**	122415	**94,25%**	7463	**5,75%**	129878	200843	**64,67%**
MG_U74Av2	70767	**49,01%**	73635	**50,99%**	136408	**94,46%**	7994	**5,54%**	144402	197037	**73,29%**
MG_U74B	50527	**51,59%**	47419	**48,41%**	95976	**97,99%**	1970	**2,01%**	97946	201514	**48,61%**
MG_U74Bv2	61653	**51,40%**	58306	**48,60%**	117670	**98,09%**	2289	**1,91%**	119959	196971	**60,90%**
MG_U74C	16164	**59,45%**	11027	**40,55%**	26505	**97,48%**	686	**2,52%**	27191	200299	**13,58%**
MG_U74Cv2	27675	**57,30%**	20627	**42,70%**	47329	**97,99%**	973	**2,01%**	48302	182488	**26,47%**
Mouse430_2	158665	**51,51%**	149387	**48,49%**	296966	**96,40%**	11086	**3,60%**	308052	490490	**62,80%**
Mouse430A_2	99246	**48,08%**	107163	**51,92%**	197309	**95,59%**	9100	**4,41%**	206409	245487	**84,08%**
MOE430A	99246	**48,08%**	107163	**51,92%**	197309	**95,59%**	9100	**4,41%**	206409	245487	**84,08%**
MOE430B	59861	**58,13%**	43114	**41,87%**	100780	**97,87%**	2195	**2,13%**	102975	247199	**41,66%**
Mu11KsubA	48328	**49,85%**	48628	**50,15%**	91735	**94,62%**	5221	**5,38%**	96956	131205	**73,90%**
Mu11KsubB	31148	**44,66%**	38591	**55,34%**	64613	**92,65%**	5126	**7,35%**	69739	118591	**58,81%**
Mouse_Gene_1.0	352685	**50,36%**	347663	**49,64%**	662327	**94,57%**	38023	**5,43%**	700348	833688	**84,01%**
Mouse_Exon_1.0	618129	**47,47%**	684002	**52,53%**	1251511	**96,11%**	50622	**3,89%**	1302131	4625878	**28,15%**
Rat											
RG_U34A	57918	**64,96%**	31235	**35,04%**	84557	**94,84%**	4596	**5,16%**	89153	140057	**63,65%**
RG_U34B	33570	**70,08%**	14332	**29,92%**	47057	**98,24%**	845	**1,76%**	47902	140293	**34,14%**
RG_U34C	37294	**70,14%**	15880	**29,86%**	51938	**97,68%**	1236	**2,32%**	53174	140252	**37,91%**
Rat230_2	101652	**67,92%**	48018	**32,08%**	145812	**97,42%**	3858	**2,58%**	149670	341442	**43,83%**
RAE230A	69932	**67,82%**	33180	**32,18%**	99912	**96,90%**	3200	**3,10%**	103112	174975	**58,93%**
RAE230B	32336	**68,08%**	15160	**31,92%**	46737	**98,40%**	759	**1,60%**	47496	168505	**28,19%**
RN_U34	9480	**60,60%**	6164	**39,40%**	15182	**97,05%**	462	**2,95%**	15644	21300	**73,45%**
RT_U34	9591	**67,07%**	4708	**32,93%**	13190	**92,24%**	1109	**7,76%**	14299	20407	**70,07%**
Rat_Gene_1.0	451267	**70,94%**	184883	**29,06%**	609761	**95,85%**	26389	**4,15%**	636150	793624	**80,16%**
Rat_Exon_1.0	572414	**60,16%**	379114	**39,84%**	919569	**96,64%**	31959	**3,36%**	951528	3997586	**23,80%**

Tabla 1.8. Número y porcentaje de sondas que mapean sobre genes y transcritos para cada modelo de microarray de *Affymetrix*, detallando cuántas de esas sondas son únicas y presentan hibridación cruzada. La tabla se muestra en inglés por mostrar directamente los datos incluidos en la aplicación bioinformática ***GATExplorer***.

La **tabla 1.7** muestra el número y porcentaje de genes codificante de proteína que son mapeados por las sondas de cada modelo de microarray de expresión de *Affymetrix*. En esta tabla se diferencia entre genes y transcritos, especificando cuántos de ellos son mapeados de forma única. Esta tabla muestra que la cobertura de genes conocidos (21281 para humano en la versión 57 de *Ensembl*) ha aumentado en con la llegada de cada modelo nuevo: *HG-U133A* 63,0%; *HG-U133 Plus 2.0* 89,0%; *Human Gene 1.0* 94,9%; *Human Exon 1.0* 98,7%. En el caso de los transcritos se ha considerado únicamente los transcritos pertenecientes a los genes codificante de proteína (100299 para humano en la versión 57 de *Ensembl*), obteniendo el mismo resultado de aumento de cobertura con la llegada de nuevos modelos de microarrays.

La **tabla 1.8** muestra el número y porcentaje de sondas que mapean sobre genes y transcritos para cada modelo de microarray, detallando cuántas de esas sondas son únicas y cuantas ambiguas (es decir, presentan hibridación cruzada). Estos datos reflejan que el modelo más eficiente sobre el organismo humano es el *Human Gene 1.0* con un 91,22% de sondas. Por ejemplo, para el caso del modelo *HG-U133A* el 16,5% de las sondas no mapean en ningún gene de la citada versión de *Ensembl*. Si además sólo se considera el número de sondas únicas (192213 para el array *HG-U133A*) la eficiencia en el mapeo es solo del 79,5% para este modelo. Todo ello indica que una proporción considerable de sondas (16-21%) pueden producir **ruido debido al mapeo incorrecto o ambiguo**, especialmente si se calcula la expresión utilizando la agrupación original proporcionada por *Affymetrix*. Este problema está también presente en el nuevo microarray de exones que muestra la eficiencia más baja, con solo un 31% de las sondas mapeando sobre exones. Estos datos indican que estos microarrays están sujetos a un alto nivel de ruido, y esto debe ser tenido en cuenta a la hora de su utilización.

1.3.4. Distribuciones del número de sondas únicas no ambiguas y del número de genes mapeados

En las estadísticas anteriores se determinó el número de sondas no ambiguas a nivel de gen, siendo por lo tanto las únicas que pueden utilizarse para los análisis de expresión génica ya que no presentan hibridación cruzada con más genes. En la **figura 1.9a** se muestra la distribución del número de sondas presentes por número de genes para dos modelos de arrays de distinto diseño: *HG-U133A* y *Human Gene 1.0*. Esta figura indica que la mayoría de las sondas detecta un único gen (en concordancia con la **tabla 1.8**) descendiendo rápidamente el número de sondas que detectan más de un gen. En la **figura 1.9b** se muestra el número de genes en función del número de sondas que los detectan. El diseño de las sondas de los antiguos modelos IVT 3' –como es el *HG-U133A*– se diseñaron definiendo grupos de 11 sondas próximas en el transcriptoma (*probesets*). Algunos genes son detectados por más de un *probeset* y esto queda reflejado en la **figura 1.9b** en forma de picos múltiplos de 11 para el array *HG-U133A* (línea negra). En el caso del modelo *Human Gene 1.0* la distribución es muy distinta mostrando un pico máximo en 25. Llama la atención el alto número de genes que son mapeados por una única sonda en ambos modelos. Esto podría ser explicado por la hibridación cruzada entre genes de la misma familia con secuencias similares (genes parálogos), o por la aparición en las bases de datos actuales de nuevos genes no conocidos en el momento del diseño de los chips. Muchos de estos genes nuevos son detectados por técnicas automáticas de análisis de secuencia y anotados como genes putativos (*genes like L*) o pseudo-genes, y su expresión muchas veces es dudosa.

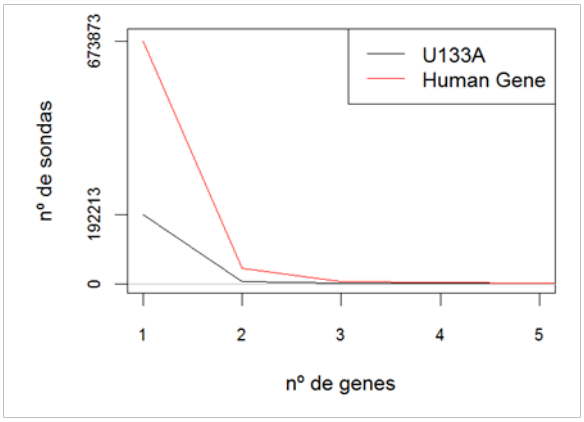


Figura 1.9a. Distribución del número de sondas mapeadas a un número de genes único (1, exclusivas) o a varios genes (>1, ambiguas) para los microarrays *HG-U133A* y *Human Gene 1.0.*

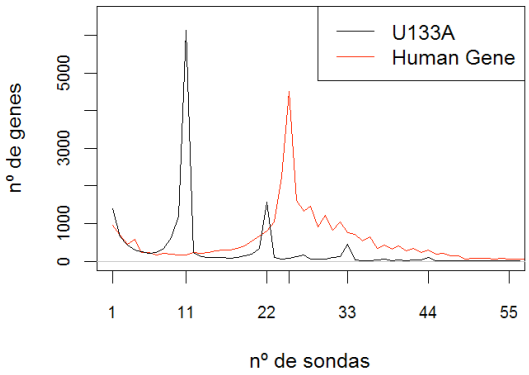


Figura 1.9b. Distribución del número de genes que son mapeados por un número concreto de sondas (11, 22, etc) para los microarrays *HG-U133A* y *Human Gene 1.0.*

1.3.5. Expresión de transcritos no codificantes de proteína (ncRNAs)

Una vez realizado el mapeo sobre ncRNA, cabe preguntarse si estas sondas situadas exclusivamente en regiones no codificantes de proteína muestran perfiles similares a las sondas que detectan genes codificantes. Según algunos estudios recientes los transcritos no codificantes muestran una expresión variable y regulada a través de distintos tejidos, lo que implica que son partes funcionales de la célula **(Mercer et al., 2009; Nakaya et al., 2007)**.

Para comprobar si las sondas de los microarrays pueden detectar realmente cambios en los transcritos no codificantes, se utilizó un *set* de datos de 353 microarrays de expresión en tejidos humanos (GEO ID GSE3526) **(Roth et al., 2006)**, de los que se seleccionaron 15 tomando 3 réplicas de 5 tejidos de regiones corporales y fisiología muy diferente: hipotálamo (tejido nervioso central), corazón (tejido muscular cardiaco), médula ósea (tejido fuente de la hematopoyesis), hígado y bazo (órganos con funciones específicas). Como prueba inicial, se comprobó si la expresión de las sondas que detectan los genes de *Ensembl*, entre los que se incluyen también algunos genes no codificantes, agrupaba las réplicas biológicas correctamente en un test de agrupamiento (*clustering*) no supervisado. La **figura 1.10a** muestra cómo se agrupan de tres en tres las distintas muestras en función de su tipo biológico, indicando su semejanza en cuanto a expresión génica. Posteriormente se procedió de la misma manera pero utilizando únicamente las sondas ubicadas en transcritos procedentes de la base de datos *RNAdb*. El resultado (**figura 10b**) muestra que estas sondas también son capaces de agrupar correctamente los distintos tejidos, aunque de una forma no tan fuerte como en el caso anterior. Las **figuras 10c** y **10d** muestran un resultado similar utilizando solo las sondas ubicadas en intrones y ubicadas en la hebra complementaria de genes respectivamente. A pesar de que las sondas intrónicas y las sondas complementarias de genes utilizadas en este test no han podido asignarse a ninguna entidad transcripcional conocida, muestran una regulación específica entre diferentes tejidos, con lo que puede inferirse que estas sondas están detectando señales biológicas y que realmente esas regiones del genoma tienen una función aún por determinar.

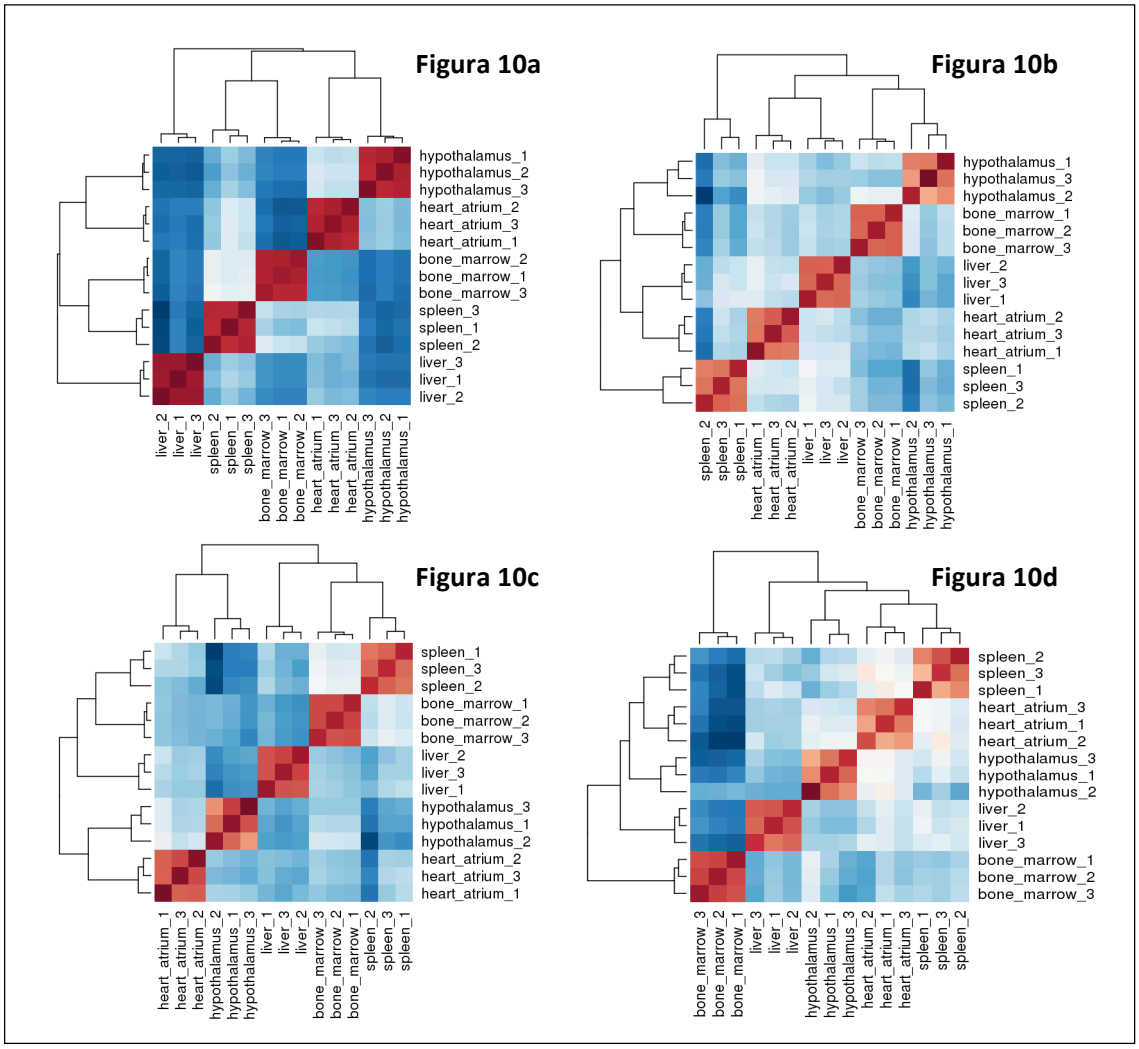


Figura 1.10. Conjunto de datos de 15 microarrays conteniendo 3 réplicas de 5 tejidos distintos: hipotálamo, corazón, médula ósea, hígado y bazo. Este conjunto de microarrays fue analizado con un método de agrupamiento jerárquico. La figura **(a)** muestra los microarrays normalizados con el mapeo a genes anotados de *Ensembl*; la figura **(b)** muestra los datos normalizados considerando únicamente la expresión de RNA no codificante (ncRNAs) derivada de *RNAdb*; la figura **(c)** muestra los datos únicamente considerando la expresión de las sondas situadas en intrones; y finalmente la figura **(d)** muestra el agrupamiento jerárquico considerando solamente las sondas situadas en la hebra complementaria a genes. A pesar de que el mapeo a genes segrega los diferentes tejidos de una forma más clara, en todos los casos la matriz de expresión resultante es capaz de agrupar correctamente las 3 réplicas.

1.3.6. Variación de la información en las diferentes actualizaciones de *GATExplorer*

Durante el periodo 2008 a 2010 la base de datos de ***GATExplorer*** fue actualizada tres veces (ver **tabla 1.4**). Con estos datos históricos se puede demostrar que la definición del genoma y transcriptoma en metazoos sufre cambios sustanciales conforme se van logrando nuevas evidencias biológicas. La **figura 1.11** muestra el número de genes y transcritos de humano, ratón y rata para las versiones de *Ensembl* v50, v53 y v57 respectivamente. Estos datos muestran una gran variación de una versión a otra, siendo el genoma humano el que más ha cambiado seguido de ratón y de rata. Además en el caso de humano y ratón el número de transcritos ha aumentado considerablemente en la versión 57 respecto a la 53, mientras que

por contraste el número de genes se ha reducido. Con esta reducción en el número de genes, el organismo humano cuenta con un número inferior al de ratón y rata. Esto puede parecer paradójico ya el organismo humano es el más complejo, y por lo tanto debería mostrar mayor complejidad a nivel genómico contando con más locus génicos. Sin embargo, el significativo aumento en el número de transcritos explica este aumento de complejidad y la reducción en el número de genes ya que muchos de ellos han podido ser solapados con el descubrimiento de nuevos transcritos. Estos datos muestran lo variante que es el genoma y transcriptoma consenso de las distintas especies, y también sugiere que ratón y humano son especies que se encuentran en un mismo nivel de conocimiento. Por otro lado, siendo ratón y rata organismos cercanos evolutivamente, muestran una gran disparidad en número de genes y transcritos, sugiriendo que el ratón es un organismo más estudiado mostrando una tendencia similar al humano.

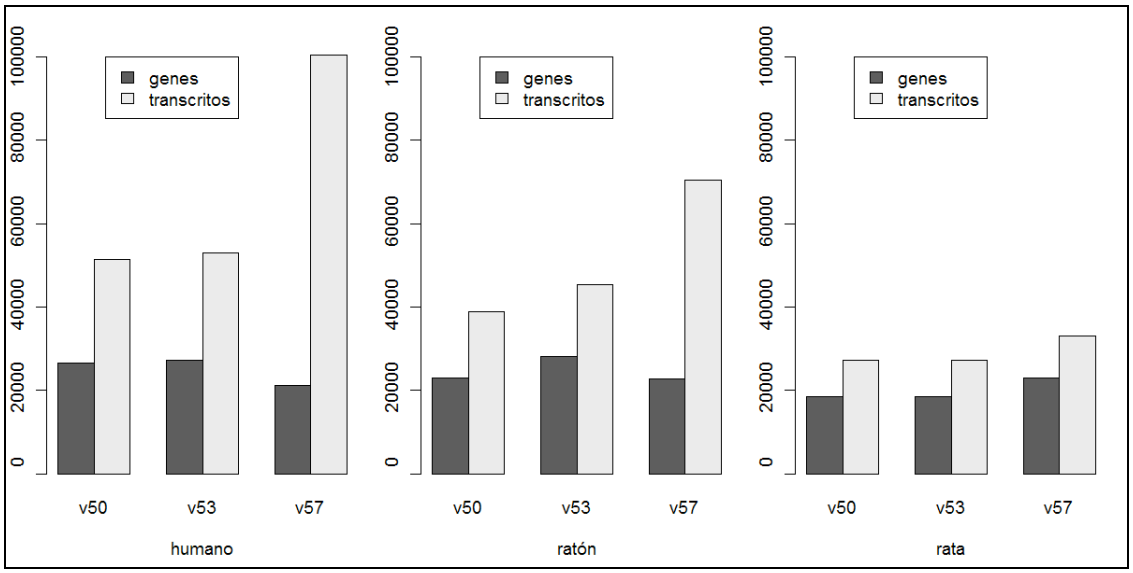


Figura 1.11. Evolución del número de genes y transcritos para humano, ratón y rata a lo largo de 3 versiones de los genomas en *Ensembl*.

Otro ejemplo de este cambio en la definición del genoma lo ilustra la **figura 1.12** en donde se muestran los genes definidos en tres años diferentes (2005, 2008, 2010) localizados en una región concreta de la banda q13.41 del cromosoma 19. En 2005 se identificaban 3 genes: ZNF528, Q96LN7_HUMAN y ZNF578. Tres años después, estos genes fueron fusionados en uno solo al que se le mantuvo el nombre ZNF528. Dos años después, en 2010, ese gen se convirtió de nuevo en 3, re-apareciendo ZNF578 y apareciendo un gen nuevo llamado ZNF534. Sin embargo, la longitud y estructura de estos genes son diferentes de la versión de 2005. Todo esto demuestra la necesidad de actualizar la interpretación de los análisis de microarrays a medida que la información biológica disponible se va corrigiendo y aumentando.

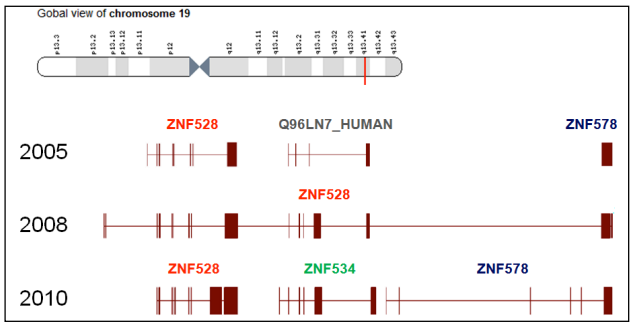


Figura 1.12. Ejemplo de las variaciones que sufre la definición de los genes en metazoos: una misma región del cromosoma 19 humano en un periodo de 5 años evidencian cambios significativos

1.3.7. Herramientas para visualización y exploración de datos incluidas en *GATExplorer*

Con el propósito de ofrecer acceso y uso de los resultados del re-mapeo de sondas de microarrays así como su visualización y posibilidad de exploración en un contexto genómico y transcriptómico, se desarrolló la plataforma bioinformática *web* interactiva accesible vía internet denominada ***GATExplorer*** (*Genomic* and *Transcriptomic Explorer*), ya citada en el **apartado 1.1.4** y descrita en parte en el **apartado 1.2** de Métodos. Dicha *web* integra la base de datos del mapeo con un navegador genómico y con datos de expresión. Esta aplicación se encuentra accesible en la dirección URL: **http://bioinfow.dep.usal.es/xgate**. La herramienta permite la búsqueda de un gen mediante cuatro tipos de acceso: **(i)** acceso por palabra clave, **(ii)** acceso por *probeset*, **(iii)** acceso por secuencia y **(iv)** acceso por coordenadas genómicas. En la **figura 1.13** se muestra, con el ejemplo del gen STAT5A, una página de ***GATExplorer***.

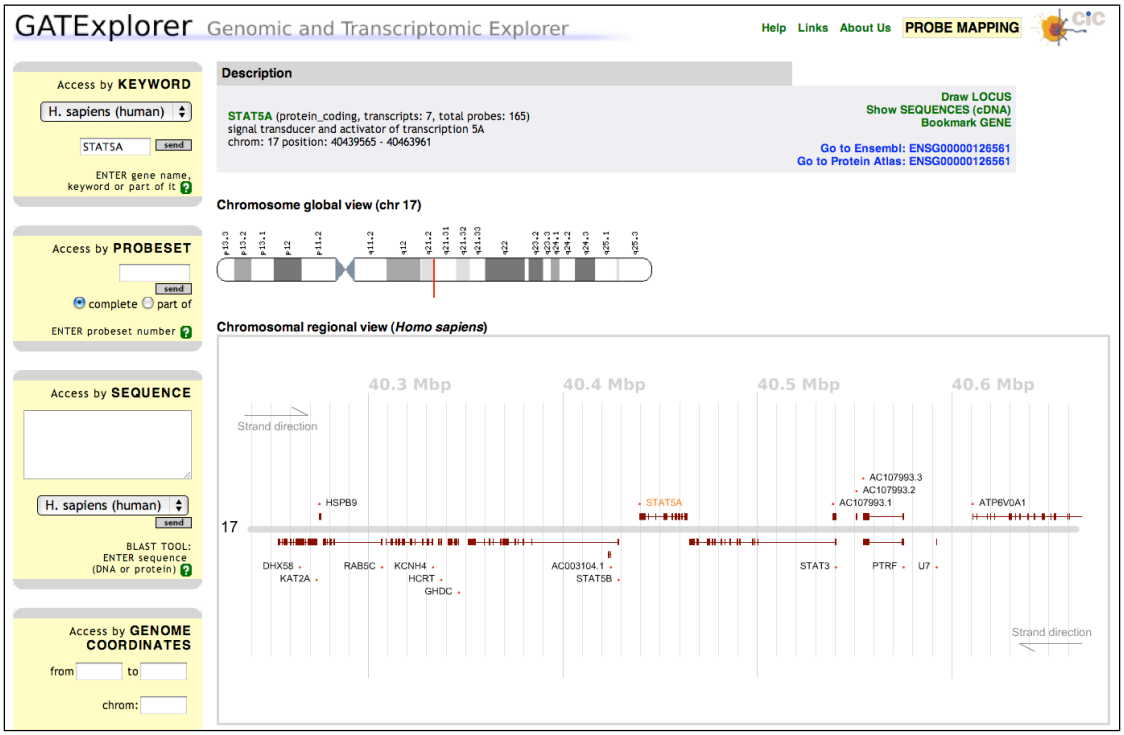


Figura 1.13. Página de la aplicación *web* ***GATExplorer*** implementada para visualizar y explorar datos genómicos y transcriptómicos y para descargar los resultados de re-mapeo de sondas de expresión de microarrays. Pantalla mostrando la ubicación cromosómica global y regional del gen humano STAT5A.

El acceso por palabra clave en la aplicación *web* permite buscar un gen por su nombre (p.ej. STAT5A) o por su descripción (p. ej. *Signal transducer and activator of transcription 5A*) (**figura 1.13**). El acceso en la aplicación por *probeset* permite localizar un gen a partir del correspondiente identificador de *Affymetrix* que hibrida con él (p. ej. 203010_at). El acceso por secuencia lanza un alineamiento a través del algoritmo BLAST sobre el genoma del organismo seleccionado. En este acceso se permite tanto secuencias de nucleótidos como de aminoácidos lanzando automáticamente el algoritmo correspondiente: *blastn* o *tblastn*. Este alineamiento se hace en tiempo real sobre ficheros de DNA genómico permitiendo, a diferencia del alineamiento realizado durante el alineamiento de las sondas, ubicar secuencias en intrones o entre *loci* génicos. Finalmente el acceso por coordenadas permite al usuario especificar una región concreta de un cromosoma con las posiciones de inicio y fin, que posteriormente se mostrarán en el navegador genómico. Una vez seleccionado el gen de interés la aplicación mostrará una serie de visores jerárquicos con sus correspondientes entidades genómicas

(cromosoma, *locus*, exones, transcritos y dominio de proteínas) con la correspondiente información del mapeo de sondas (ver **figura 1.13**). Antes de estos visores, en primer lugar, se muestra una caja con el nombre y descripción del gen, que proporciona también enlaces específicos para dicho gen a las bases de datos biomoleculares *Ensembl* y *Protein Atlas* (**http://www.proteinatlas.org/**) **(Uhlen, 2005; Uhlen et al., 2010)**, así como la posibilidad de guardar la búsqueda actual mediante un botón para marcar el gen (*Bookmark GENE*), que da la posibilidad de volver nuevamente al gen seleccionado en cualquier momento de la navegación. También muestra un enlace a otra ventana donde se muestran todas las secuencias (*Show SEQUENCES*) de nucleótidos para los transcritos (cDNAs) del gen, así como las correspondientes secuencias de aminoácidos de las proteínas (para todas las isoformas del gen). En el siguiente apartado, *Chromosomal global view*, se muestra una imagen del cromosoma correspondiente a ese gen indicando su ubicación exacta en el cromosoma a través de una línea roja que lo marca. En la sección *Chromosomal regional view* se muestra el gen seleccionado (en el ejemplo STAT5A) en su contexto genómico mostrando también los genes vecinos. En el siguiente visor que presenta la aplicación se muestra el locus y su estructura de transcritos conocidos (ver **figura 1.14**).

Figura 1.14. Ventana de ***GATExplorer*** que muestra la estructura de transcritos del *locus* del gen STAT5A.

En la parte superior existe un menú con los nombres distintos modelos de microarrays de expresión de *Affymetrix* que, una vez activados, mostrarán sus sondas en los lugares del locus donde estén ubicadas. También existe la posibilidad de hacer *zoom* y moverse a lo largo del *locus* identificando cada exón o sonda posicionando el cursor sobre cada uno de los elementos. De esta manera el investigador puede saber en qué zona del gen están ubicadas

cada una de las sondas, haciendo particularmente interesante la exploración de eventos de *splicing* alternativo cuando se trabaja con microarrays de exones.

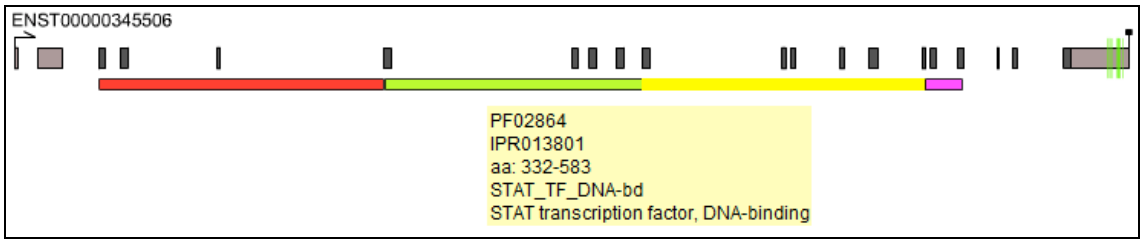


Figura 1.15. Integración de dominios de proteína en coordinadas genómicas en ***GATExplorer***.

También se han incorporado unos botones que permiten la visualización de los dominios de proteínas correspondientes a ese gen, obtenidos de las bases de datos *Pfam* y *Prosite* (**http://pfam.sanger.ac.uk/**; **http://prosite.expasy.org/**) **(Punta et al., 2012; Sigrist et al., 2010).** Estos dominios aparecerán alineados con el locus génico asociando cada exón con su dominio correspondiente (ver **figura 1.15**). Finalmente, el último gráfico de ***GATExplorer*** muestra un perfil de expresión del gen seleccionado medida en un *set* de 67 tejidos en el caso de humanos, en *sets* de 43 y 29 tejidos para los organismos ratón y rata, respectivamente, procedentes del set de datos *GeneAtlas* **(Su et al., 2004)**. Este visor (**figura 1.16**) promedia dos microarrays para cada tejido, distinguiendo además la expresión individual de cada una de las sondas para permitir conocer la aportación de cada una de ellas a la expresión global del gen.

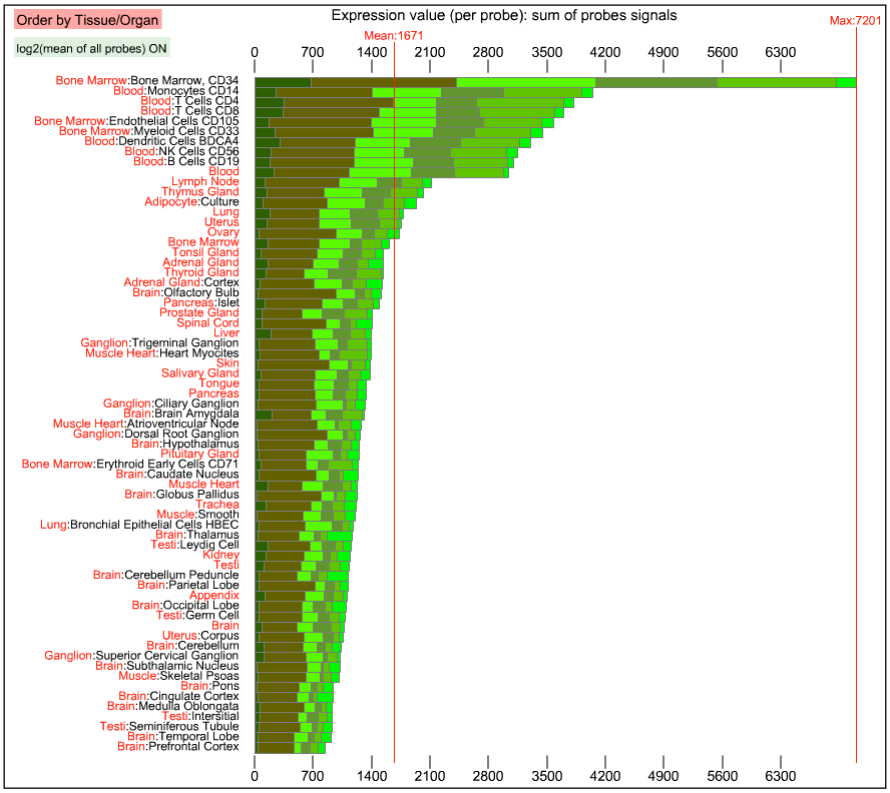


Figura 1.16. Visor de expresión de ***GATExplorer*** mostrando el perfil del gen STAT5A en 67 tejidos humanos.

Existen también dentro del visor de expresión un botón (*Order by Tissue/Organ*) con el que el usuario puede cambiar el orden de los tejidos eligiendo entre orden alfabético o nivel de expresión, y otro botón (*log2(mean of all probes) ON*) con el que el usuario puede cambiar entre ver el promedio logarítmico de las sondas en una sola barra o ver todas ellas como barras independientes sumadas y mostrando el nivel de expresión crudo de cada una (ver

figura 1.16). Además de todo esto, se proporcionan distintas tablas con el listado completo de las sondas mapeadas en el gen desglosadas por modelo de microarray e identificador de *probeset* (ver **figura 1.17**). Cada una de las filas de las tablas contiene la información de una sola sonda, incluyendo su posición X e Y en la matriz del microarray, la secuencia del oligo y el contenido GC (%). Además se detalla el número de genes y de transcritos detectados por la sonda, proporcionando un link al listado de genes o transcritos si el número es mayor que uno, es decir, son ambiguos a nivel de gen o de transcrito. En base a estos números, se ha construido un código con los colores **verde**, **amarillo**, **rojo** y **negro**. El color **verde** indica sonda específica (número=1) a nivel de genes y transcritos, el color **amarillo** especificidad a nivel de genes y ambigüedad (número >1) a nivel de transcritos, el color **rojo** ambigüedad a nivel de genes y por lo tanto de transcritos y, finalmente, el color **negro** indica que la sonda se ubica únicamente en intrones, ya sea en uno o varios genes. En el caso de los arrays de exones el código de colores está en función de su ambigüedad a nivel de exones en lugar de genes y transcritos.

Probesets table: Affymetrix **Probesets** which map on **MEST** — Download PROBESETS (.txt)

Chip	Probesets
HG_U133A	202016_at
HG_U133A_2	202016_at
HG_U133_Plus_2	202016_at
HG_U95A	37749_at
HG_U95Av2	37749_at
HG_Focus	202016_at
U133_X3P	g4505154_3p_at
Human_Gene_1.0	8136248
Human_Exon_1.0	2764910, 3024026, 3024027, 3024028, 3024029, 3024030, 3024031, 3024032, 3024033, 3024034, 3024035, 3024036, 3024037, 3024038, 3024039, 3024040, 3024041, 3024042, 3024043, 3024044, 3024045, 3024046, 3024047, 3024048, 3024049, 3024050, 3024051, 3024052, 3024053, 3024054, 3024055, 3024056, 3024057, 3024058, 3024059, 3024060, 3024061, 3627223

Probesets table: Affymetrix **Probesets** which map on the **antisense** of **MEST** — Download PROBESETS (.txt)

Chip	Probesets
Human_Exon_1.0	3072607, 3072609, 3072625, 3072631

Probes table: Affymetrix **Probes** which map on **MEST** — Download PROBES (.txt)

Chip	Probeset	x,y position	Probe Sequence (% GC content)	Map to distinct Locus	Map to distinct Transcript
HG_U133A	202016_at	(669,131)	AACAGTCTGAGACTCCTCATACCTC (48% GC)	1	4
		(348,37)	ATGTACACTGATTCCTTTATGATGA (32% GC)	1	4
...					
Human_Gene_1.0	8136248	(887,83)	AAGAAGTTCATCAGTCGTGTGAGGA (44% GC)	1	4
		(702,436)	ACTAAGTTCCCGTCATTGTTGCGGA (48% GC)	1	3
		(881,137)	ATTGTTGCGGATCCCTGCCCACATG (56% GC)	1	3
...		(21,992)	CAAGCTAGTAAACAGATTCATGGGA (40% GC)	1	1

Figura 17. Tabla de sondas proporcionada en ***GATExplorer*** indicando donde mapea cada una de las sondas de oligonucleótidos de cada microarray así como su posición en el array, su secuencia y su contenido en GC

Para que esta información sea útil a la hora de combinarla con otros datos del investigador, ***GATExplorer*** ofrece la posibilidad de descargar toda la información presente en la aplicación mediante ficheros de texto: el listado de las sondas, las sondas específicas a nivel de transcrito y a nivel de exón o los valores de expresión para el *set* de datos de *GeneAtlas*. ***GATExplorer*** también facilita los mapeos completos (todos los genes) para cada microarray a través del menú *PROBE MAPPING*. Este link abrirá un pequeño portal con varias opciones en un menú lateral ofreciendo los mapeos en dos formatos, ficheros de texto o paquetes de R. Además ofrece unos ficheros de anotación para asociar cada identificador de *Ensembl* con toda la información del gen (nombre, descripción, localización en el genoma, etc). También incluye una sección de estadísticas, una comparativa con otras aplicaciones similares y la información detallada de las versiones de las bases de datos utilizadas.

1.3.8. Análisis comparativo de *GATExplorer* con el mapeo original de *Affymetrix* y con otras plataformas de mapeo alternativo

Hay otros estudios bioinformáticos que han abordado la realización de un mapeo alternativo para los microarrays de *Affymetrix* (ver **tabla 1.9**). Algunas de estas aproximaciones están limitadas a sólo algunos de los modelos de microarrays no incluyendo la mayoría los modelos nuevos tipo WT. Además ninguno de ellos aplica al re-mapeo a regiones intrónicas o transcritos ncRNA para extender la cobertura de los microarrays más allá del mRNA. Tampoco ninguno de estos métodos ofrece un servidor *web* donde las sondas puedan ser encontradas de forma interactiva integrando los datos del mapeo con su ubicación en el genoma. En esta sección hemos comparado el trabajo realizado en ***GATExplorer*** con otras 4 publicaciones: **(Gautier et al., 2004)**, **(Harbig et al., 2005)**, **(Dai et al., 2005)** y **(Liu et al., 2007)**.

METHOD / TOOL	Gautier *et al.*	Harbig *et al.*	Dai *et al.* (BrainArray)	Liu *et al.* (AffyProbeMiner)	Risueño *et al.* (GATExplorer)
YEAR	2004	2005	2005 (updated 2010)	2007	2009
SOURCE DATABASES	RefSeq	RefSeq	UniGene RefSeq DoTS Entrez gene Ensembl	RefSeq GenBank	Ensembl
ORGANISMS	human	human	human mouse rat others	human mouse rat others	human mouse rat
MICROARRAY MODELS	-Expression Arrays •HG_U133A •HG_U95Av2	-Expression Arrays •HG_U133_Plus_2	-Expression Arrays -Exon Arrays -Tiling Arrays -Promoter Arrays	-Expression Arrays •U95 serie •U133 serie	-Expression Arrays -Exon Arrays
Minimal N° of probes in a probeset	1	NA	3	5	3
Type of mapping data provided	R objects	Text Files Normalization Tool	R Packages (CDFs)	R Packages (CDFs)	R Packages (CDFs) Text files (.txt)
Biomolecular entities mapped	genes	genes	genes transcripts exons	genes transcripts	genes transcripts exons ncRNAs
N° of human genes (mapped with unambiguous probes)	11640 [HG_U133A]	(20415) (only done for U133_Plus_2)	11853 [HG_U133A]	12550 [HG_U133A]	12576 [HG_U133A]
Mapping to ncRNA	NO	NO	NO	NO	YES
Web page with data	YES	YES	YES	YES	YES
Integration of the mapping in a GENOMIC WEB SERVER	NO	NO	NO	NO	YES

Tabla 1.9. Análisis comparativo de ***GATExplorer*** con otras publicaciones relacionadas.

La **tabla 1.9** muestra una comparativa de estas publicaciones comparándola con ***GATExplorer*** a través de una serie de características que hemos considerado interesantes tanto por la información que aportan al investigador, como la facilidad de uso de la misma. El primer trabajo publicado de re-mapeo de sondas de los microarrays de *Affymetrix* de acuerdo con la versión actualizada de los genes humanos fue llevado a cabo por Gautier y colaboradores **(Gautier et al., 2004)**. A partir de este trabajo, fueron varios los autores que proporcionaron su propio re-mapeo incluyendo diferentes maneras de ofrecer los resultados al investigador. En el año 2005 el grupo de Dai **(Dai et al., 2005)** desarrolló el que probablemente sea el trabajo más

exhaustivo en este ámbito, proporcionando una redefinición de los arrays de *Affymetrix* para varias especies. Nuestro grupo de investigación del Centro del Cáncer inició los trabajos de re-mapeo de sondas de microarrays también en torno a 2005, aunque nos propusimos profundizar más en el problema del análisis de la expresión génica global estudiando a fondo los algoritmos de análisis y abordando de modo más integrado el sentido y las ventajas que tiene re-mapear las sondas de los microarrays. Por ello, emprendimos un trabajo ambicioso que incluye no sólo la re-anotación de sondas y el re-mapeo a genes conocidos, sino también la búsqueda de otras entidades transcriptómicas a las que los miles de sondas no asignadas a genes codificantes se pudieran reasignar. Por ello, se realizó el re-mapeo a la base de datos de ncRNAs (ver **figura 1.8**) que es genuino de nuestro estudio (como se indica también en la **tabla 1.9**). Finalmente, ***GATExplorer*** también incluye la herramienta de visualización de sondas integrada en una plataforma *web*, navegador genómico transcriptómico, que no se ha desarrollado en ninguno de los otros trabajos publicados antes y que proporciona una gran ayuda para la mejor comprensión de resultados de estudios de expresión.

Tras el desarrollo del algoritmo de remapeo incluido en ***GATExplorer*** realizamos un análisis comparativo con el mapeo original de *Affymetrix* combinando distintos algoritmos de normalización, utilizando una colección de microarrays que compara la expresión global de ratones *knock-out* (a los cuales les falta un gen concreto) con la expresión en los ratones normales (*wild-type*). Este análisis busca demostrar con un ejemplo práctico la eficiencia de aplicar un re-mapeo actualizado en base a conocimiento actual de los genes para cualquier análisis de expresión dado.

En la **tabla 1.10** se presentan los resultados del estudio comparativo para identificar expresión diferencial en varios *sets* de microarrays para los que se utilizan tanto la anotación original de *Affymetrix* (*probesets*) como con el remapeo nuevo de ***GATExplorer***. Los análisis se han hecho usando 3 algoritmos de normalización: **MAS5.0 (Hubbell et al., 2002; Liu et al., 2002)**, **FARMS (Hochreiter et al., 2006)** y **RMA (Irizarry et al., 2003a)**. Los tres algoritmos se han probado utilizando la anotación original de *Affymetrix*; y además el algoritmo **RMA** es el que se ha utilizado para obtener la señal de expresión con el re-mapeo de ***GATExplorer***. El uso de ambos mapeos se realiza mediante *Chip Definition Files* (CDFs) utilizando R y los paquetes de *BioConductor* **(Gentleman et al., 2004; Smyth, 2005)**. Como algoritmo de detección de expresión diferencial utilizamos **SAM (Tusher et al., 2001)**. Las muestras comparadas son una colección de 5 series de microarrays de 6 muestras cada una. Cada *set* incluye 3 réplicas biológicas de ratones *knock-out* (KO) para un gen específico que es comparado con las 3 réplicas biológicas correspondientes con el *wild-type* (WT). Estos 5 genes son: **APOE-/-**, **IRS2-/-**, **NRAS-/-**, **SCD1-/-** y **ENG+/-** (ver **tabla 1.10**). El objetivo de estas comparaciones es identificar la posición del gen no presente (i.e. eliminado por *knock-out*) en el *ranking* de genes cuya expresión está alterada entre ambos grupos de ratones. En condiciones ideales, el gen que no está presente en los ratones KO debería ser el que experimentase la mayor diferencia, mostrando la infra-expresión máxima comparada con los ratones WT. En esta comparación se trata de medir cómo afectan los métodos y el re-mapeo al resultado de la expresión diferencial y qué análisis sitúa al gen KO en una posición más alta en el *ranking*. Los resultados indican que el uso del mapeo de ***GATExplorer*** funciona en general mejor que el uso de *probesets* originales de *Affymetrix*, ya que en todos los casos el gen KO se encuentra entre los primeros de la lista de genes infra-expresados. Cada uno de los genes testados tiene funciones diferentes, están implicados en distintos mecanismos moleculares y tienen distintos genes asociados que pueden verse afectados por el KO. Por lo tanto, no se puede asumir que el gen KO debe ser siempre identificado como el más reprimido, pero sí debe ser uno de los más reprimidos y, sin duda, debe ser estadísticamente significativo. No obstante, podemos encontrar diferencias significativas entre los 4 métodos utilizados, diferencias que no pueden deberse a la biología – que es idéntica por usarse las mismas muestras en la comparación– sino a los diferentes

algoritmos para calcular la expresión de los genes y al tipo de mapeo que se utiliza. Como ya se ha indicado, en la primera parte del estudio se aplicaron los algoritmos **MAS5.0**, **FARMS** y **RMA** utilizando en todos el mapeo a *probesets*. Los resultados de esta parte nos confirman lo observado por otros autores **(Barash et al., 2004; Bolstad et al., 2003)**, que identifican **RMA** como un método de normalización y obtención de señal que resulta muy eficaz para la posterior identificación correcta de genes diferencialmente expresados. Por ello, en nuestro estudio concreto sobre le re-mapeo utilizamos el algoritmo **RMA** en las dos condiciones de contraste: mapeo de *Affymetrix* a *probesets* frente a mapeo nuevo de ***GATExplorer*** a genes.

		MAS5 with CDF to *Affymetrix* probesets	FARMS with CDF to *Affymetrix* probesets	RMA with CDF to *Affymetrix* probesets	RMA with CDF to genes (using *GeneMapper*)
APOE -/-	**Entity** (1)	apolipoprotein e		1432466_a_at [1]	ENSMUSG00000002985
3 WT vs 3 KO	Rank (in DOWN) (2)	38	17	2	17
mouse4302	Rank (in ALL) (3)	250	72	22	47
45101 gp	p-value (4)	0.00128600	0.00085227	0.00005321	0.00043362
	d-value (5)	-4.76	-9.21	-8.54	-9.31
	nº gn loci q-value<0.10 (6)	2	208	1350	1564
	nº gn loci total (7)	24100	24100	24100	16835
	% (nº glc sig / nº glc total) (8)	0.01	0.86	5.60	9.29
IRS2 -/-	**Entity** (1)	insulin receptor substrate 2		1443969_at [1]	ENSMUSG00000038894
3 WT vs 3 KO	Rank (in DOWN) (2)	45	82	10	4
moe430a	Rank (in ALL) (3)	99	137	19	8
22690 gp	p-value (4)	0.00374174	0.01178727	0.00011018	0.00009661
	d-value (5)	-3.41	-4.40	-3.82	-3.94
	nº gn loci q-value<0.10 (6)	2	0	12	2
	nº gn loci total (7)	13702	13702	13702	12421
	% (nº glc sig / nº glc total) (8)	0.01	0.00	0.09	0.02
NRAS -/-	**Entity** (1)	neuroblastoma ras oncogene		94362_at [1] & 160925_at [2]	ENSMUSG00000027852 *
3 WT vs 3 KO	Rank (in DOWN) (2)	1 & >200	1 & 22	1 & 17	1
mgu74av2	Rank (in ALL) (3)	4 & >200	6 & 48	1 & 53	2
12488 gp	p-value (4)	0.00006406	0.00016576	0.00000801	0.00002552
	d-value (5)	-6.97	-15.53	-15.48	-10.83
	nº gn loci q-value<0.10 (6)	2	0	10	22
	nº gn loci total (7)	9557	9557	9557	7837
	% (nº glc sig / nº glc total) (8)	0.02	0.00	0.10	0.28
SCD1 -/-	**Entity** (1)	stearoyl-Coenzyme A desaturase 1		94056_at [1] & 94057_g_at [2]	ENSMUSG00000037071
3 WT vs 3 KO	Rank (in DOWN) (2)	2 & 1	18 & 1	2 & 1	1
mgu74a	Rank (in ALL) (3)	2 & 1	22 & 1	2 & 1	1
12654 gp	p-value (4)	0.00000790	0.00001486	0.00000791	0.00001407
	d-value (5)	-15.61	-33.45	-16.60	-11.32
	nº gn loci q-value<0.10 (6)	2	2414	2589	2049
	nº gn loci total (7)	9662	9662	9662	7122
	% (nº glc sig / nº glc total) (8)	0.02	24.98	26.80	28.77
ENG +/-	**Entity** (1)	endoglin		1417271_a_at [1]	ENSMUSG00000026814
3 WT vs 3 KO	Rank (in DOWN) (2)	32	28	2	2
moe430a	Rank (in ALL) (3)	40	32	2	3
22690 gp	p-value (4)	0.00174967	0.00694714	0.00004847	0.00004025
	d-value (5)	-5.34	-4.74	-3.21	-3.62
	nº gn loci q-value<0.10 (6)	0	0	1	0
	nº gn loci total (7)	13702	13702	13702	12421
	% (nº glc sig / nº glc total) (8)	0.00	0.00	0.01	0.00

Entity (1)	name of the KO gene; probesets for this gene in the microarray; ENSEMBL gene ID
Rank (in DOWN) (2)	rank of the KO gene in the list of DOWN-regulated significant genes ordered by p-value
Rank (in ALL) (3)	rank of the KO gene in the list of ALL significant genes ordered by p-value
p-value (4)	p-value of the KO gene in the analysis with SAM
d-value (5)	d-value of the KO gene in the analysis with SAM
nº gn loci q-value<0.10 (6)	number of significant gene loci with a q-value lower than 0.10 in the analysis with SAM
nº gn loci total (7)	total number of mouse gene loci assigned within the microarray
% (nº glc sig / nº glc total) (8)	percentage of significant gene loci with respect to the total
*	this gene (NRAS) correspond to ENSEMBL v47, all the rest to v53

Tabla 1.10. Comparación entre distintas combinaciones de métodos de obtención de señal y mapeos aplicados a 5 *sets* de datos de genes *knock-out*: APOE-/-, IRS2-/-, NRAS-/-, SCD1-/-, ENG+/- . En la tabla se indican los siguientes parámetros de cada uno de estos genes KO: **(1)** nombre completo, **(2)** posición en el *ranking* de genes infra-expresados, **(3)** posición en el *ranking* de todos los genes alterados, **(4)** p-valor de SAM, **(5)** d-valor de SAM, **(6)** número de genes con un valor de significación q-valor <0.1, **(7)** número total de genes/*probesets* asignados en el microarray, **(8)** porcentaje de genes significativos respecto del total. Los mejores resultados para cada parámetro se resaltan en amarillo. Las últimas 2 columnas bordeadas por un marco negro muestran la comparación métodos idénticos para el cálculo de la expresión (RMA) y para la expresión diferencial (SAM) cambiando únicamente los CDFs, es decir el tipo de mapeo.

Finalmente, respecto a la visualización de sondas en contexto genómico, existen navegadores genómicos como el desarrollado por UCSC (**http://genome.ucsc.edu**) o el bien conocido navegador de *Ensembl* (**http://www.ensembl.org**) que incluyen en sus bases de datos gran cantidad de información, también la ubicación de los *probesets* de microarrays. Sin embargo,

su presentación y detalle no es igual a ***GATExplorer*** y resulta muchas veces difícil de localizar y de analizar. Otra aplicación *web* llamada *X:map* **(Yates et al., 2008)** proporciona anotación y visualización de los *probesets* y sondas del microarray de exones de *Affymetrix*. También existen paquetes en *BioConductor* como *GenomeGraphs* **(Durinck et al., 2009)** que realizan gráficos de regiones cromosómicas consultando la base de datos de *Ensembl*. Incluso *Affymetrix* realizó su propio programa de visualización y exploración de genes llamado *Integrated Genome Browser* (IGB). Aunque todas estas aplicaciones son útiles, tienen unos propósitos diferentes del de ***GATExplorer*** que mantiene una coherente integración de la información sobre genes, transcritos y exones con la información sobre el re-mapeo de sondas y la visualización de datos de expresión. Además como plataforma bioinformática interactiva también permite el uso por parte de los investigadores de herramientas BLAST para localización de secuencias concretas de oligos o de genes en dicho contexto genómico.

APOE	GEO GSE2372 - Plataforma GPL1261 – Identificadores de array: GSM44658, GSM44663, GSM44659, GSM44660, GSM44661, GSM44662.
NRAS	GEO GSE14829 - Plataforma GPL81 – Identificadores de array: GSM371168, GSM371169, GSM371170, GSM371174, GSM371175, GSM371176.
SCD1	GEO GSE2926 - Plataforma GPL32 – Identificadores de array: GSM63851, GSM63852, GSM63853, GSM63856, GSM63857, GSM63858.
IRS2	Arrays no publicados hasta la fecha (FONT DE MORA J. et al. 2010).
ENG	Arrays no publicados hasta la fecha (RODRIGUEZ-BARBERO A. et al. 2010).

Tabla 1.11. *Set* de datos de microarrays de ratones *knock-out* utilizados en la comparativa de la **tabla 1.10**.

1.3.9. Paquetes de R y ficheros de texto proporcionados en *GATExplorer*

Con el objetivo de incorporar el re-mapeo desarrollado en el presente trabajo al análisis de cualquier *set* de datos de microarrays de expresión de *Affymetrix,* ***GATExplorer*** recoge toda la información de dicho re-mapeo en ficheros descargables con dos formatos distintos: ficheros de texto plano y paquetes de R. En estos ficheros está presente la información de los *probesets* y sondas "no ambiguos" a nivel de genes, de transcritos y de exones. Es decir, los oligos que sólo mapean en una única entidad transcripcional y, por lo tanto, no presentan hibridación cruzada; siendo los que deben usarse a la hora de calcular la expresión de los genes, transcritos y exones.

Por cada modelo de microarray de expresión de *Affymetrix* existen en ***GATExplorer*** 4 tipos distintos de ficheros de texto:

- ***probesets2genes***: mapeo de *probesets* no ambiguos a nivel de genes.
- ***probes2genes***: mapeo de sondas no ambiguas a nivel genes.
- ***ambigprobes2genes***: mapeo de sondas ambiguas a nivel genes.
- ***probes2transcripts***: mapeo de sondas no ambiguas a nivel de transcritos.

Para los microarrays de exones existe además el fichero:

- ***probes2genesplustranscripts***: mapeo de sondas no ambiguas a nivel de genes, mapeando en 1 o más transcritos del mismo gen.

Estos ficheros de texto plano son genéricos y pueden ser manejados por cualquier tipo de aplicación, sin embargo, para aumentar la usabilidad del re-mapeo de ***GATExplorer***, se incluyó también la construcción de paquetes CDFs para utilizar con R.

R es un lenguaje de programación estadístico de software libre que ha alcanzado gran popularidad en los últimos años. Gracias a su sencillez y potencia su uso, R se ha convertido en una de las principales herramientas de cálculo y análisis estadístico para muchos investigadores. A medida que su uso se ha ido incrementando y al ser de código libre, la creciente comunidad de usuarios de R contribuye a su vez en la creación de librerías o paquetes aumentando su potencial. Cabe destacar especialmente el proyecto *Bioconductor* (**http://www.bioconductor.org/**) que ofrece multitud de paquetes gratuitos orientados al manejo de datos biomoleculares y cuyo uso está generalizado en el campo de la biología computacional.

Los paquetes de R disponibles en ***GATExplorer*** para utilizar el re-mapeo con cada modelo de microarray de expresión génica son:

- ***GeneMapper***: grupo de sondas no ambiguas que mapean en un gen específico.
- ***TranscriptMapper***: grupo de sondas no ambiguas que mapean en un transcrito específico.
- ***ncRNA-Mapper***: grupo de sondas no ambiguas que mapean en un ncRNA y que no mapean en ningún mRNA codificante.

Para los microarrays de exones existe además el paquete específico:

- ***ExonMapper***: grupo de sondas no ambiguas que mapean en un exón específico.

R puede descargarse gratuitamente desde la página oficial (**http://www.r-project.org/**). Todos los paquetes de R y ficheros de texto citados en este apartado pueden bajarse libremente de la web de ***GATExplorer*** en su sección *PROBE MAPPING*.

1.4. Discusión y posible trabajo futuro

En los últimos años –i.e. primera década de este siglo–, el desarrollo de las técnicas de alto rendimiento, capaces de estudiar el transcriptoma completo de un determinado tipo celular, ha crecido exponencialmente gracias al exitoso uso de la tecnología de microarrays de expresión. El presente trabajo tiene como propósito mejorar la precisión y cobertura de los análisis transcriptómicos derivados de microarrays de expresión y también ayudar a una mejor comprensión por parte de los investigadores de los resultados de este tipo de técnicas "ómicas". Para ello, el trabajo se ha dividido en dos partes bien diferenciadas.

La **primera** es la construcción de una base de datos conteniendo un mapeo alternativo de las sondas de los microarrays de expresión más usados en estudios biomédicos (manufacturados por *Affymetrix*). El re-mapeo realizado aquí supone una mejora respecto al original de *Affymetrix*, ya que se basa en la re-definición de sus sondas en función al conocimiento biológico más actualizado. Desde el diseño original de estos microarrays, el escenario de la genómica y la transcriptómica se han ido revelando más y más complejos gracias a un espectacular aumento del conocimiento tanto de genes codificantes de proteína como de *locus* génicos no codificantes pero activos (ncRNAs, en su mayoría del tipo *long non-coding* RNA) **(Amaral et al., 2008; Carninci et al., 2005)**. Esto ha dejado patente la necesidad de adaptar los análisis de microarrays a esta nueva realidad biológica. ***GATExplorer*** realiza esta

adaptación para los microarrays de expresión de *Affymetrix* mediante el agrupamiento las sondas en base a las nuevas entidades transcripcionales conocidas, incorporando de forma novedosa en estos análisis la posibilidad de medir expresión de ncRNA. Una posterior eliminación de sondas que presentan hibridación cruzada, aseguran además, minimizar el ruido y mejorar la señal a la hora de calcular la expresión de los genes.

La **segunda** parte del trabajo es el desarrollo de una plataforma *web* interactiva que sirva como instrumento de visualización, información y descarga de archivos, ofreciendo toda la información del re-mapeo al investigador. Esta plataforma incluye un navegador genómico en el que se pueden buscar los genes de interés, integrando el mapeo de sondas tanto mediante su ubicación en la región del *locus* génico correspondiente, como mediante su listado en tablas. Además, la herramienta desarrollada ofrece toda la información lista para descargarse en distintos ficheros lo cual hace factible la incorporación de este conocimiento a los análisis de microarrays de expresión.

Como trabajo futuro se determina la necesidad de realizar actualizaciones periódicas de la base de datos de ***GATExplorer***, ya que el conocimiento de la biología molecular avanza deprisa y adaptar a esos cambios los análisis genómicos de expresión de escala global (*genome-wide*) es la base científica de este trabajo.

Otro ámbito interesante en el que podría seguir desarrollándose la investigación en el futuro, es el del análisis de expresión de ncRNAs. La expresión de estos tipos de genes no codificantes de proteína no han sido muy estudiados hasta la fecha debido al reciente descubrimiento de muchos de ellos. Los paquetes *ncRNA-Mapper*, desarrollados en ***GATExplorer***, permiten la medición de la expresión de algunos de ellos en gran cantidad de muestras almacenadas en repositorios públicos, lo cual podría ser utilizado para tratar de descubrir funcionalidad de algunos de estos transcritos en tejidos sanos y su papel en enfermedades como el cáncer.

Capítulo 2
Análisis de expresión diferencial de genes y microRNAs para la detección de biomarcadores en muestras de leucemia y mieloma múltiple

2.1. Introducción

El cáncer es una enfermedad causada por alteraciones genéticas en oncogenes, genes supresores de tumores y en otros productos de transcripción como microRNAs **(Croce, 2008; Weinberg, 2007)**. Estas alteraciones comienzan con alteraciones cromosómicas (trisomías, deleciones, o translocaciones) o mutaciones puntuales que afectan a la estructura y expresión de determinados genes **(Knudson, 2001)**. Estos genes alterados provocan una desregulación en el ciclo celular y muerte celular programada (apoptosis), confiriendo a la célula una actividad proliferativa descontrolada que termina por desplazar al tejido sano. El término cáncer engloba en realidad diversas enfermedades, que se pueden categorizar en función de su localización en el organismo, el tipo celular del que proceden, las alteraciones genómicas concretas que muestran y el estadio en el que se encuentren. Los trabajos actuales enfocados en la obtención de biomarcadores pretenden mejorar el conocimiento molecular sobre los mecanismos celulares específicos que causan o impulsan la transformación tumoral dentro de la enorme complejidad que supone el cáncer. A esta búsqueda están ayudando enormemente las modernas técnicas genómicas de alto rendimiento, como los microarrays o la secuenciación masiva, que permiten medir miles de características de una determinada muestra en un solo experimento proporcionando una gran cantidad de variables con las que trabajar. El hallazgo de patrones entre las comparaciones de tejido sano y distintos tipos de cáncer, permite refinar y encontrar nuevos subtipos de enfermedad que antes se consideraban una sola, proporcionando de esta manera un trato más personalizado al paciente. Esta estratificación de pacientes conlleva una mejora a nivel de diagnóstico, de prognosis y de tratamiento, así como un avance para el desarrollo y utilización de fármacos más específicos y eficaces.

La bioinformática es clave en los estudios genómicos en cáncer, como herramienta esencial para lograr análisis robustos de datos complejos. Existen numerosas técnicas computacionales de aprendizaje automático que permiten extraer información relevante de entre la cantidad masiva de datos proporcionada por experimentos de alto rendimiento. Estas técnicas se

pueden agrupar en: técnicas de aprendizaje no supervisado, técnicas de aprendizaje supervisado y técnicas de aprendizaje semi-supervisado.

2.1.1. Análisis de datos genómicos por técnicas de aprendizaje no supervisado

El aprendizaje no supervisado trata de proporcionar información basándose en los datos de múltiples muestras sin haber sido previamente categorizadas o etiquetadas por algún tipo de característica o tipología. Este aprendizaje se basa en encontrar similitudes y diferencias entre las distintas muestras, para lo cual el cálculo de las distancias o medida de disimilitud entre ellas es una tarea crítica. Este tipo de aprendizaje tiene distintas utilidades en el ámbito de la biomedicina como en la categorización automática de enfermedades agrupando pacientes con perfiles similares partiendo de una población supuestamente homogénea. También se ha utilizado con éxito para encontrar relaciones entre genes partiendo de sus perfiles de expresión genómica, revelando la existencia de grupos que se asocian con funciones biológicas diferentes. Entre las técnicas no supervisadas figuran los métodos de **agrupamiento jerárquico** que representan las distintas muestras en función de su cercanía conformando una estructura de árbol llamada dendrograma. Estos métodos son muy utilizados para representar las similitudes y diferencias existentes entre las distintas muestras de un estudio. También es común combinar dos dendrogramas formando un mapa bidimensional (llamado mapa de calor o *heatmap*) que, ayudado de un código de colores, permite una mejor identificación visual de los distintos grupos de muestras encontrados. Existen otros muchos métodos de particionamiento que dividen las muestras en distintos grupos, como el algoritmo **k-medias** **(MacQueen, 1967)** y otro tipo de métodos de agrupamiento difuso, en donde la clasificación de las distintas muestras no es excluyente entre los distintos grupos identificados **(Gath and Geva, 1989; Xie and Beni, 1991)**.

Debido a la alta dimensionalidad de muchos de los conjuntos de datos genómicos, existen métodos que tratan de reducir el número de variables con la menor pérdida de información posible. Entre estas técnicas se encuentra el llamado **análisis de componentes principales** (PCA), que transforma los datos creando unas nuevas variables llamadas componentes principales, calculada cada una a partir una aplicación lineal de las variables originales **(Jolliffe, 1986)**. Estas componentes principales están ordenadas en base a la importancia de la información que contienen y tienen el poder de representarla en una dimensionalidad muy inferior a la original. Esto supone una ventaja al plantear una estrategia de clasificación en donde muchos métodos tienen problemas al manejar los datos genómicos de alta dimensionalidad de los microarrays **(Pochet et al., 2004)**. Esta reducción de dimensionalidad también es muy utilizada para representar datos de forma gráfica **(Geng et al., 2005)**.

2.1.2. Análisis de datos genómicos por técnicas de aprendizaje supervisado y semi-supervisado.

Las técnicas de aprendizaje supervisado son aquellas en las que se utilizan etiquetas para marcar las distintas clases que componen los datos. En este tipo de aprendizaje, se trata de entrenar un sistema para obtener información que permita clasificar *a posteriori* las muestras de acuerdo a sus categorías. Distintas técnicas de este tipo de aprendizaje son las **máquinas de vector soporte** (SVM), **modelos de mixtura**, **redes neuronales** o el algoritmo de **k-vecinos** más cercanos **(Coomans and Massart, 1982; Cover and Hart, 1967)**. Todas estas técnicas se han

aplicado con éxito en el reconocimiento de patrones sobre datos de microarrays de expresión, pero probablemente el análisis más común es el llamado "**expresión diferencial**". Este tipo de test consiste en seleccionar los genes que presentan una expresión significativamente diferente, mayor o menor, entre dos categorías distintas previamente definidas. Existen distintas distintos algoritmos de expresión diferencial, uno de los más utilizados se basa en **modelos lineales** **(Smyth, 2005)** y está implementado en el paquete *limma* de R **(Smyth et al., 2012)**. Otro de los más citados en la literatura es **SAM** (*Significance Analysis of Microarrays*) **(Tusher et al., 2001)** implementado en el paquete *siggenes* **(Schwender, 2012)** también de R. Estos algoritmos realizan un test por cada gen de la matriz de expresión asignando un valor de *R-fold* y un valor de probabilidad p-valor, con una posterior corrección para test múltiples. Probablemente el método más popular en el ámbito de la bioinformática para la corrección del p-valor sea el método llamado *False Discovery Rate* (FDR) **(Benjamini et al., 2001)**. También son utilizados otros métodos como el de *Bonferroni*, el de *Holm* **(Holm, 1979)** o el de *Hochberg* **(Hochberg, 1988)**. Una vez elegido un punto de corte sobre el p-valor corregido (que normalmente se sitúa entre 0.01 y 0.05) se obtiene un grupo de genes que están cambiados de modo estadísticamente significativo y que por ello se supone tiene una regulación distinta entre las categorías comparadas.

El aprendizaje semi-supervisado consiste en una mezcla de métodos supervisados y no supervisados. Un ejemplo típico es la aplicación de un método de agrupamiento de variables no supervisado –por ejemplo, un agrupamiento o *clustering* jerárquico– a partir de una matriz de datos que incluye únicamente variables significativas que han sido previamente seleccionadas por un método de aprendizaje supervisado. Este tipo de aproximación permite reducir el tipo de variables a las únicamente significativas, basándose en los tipos o categorías de muestras que se conocen *a priori*, y consigue que el método de agrupamiento o clusterización no supervisado clasifique bien las muestras –es decir, los pacientes o individuos estudiados– y permita explorar con mayor precisión el agrupamiento de las variables –es decir, de los genes en el caso de datos de expresión–.

2.1.3. Análisis genómicos de dos tipos de hemopatías malignas: CLL, MM.

El siguiente trabajo aquí descrito se centra en el descubrimiento de biomarcadores en datos de expresión para diferentes subtipos de dos enfermedades hematológicas: **leucemia linfocítica crónica** (CLL) y **mieloma múltiple** (MM). Estas enfermedades serán categorizadas en función a sus diferentes alteraciones cromosómicas, cada una de las cuales tiene asociada un pronóstico distinto.

La CLL es el tipo más frecuente de leucemia en los países occidentales y se caracteriza por una expansión clonal de linfocitos B en la sangre, médula ósea, nódulos linfáticos y bazo **(Rozman and Montserrat, 1995)**. La deleción de brazo largo del cromosoma 13 (13q-) es una de las alteraciones más frecuentes en esta enfermedad y, en general, está considerada como una aberración de buen pronóstico **(Mehes, 2005)**. Estudios recientes, sin embargo, sugieren que el pronóstico puede variar en los pacientes con 13q- dependiendo del número de células que muestran esta anormalidad como única aberración **(Dal Bo et al., 2011; Hernandez et al., 2009)**. Los casos que presentan un alto porcentaje de células 13q- (13q-H) tienen una esperanza de vida media inferior que los casos con bajo porcentaje de 13q- (13q-L), que es muy similar a los casos que presentan un cariotipo normal. Por otro lado, pérdidas en otros cromosomas como 17p y 11q –i.e., del(17p) y/o del(11q)– que afectan a genes como TP53 y ATM también están relacionadas con mal pronóstico **(Catovsky et al., 2007; Krober et al., 2002)**. Uno de los propósitos de este trabajo es caracterizar los distintos subtipos de

enfermedad a través de un análisis de expresión génica global detectada con microarrays, buscando lograr una interpretación biológica más adecuada sobre las diferencias entre los subgrupos 13q- en CLLs.

El MM es una neoplasia de células plasmáticas, también denominadas plasmocitos, que son linfocitos B diferenciados que se originan en la médula ósea y pertenecen al sistema inmunitario, consistiendo su principal papel en la secreción de grandes cantidades de anticuerpos. El MM es una hemopatía maligna en la que se da una proliferación anormal de células plasmáticas que también se puede categorizar atendiendo a las diferentes alteraciones citogenéticas que presenta **(Bergsagel and Kuehl, 2005; Chng et al., 2007a)**. Algunos trabajos han tenido éxito en caracterizar molecularmente las principales categorías o clases de MM mostrando sus diferencias más importantes **(Chng et al., 2007b; Zhan et al., 2002; Zhan et al., 2006; Zhan et al., 2003)**. Sin embargo, a pesar de estos esfuerzos, existe una parte de la biología que se escapa a la señal de los genes codificantes clásicos que no ha sido explorada convenientemente en trabajos previos. Entre esas nuevas señales transcriptómicas destaca el papel de los miRNAs en la enfermedad y cómo se relacionan con los genes codificantes de proteína. Es conocida la actividad regulatoria que estos pequeños fragmentos de RNA no codificante tienen sobre ciertos genes diana, procediendo a silenciarlos mediante la unión específica y marcaje para la degradación de sus mRNAs **(Bushati and Cohen, 2007)**. Estos miRNAs desempeñan un papel importante en la regulación de los diferentes procesos biológicos en células sanas **(Alvarez-Garcia and Miska, 2005; Cheng et al., 2005)**, pero también se ha demostrado su papel en procesos tumorales y de cancerogénesis **(Croce, 2009; Osada and Takahashi, 2007)**. Además, se ha comprobado que los perfiles de expresión de miRNAs tienen capacidad para clasificar tumores **(Calin and Croce, 2006; Lu et al., 2005)**. En este trabajo, se realizaron los análisis de la señal de expresión de 365 miRNAs medidos en distintos tipos de MM clasificados por las aberraciones cromosómicas más comunes en este tipo de enfermedad. En concreto, se establecieron 5 categorías de MM: t(14;16), t(11;14), t(4;14), deleción de retinoblastoma (delRB) y FISH normal; añadiendo como otra categoría control células plasmáticas normales. El propósito de este estudio fue caracterizar los distintos subtipos de MM a través de la expresión de miRNAs encontrando marcadores para cada clase y tratar de relacionarlos con la expresión de los genes codificantes de proteína. Esta asociación entre datos biológicos de expresión y datos clínicos de subtipos patológicos es un paso clave para ayudar a encontrar tratamientos más personalizados para el MM.

El trabajo descrito en este capítulo se enmarca dentro de proyectos de colaboración con otros grupos del Centro de Investigación del Cáncer de la Universidad de Salamanca y del Hospital Clínico Universitario de Salamanca. Por este motivo, el trabajo realizado sobre muestras de CLL y MM que se presenta en esta Tesis Doctoral se encuentra restringido al desarrollo y aplicación de análisis bioinformáticos de datos procedentes de técnicas genómicas, sin presentar detalle sobre el diseño de experimentos ni sobre la validación biológica-funcional de biomarcadores encontrados.

2.2. Materiales y métodos

2.2.1. Muestras de *Leucemia Linfocítica Crónica* y métodos aplicados

El análisis para la identificación de biomarcadores en CLL, se realizó sobre datos de expresión genómica obtenidos utilizando el microarrays de alta densidad de *Affymetrix*. Las muestras

procedentes del Hospital Clínico Universitario de Salamanca fueron procesadas e hibridadas en la plataforma de microarrays por el grupo de investigación del laboratorio 12 del Centro de Investigación del Cáncer de Salamanca. Estas muestras incluyen un total de 102 perfiles de expresión que contienen un grupo de estudio, un grupo de validación y un grupo de muestras sanas. El grupo de estudio lo componen 70 muestras procedentes de células mononucleares de sangre periférica (PBMCs) aisladas por gradiente de Ficoll. El grupo de validación, compuesto de 32 muestras, fue obtenido a partir de células CD19 positivas (linfocitos B, CD19+) purificadas al 98% mediante la técnica de clasificación de células activadas magnéticamente (MACS). Finalmente el grupo de control está formado por 5 muestras procedentes de donantes sanos también purificadas al 98% mediante MACS.

La clasificación de estas muestras se hizo en función de sus aberraciones cromosómicas. Los casos 13q- se dividieron en 13q-H (A de porcentaje Alto) cuando el porcentaje de células presentando la alteración 13q- era superior al 80%, y 13q-L (B de porcentaje Bajo) en el caso contrario. Los casos de pérdida en 17p y/o de 11q –del(17p/11q)– fueron agrupados en la misma categoría por tener características clínicas similares. También se incluyeron muestras de CLL que, mediante la técnica de hibridación con fluorescencia *in situ* (FISH), mostraron un cariotipo normal (denominadas CLL-nk ó CLL 13q-N). La última categoría la constituyen las muestras de linfocito B sanos como grupo control.

Al ser un estudio amplio y planteado en varias fases, la mayoría de las muestras del grupo de estudio fueron utilizadas para ser analizadas con PCR cuantitativa de genes y de miRNAs, mientras que otra parte (27 muestras) fueron hibridadas con el chip *Human Exon 1.0* para obtener la firma de expresión génica global. El grupo de validación fue íntegramente analizado con *Human Exon 1.0* (ver **tabla 2.1**). El trabajo de detección de biomarcadores se centró únicamente en el análisis de las muestras hibridadas con microarrays de exones para los que se tenía la señal de expresión global.

Tipos de muestras	**Grupo de estudio** (PBMC)	**Subgrupo de estudio:** muestras hibridadas con *HEx1* (células mononucleares PBMC)	**Grupo de validación:** hibridadas con *HEx1* (células CD19+)
CLL 13q-H	25	7	7
CLL 13q-L	27	6	11
CLL-nk (cariotipo normal)	8	8	9
CLL del(17p/11q)	10	6	–
Células CD19+ sanas (control)	0	–	5
TOTAL	**70**	**27** (incluidas en el grupo de 70)	**32**

Tabla 2.1. Número de muestras por cada subtipo de CLL para el grupo de estudio y el grupo de validación.

Dado el distinto contenido celular (muestras PBMCs y muestras CD19+), dependiente del método de purificación utilizado, los microarrays hibridados con las muestras del grupo de estudio y los hibridados con la muestras del grupo de validación fueron analizados por separado. En ambos grupos se utilizó el algoritmo RMA para la normalización y cálculo de señal en combinación con el paquete CDF *GeneMapper* de ***GATExplorer*** en su versión para *Human Exon 1.0* (*Ensembl* v53 – NCBI36). La utilización de la herramienta descrita en el **capítulo 1** en lugar de los *probesets* originales de *Affymetrix* conlleva todas las ventajas descritas también en dicho capítulo.

Para encontrar genes que marcadores de las distintas categorías anteriormente especificadas se utilizó el algoritmo SAM, proporcionando genes con una diferencia de expresión significativa, situando el punto de corte del p-valor corregido en FDR < 0,05. La capacidad de

los genes encontrados para diferenciar las distintas categorías se comprobó mediante métodos de agrupamiento jerárquicos representados como *heatmaps*. Para realizar un análisis más detallado de la firma génica significativa y ver si esta permitía discernir y separar bien las distintas categorías en base a sus semejanzas y diferencias de expresión, se utilizó un análisis de componentes principales (PCA). Basados en el resultado de este análisis se realiza una representación tridimensional de las muestras sobre los valores de las tres primeras componentes proporcionadas por el PCA a partir de la matriz de expresión normalizada. Dicha matriz de expresión fue filtrada previamente eliminando el 25% de los genes que menos variaban su expresión (calculado con el rango intercuartil, IQR) para permitir reducir ruido al eliminar genes no informativos. A continuación se calcula la mediana de la expresión de cada gen por cada una de las categorías y se introducen estos valores en la siguiente fórmula:

$$Y_{ij} = \frac{X_{ij} - median(ik)}{sd(ik) + \beta} + median(ik)$$

Esta fórmula fue diseñada para calcular los valores de expresión por gen y muestra considerando su variabilidad dentro de su categoría; siendo Y_{ij} la matriz de expresión utilizada para el PCA, X_{ij} la matriz de expresión original, i el gen, j la muestra, k la categoría y β una constante positiva con valor 2, añadida al denominador para asegurar que la varianza de Y_{ij} es independiente de la desviación estándar de los genes. Esta fórmula representa una estrategia eficaz para calcular la dispersión de las muestras, réplicas biológicas, basada en su mediana en cada categoría. De esta forma se pueden representar las diferencias entre categorías atenuando la variación entre muestras individuales. El cálculo del PCA se realizó mediante la función *prcomp* del paquete *stats* **(R_Development_Core_Team, 2010)** y la representación visual del mismo mediante el paquete *rgl* **(Alder and Murdoch, 2011)**, ambos de R.

Por último, para el grupo de validación se realizó una selección de las muestras de CLL que marcan de modo más coherente las categorías CLL 13q-H y CLL 13q-L aplicando un algoritmo de análisis de respuesta (*outcome*) denominado *Global Test* que se describe en el siguiente apartado.

2.2.2. Muestras de *Mieloma Múltiple* y métodos aplicados

Los análisis sobre MM se realizan sobre un conjunto de muestras de médula ósea tomadas de 60 pacientes y de 5 controles sanos (obtenidas por el grupo de Hematología del Hospital Clínico Universitario de Salamanca).

Alteraciones citogenéticas	Número de muestras
MM t(4;14)	17
MM t(11;14)	11
MM t(14;16)	4
MM delRB (con deleción de RB)	14
MM delRBp53 (deleción RB y TP53)	1
MM con FISH normal	13
Células plasmáticas sanas (control)	5
TOTAL	**65**

Tabla 2.2. Características citogenéticas de 60 pacientes diagnosticados con MM incluyendo 5 muestras de células sanas como controles.

En estas muestras se aislaron las células plasmáticas utilizando el marcador CD138+ mediante

el sistema *AutoMACs automated separation system* (Milteyi-Biotec, Auburn CA, USA) elevando la pureza por encima del 90%. La clasificación de los pacientes se hizo en función de las aberraciones cromosómicas, obteniendo seis subtipos de MM que representan las alteraciones citogenéticas y deleciones más recurrentes en esta enfermedad (ver **tabla 2.2**).

MM subtipo	miRNAs alterados	Localización cromosómica
t(4;14)		
	hsa-miR-203	14q32
	hsa-miR-155	21q21
	hsa-miR-650	22q11
	hsa-miR-375	2q35
	hsa-miR-196b	7p15
	hsa-miR-342	14q32
	hsa-miR-214	1q24
	hsa-miR-193a	17q11
	hsa-miR-135b	1q32
	hsa-miR-146a	5q33
	hsa-miR-133b	6p12
t(11;14)		
	hsa-miR-650	22q11
	hsa-miR-125a	19q13
	hsa-miR-375	2q35
	hsa-miR-184	15q25
	hsa-miR-214	1q24
	hsa-miR-95	4p16
	hsa-miR-199a	19p13
t(14;16)		
	hsa-miR-1	18q11
	hsa-miR-449	5q11
	hsa-miR-133a	18q11
	hsa-miR-196b	7p15
	hsa-miR-135b	1q32
	hsa-miR-214	1q24
	hsa-miR-375	2q35
	hsa-miR-642	19q13
delRB		
	hsa-miR-196a	17q21
	hsa-miR-486	8p11
	hsa-miR-375	2q35
	hsa-miR-501	Xp11
	hsa-miR-320	8p21
	hsa-miR-20a	13q31
	hsa-miR-133b	6p12
	hsa-miR-135b	1q32
	hsa-miR-126	9q34
	hsa-miR-650	22q11
	hsa-miR-214	1q24
	hsa-miR-19b	13q31
	hsa-miR-10a	17q21
	hsa-miR-15a	13q14
	hsa-miR-133a	18q11
	hsa-miR-139	11q13
	hsa-miR-197	1p13
	hsa-miR-10b	2q31
	hsa-miR-95	4p16
	hsa-miR-126	9q34
	hsa-miR-186	1p31
	hsa-miR-19a	13q31
	hsa-miR-451	17q11
	hsa-let-7b	22q13
	hsa-miR-140	16q22
	hsa-miR-125a	19q13
	hsa-miR-362	Xp11
	hsa-miR-33	22q13
	hsa-miR-223	Xq12
	hsa-miR-224	Xq28
	hsa-miR-221	Xp11
	hsa-miR-30e	1p34
	hsa-miR-374	Xq13
	hsa-let-7c	21q21
	hsa-miR-99b	19q13
	hsa-miR-130a	11q12
	hsa-miR-193a	17q11
FISH normal		
	hsa-miR-135b	1q32
	hsa-miR-375	2q35
	hsa-miR-155	21q21
	hsa-miR-650	22q11
	hsa-miR-572	4p15
	hsa-miR-152	17q21
	hsa-miR-362	Xp11
	hsa-miR-486	8p11
	hsa-miR-95	4p16
	hsa-miR-214	1q24
	hsa-miR-501	Xp11
	hsa-miR-196a	17q21
	hsa-miR-642	19q13
	hsa-miR-10a	17q21
	hsa-miR-452	Xq28
	hsa-miR-342	14q32
	hsa-let-7c	21q21
	hsa-miR-203	14q32

(Continuación tabla)

Tabla 2.3. Listado de los miRNAs cuya expresión ha sido encontrada significativamente desregulada en distintos subtipos de MM: **t(4;14)**, **t(11;14)**, **t(14;16)**, **delRB** y **FISH normal**.

Tras la obtención y purificación de muestras se procedió a los análisis del perfil de expresión de miRNAs. Primero se aisló de cada muestra el correspondiente RNA y se realizó transcripción reversa de RNA a cDNA usando el protocolo específico *TaqMan miRNA Reverse Transcription Kit* (PE Applied Biosystems, Foster City, CA, USA). Posteriormente se utilizaron arrays *TaqMan* de baja densidad que permiten la cuantificación de 365 miRNAs humanos por PCR en tiempo

real (RT-PCR). Tras la obtención de los datos crudos de expresión de todos los miRNAs, se procedió a la elaboración de una lista de miRNAs desregulados/alterados significativamente. Dicha lista fue obtenida mediante la comparación de la expresión de los miRNAs de cada alteración citogenética en relación a las células plasmáticas sanas, utilizando para ello el algoritmo SAM con un punto de corte por p-valor corregido: FDR < 0.001. De este modo, se identificaron 11, 7, 8, 37 and 18 miRNAs para los pacientes con t(4;14), t(11;14), t(14;16), delRB y FISH normal respectivamente (ver **tabla 2.3**).

El trabajo principal realizado en la presente Tesis Doctoral con los datos de MM, se centra en aplicar un modelo predictivo para estimar el nivel de asociación entre las distintas categorías de pacientes y el conjunto de miRNAs alterados, para lo cual se utilizó el algoritmo *Global Test* **(Goeman et al., 2004)**. Este algoritmo permite predecir la influencia de uno o varios factores predefinidos –que en este caso son un conjunto de miRNAs con sus perfiles de expresión– sobre una variable de respuesta o de salida (*outcome*) que es testada –que en este caso son los subtipos de enfermedad MM definidos–. La hipótesis nula correspondería al hecho de que el perfil de expresión del conjunto de factores testado no está asociado a la variable de respuesta. En el caso de MM analizamos la expresión de los miRNAs para identificar cuales son los que mejor marcan o predicen las distintas alteraciones del cariotipo usadas como respuesta. Este algoritmo se utilizó en su versión para R con el paquete llamado *globaltest* **(Goeman and Oosting, 2009)** incluido en *Bioconductor.*

2.3. Resultados

2.3.1. Análisis de muestras de *Leucemia Linfocítica Crónica*

El análisis de expresión diferencial aplicado a las muestras del grupo de estudio de CLL 13q-H contra las muestras CLL 13q-L devolvió un total de 3.450 genes significativos de los cuales 1.244 estaban sobre-expresados y 2.206 reprimidos en 13q-H. Esta diferencia importante indica que muestras que presentan la misma enfermedad con una misma alteración citogenética tienen un patrón de expresión diferente atendiendo al porcentaje de blastos encontrados con dicha pérdida de 13q. Esto significa que las categorías patológicas definidas por los datos clínicos (de mejor o peor pronóstico) presentan también diferencias a nivel biológico transcriptómico, lo que confirma la principal hipótesis de trabajo. Una posterior comprobación por RT-PCR validó varios de los genes encontrados como alterados: GAS7, E2F1 y FCRL2.

Otras de las hipótesis de trabajo de este estudio es que los grupos de peor pronóstico, como son CLL 13q-H y del(17p/11q), comparten un perfil de expresión más cercano que las de mejor pronóstico, CLL 13q-L y CLL-nk. La utilización de un método de agrupamiento no supervisado utilizando todos los genes detectables por los microarrays no fue de utilidad debido a que la gran variabilidad entre pacientes por muchos factores posibles –ocultos o no definidos– hace imposible el agrupamiento entre las muestras del mismo tipo según las categorías dadas (datos no presentados). De esta manera, se realizó un análisis semi-supervisado calculando la distancia entre las distintas categorías utilizando los 3450 genes significativos que diferencian las CLLs 13q-H de las 13q-L. La **figura 2.1** muestra un *heatmap* de este análisis de clasificación semi-supervisado presentando los correspondientes agrupamientos jerárquicos a nivel de genes y a nivel de muestras. Las distancias en ambos casos se miden utilizando la correlación de *Pearson* (*d=1-cor*) y el tipo "*complete*" para calcular la disimilaridad con un método

aglomerativo. El propósito de este test no fue medir las diferencias específicas entre 13q-H y 13q-L, ya que utilizando únicamente los genes con expresión diferencial significativa entre estas categorías ya observamos una gran diferencia. El propósito era determinar cuál de los dos grupos es más parecido al grupo del(17p/11q) de mal pronóstico. Es manifiesto en el análisis que las categorías 13q-H (**rojo**) y del(17p/11q) (**magenta**) quedan más cercanas entre sí que 13p-B (**azul**), lo cual puede apreciarse en la primera división del dendrograma. Esto confirma de nuevo que las semejanzas y diferencias en las variables clínicas (grupos de peor pronóstico) de los distintos subtipos de CLL, se reflejan en sus características transcriptómicas.

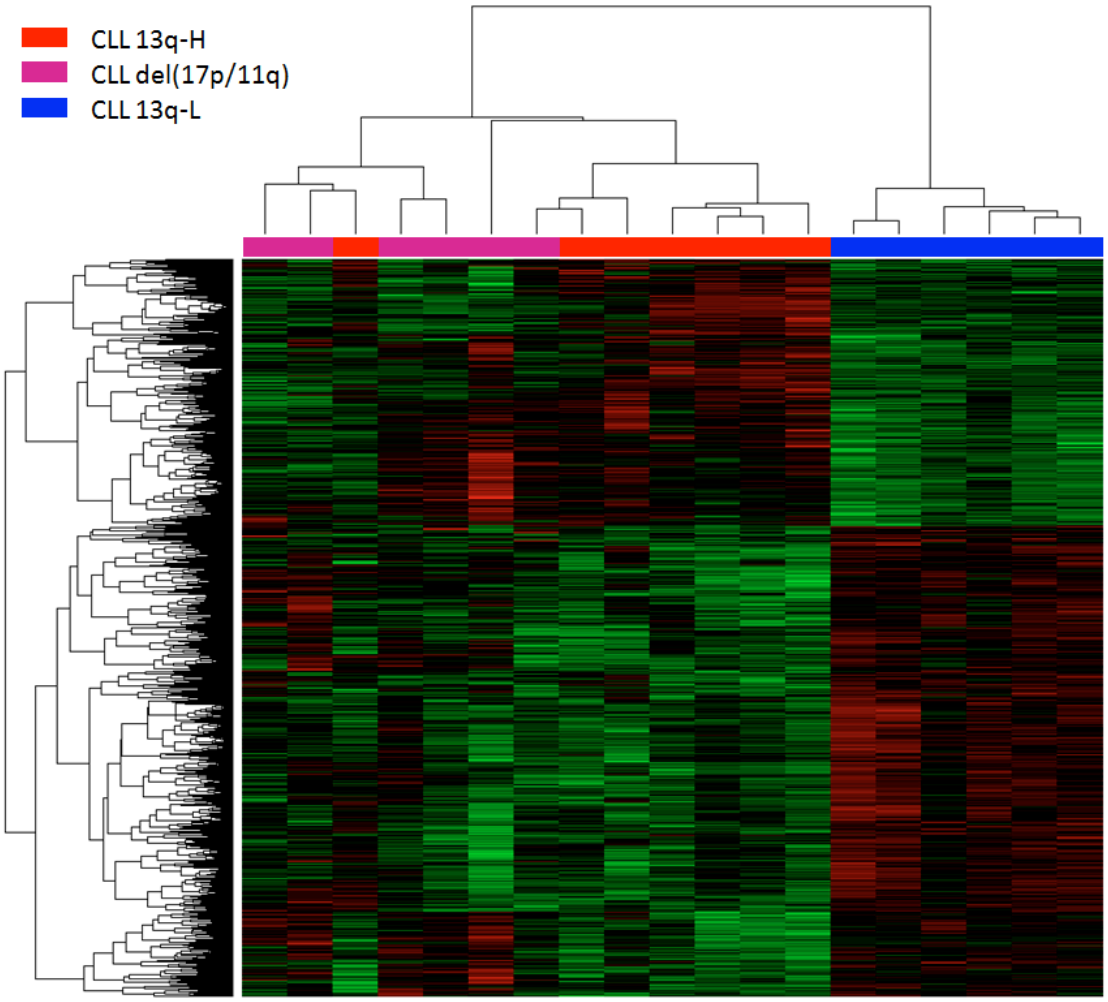


Figura 2.1. *Heatmap* de las muestras de CLL 13q-H (n=7), 13q-L (n=6) y del(17p/11q) (n=6) etiquetadas en **rojo**, **azul** y **magenta** respectivamente. Se utilizaron los 3450 genes encontrados en el test de expresión diferencial entre 13q-H y 13q-L. La figura muestra dos grupos principales: uno con las muestras 13q-L; otro con las muestras 13q-H y del(17p/11q). El resultado es consistente con datos clínicos de prognosis que es diferente entre ambos grupos.

Para comprobar si las muestras del grupo de validación (32 pacientes diferentes, ver **tabla 2.1**) replicaban el comportamiento de las muestras del grupo de estudio (27 muestras) se obtuvo para este grupo un conjunto más fiable de genes marcadores reduciendo el punto de corte del contraste de CLLs 13q-H contra 13q-L (de FDR < 0,05 a < 0,01), lo cual proporcionó 1030 genes. A continuación, utilizando el algoritmo *Global Test*, se analizó la influencia de este conjunto de genes para poder discernir las dos categorías A y B en el grupo de validación (que incluye 7 muestras de CLL 13q-H y 11 muestras de CLL 13q-L). Este análisis dio un p-valor significativo de 0,0935 que confirma que los genes encontrados en el grupo de estudio sirven también para distinguir entre las categorías 13q-H y 13q-L en el grupo de validación para la mayoría de las

muestras (ver **figura 2.2**). No obstante 2 muestras 13q-H y 3 muestras 13q-L presentaron influencia negativa contraria a la influencia general, indicando que tienen un comportamiento no equivalente a las otras muestras, por lo que fueron descartadas para el análisis posterior. También, en la serie de validación tuvo que ser descartada una muestra del tipo CLL-nk por descubrirse que tenía otras alteraciones cromosómicas y no cumplía los criterios de FISH normal a los que previamente se había asociado. Con estos recortes la serie de validación quedó con el siguiente número de muestras: CLL 13q-H (n=7-2=5); CLL 13q-L (n=11-3=8); CLL-nk (n=9-1=8); células sanas controles (n=5).

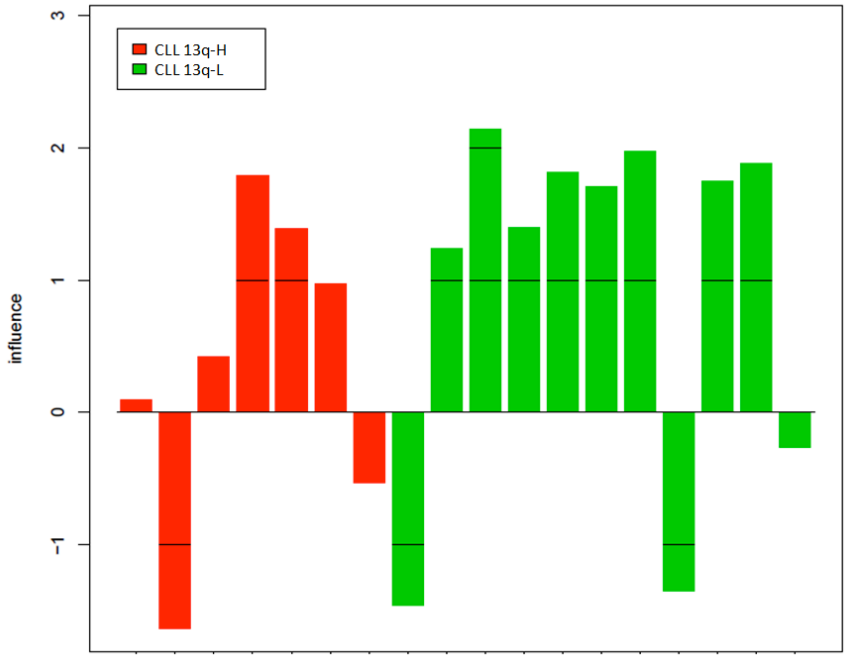


Figura 2.2. *Global Test* sobre muestras CLL 13q-H y CLL 13q-L del grupo de validación utilizando un grupo de genes diferencialmente expresados entre ambas categorías en el grupo de estudio (p-valor=0,0935).

Para confirmar estos resultados se realizó un último análisis de clasificación no supervisado sobre el grupo de validación. Debido a la pureza superior de las células aisladas CD19+ que son utilizadas en este grupo de validación, esta vez se observó que sí era posible realizar comparaciones entre categorías a nivel ómico (i.e. transcriptómico) de una forma no supervisada. En este caso se optó por un nuevo análisis de componentes principales (PCA) para medir las distancias entre las categorías CLL 13q-H y 13q-L, añadiendo además CLL-nk y muestras sanas control para dar más perspectiva a las diferencias entre los 2 tipos de 13q- estudiados. Después de procesar la matriz globales de expresión según lo descrito en la sección de "Materiales y métodos", 28806 genes fueron utilizados para el cálculo de componentes principales (es decir, se usaron todos los genes detectados por el array *Human Exon 1.0*). El resultado de este análisis fue representado en tres dimensiones acorde con las 3 componentes de mayor peso, que acumulaban una proporción de la varianza de 48,3%, 60,9% y 68,3% respectivamente (ver **figura 2.3**). Esto quiere decir que, la representación en un espacio tridimensional de estas muestras, refleja una alta proporción de la varianza contenida en la matriz de expresión original que tiene una dimensionalidad mucho mayor, siendo por lo tanto un análisis adecuado para estudiar visualmente las diferencias entre categorías. En la **figura 2.3** se escalaron los ejes X, Y y Z en función al peso de las componentes principales 1, 2 y 3, de manera que dichos ejes muestran diferencias en su longitud. El eje X correspondiente a la componente principal de mayor peso (PC1), separa las muestras procedentes de donantes sanos de los pacientes enfermos. El eje Y, correspondiente a la componente principal número

2 (PC2), separa las muestras de peor pronóstico (CLL 13q-H) de las de mejor pronóstico (CLL 13q-L y CLL-nk). Por último, el eje Z que corresponde a la componente principal número 3 (PC3), tiene la capacidad de separar los tipos de muestras más parecidos entre sí, CLL 13q-L y CLL-nk. Por lo tanto, se puede concluir que este PCA recoge en una sola imagen de tres dimensiones la influencia de todos los genes de una manera coherente con los datos clínicos.

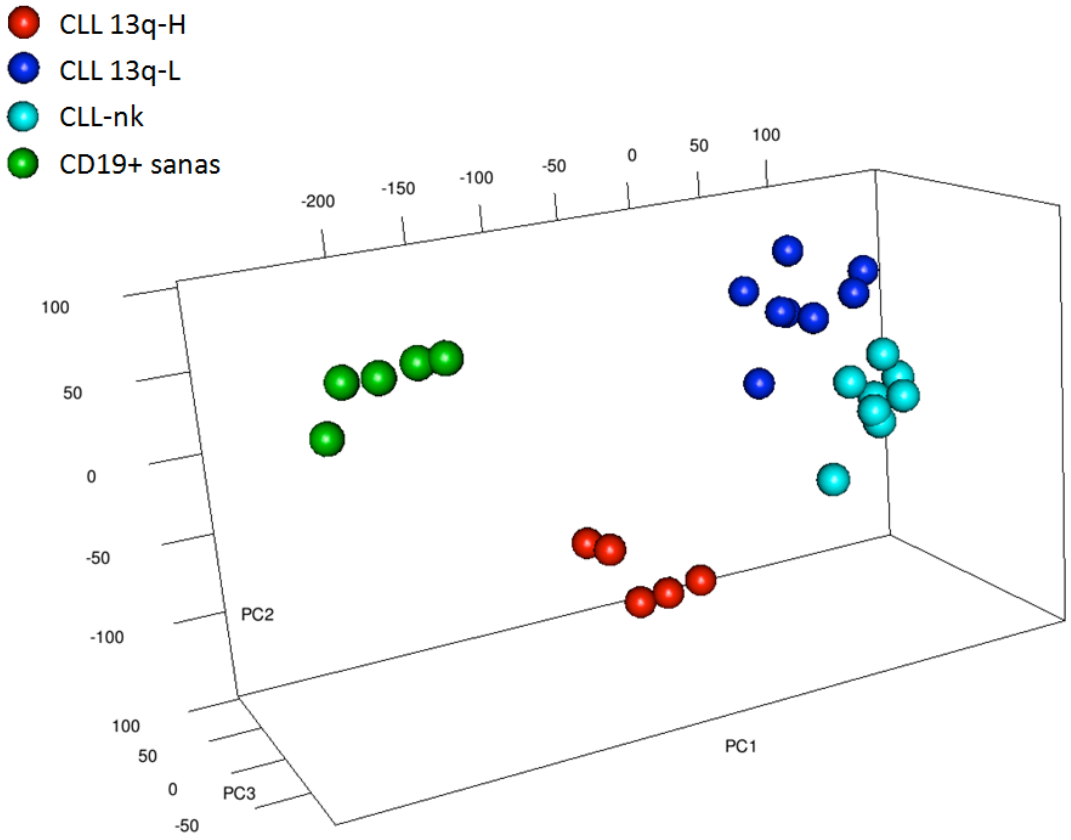


Figura 2.3. Análisis de componentes principales mostrando las distancias entre CLLs 13q-H, CLLs 13q-L, CLL-nk (ó CLL 13q-N) y muestras CD19+ controles. La matriz de expresión fue previamente filtrada eliminando el 25% de genes menos variante y transformada de acuerdo a la fórmula descrita en "Materiales y métodos". La cercanía entre los grupos de mejor pronóstico 13q-L y 13q-N, sugiere que comparten un perfil de expresión génica más parecido que a la categoría de peor pronóstico 13q-H.

En análisis de expresión diferencial con SAM devolvió 15332 genes al comparar CLLs 13q-L con muestras sanas controles CD19+, y 16754 genes en la comparación de CLL-nk con controles CD19+. Al hacer la intersección entre ambos grupos se obtuvo que 13749 genes estaban presentes en ambos contrastes y 10425 compartían además el mismo sentido de sobre-expresión o infra-expresión, indicando una fuerte similitud en los genes desregulados en este tipo de cáncer: 56,85 % de genes comunes con idéntico sentido de alteración. Sin embargo, el análisis de expresión diferencial con SAM entre las muestras CLL 13q-H y las CD19+ controles devolvió sólo 6775 genes, la mayoría de los cuales (6339 genes: 93,56 %) eran comunes con CLL 13q-L y CLL-nk (es decir, eran marcadores de la patología general CLL común a todos los subtipos). Este menor número encontrado al comparar el subtipo de enfermedad de peor pronóstico contra células procedentes de pacientes sanos puede parecer una contradicción, pero también puede explicarse si este subtipo de enfermedad es más heterogéneo o variable que los subtipos de mejor pronóstico CLL 13q-L y CLL-nk. Para comprobar esta hipótesis se midió el parecido entre sí de las muestras de cada tipo utilizando la correlación de *Pearson* sobre todos los pares posibles. La media de las correlaciones entre las muestras de cada

subtipo fueron: 0.971 para controles sanos CD19+; 0.954 para CLL-nk; 0.950 para CLL 13q-L y 0.930 para CLL 13q-H. Esto indica que las muestras sanas son más homogéneas que las muestras CLL de mejor pronóstico, que a su vez son más parecidas entre sí que las muestras de CLL de peor pronóstico. Esto confirma la idea de que las muestras de CLL 13q-H presentarían una mayor variabilidad interna. Debido a que los algoritmos de expresión diferencial buscan estabilidad dentro de cada categoría y diferencias inter-categorías, puede inferirse que el hecho de encontrar un menor número de genes en CLL 13q-H cuando se compara con el control refleja, en este caso, una menor estabilidad interna en los perfiles transcriptómicos de los pacientes con CLL 13q-H.

Finalmente, respecto a los genes específicos identificados para las categorías CLL 13q-H y CLL 13q-L, aunque la firma génica global es muy grande, se pueden asociar principalmente con procesos de proliferación celular, apoptosis y señalización celular. Se pueden citar SYK, BLNK y PRKCB1 que han sido validados por PCR semicuantitativa y se relacionaron funcionalmente a señalización celular vía receptor de célula B. También se observó un desequilibrio entre proliferación y apoptosis, en donde se identificaron sobre-expresados genes anti-apoptóticos como BCL2 y reprimidos genes pro-apoptóticos como RASSF5, BAD, CASP8, CASP10, FAS. Además de esto, genes relacionados con la proliferación celular como LEF1, E2F5 y RRAS2 pudieron ser también validados experimentalmente **(Rodriguez et al., 2012)**.

2.3.2. Análisis de muestras de *Mieloma Múltiple*

Como se ha indicado en la **tabla 2.2** la serie de muestras de MM incluye 6 subtipos de la enfermedad. Sobre estos datos –eliminando previamente la categoría que presenta la doble deleción de RB y de TP53, por ser una única muestra– se aplicó el algoritmo *Global Test*, para comprobar en qué medida la firma de expresión basada en miRNAs, e identificada por métodos de expresión diferencial, marca bien cada una de las 59 muestras de pacientes de MM analizados.

Según lo indicado anteriormente, para realizar el *Global Test* como factores de entrada *a priori* se utilizaron los datos de expresión de los miRNAs identificados como desregulados que, después de eliminar los repetidos, sumaron 49. El resultado de dicho test devolvió un p-valor inferior a 0,001, demostrando una fuerte relación entre expresión y categorías citogenéticas de MM. La **figura 2.4** indica cómo las diferentes muestras de cada categoría de MM se identifican con el perfil de expresión de los miRNAs. Dicha figura fue realizada con la función llamada *sampleplot*, que representa la influencia de la expresión de los miRNAs seleccionados respecto a cada una de las muestras. Cada barra representa por tanto un paciente y su sentido arriba o abajo, indica evidencia en contra o a favor de la hipótesis nula, es decir, de poder categorizar o no cada muestra individual dentro de su grupo. Si la barra tiene valor positivo (hacia arriba), su perfil de expresión es similar a las otras muestras de su categoría y relativamente distinto al resto de categorías. Si la barra es negativa (hacia abajo), la muestra tiene un perfil de expresión poco parecido a las de su categoría. En caso de cumplirse la hipótesis nula, la influencia esperada sería 0. Las marcas de las barras indican la desviación estándar de la influencia de la muestra bajo la hipótesis nula. El subtipo de MM que mostró mayor influencia fue la clase t(14;16), siendo por lo tanto el que mejor se identifica con el perfil de expresión de los 49 miRNAS seleccionados. El subtipo de MM que mostró la menor influencia resultó ser el grupo de FISH normal.

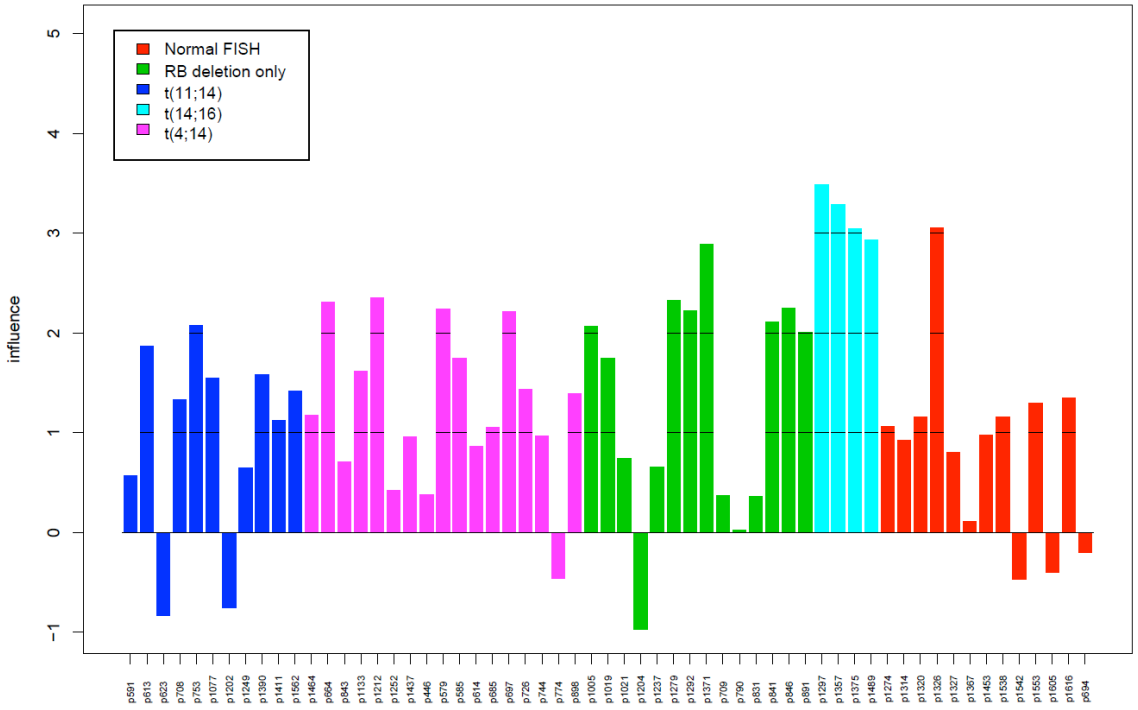


Figura 2.4. Influencia del perfil de expresión de 49 miRNAs desregulados en 5 subtipos de MM. Cada muestra es representada con una barra cuya altura indica la influencia que tiene la expresión para asociar cada muestra con su grupo (p-valor < 0,001).

Con estos resultados, a modo de resumen, se podría decir que existe una correspondencia entre la expresión de los miRNAs y las anormalidades citogenéticas, evidenciando así la capacidad de este grupo de miRNAs para diferenciar los distintos tipos de MM. Dicho de otra manera, los distintos tipos de MM muestran distintos patrones de expresión de miRNAs.

2.4. Discusión y posible trabajo futuro

Las enfermedades complejas, como es el cáncer, muestran una gran disparidad entre distintos pacientes, pudiendo establecerse entre los enfermos subcategorías en función no solo de sus parámetros clínicos y patológicos, del grado de afectación y del estadio en el que se encuentran, sino también por diferencias más sutiles a nivel molecular. La identificación de las características moleculares que definen cada subtipo de enfermedad es clave para el avance de la medicina moderna y para que se pase de las estrategias de tratamientos basados en extirpación –muy frecuentes en cáncer– a tratamientos personalizados de fundamento molecular que apuntan a las causas.

En el caso de cáncer hematológico –i.e. hemopatías malignas– es muy frecuente la clasificación de subtipos patológicos por las aberraciones cromosómicas que identifican en las células tumorales de cada paciente. Estas clasificaciones tienen una clara base molecular y se han demostrado de gran ayuda para conseguir diagnósticos y pronósticos precisos. Sin embargo, detrás de cada tipo de alteración cromosómica subyacen alteraciones biológicas moleculares y mecanismos celulares que están todavía en muchos casos por descubrir y definir.

En este trabajo se han manejado muestras de distintos subtipos de CLL agrupadas en función de sus alteraciones citogenéticas, encontrando correlación entre el pronóstico observado en la clínica y el perfil molecular obtenido a partir de datos de expresión génica. También se ha ayudado a validar un grupo de miRNAs como útiles para caracterizar y clasificar distintos subtipos de MM. Sin duda, como se ha indicado, la identificación de biomarcadores específicos para los subtipos de una enfermedad, ayuda a revelar y descubrir las características biológicas moleculares y celulares que los diferencian, y son un claro camino para testar nuevas dianas terapéuticas permitiendo el desarrollo de una medicina más personalizada.

El mapeo alternativo, orientado a genes, de datos de expresión global obtenidos con microarrays de alta densidad (realizado y descrito en el **capítulo 1**), ha sido aplicado con éxito al estudio de dos patologías concretas en este **capítulo 2**. La coherencia de los resultados y la validación experimental de algunos genes identificados como desregulados, puede verse como una validación en experimentos reales –trabajando con datos de pacientes– de las herramientas bioinformáticas desarrolladas. Las ventajas de apuntar directamente a genes a la hora de mapear y medir la señal transcriptómica de los microarrays han sido patentes a la hora de identificar dianas moleculares, evitando pasos intermedios artificiales. Las mejoras en cuanto a eliminación de hibridación cruzada, y una mejor y más actualizada definición de las entidades génicas, han tenido también un papel importante en los resultados finales.

Como perspectivas para el futuro, parece claro que el proceso de dividir las enfermedades en distintas subcategorías seguirá refinándose. En esta labor, el uso de técnicas genómicas y transcriptómicas de alto rendimiento seguirá teniendo un peso importante, ya que permiten identificar variables moleculares distintas y específicas por cada muestra que, en combinación con distintas estrategias de aprendizaje, pueden hallar nuevos subgrupos de enfermedades con una repercusión clínica importante de cara a mejorar los diagnósticos, los tratamientos y los pronósticos.

Capítulo 3
Diseño, construcción y validación de un algoritmo para detección de *splicing* alternativo

3.1. Introducción

3.1.1 *Splicing* alternativo: papel biológico e implicaciones en cáncer

El *splicing* alternativo es el proceso que sufre el pre-mRNA en la fase de maduración a mRNA por el cual se seleccionan ciertos exones descartando otros, dando lugar a distintas posibles proteínas ó isoformas de una proteína a partir del mismo *locus* génico (ver **introducción**). Este proceso incrementa notablemente el tamaño y diversidad del proteoma en eucariotas sin necesidad de aumentar en la misma proporción el número de *locus* génicos en un genoma **(Graveley, 2001)**. El número de genes –considerando el concepto clásico de gen como una entidad genética funcional– necesario para el desarrollo y mantenimiento de un ser complejo como un mamífero, se calcula en torno a 100.000, mientras que el número encontrado en el genoma suele ser cuatro veces inferior **(Nilsen and Graveley, 2010)**. Este cambio de números se debe a la multiplicidad de productos génicos que se suelen derivar de cada *locus* génico en organismos superiores. En el caso de los humanos, se ha estimado que alrededor del 95% de los genes con más de un exón sufren *splicing* alternativo dando lugar a cerca de 100.000 eventos de *splicing* en los diferentes tejidos **(Pan et al., 2008)**. El proceso de *splicing* alternativo no ha empezado a entenderse bien hasta los últimos años, cuando se está descubriendo la gran complejidad de su regulación y la compleja maquinaria molecular que lo controla, donde se incluye el *spliceosoma* **(Smith and Valcarcel, 2000; Wahl et al., 2009)**. Además, el proceso celular de *splicing* no solo está relacionado con estados biológicos normales, sino que su desregulación y alteración también se ha asociado con el desarrollo de enfermedades como el cáncer **(Grosso et al., 2008)**. Mutaciones producidas en puntos de *splicing* situados en intrones de genes supresores de tumores pueden provocar la eliminación de exones produciendo proteínas truncadas **(Venables, 2004)**. También, según se ha descrito, la ganancia de algunas isoformas alternativas puede contribuir al desarrollo tumoral **(Kalnina et al., 2005)**. Así se entiende que debido a la relevancia de este sistema de procesamiento post-transcripcional del mRNA –que es el *splicing* alternativo– el avance en el conocimiento de la actividad celular y el abordaje de datos transcriptómicos requerirá su estudio más allá de la simple medida de expresión de los diferentes genes. La medida de expresión de un *locus* génico de modo global no apunta directamente a las distintas isoformas posibles producidas a partir de ese *locus*, perdiendo así información de gran relevancia. Por otro lado, una descripción y medida precisa de la expresión de los distintos transcritos derivados de un *locus*

permitiría, sin duda, estudiar de modo más directo y específico las distintas isoformas proteicas que de dicho *locus* pueden surgir en distintos tipos celulares o en distintos estados funcionales.

3.1.2 Técnicas de detección de *splicing* alternativo

La detección de eventos de *splicing* se puede realizar mediante técnicas de biología molecular como es la PCR en tiempo real (RT-PCR). También la información acumulada en bases de datos biológicas como en librerías de EST u otras como ASD **(Thanaraj et al., 2004)** o ASTD **(Koscielny et al., 2009)**, pueden proporcionar pistas sobre la transcripción de isoformas en distintos tipos de tejidos sanos y también en tejidos enfermos o estados patológicos. Sin embargo, para obtener una visión global a nivel genómico/transcriptómico en un solo análisis, se precisa el uso de técnicas de alto rendimiento de gran escala (técnicas "ómicas").

Como se ha explicado en el **capítulo 1**, los microarrays basados en tecnología de amplificación 3' IVT, solo incluyen sondas en regiones cercanas al 3' del *locus*. No en todos los *loci* estas sondas se sitúan en el último exón, pudiendo encontrarse también en exones anteriores o incluso entre dos exones (sondas inter-exónicas). Algunos autores han aprovechado estos hechos para inferir eventos de *splicing* alternativo en determinados genes **(Stalteri and Harrison, 2007)** o incluso diseñar alguna herramienta de predicción **(Rambaldi et al., 2007)**. Sin embargo, estos microarrays no tienen una buena cobertura de todos los exones de los genes que mapean. Existen por otro lado microarrays diseñados para el estudio del *splicing* alternativo que ubican sus sondas entre exones, llamados "*exon junction microarrays*" (EJM). Este tipo de tecnología ha sido quizás la más utilizada para estudios de *splicing* antes de la llegada de los microarrays específicos de exones y de la secuenciación masiva, y ha dado buenos resultados según reportan algunas publicaciones **(Anton et al., 2008; Johnson et al., 2003)**. Incluso más recientemente se han seguido diseñando y comercializando nuevos modelos de este tipo de microarrays, así como haciendo esfuerzos por plantear estrategias para su análisis, como es el caso de los métodos computacionales GenASAP **(Shai et al., 2006)** o MADS+ **(Shen et al., 2010)**. Con la llegada de la técnica *random priming* para la copia y amplificación de material genético, se hizo posible situar sondas a lo largo de un *locus* génico de modo uniforme sin sufrir la pérdida de señal que provocaba la lejanía al 3' con las técnicas clásicas. De esta manera se pudieron diseñar modelos de microarrays como el *Human Exon 1.0* de *Affymetrix.* Este microarray permite medir no solo la expresión del gen, sino también la expresión de cada uno de sus exones (ver **capítulo 1**). Esta forma de poder medir la expresión a dos niveles (exón y *locus* completo) constituye la base de todos los algoritmos de medida de *splicing* que se detallan en la sección "Materiales y métodos". Este sistema ofrece además la ventaja de no depender del conocimiento que se tenga de los distintos exones que componen el genoma en el momento del diseño del microarray, como ocurre en el caso de los EJA **(Clark et al., 2007)**. El diseño de sondas para todos los exones permite reagrupar las sondas en transcritos alternativos que pueden cambiar conforme avanza el conocimiento de un *locus*, permitiendo así descubrir isoformas no anotadas en el momento del diseño del microarray. La secuenciación masiva de RNA (RNA-seq) es capaz de identificar y cuantificar las secuencias de regiones genómicas expresadas en una determinada muestra, a partir de su mRNA total, independientemente de que dichas regiones estén definidas como un exón **(Mortazavi et al., 2008; Wang et al., 2009).** El posterior alineamiento de estas secuencias sobre el genoma, revela todos los exones que son transcritos para cada gen, lo cual permite la identificación de eventos de *splicing* mediante la comparación entre muestras.

El trabajo previo con ***GATExplorer***, presentado en el **capítulo 1**, proporcionó un mapeo de las

sondas de microarrays de alta densidad de oligonucleótidos a nivel de exones y a nivel de genes. El hecho de que entre estos microarrays figure el modelo *Human Exon* de *Affymetrix*, proporciona una buena base para plantear el desarrollo de un método de predicción de *splicing* alternativo con esta plataforma. Partiendo del remapeo que hemos logrado en ***GATExplorer*** de todas las sondas de cada microarray a todos los *loci* conocidos del genoma humano se procedió, en esta nueva fase del trabajo, a realizar un análisis de la expresión más profundo diseccionando la señal procedente de cada *locus* génico en las señales específicas proporcionadas por cada exón.

3.2. Materiales y métodos

3.2.1 Datos de expresión de exones y datos de validación de *splicing*

Utilizamos un *set* de microarrays humanos de exones de *Affymetrix* (obtenidos de la *web* de la compañía: **www.affymetrix.com**) como datos de expresión para realizar las diferentes pruebas de desarrollo del nuevo algoritmo y su comparación con otros métodos previamente publicados. Este set se compone de 33 muestras de tejidos humanos sanos hibridadas con el chip *Human Exon 1.0* correspondientes a 3 muestras de 11 tejidos distintos (ver **tabla 3.1**).

Tejidos y líneas celulares (Wang *et al*.)	Set de datos público (*Affymetrix*)
BT474	**breast**
HME	**cerebellum**
T47D	**heart**
MB435	kidney
MCF7	**liver**
adipose	**muscle**
brain	pancreas
breast	prostate
cerebellum	spleen
colon	**testes**
heart	thyroid
liver	
lymph node	
muscle	
testes	

Tabla 3.1. Tipo de tejidos y líneas celulares utilizados en el trabajo de Wang *et al*. y tejidos utilizados presentes en el set de datos de microarrays *Human Exon* de *Affymetrix*. Los tejidos comunes, resaltados en negrita, serán utilizados para probar y validar distintas estrategias de detección de *splicing* alternativo.

Para validar los resultados obtenidos en los diferentes análisis sobre un *set* de genes humanos para los que se conocen eventos de *splicing*, se utilizó un conjunto de eventos de *splicing* en genes humanos reportados por **(Wang et al., 2008)**. Este trabajo se basa en la identificación de secuencias expresadas en diferentes tejidos y líneas celulares –utilizando secuenciación masiva– que son mapeadas sobre diferentes zonas de los *loci* del genoma: sobre exones específicos, zonas unión de exones, regiones de corte y poliadenilación, etc. Tras la identificación y mapeo, cada uno de los eventos de *splicing* queda asociado a una posición concreta del genoma y a dos tejidos, a cada uno de los cuales se les asigna un valor de

inclusión entre 0 y 1 calculado mediante una aproximación bayesiana. La diferencia entre los valores de inclusión entre los dos tejidos es lo que da el nivel de confianza (*score*) de estar realmente ante un evento de *splicing* alternativo. Finalmente para nuestro estudio en esta Tesis Doctoral será considerado únicamente el conjunto de tejidos comunes a los dos *sets* de datos citados: el *set* de microarrays de *Affymetrix* y el trabajo de Wang *et al*. (ver **tabla 3.1**).

El conjunto de tejidos comunes a ambos grupos corresponde con 6 tejidos distintos: **mama**, **cerebelo**, **corazón**, **hígado**, **músculo** y **testículo**. La combinación de estos 6 tejidos cuando se comparan de 2 en 2 proporciona un total de 15 pares distintos. El número total de pares (gen :: combinación de tejidos) suma 282, mientras que el número de genes distintos es de 270. El número de genes validados por cada par se describe en la **tabla 3.2**.

	breast	**cerebellum**	**heart**	**liver**	**muscle**	**testes**
breast	–	75	11	5	8	16
cerebellum		–	22	12	38	52
heart			–	2	13	7
liver				–	2	5
muscle					–	14
testes						–

Tabla 3.2. Número de genes validados por Wang *et al.* **(Wang et al., 2008)** en cada una de las combinaciones de tejido. La suma total de genes es de 282 y el número de genes distintos es 270.

3.2.2 Descripción de algoritmos y métodos para análisis de *splicing* previamente publicados

La estrategia común a todos los algoritmos de detección de *splicing* alternativo previamente publicados es la de comparar la expresión global del gen contra la expresión individual de cada uno de los exones. La hipótesis fundamental es que, en ausencia de *splicing* alternativo, un cambio de expresión del gen debe suponer un cambio de cada uno de sus exones en la misma proporción y sentido. En este punto, los genes que se desvían de esa norma se interpretan como *splicing*. El reto de los diferentes algoritmos es calcular con precisión la expresión de cada una de las partes (gen y exones) y medir sus variaciones asignado un valor de probabilidad.

Los primeros métodos fueron propuestos por parte de la propia compañía que comercializa los chips. En un artículo llamado *"Alternative Transcript Analysis Methods for Exon Arrays"* de la documentación técnica publicada por *Affymetrix* **(Affymetrix, 2005a)**, se describen 5 métodos entre los que figura el popular y sencillo *Splicing Index.* Una vez estos arrays entraron en el mercado y empezaron a usarse, distintos grupos de investigación desarrollaron sus propuestas. A continuación se revisan brevemente las más relevantes:

- **Splicing Index** **(Affymetrix, 2005a)**: Es el método más simple. En un primer paso propone normalizar la expresión del exón dividiendo su señal por la expresión del gen. En un segundo paso se realiza la media de la expresión normalizada por cada grupo de estudio (p. ej. tejido sano y tumor) calculando su relación o *ratio*.

- **PAC** **(Affymetrix, 2005a)**: Se deriva de *Splicing Index* y asume que en ausencia de

splicing el ratio entre el exón y el gen permanece constante. De esta manera debe de existir una correlación entre la expresión del exón y la expresión del gen, en caso contrario se interpretaría que existe *splicing* alternativo.

- **MIDAS (Affymetrix, 2005a)**: Este algoritmo se basa en la misma idea de *Splicing Index* y PAC, con la diferencia de utilizar un análisis de varianza (ANOVA) para encontrar diferencias entre distintos grupos de estudio utilizando el *ratio* entre el exón y el gen añadiendo una constante para estabilizar la varianza.

- **ANOSVA (Affymetrix, 2005a; Cline et al., 2005)**: Método en donde se propone un modelo lineal cuya hipótesis nula es la no varianza entre genes y exones. La significación de los desvíos de los residuales entre grupos de muestras se calculan mediante un ANOVA.

- **DECONV (Affymetrix, 2005a; Wang et al., 2003)**: Se basa en la estructura del gen con sus diferentes exones para tratar de cuantificar la cantidad relativa de cada una de las isoformas mediante su deconvolución. Requiere conocer *a priori* el número y exones utilizados en de cada una de las isoformas. Es decir, requiere conocer los distintos transcritos alternativos que se pueden generar en un *locus* concreto.

- **FIRMA (Purdom et al., 2008)**: Extiende el modelo aditivo del algoritmo RMA introduciendo nuevos parámetros que representan el valor real del exón y sus discrepancia o desvío frente al valor esperado. Este método calcula este desvío para cada exón y muestra independientemente de las categorías biológicas predefinidas, lo cual permite hacer las comparaciones pertinentes de la manera más conveniente una vez realizados todos los cálculos (p. ej. análisis con muestras pareadas).

- **COSIE (Gaidatzis et al., 2009)**: Este trabajo aborda el problema del "efecto sonda", describiendo cómo se producen falsos positivos en ciertos métodos predictivos por asumir que las diferentes sondas se comportan todas de la misma manera. Para solucionar este problema proponen un método de corrección sonda a sonda en base a un entrenamiento previo con datos procedentes de repositorios públicos. Finalmente utilizan *Splicing Index* para mostrar la mejora introducida por dicha corrección.

- **ARH (Rasche and Herwig, 2010)**: En este algoritmo se utilizan los fundamentos de entropía para calcular la probabilidad de que un exón sufra *splicing* entre dos fenotipos distintos. Mide el desvío de cada exón respecto a la expresión global del gen tratando de determinar si la probabilidad es similar entre ellos, o por el contrario es dominada por uno, o unos pocos exones.

- **SPACE (Anton et al., 2010):** Basado en una versión anterior para arrays con sodas *"exon-junction"* (**Anton et al., 2008)**, el nuevo SPACE pretende mejorar su rendimiento mediante una adaptación a *Human Exon Array*. Este programa se basa en anotar las sondas a nivel de transcritos de *Ensembl* (ENSTs), para calcular su expresión mediante factorización de matrices no negativas.

Algunos de estos métodos, en concreto los propuestos por *Affymetrix*, han sido claramente superado por los métodos publicados posteriormente, o incluso han ofrecido malos resultados desde su origen, como es el caso de ANOSVA, en donde el propio escrito original **(Affymetrix, 2005a)** lo critica fuertemente. Los métodos como DECONV y SPACE, basados en la estimación de la cantidad de cada uno de los transcritos expresados, requieren de un conocimiento previo

exacto del número de exones pertenecientes a cada isoforma. Esta es una clara limitación debido a que el número y definición de los transcritos conocidos es muy variable entre distintas versiones de las bases de datos de referencia (ver **apartado 1.3.6**) y a que a nivel global del genoma esta información es bastante parcial para muchos *loci* génicos. Además, el rápido aumento en el número de transcritos identificados en las nuevas entregas de las bases de datos biológicas (p. ej. *Ensembl*) aumenta notablemente la ambigüedad de resultados basados en estructuras obsoletas complicando su interpretación. Por todo ello, no presuponer unas isoformas concretas y apuntar a exones y genes directamente en lugar de transcritos es una solución más acertada que se basa en una evidencia biológica más sólida y en una información más estable.

Respecto a la implementación de las diferentes estrategias, predominan los desarrollos y métodos hechos en R (como es el caso de FIRMA, COSIE y ARH). Sin embargo ARH no proporciona un programa completo en R, ya que es necesario descargar y utilizar un código escrito en C++, Python, Perl y R llamado *"MAT background correction"* **(Kapur et al., 2007)**. Rasche y Herwing utilizan este programa para corregir el *background* y normalizar las muestras, pero su falta de integración total con R y la necesidad de comunicar los distintos pasos del análisis mediante ficheros de texto hacen de ARH una herramienta de uso tedioso y poco eficaz. Además el usuario debe proporcionar por su cuenta los distintos ficheros de anotación, como el nombre de los identificadores de los exones y a qué genes pertenecen.

Nuestra propuesta para un nuevo algoritmo de análisis de *splicing* se centra en estudiar la relación entre la expresión global del gen y la expresión individual de cada exón, como en la mayoría de algoritmos revisados anteriormente. La estrategia novedosa que planteamos es estimar la expresión de cada exón en función de los valores de expresión de los otros exones del gen utilizando modelos lineales y calcular también su desvío sobre la expresión global esperada, asignándole un valor de probabilidad *p*.

3.2.3 El efecto sonda y su papel en los microarrays de exones

Como es sabido, no todas las sondas de oligos (de 25 nucleótidos) de un microarray reflejan la cantidad de su RNA diana de la misma manera. Análisis de series de arrays hibridadas con concentraciones crecientes de RNA (llamados experimentos de *spiked-in* con arrays) han revelado características variables en las sondas respecto a la señal de hibridación que muestran ante la misma cantidad de RNA y la señal de aumento de expresión ante un mismo aumento de concentración de RNA **(Irizarry et al., 2003b)**. En la **figura 3.1**, Irizarry *et al*. muestran 20 sondas de control (*perfect match* del *probeset* AFFX-BioB-5) del modelo del microarray *Affymetrix* U95A.

Estas sondas de control no mapean sobre un mRNA humano, sino que tienen como objetivo servir de medida de calidad hibridando genes de *E. coli* que se añaden como controles en concentración conocida junto a la muestra a estudiar (siguiendo el protocolo experimental de *Affymetrix*). Este RNA de *E. coli*, al ser añadido en concentraciones conocidas crecientes, sirve para estudiar el comportamiento de las sondas que detectan su expresión. En la **figura 3.1** se ve la tendencia general de todas las sondas a aumentar su intensidad de manera dependiente de la concentración de RNA, sin embargo, no todas se sitúan al mismo nivel ni tienen la misma pendiente. Esto es lo que se ha denominado el "efecto sonda", e implica que la misma sonda es comparable entre distintas muestras, pero no se pueden comparar distintas sondas de la misma muestra de forma directa. Por extensión, este efecto se traslada igual cuando se consideran grupos o conjuntos de sondas (i.e. *probeset*s).

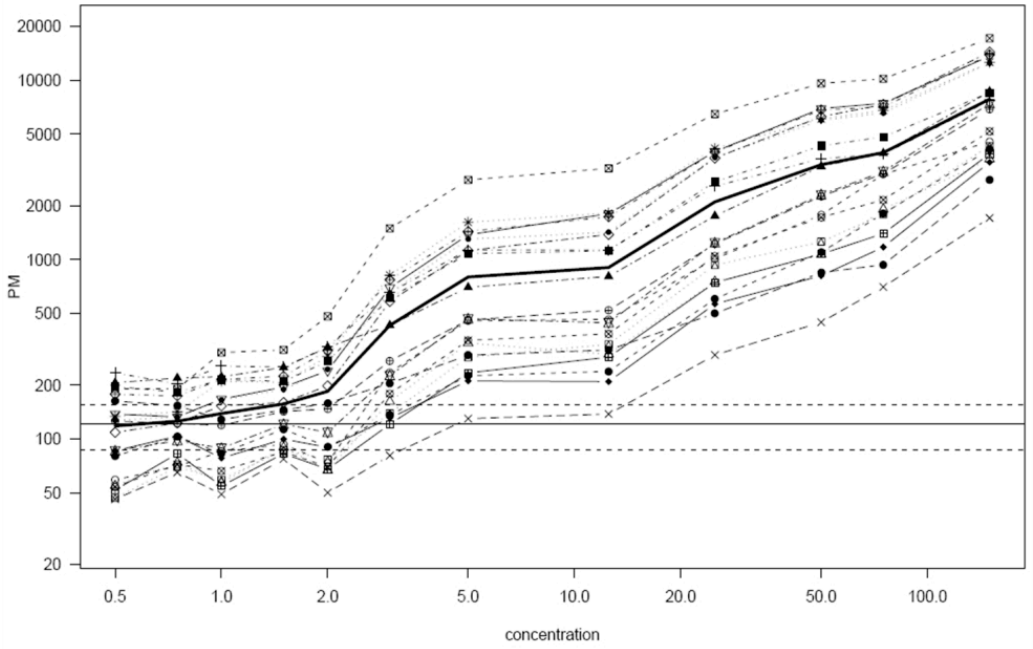


Figura 3.1. El nivel de intensidad y respuesta ante concentraciones crecientes de RNA no es el mismo para cada una de las 20 sondas pertenecientes al grupo de control AFFX-BioB-5 **(Irizarry et al., 2003b)**.

Las causas de estas diferencias de reactividad entre sondas no son bien conocidas. Una de ellas puede ser el distinto porcentaje de guanina, citosina (%GC) entre distintas sondas. La **figura 3.2** muestra un diagrama de dispersión en el que cada punto representa una sonda del array *Human Exon 1.0*, ubicadas espacialmente en función de su mediana (eje x) y su rango (eje y) de expresión en un set de datos de 33 muestras (11 tejidos) publicado por *Affymetrix*.

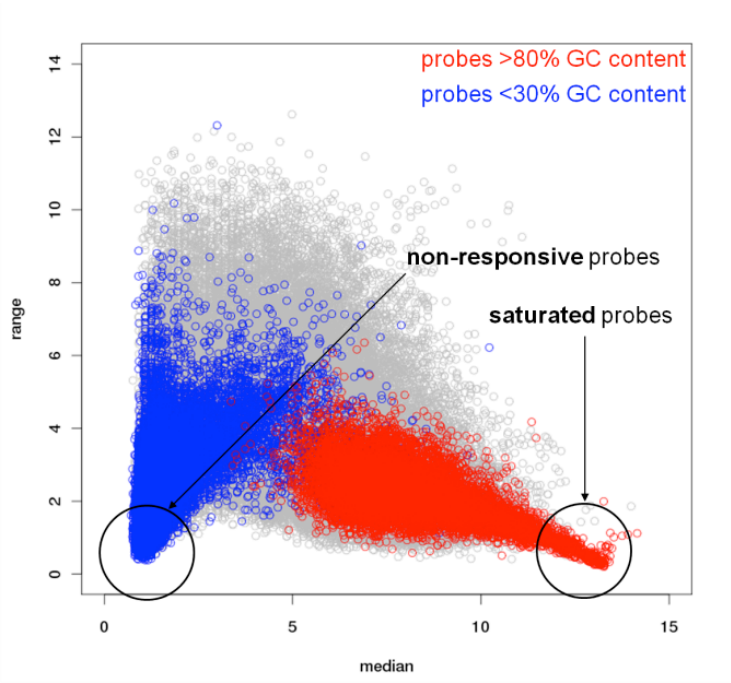


Figura 3.2. Sondas del array *Human Exon 1.0* situadas en función de su mediana y rango de expresión según el *set* de 33 microarrays publicado por *Affymetrix* (11 tejidos x 3 réplicas). Las sondas en **rojo** y en **azul** se corresponden con alto y bajo %GC respectivamente. Ambos grupos muestran características diferentes.

Los valores extremos de %GC se muestran en **rojo** para >80% y **azul** para <30%. La diferencia entre ambos grupos es claramente apreciable mostrando dos grupos separados con un solapamiento muy pequeño. Las sondas con alto %GC muestran una mayor expresión media, presentando una zona de saturación en donde la expresión es muy alta e invariante. Por el contrario, las sondas con bajo %GC tienen menor expresión y un rango mucho más variable, abarcando desde una zona de no respuesta, hasta niveles relativamente altos comparados con las de alto %GC. A nivel molecular los enlaces **G-C**, son más fuertes que los **A-T** debido a que su unión la componen 3 puentes de hidrógeno en lugar de los 2 de las uniones **A-T** **(Lewin, 2004).** Esta información combinada con los resultados del análisis de la **figura 3.2**, significaría que cuanto mayor %GC (zona **roja**) se observa una menor reactividad en el rango de señal y una mayor inespecificidad de las sondas con su *perfect match,* hibridando quizás más fácilmente con secuencias de RNA similares.

Un mismo gen puede tener sondas con una variación muy grande de % GC. En el ejemplo de la **figura 3.3** se muestran 2 probesets del array *Human Exon* (probesets número 2541755 y 2541739) que mapean sobre distintos exones del gen FAM49A. Ambos tienen sondas con un %GC muy diferente, por lo que sería de esperar que ambos exones muestren un comportamiento muy distinto en la señal del microarray.

```
Probeset 2541755
(110,1332) CCGCTCCGCCGCCGCCGAGTACGCC (84% GC)
(544,355)  CGAGTCGCCGCCGCCCTGCTCTGCC (80% GC)
(816,1064) GCCCGCTCCGCCGCCGCCGAGTACG (84% GC)
(20,1875)  GCCGCCCTGCTCTGCCGCCCGCTCC (84% GC)

Probeset 2541739
(964,702)  AAGTCAAGGTATCTAGCTGAAAAAG (36% GC)
(606,798)  CAGAAAAGTCAAGGTATCTAGCTGA (40% GC)
(726,644)  TGTGCTTAGCTTCGTGAAATGCTTC (44% GC)
(364,1048) TTCTGTCCTGTGCTTAGCTTCGTGA (48% GC)
```

Figura 3.3. Dos probesets del array *Human Exon 1.0* mapeando sobre el gen FAM49A muestran %GC muy dispares. Esto significa que los exones detectados tendrán un comportamiento muy distinto en la señal del microarray.

La **figura 3.4** muestra el perfil de expresión del gen FAM49A a lo largo del *set* de datos de 33 muestras de 11 tejidos distintos (procedentes de *Affymetrix*). En este gráfico, se puede ver el comportamiento de cada uno de los exones (en **gris** y **negro**) cuya expresión fue normalizada individualmente con *ExonMapper*, y el comportamiento global del gen (en **rojo**) normalizado con *GeneMapper*. La mayoría de los exones muestran una fuerte correlación entre sus perfiles, aunque se aprecian diferencias sustanciales en el nivel de expresión de unas muestras a otras. Sin embargo el exón ENSE00001488214 (en **negro**) muestra un perfil muy distinto, teniendo una expresión muy alta y totalmente plana, es decir, muy poco cambiante. Este exón está siendo detectado por el probeset 2541755 de la **figura 3.3** en donde todas sus sondas tienen un contenido GC >= 80%. Este ejemplo es un caso extremo de "efecto sonda", pero ilustra muy bien cómo puede influir en la expresión asignada a cada uno de los exones de un mismo gen.

Se ha observado que el %GC varía bastante dependiendo de la ubicación en el *locus* génico, siendo mayor en regiones próximas a 5' y haciéndose menor conforme se avanza al 3', probablemente debido a los mecanismos de regulación epigenética en el promotor de ciertos genes **(Bemmo et al., 2008)**.

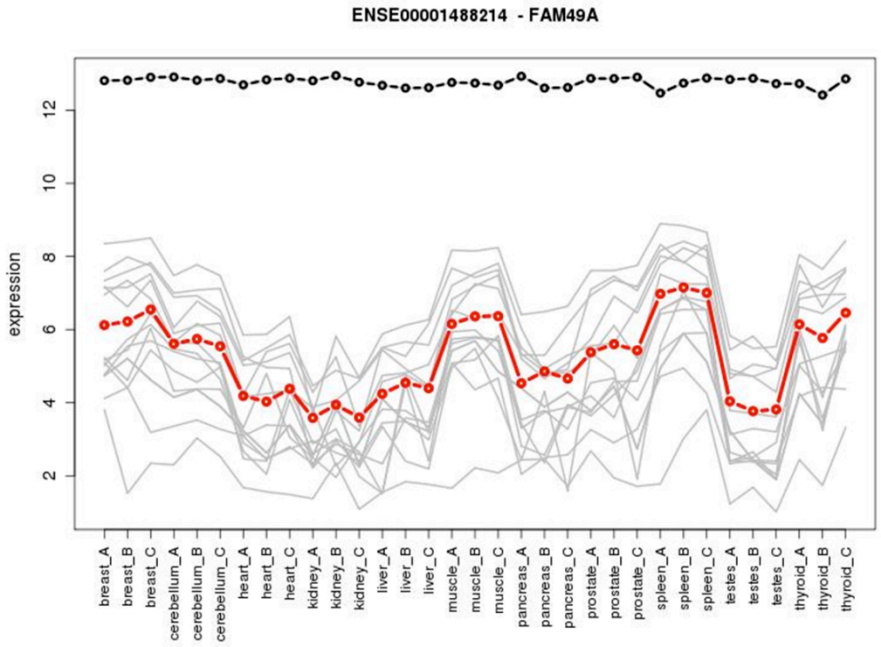


Figura 3.4. Perfil de expresión del gen FAM49A (**rojo**), de su exón ENSE00001488214 (**negro**) y del resto de exones (**gris**), sobre el set de datos de *Affymetrix*. La expresión del gen fue calculada con GeneMapper y la sus exones con ExonMapper. La mayoría de exones muestra una fuerte correlación aunque diferencias entre su nivel de expresión, sin embargo la línea **negra** muestra un exón con un comportamiento muy distinto. Este exón es detectado por sondas con un alto %GC lo cual satura su expresión.

Otra causa conocida del "efecto sonda" es la distancia del fragmento de transcrito reconocido por la sonda al extremo 3'. Esto es un problema reportado en los antiguos modelos IVT 3' y asociado a la tecnología de estos microarrays en los que las amplificaciones se realizan empezando por la cola poli-A (ubicada en 3'), lo cual puede provocar diferencias en la expresión detectada dependiendo de la distancia a 3' **(Auer et al., 2003)**. Aunque este desvío hacía el 3' no debería ser un problema para la tecnología de *Random Priming* utilizada en los microarrays de exones, el efecto de la amplificación podría introducir ciertas variaciones entre las distintas sondas.

La hibridación cruzada también puede ser causa de diferencias de comportamiento entre sondas. El uso de herramientas como ***GATExplorer***, asegura minimizar este efecto eliminando las sondas ambiguas que mapean en más de un gen.

Todo estos efectos laterales posibles asociados a las sondas hacen que la señal detectada por cada una de las sondas de los microarrays no sea producto únicamente de la cantidad de RNA presente en la muestra, que es lo que realmente se quiere medir. Por ello, es importante que los métodos de análisis de expresión comparen siempre las mismas sondas o conjuntos sumarizados de sondas entre las distintas muestras, pero eviten hacer inferencias sobre la expresión mezclando distintos tipos de sondas.

Métodos como el del algoritmo RMA y su extensión adaptada a la predicción de *splicing* alternativo (FIRMA) tienen en cuenta el efecto de la sonda para estimar el nivel de expresión del gen y de sus exones. Estos modelos funcionan para realizar comparaciones entre muestras gen a gen o exón a exón, pero pueden ser bastante erróneos al comparar ratios entre gen y exones ya que asumen que ambas señales tienen el mismo comportamiento. Como se ha

dicho, la comparación de entidades detectadas por distintas sondas puede ser bastante problemática y el único algoritmo reportado diseñado para afrontar esta dificultad es COSIE **(Gaidatzis et al., 2009)**.

3.2.3 Un nuevo método de análisis de *splicing*: *Exon Splicing using Linear Modeling* (ESLiM)

Como se ha descrito, el "efecto sonda" puede tener un peso grande en los datos procedentes de microarrays, por ello, las alteraciones que pueda producir deberían de ser tenidas en cuenta a la hora de analizar e interpretar datos de expresión. Sin embargo, esto no parece haber sido debidamente tratado en las publicaciones previas en el ámbito de detección de *splicing* alternativo con arrays de exones. El diseño de una estrategia para corregir dicho problema podría ayudar a mejorar los resultados obtenidos con este tipo de tecnología y ha sido la principal motivación para la realización de esta parte del presente trabajo de Tesis Doctoral.

La **figura 3.1** puso de manifiesto cómo las sondas reaccionan de distinta forma ante concentraciones crecientes de RNA mostrando distintas pendientes. De la misma forma, se puede tratar de hacer una estimación del efecto exón para la detección de *splicing* alternativo. Dado que en un *set* de datos cualquiera a analizar, no se puede conocer la cantidad exacta de RNA de cada exón presente en cada una de las muestras, este dato se puede aproximar mediante la expresión global del gen.

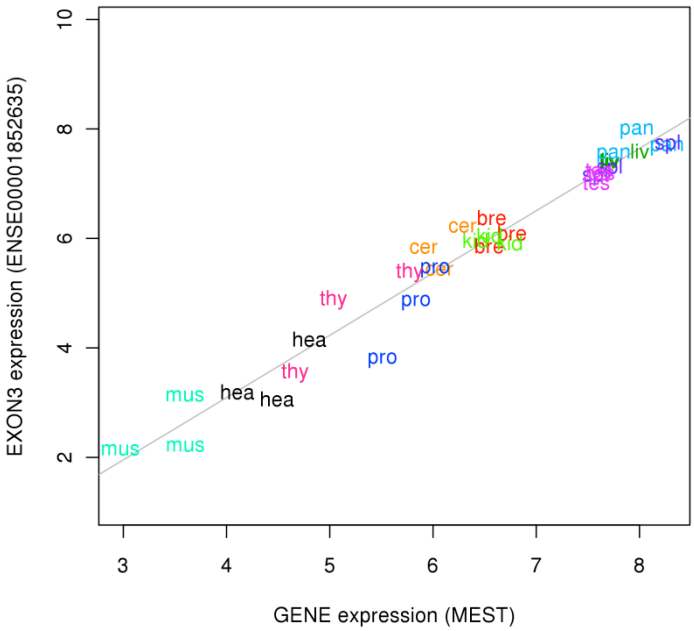


Figura 3.5. Relación lineal existente entre la expresión del gen MEST con la expresión de uno de sus exones (ENSE00001852635). Cada muestra del set de 33 microarrays de *Affymetrix* se representa por la abreviatura del nombre del tejido en inglés: *breast* (bre), *cerebellum* (cer), *heart* (hea), *kidney* (kid), *liver* (liv), *muscle* (mus), *pancreas* (pan), *prostate* (pro), *spleen* (spl), *testes* (tes) y *thyroid* (thy).

La **figura 3.5** muestra un diagrama de dispersión en donde se han ubicado los 33 microarrays del set de *Affymetrix* en función de la expresión del gen MEST y su exón número 3 (*Ensembl* Id: ENSE00001852635) calculadas con el algoritmo RMA utilizando *ExonMapper* y *GeneMapper*

por separado. Como se puede ver, la relación entre ambas expresiones es lineal correspondiendo más cantidad de exón cuando existe más cantidad de gen. De esta manera, se puede obtener la relación entre cada exón y su gen, lo cual es básico para las estrategias de *splicing* alternativo, independientemente de las características propias de sus sondas. Sondas más o menos reactivas tendrán más o menos pendiente pero guardarán una relación lineal concreta entre las distintas muestras reflejadas en gráficos como el presentado en la **figura 3.5**. A partir de estos modelos lineales entre cada exón y gen se puede observar si existen casos de desvíos de dicha línea, que se podrán interpretar como eventos de *splicing* alternativo.

De este modo, nuestra estrategia de detección de *splicing* alternativo se va a basar en encontrar desvíos sobre modelos lineales entre la expresión de cada exón con su gen. La ecuación de una línea recta se describe como:

$$y = a \cdot x + b$$

donde la variable y se calcula en función de la variable x multiplicada por una constante a que es el valor de la pendiente de la recta mas una constante b que es el valor de y cuando x es 0. Para modelar los datos procedentes de una observación como una línea se realiza lo que se llama una regresión lineal. y se aproxima en función de x calculando a como:

$$a = \frac{\sum (x_i - \bar{x}) \cdot (y_j - \bar{y})}{\sum (x_i - \bar{x})^2}$$

y b como:

$$b = \bar{y} - a \cdot \bar{x}$$

Anotando las variables acorde con sus categorías biológicas podemos estimar la expresión del exón como:

$$\hat{e}_{ijk} = s_{ij} \cdot g_{jk} + b_{ijk}$$

siendo $\hat{e}$ la expresión estimada del exón i perteneciente al gen j y a la muestra k. g es el valor de expresión del gen que es multiplicada a una pendiente dada para cada par gen, exón. Los desvíos sobre la línea de regresión se calculan como la expresión observada del exón menos la expresión estimada:

$$r_{ijk} = e_{ijk} - \hat{e}_{ijk}$$

de esta manera el valor "residual" r es la diferencia entre el valor de expresión real del exón e y su valor estimado $\hat{e}$. La **figura 3.6** muestra un ejemplo en donde un grupo de muestras pertenecientes a cerebelo, muestran una distancia vertical a la línea de regresión significativamente mayor que el resto. En este caso, el valor de los residuales para las muestras de cerebelo podría indicar presencia de *splicing* alternativo.

La ecuación final que modela la relación entre los valores de expresión del gen y los valores de expresión de los exones es la siguiente:

$$e_{ijk} = s_{ij} \cdot g_{jk} + b_{ijk} + r_{ijk}$$

Este modelo se hace de forma totalmente no supervisada teniendo posteriormente que realizar un test supervisado para encontrar diferencias significativas entre los residuales de las distintas categorías biológicas. Debido a sus fundamentos matemáticos, este método fue bautizado con el nombre en inglés de "***Exon Splicing by Linear Modeling Analysis***" (**ESLiM**).

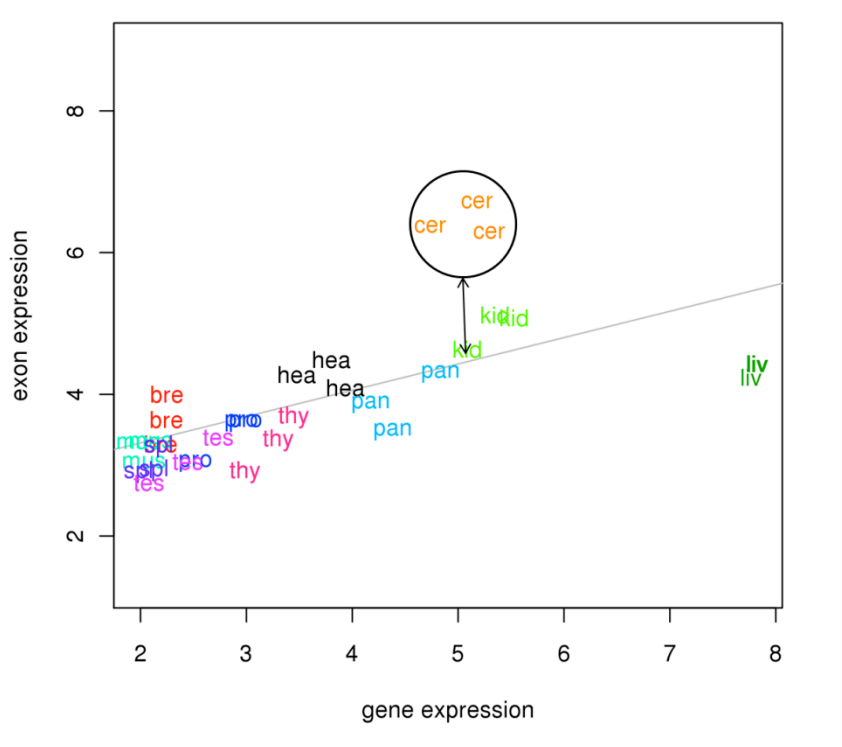


Figura 3.6. Los desvíos significativos entre el valor de expresión observado de los exones respecto a su valor estimado (línea de regresión) se interpretan como eventos de *splicing* alternativo.

3.2.4 Cálculo robusto de la expresión del gen

En las bases de datos genómicas, las estructuras de transcritos definidas para los distintos genes humanos a aumentado su complejidad enormemente en los últimos años (ver **capítulo 1**). Esto ha hecho que abunden transcritos de longitudes muy dispares y de distinta naturaleza como transcritos codificantes de proteínas y "pseudo-transcritos" conviviendo en el mismo *locus* génico. La estrategia definida en la sección anterior requiere de una medida de expresión del gen lo más estable posible ante eventos de *splicing* alternativo. Esto lleva a la necesidad de hacer una selección de las sondas conservadas y centrales que se utilizarán para calcular la expresión del gen en lugar de utilizar la totalidad que mapean en el *locus*.

Las sondas ubicadas en exones exclusivos de un solo transcrito –no presentes en otros transcritos del mismo *locus*– pueden hacer la expresión global del gen muy dependiente de la transcripción alternativa de dicho transcrito, lo cual resulta ser negativo para el cálculo de residuales con ESLiM ya que no dan una medida estable global del *locus* génico y comportan variabilidad. Por este motivo se decidió establecer el uso de un núcleo de sondas para cada gen con aquellas que mapean en todos sus transcritos reportados para dicho gen (según la base de datos *Ensembl*). Esta aproximación, a la que llamamos **ESLiM total** (**ESLiMt**), debería ser robusta pero tiene el inconveniente de que puede ser demasiado restrictiva, pudiendo resultar que muchos genes quedarían desiertos de sondas comunes a todos sus transcritos.

Para solventar el problema de falta de cobertura, posible en muchos *loci*, se estableció una variante del método con un criterio adicional menos restrictivo en el que se consideraron las sondas comunes a los transcritos incluyendo solamente los transcritos largos que cubren más del 60% del *locus*. Esta variación del método la llamamos **ESLiM *core*** (**ESLiMc**).

Recientemente se han incorporado al transcriptoma humano muchos transcritos cortos del tipo "procesados" o no codificantes de proteínas, con una baja evidencia biológica de existencia real. Este tipo de transcritos cortos no incluye muchas sondas ubicadas en exones bien anotados, y se consideran para calcular la señal del *locus* utilizando el método **ESLiMt** dan lugar a una pérdida importante de cobertura. Por ello en **ESLiMc** simplemente se han obviado estos transcritos eliminando aquellos menores a un 60% del tamaño del *locus*.

La **figura 3.7** muestra un ejemplo de los transcritos definidos en *Ensembl* v57 para el gen RGN. De los 6 transcritos, 4 son codificantes de proteína y 2 son transcritos procesados. De acuerdo con **ESLiMt** para el cálculo de la señal de expresión de este gen se considerarían únicamente algunas de las sondas ubicadas en el cuarto exón, mientras que con **ESLiMc** se incluirían todos los exones marcados en verde, lo cual supone incluir la mayoría de los exones codificantes, exceptuando uno, mas el UTR del extremo 3'.

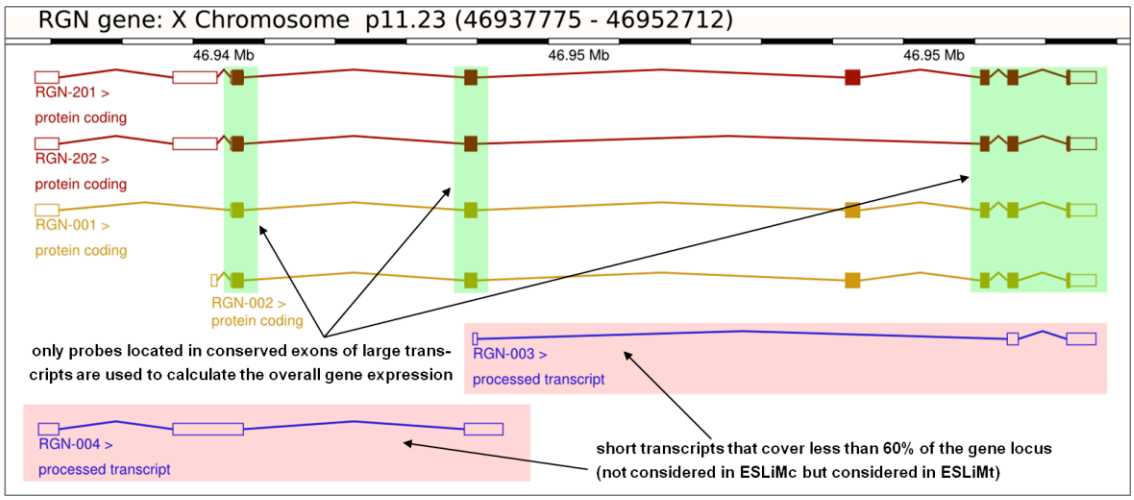


Figura 3.7. Transcritos definidos en la versión 57 de *Ensembl* para el gen RGN. Dos de los 6 transcritos son de longitud <60% del tamaño total del locus –pintados en **azul**– y no son considerados por el método **ESLiMc**, pero sí por el método **ESLiMt**. Los exones resaltados en verde son los que en este caso permitirían la selección de sondas según **ESLiMc** ya que son los exones comunes y conservados en los otros cuatro transcritos.

3.2.5 Minimización de falsos positivos producidos por el efecto sonda

Los exones, debido a que normalmente son de pequeño tamaño, están mapeados por un número limitado de sondas, normalmente 4 en el caso de los microarrays de exones de *Affymetrix*. Esto hace que la expresión de los exones detectada por los arrays esté muy expuesta al "efecto sonda" y aun no habiendo diferencias en la concentración de RNA real se pueda observar un comportamiento muy distinto entre ellos, haciendo también que el ratio entre un exón y su gen no sea constante cuando el gen cambia de expresión. Este efecto –no tenido en cuenta por ningún algoritmo previamente publicado salvo COSIE– puede producir una gran cantidad de falsos positivos en la detección de procesos de *splicing* alternativo.

En la **figura 3.8** se presenta un ejemplo de esta posible variabilidad no debida a cambios en la concentración de RNA mostrando un exón del gen MEST (con identificador de *Ensembl* ENSE00001607748) que no varía su expresión de forma significativa a lo largo de todo el conjunto de muestras (**figura 3.8.a**) como se demuestra por el *residual regression score* (**figura**

3.8.b). Sin embargo la expresión global del gen MEST sufre grandes variaciones en los distintos tejidos (**figura 3.8.c**), afectando a la relación de expresión entre gen y exón (**figura 3.8.d**). Ante esto, cualquier estrategia de análisis de *splicing* que no tenga en cuenta el desacuerdo en la dinámica de expresión de las distintas entidades transcripcionales (en este caso entre el exón y el gen), estará expuesta a la detección de falsos positivos.

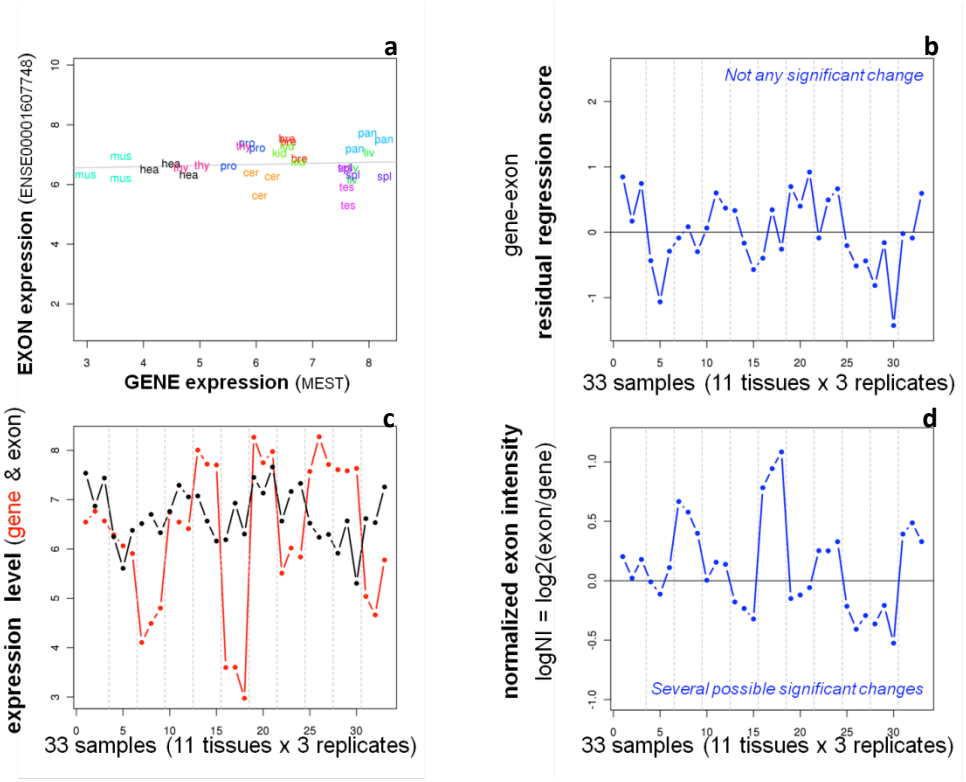


Figura 3.8. Resultados obtenidos siguiendo dos estrategias distintas en el análisis del exón ENSE00001607748 del gen MEST. El exón muestra un nivel de expresión sin diferencias significativas a lo largo de los 11 tejidos **(a)**, lo cual se refleja en sus residuales calculados con **ESLiMc (b)**. Cuando se analiza el ratio entre la expresión del exón y la del gen (c) –líneas **negra** y **roja** respectivamente–, se produce de forma artificial una modulación significativa entre tejidos **(d)**, debido a las variaciones de expresión del gen. Sin embargo, esto no significa que son debidas a un *splicing* alternativo en el exón pues la variación del mismo en los distintos tejidos no es significativa.

3.2.6 Detección de cambios específicos debidos a *splicing*

Como se ha indicado, el método **ESLiM** basa la detección de cambios específicos debidos a *splicing* alternativo detectando desvíos significativos entre la expresión observada de un exón en un tejido o muestra concreta respecto al comportamiento global de expresión relativa exón/gen en todas las muestras medido mediante un modelo de regresión lineal.

En la **figura 3.9** se presenta como el método **ESLiMc** siguiendo el modelo de regresión lineal (**figura 3.9.a**) detectó diferencias significativas entre el hígado y el cerebelo en el exón ENSE00001527616 del gen RGN (**figura 3.9.b**), mientras que en los otros tejidos no detecta ningún cambio significativo que refleje un evento de *splicing*. Por otro lado, mediante la comparación simple entre la expresión del gen y el exón (**figura 3.9.c**) al comparar los tejidos dos a dos se encuentran varios cambios importantes que reflejan posibles eventos de *splicing* (**figura 3.9.d**) y no queda claro cuáles son los más significativos. Como demuestra la figura, los resultados producidos por **ESLiMc** son menos ruidosos que los calculados con el logNI, y

muestran una señal clara que además es coherente con datos biológicos previamente publicados, en donde a RGN se le ha atribuido un evento de *splicing* diferencial entre los tejidos de hígado y cerebelo **(Wang et al., 2008)**.

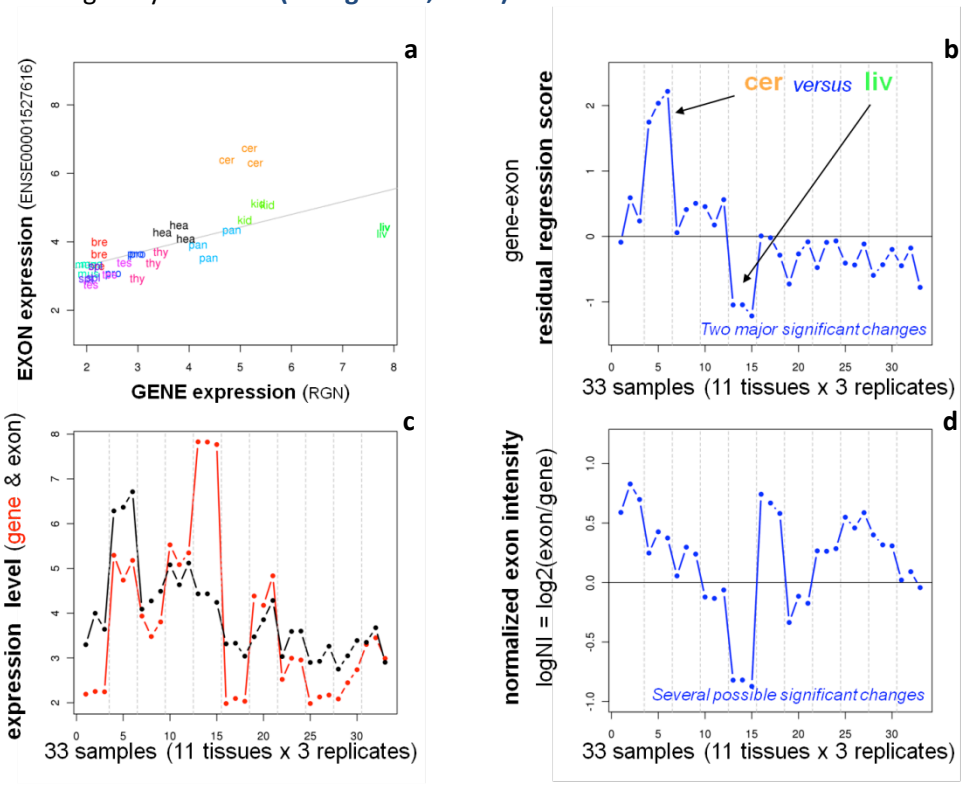


Figura 3.9. Resultados obtenidos siguiendo dos estrategias distintas en el análisis del exón ENSE00001527616 del gen RGN. El exón muestra desvíos en hígado y cerebelo produciendo las mayores diferencias cuando se comparan estos dos tejidos entre sí utilizando los residuales obtenidos por ESLiMc (a y b). Al analiza el ratio entre la expresión del exón y la del gen (c - líneas negra y roja respectivamente), se produce un resultado diferente mostrando diferencias entre varios pares de tejidos.

Una vez obtenidos por el método **ESLiM** los residuales para cada exón y cada muestra, se puede aplicar cualquiera de los test estadísticos clásicos para comprobar si las desviaciones observadas implican cambios significativos, y poder así asignar un valor de probabilidad (p-valor) a la hipótesis alternativa por la que se encuentran diferencias. Con este último paso, el método identifica eventos de *splicing* de modo robusto.

3.3. Resultados

Para poner a prueba el rendimiento del nuevo algoritmo descrito anteriormente en sus dos versiones **ESLiMt** y **ESLiMc**, se comparó en igualdad de condiciones con los algoritmos FIRMA, COSIE y ARH. Elegimos estos algoritmos por ser más modernos que los originales de *Affymetrix* y por proporcionar mejores resultados, tal y como se describe en sus publicaciones.

3.3.1 Implementación del algoritmo ESLiM

La implementación del algoritmo se realizó en R. Se realizaron versiones modificadas de *GeneMapper* seleccionando las sondas según lo descrito en **ESLiMt** y **ESLiMc** para el array

Human Exon 1.0. La lógica del algoritmo se implementó en un paquete llamado **ESLiM** que incluye las siguientes funciones:

- ***doLinearModel***: Realiza el cálculo de residuales a partir de las matrices de expresión a nivel de genes (medidos con la estrategia "total" de **ESLiMt** o por la estrategia "*core*" de **ESLiMc** ya explicadas) y a nivel de exones. Para este cálculo también se necesita el fichero de anotación de exones presente en ***GATExplorer*** donde se relaciona cada identificador de exón (ENSE) con el identificador del gen (ENSG).

- ***removeRedundantExons***: Elimina los exones de *Ensembl* que son redundantes, ya que –pese a ser casi solapantes y estar localizados en la misma región cromosómica y *locus* génico– tienen identificadores (ENSE ids) diferentes por variar su longitud en una pocas bases (bp) o tener distintos puntos de inicio o final UTR. Para eliminar esta redundancia, todos los identificadores de exones mapeados por el mismo conjunto de sondas de un microarray son agrupados en uno solo, tomándose como identificador único el primero (siguiendo el orden alfabético). Esta función elimina así muchos resultados que son totalmente redundantes (debido al problema de los ids descrito) que analizan exactamente la misma región exónica codificante. Esta función toma como entrada el listado generado por la función anterior: ***doLinearModel.***

- ***geneOriented***: Agrupa los distintos exones significativos colapsándolos en genes. Estos exones significativos han sido identificados previamente a partir de los residuales obtenidos por la función ***doLinearModel*** y tras el paso de eliminación de redundancias. La función toma como entrada los identificadores de exones, sus p-valores y el fichero de anotación de exones de ***GATExplorer***. Como salida esta función proporciona el porcentaje de exones alterados para cada gen respecto del total de exones detectables por el microarray, haciendo distinción entre exones codificantes y exones no codificantes de proteína. Además proporciona 2 valores de probabilidad: el p-valor más bajo de todos sus exones y la mediana de todos ellos.

Las listas procedentes de análisis de expresión diferencial a nivel de genes suelen proporcionar varios cientos o miles de entradas. A pesar de existir herramientas para realizar interpretaciones automáticas de estos listados, a menudo puede ser demasiada información para el investigador. Este problema se agrava con el manejo de listas de exones, que multiplica en más de 10 el tamaño de las listas de genes (ver **introducción, tabla 1**)**.** La orientación a genes que proporciona el paquete **ESLiM** en su salida final, permite una reducción del número de resultados significativos ya que apunta directamente a genes específicos como los mejores candidatos a sufrir *splicing*. Los resultados amplios –muchas veces masivos– obtenidos exón a exón también son proporcionados por el algoritmo, pero son puestos en un segundo plano. En todo caso la salida puede ser reordenada por el usuario por significación acorde a los distintos parámetros proporcionados por el algoritmo.

3.3.2 Comparativa de ESLiMt y ESLiMc con otros algoritmos para la búsqueda de *splicing* previamente publicados

La eficiencia de cada algoritmo fue medida utilizando como *set* de datos de comparación el *set* de *Affymetrix* de microarrays de exones de 11 tejidos humanos –con tres réplicas cada uno– combinado con los datos de *splicing* validados respecto a una serie de genes humanos en distintos tejidos **(Wang et al., 2008)**. Esta combinación suministró un conjunto final de 6 tejidos (ver **apartado 3.2.1**). Estos 6 tejidos fueron comparados 2 a 2 por cada uno de los 5

métodos analizados produciendo un total de 15 contrastes. Los algoritmos **ESLiMc**, **ESLiMt**, FIRMA y COSIE devuelven un valor por muestra y exón, por lo que solamente hizo falta ejecutar una vez cada uno de ellos utilizando posteriormente el paquete *limma* **(Smyth et al., 2012)** para realizar el ranking de exones por cada par de tejidos. ARH devuelve directamente un valor de significación por gen (ARH *values*), por lo que debió ejecutarse 15 veces, uno por cada contraste.

La cobertura sobre los genes con *splicing* conocido validado experimentalmente **(Wang et al., 2008)** por cada par de tejidos es muy similar en todos los algoritmos, exceptuando **ESLiMt** que, debido a lo restrictivo del método de selección de sondas, solamente tiene la capacidad de medir aproximadamente la mitad de los genes (ver **tabla 3.3**). La comparación entre hígado y músculo (LIV-MUS) no pudo medirse con **ESLiMt** ya que ninguno de los 2 genes validados en este contraste pudo mapearse según en el criterio de esta versión del algoritmo.

	ESLiMc	**ESLiMt**	**FIRMA**	**ARH**	**COSIE**	**Nº total de genes con *splicing* validado**
BRE-CER	**71**	**32**	75	75	73	75
BRE-HEA	**10**	**5**	11	11	9	11
BRE-LIV	**4**	**2**	5	5	5	5
BRE-MUS	**8**	**4**	8	8	8	8
BRE-TES	**16**	**9**	16	16	15	16
CER-HEA	**21**	**15**	22	22	21	22
CER-LIV	**12**	**5**	12	12	12	12
CER-MUS	**37**	**19**	38	38	38	38
CER-TES	**48**	**22**	52	52	51	52
HEA-LIV	**2**	**2**	2	2	2	2
HEA-MUS	**12**	**5**	13	13	13	13
HEA-TES	**7**	**5**	7	7	7	7
LIV-MUS	**2**	**0**	2	2	2	2
LIV-TES	**5**	**4**	5	4	4	5
MUS-TES	**14**	**4**	14	14	14	14
Nº de genes validados detectados	**269**	**133**	282	281	274	282
Nº total de genes	**37388**	**32039**	37567	39316	21504	
Nº de genes codificantes de proteína	**19346**	**14351**	20812	19871	17926	

Tabla 3.3. Número de genes validados por cada algoritmo y par de tejidos. Los números de los nuevos métodos presentados en este trabajo (**ESLiMc** y **ESLiMt**) figuran en **negrita**. **ESLiMt** es el método que menor cobertura presenta debido a una selección muy restrictiva de sondas para calcular la expresión del gen.

Para comparar la precisión de cada uno de los métodos se utilizaron curvas ROC (*Receiver Operating Characteristic*) para cada uno de los pares de tejido comparados. Las curvas ROC miden y comparan de modo gráfico la "tasa de verdaderos positivos" (TPR) o "sensibilidad" (VP/(VP+FN)) frente a la "tasa de falsos positivos" (TFP) o "1-especificidad" (FP/(FP+VN)) sobre un clasificador binario **(Draghici, 2003).** La comparación entre dos curvas se mide mediante el área bajo la curva (AUC), que varía entre 0.5 en caso de total aleatoriedad y 1.0 en caso de clasificación perfecta. Este análisis se realizó mediante el uso del paquete de R llamado ROCR **(Sing et al., 2005)**. La **figura 3.10** presenta en 15 paneles consecutivos las curvas ROC correspondientes a los contrastes de los 15 pares de tejidos comparados, incluyendo cada panel la curva correspondiente a los 5 métodos: **ESLiMc**, **ESLiMt**, FIRMA, ARH y COSIE.

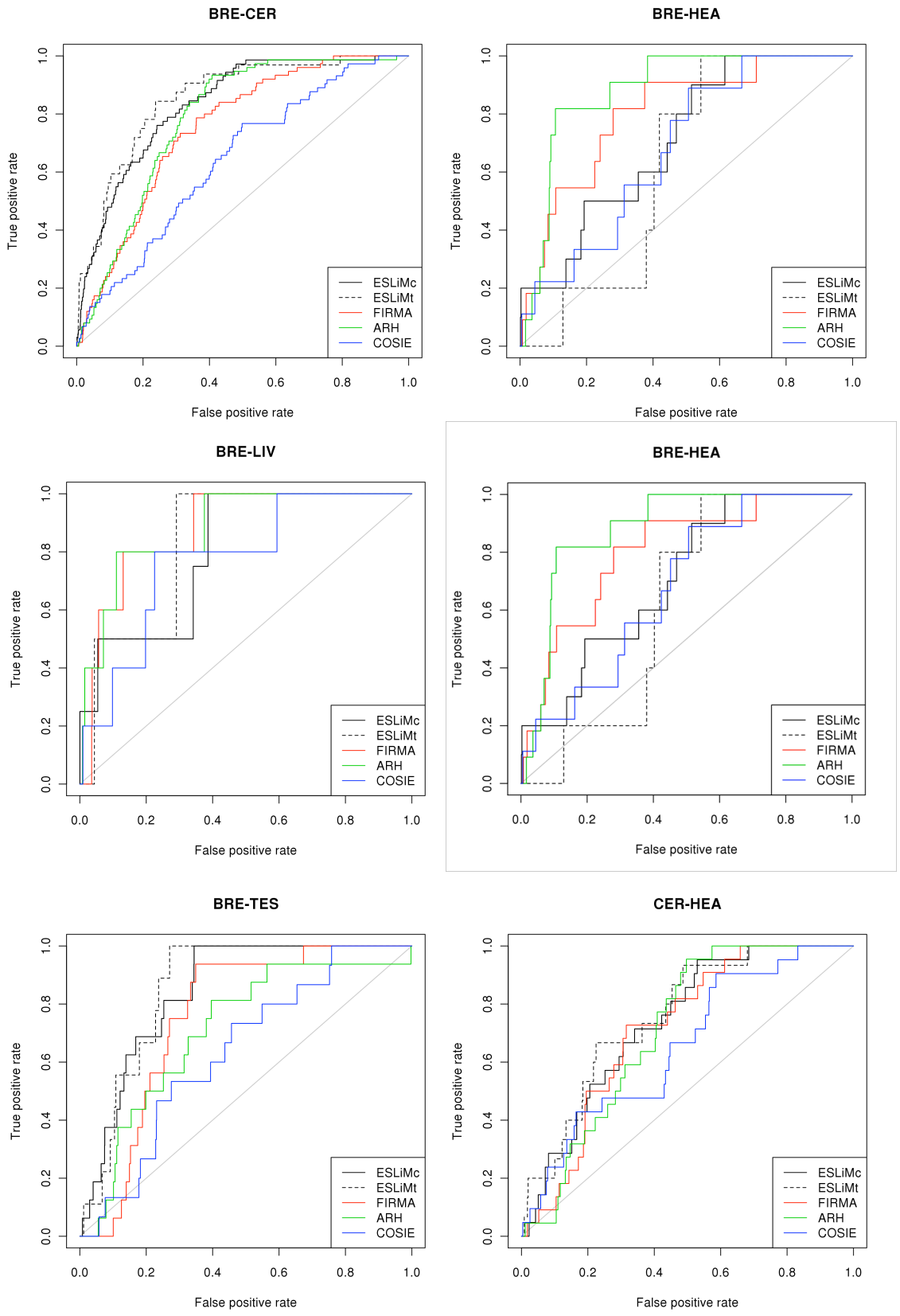


Figura 3.10. Curvas ROC comparando 5 métodos de detección de *splicing* alternativo.

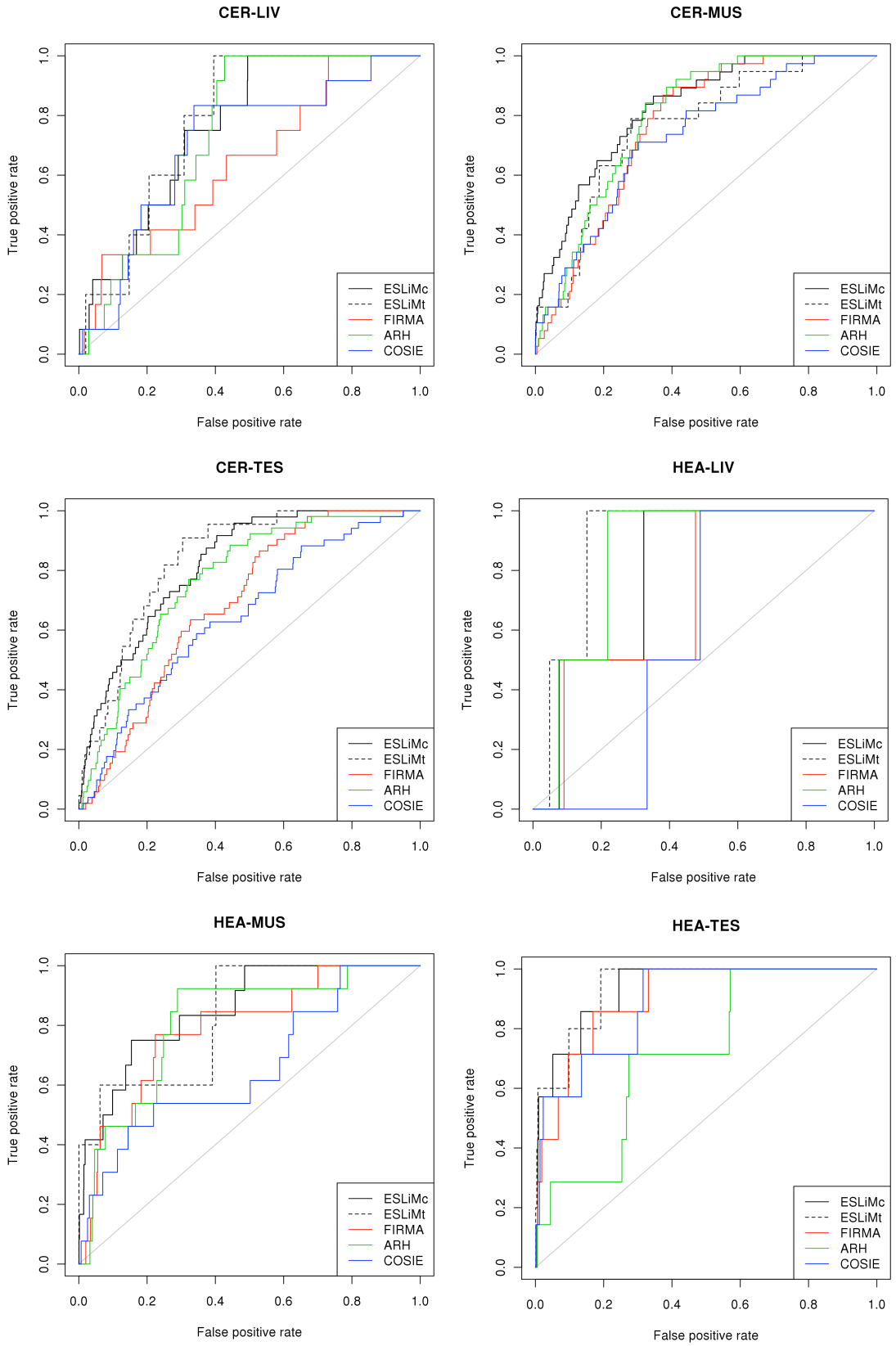


Figura 3.10 (continuación). Curvas ROC comparando 5 métodos de detección de *splicing* alternativo.

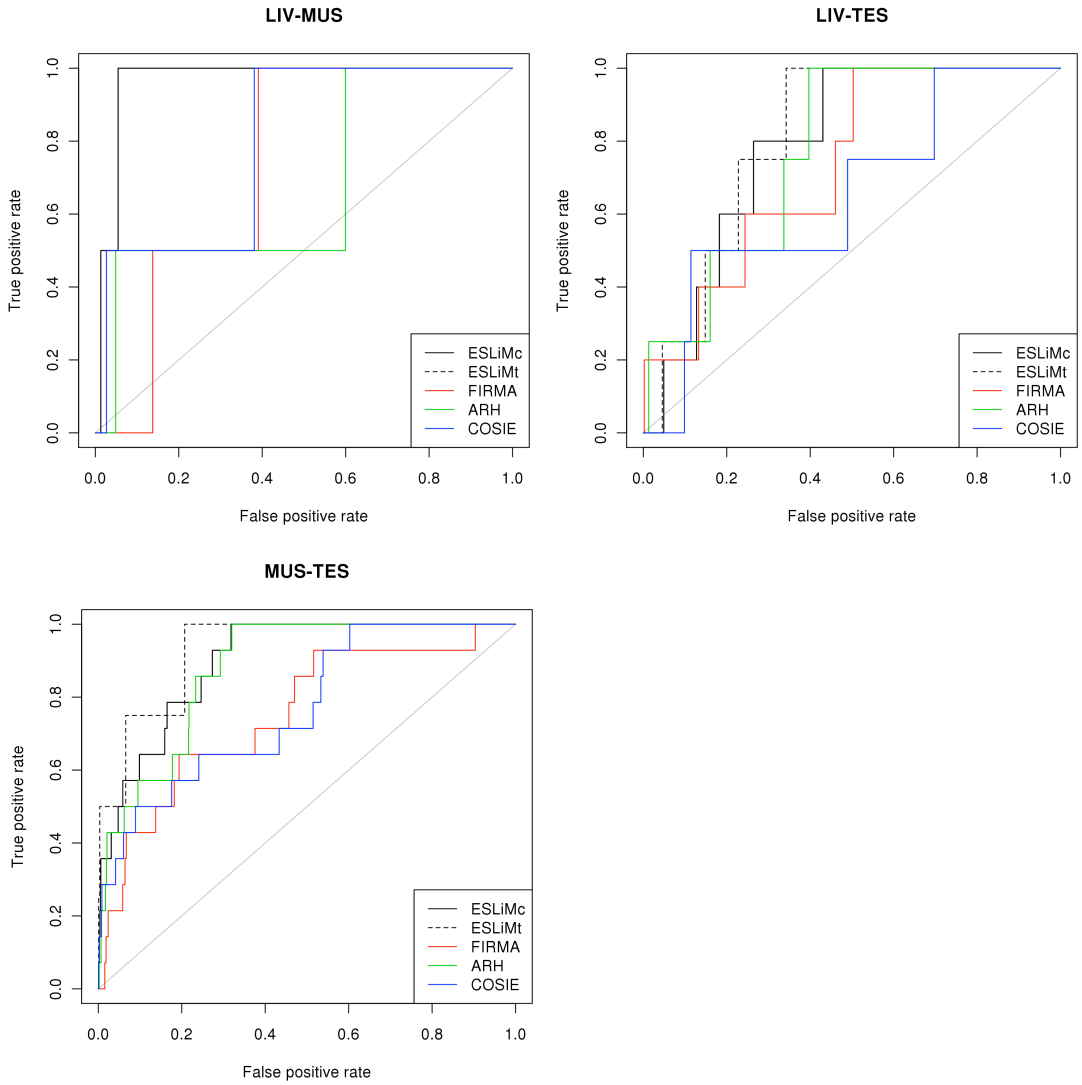


Figura 3.10 (continuación). Curvas ROC comparando 5 métodos de detección de splicing alternativo.

Las áreas debajo de la curva (AUCs) de cada comparación determinan qué método demuestra un mejor rendimiento. Los algoritmos **ESLiMc** y **ESLiMt** son los que muestran un mayor valor de AUC medido a partir del promedio de todos los pares de tejido. Este promedio se hizo de dos maneras distintas: la media simple y la media ponderada por número de genes validados. La media ponderada da mayor peso en el cálculo del promedio a los pares de tejido con un mayor número de positivos como BRE-CER (n=75), que a los de menor número de positivos como HEA-LIV (n=2); y quizás por ello sea una medida de comparación más adecuada. Ambos métodos de promediado dan un resultado similar (ver **tabla 3.4**). En todos los casos los algoritmos **ESLiMc** y **ESLiMt** resultan mejores que cualquiera de los otros tres comparados.

La **figura 3.11** muestra un gráfico de barras en el que se representa la media ponderada de las AUCs por cada método. Los dos mejores algoritmos son **ESLiMc** y **ESLiMt** con un rendimiento muy similar en torno 0,82. El siguiente algoritmo es ARH –con un AUCs promediada 0,782– seguido por FIRMA con 0,746 y finalmente más lejos por COSIE con 0,675.

	ESLiMc	ESLiMt	FIRMA	ARH	COSIE
BRE-CER	**0,834**	**0,854**	0,752	0,784	0,640
BRE-HEA	**0,708**	**0,625**	0,802	0,882	0,681
BRE-LIV	**0,804**	**0,833**	0,879	0,884	0,775
BRE-MUS	**0,894**	**0,872**	0,801	0,858	0,647
BRE-TES	**0,844**	**0,856**	0,756	0,708	0,636
CER-HEA	**0,737**	**0,758**	0,715	0,711	0,656
CER-LIV	**0,763**	**0,785**	0,645	0,735	0,705
CER-MUS	**0,825**	**0,763**	0,767	0,795	0,728
CER-TES	**0,820**	**0,840**	0,690	0,768	0,647
HEA-LIV	**0,799**	**0,897**	0,716	0,853	0,588
HEA-MUS	**0,854**	**0,829**	0,790	0,807	0,656
HEA-TES	**0,935**	**0,940**	0,901	0,717	0,886
LIV-MUS	**0,966**	**NA**	0,736	0,676	0,796
LIV-TES	**0,789**	**0,809**	0,732	0,773	0,650
MUS-TES	**0,899**	**0,931**	0,751	0,880	0,768
Media	**0,832**	**0,828**	0,762	0,789	0,697
Media ponderada	**0,823**	**0,818**	0,746	0,782	0,675

Tabla 3.4. Comparativa de AUCs de las ROCs alcanzada por cada par de tejidos y método de detección de *splicing* alternativo estudiados. Se han utilizado dos métodos para promediar los 15 contrastes: media simple y media ponderada por número de positivos (genes con *splicing* validado). Los dos métodos desarrollados en este trabajo (**ESLiMc** y **ESLiMt**) muestran un mejor rendimiento que algoritmos previamente publicados (FIRMA, ARH y COSIE). **ESLiMt** presenta un "NA" en la comparación LIV-MUS debido a que ninguna sonda mapea en los genes validados para par de tejidos (ver **tabla 3.3**).

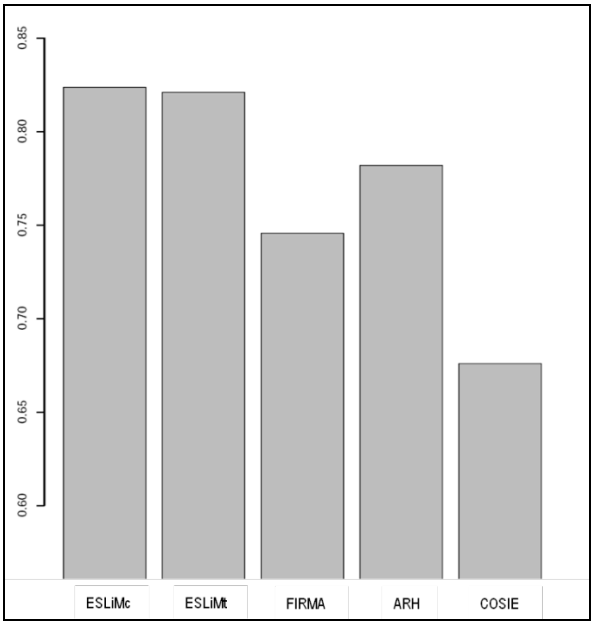


Figura 3.11. Media ponderada por número de genes validados de las AUCs provenientes de las curvas ROC de los 15 pares de tejidos. Los dos métodos desarrollados en este trabajo (**ESLiMc** y **ESLiMt**) muestran un mejor rendimiento que algoritmos FIRMA, ARH y COSIE.

Las curvas ROC son una buena técnica para comparar métodos, ya que dan una visión global clasificando todos los datos disponibles (en este caso todos los genes). Sin embargo, al tratar con datos de genómicos, la dimensión de las matrices es muy alta y sólo una pequeña parte

del inicio del ranking por significación será considerada por el investigador. Esto significa que la información realmente útil es la proporcionada a tasas bajas de falsos positivos. Por este motivo hemos hecho un estudio de la cantidad de verdaderos positivos eligiendo un criterio fijo para todos los métodos y pares de tejidos, como es considerar únicamente los primeros n genes más significativos eligiendo un umbral arbitrario de 1000 genes. Los resultados fueron diferentes a los obtenidos con las curvas ROC. En este caso los métodos **ESLiMc** y **ESLiMt** mostraron un comportamiento bastante distinto entre si, encontrando 61 y 29 genes respectivamente. El orden entre FIRMA, ARH y COSIE también se alteró, mostrando COSIE mostró mejor resultado que los otros dos métodos (ver **tabla 3.5** y **figura 3.12**).

	ESLiMc	ESLiMt	FIRMA	ARH	COSIE
BRE-CER	**17**	**8**	6	6	10
BRE-HEA	**2**	**0**	2	1	2
BRE-LIV	**1**	**0**	0	2	1
BRE-MUS	**3**	**1**	2	3	2
BRE-TES	**1**	**1**	0	0	0
CER-HEA	**1**	**3**	1	1	2
CER-LIV	**1**	**1**	1	0	1
CER-MUS	**10**	**3**	3	5	6
CER-TES	**10**	**5**	1	4	3
HEA-LIV	**0**	**0**	0	0	0
HEA-MUS	**5**	**2**	1	0	3
HEA-TES	**4**	**3**	3	1	4
LIV-MUS	**1**	**0**	0	0	1
LIV-TES	**0**	**0**	1	1	0
MUS-TES	**5**	**2**	3	6	5
Suma	**61**	**29**	24	30	40

Tabla 3.5. Número de genes con *splicing* validados en las distintas comparaciones entre tejido detectados en los 1000 genes más significativos. Se comparan 5 métodos de detección de *splicing* en 15 comparaciones binarias.

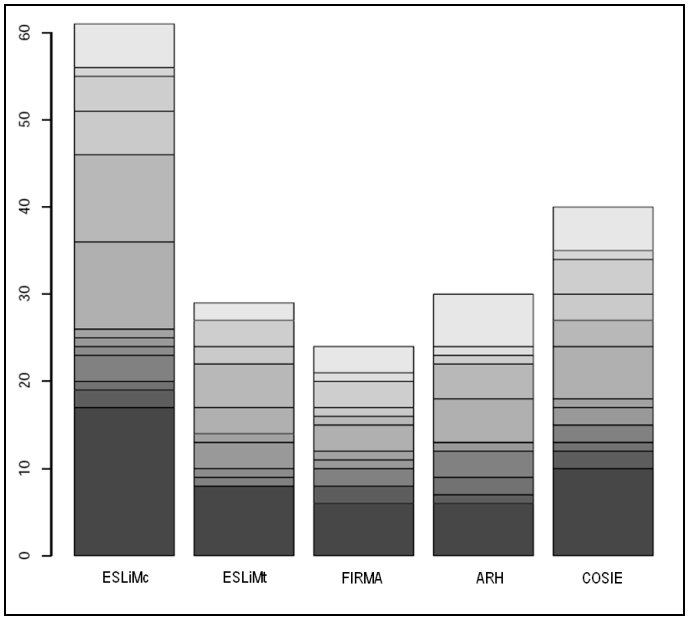


Figura 3.12. Diagrama de barras indicando el número de genes con *splicing* validados en las distintas comparaciones entre tejido detectados en los 1000 genes más significativos. Se comparan 5 métodos de detección de *splicing* en 15 comparaciones.

3.4. Discusión y posible trabajo futuro

El *splicing* alternativo es un proceso biológico fundamental para el entendimiento del funcionamiento de los genes en los distintos tipos celulares y de las distintas funciones que pueden tener las isoformas de genes y proteínas en un organismo.

En el trabajo del **capítulo 1** de la presente Tesis Doctoral, se trató de mejorar los análisis de microarrays de oligos de alta densidad con ***GATExplorer***. En el **capítulo 2**, se analizaron datos de expresión génica y de miRNAs para obtener biomarcadores en distintos tipos de cáncer. Después de esto, en este **capítulo 3**, se ha intentado avanzar en los análisis de genómicos yendo más allá de medir la expresión global del gen y tratando de desarrollar un algoritmo aplicado a la detección de *splicing* alternativo a partir de datos de microarrays de exones.

Al utilizar diversos algoritmos destinados a la detección de *splicing* basados en la tecnología de microarrays de exones, se identificó un problema no resuelto de forma satisfactoria por la mayoría de ellos. Las características propias de cada sonda de oligonucleótidos hacen que no pueda compararse directamente la expresión individual de cada exón con la expresión global del *locus* génico que lo contiene, produciendo falsos positivos. El nuevo método diseñado en este **capítulo 3**, **ESLiM**, utiliza la totalidad de las muestras para trazar una regresión lineal que enfrenta la expresión de cada exón con la de su gen, permitiendo calcular el comportamiento de la expresión del exón ante los cambios de expresión del gen. Al utilizar todas las muestras para el cálculo de residuales, se esperan resultados más fiables en un *set* de datos con un número alto de muestras y tipos biológicos diferentes. Además, la utilización del mapeo realizado en ***GATExplorer***, asegura una selección de sondas biológicamente coherente para calcular la expresión de genes y exones.

La comparativa entre algoritmos utilizando un conjunto de genes con *splicing* previamente validado determinó que el método **ESLiM** supera en precisión a las otras estrategias previamente publicadas. Las curvas ROC mostraron que las dos estrategias diseñadas (**ESLiMc** y **ESLiMc**) obtuvieron un promedio de AUC muy similar entre ellos y superior al de los otros algoritmos. Sin embargo al restringir el conjunto de genes a los de mayor significación estadística, en lugar de analizar su totalidad, los resultados fueron distintos. Esta vez **ESLiMt** se mostró inferior a **ESLiMc**, detectando aproximadamente la mitad de verdaderos positivos. El hecho de que **ESLiMt** sea un método muy restrictivo de selección de sondas, hace que no sea posible la detección de la expresión de multitud de genes. **ESLiMc** mapea aproximadamente el mismo número de genes que los otros algoritmos publicados, pero encuentra un número superior de verdaderos positivos entre los genes detectados a los que asigna mayor significación estadística. Esto es particularmente importante para el investigador, que habitualmente trabaja únicamente con los genes que muestran los mejores p-valores, y que tratará de validar experimentalmente los resultados.

El trabajo futuro en este tema podría ir encaminado en la interpretación biológica del proceso de *splicing* medido a nivel genómico, tratando de identificar los procesos en donde el *splicing* alternativo es más relevante y viendo si el grado de *splicing* está uniformemente distribuido entre todos los genes. El análisis del *spliceosoma* mediante la correlación entre los exones que sufren *splicing* y la expresión de los genes regulatorios llamados *splicing factors*, puede ser también útil para profundizar en el entendimiento de este proceso. Finalmente, un análisis de las secuencias de los intrones regulados bajo un mismo gen, o conjunto de genes reguladores, podría identificar motivos conservados que funcionan como regiones de unión al DNA.

Capítulo 4
Análisis de coexpresión de genes y estudio evolutivo de genes específicos de tejido y genes *housekeeping* en tejidos humanos sanos y en cáncer

4.1. Introducción

Los genes son en ocasiones tratados como entidades independientes pero no trabajan de forma aislada, sino que cooperan unos con otros para el correcto desarrollo de los distintos procesos biológicos que tienen lugar en la célula **(Nowak, 2006)**. Las relaciones biomoleculares entre genes suelen presentar una presión evolutiva común que se traduce un alto grado de conservación **(Jordan et al., 2004; Tirosh et al., 2006)** que, normalmente, se corresponde con el mantenimiento de funciones celulares o fisiológicas importantes. Los productos de los genes –i.e. las proteínas– son máquinas moleculares que interaccionan físicamente entre sí de modo específico produciendo efectos biológicos concretos de activación, inhibición, inducción, etc. Muchas veces varias proteínas forman complejos macromoleculares implicados en procesos celulares concretos **(Uetz et al., 2000)** como son división celular, diferenciación, señalización, transcripción génica, traducción de proteínas, etc. La co-regulación entre genes puede ser llevada a cabo por factores de transcripción comunes que dirigen la expresión de un conjunto de genes para realizar una tarea determinada **(Spiegelman and Heinrich, 2004)**. Los análisis de coexpresión proporcionan información de genes que están bajo una misma regulación, y que pueden variar según las condiciones del medio o el tipo celular donde están, haciendo posible la identificación de relaciones transcripcionales entre ellos. La construcción de redes de coexpresión calculadas a nivel genómico de escala global se ha demostrado muy útil a la hora de descubrir interacciones entre genes, y permite asociar grupos de genes a una determinada función y mejorar la caracterización de distintos procesos biológicos **(D'Haeseleer et al., 2000; van Noort et al., 2004)**.

El primer propósito de esta parte del trabajo es obtener una red de coexpresión de genes humanos utilizando datos transcriptómicos que permita identificar distintos grupos de genes con una función común. Para minimizar el nivel de ruido se seleccionará de forma rigurosa un *set* de datos de tejidos sanos hibridados con la tecnología de microarrays de expresión –ya descrita en los anteriores capítulos de esta Memoria–, que serán analizados mediante una combinación de técnicas robustas utilizando el re-mapeo obtenido en ***GATExplorer*** y descrito en el **capítulo 1** de la presente Tesis Doctoral.

4.1.1. Genes específicos de tejido (TSG) y genes *housekeeping* (HKG)

Desde el punto de vista transcripcional y de los perfiles de expresión en las distintas partes de un organismo complejo (i.e. en metazoos pluricelulares), existen dos tipos principales de genes que son antagónicos: **(i)** los genes específicos de tejido (***tissue-specific genes,*** **TSG**) y **(ii)** los genes *housekeeping* (***housekeeping genes,*** **HKG**) (ver **figura 4.1**). Los TSG son aquellos que se transcriben en un único tejido, manteniéndose silenciados en el resto de tejidos de un organismo pluricelular. Es de esperar que estos genes estén presentes solamente en células muy diferenciadas del organismo específicas de tipos celulares concretos que cumplen funciones propias de dichas células especializadas (p. ej. gen de la insulina producido por las células *beta* de los islotes de *Langerhans* del páncreas). Por el contrario, los HKG se encuentran transcritos bajo cualquier condición y en cualquier célula del organismo. Estos genes desempeñan funciones básicas esenciales y necesarias para el mantenimiento celular y para la supervivencia de la célula (p. ej. genes ribosomales o genes del citoesqueleto que están presentes en todas las células) **(Butte et al., 2001)**. La diferencia entre estos dos tipos de genes y sus características será estudiada en el presente trabajo mediante la construcción y análisis de perfiles de expresión detallados en un amplio *set* de tejidos/órganos, así como la construcción de redes transcripcionales y el análisis de su ubicación y topología en dichas redes.

4.1.2. Conservación y evolución de los genes

Con el paso del tiempo, de generación en generación y a través de la evolución de las distintas especies, los genomas van acumulando mutaciones en su secuencia. Sin embargo, existe una presión evolutiva para mantener conservadas las estructuras funcionales del genoma que se expresan –i.e. genes– que dan lugar a productos génicos activos en las células, así como los sitios de unión de factores de transcripción y otras regiones genómicas de regulación. A nivel de proteína el grado de conservación de las secuencia de aminoácidos tiende a ser mayor que en el caso del DNA, pudiendo permanecer sin cambios incluso cuando cambia la secuencia de DNA codificante –en las llamadas mutaciones silenciosas– o cuando los cambios son a aminoácidos de las mismas características –mutaciones sinónimas–. El mayor grado de conservación suele darse para los dominios de las proteínas y su estructura terciaria, que puede mantenerse invariable existiendo aún cambios leves en la estructura primaria –es decir, en la secuencia de aminoácidos–. De esta manera, los genes con funciones similares tienden a mantener parecido en sus secuencias entre las distintas especies indicando un posible origen común (genes ortólogos). En este trabajo se analiza el grado de conservación evolutiva de los distintos genes humanos para medir la relación entre estructura y funcionalidad, obteniendo diferencias entre los TSG y los HKG, y analizando su comportamiento en la red de coexpresión.

4.1.3. Conservación y evolución en los genes alterados en cáncer

En el desarrollo del cáncer y la transformación de las células normales en células tumorales malignas sucede un proceso de acumulación de mutaciones en los genes similar al descrito en los estudios de evolución de genes y familias génicas. Por este motivo también se estudiarán los genes alterados en cáncer desde un punto de vista evolutivo. Aunque las mutaciones se acumulan en el DNA, estos cambios tienen repercusión en el número y tipo de genes desregulados que aumentan o inhiben su expresión de forma significativa. Un análisis de los genes perdidos y ganados durante el proceso de carcinogénesis considerando datos genómicos transcriptómicos e incluyendo la perspectiva evolutiva del distinto grado de conservación de

los genes y la perspectiva de especificidad o esencialidad de dichos genes, puede ayudar a encontrar nuevas explicaciones respecto a las causas de esta enfermedad compleja, cuyo estudio a nivel molecular suele limitarse en analizar la funcionalidad asociada a genes o grupos de genes relacionados con cáncer (i.e. oncogenes o genes supresores).

4.2. Materiales y métodos

4.2.1. Conjunto de datos seleccionado para perfiles de expresión

En varios meta-estudios previos realizados en el campo de la expresión y relación entre genes usando datos genómicos, es habitual la utilización de múltiples *sets* de datos analizados conjuntamente abarcando muestras de distintos tipos celulares y distintos tejidos, incluyendo réplicas biológicas por cada tipo o subtipo e incluso incluyendo muestras de tejidos patológicos **(Griffith et al., 2005; Lee et al., 2004)**. Como es sabido, en enfermedades complejas como el cáncer en el que la genética juega un papel determinante, es común encontrar genes cuya expresión está desregulada y se muestra altamente variable de un paciente a otro, lo cual introduce un ruido significativo en los análisis y hace que la señal específica del estado biológico en estudio se pueda difuminar. En este trabajo, se trató de evitar dicho ruido mediante una cuidada selección de las muestras biológicas que se van a comparar, empezando por seleccionar primeramente sólo muestras procedentes de tejidos sanos. El *set* de datos utilizado contenía originalmente 353 microarrays modelo *Affymetrix* U133 Plus 2.0 (GEO ID GSE3526) **(Roth et al., 2006)** incluyendo 65 tejidos distintos, 20 de los cuales pertenecían a tejidos del sistema nervioso central y 45 a otro tipo de tejidos. El propósito de los autores de este trabajo era caracterizar el tejido nervioso obteniendo una firma molecular común. Sin embargo, para la construcción de una red de coexpresión de genes humanos, esta proporción de tejido nervioso podría introducir un sesgo importante hacia genes expresados y no expresados en este tipo de tejidos respecto al resto. Para evitar esto, se hizo una selección de las muestras equilibrada, tratando de evitar los sesgos hacia un tipo tisular determinado, lo que derivó en un subconjunto de 32 tejidos (ver **tabla 4.1**). Por cada uno de esos 32 tejidos se seleccionaron 3 réplicas haciendo un total de 96 microarrays. La selección de estas muestras se hizo de tal manera que las réplicas biológicas se agrupasen correctamente en un test de agrupación no supervisado. La **figura 4.1** muestra un *heatmap* normalizado con RMA en donde las muestras están agrupadas de 3 en 3, como corresponde a las réplicas biológicas. Esto asegura que el *set* de datos seleccionado es óptimo para la realización de un análisis de coexpresión en el que se ha minimizado tanto el ruido biológico como el ruido técnico.

Adipose tissue omental	*Heart ventricle*	*Oral mucosa*	*Stomach fundus*
Adrenal gland cortex	*Kidney cortex*	*Ovary*	*Stomach pyloric*
Bone marrow	*Kidney medulla*	*Pituitary gland*	*Testes*
Bronchus	*Liver*	*Prostate gland*	*Thyroid gland*
Cerebellum	*Lung*	*Salivary gland*	*Tongue*
Cerebral cortex	*Lymph nodes*	*Saphenous vein*	*Tonsil*
Esophagus	*Mammary gland*	*Skeletal muscle*	*Trachea*
Heart atrium	*Medulla*	*Spleen*	*Urethra*

Tabla 4.1. Relación de tejidos/órganos humanos utilizados en esta parte del trabajo procedentes de **(Roth et al., 2006)**. Para eliminar un posible sesgo respecto a tejidos y sistema nervioso –que es ampliamente estudiado en el trabajo de Roth *et al.*–, se seleccionó un subconjunto de 32 tejidos/órganos respecto al *set* completo original.

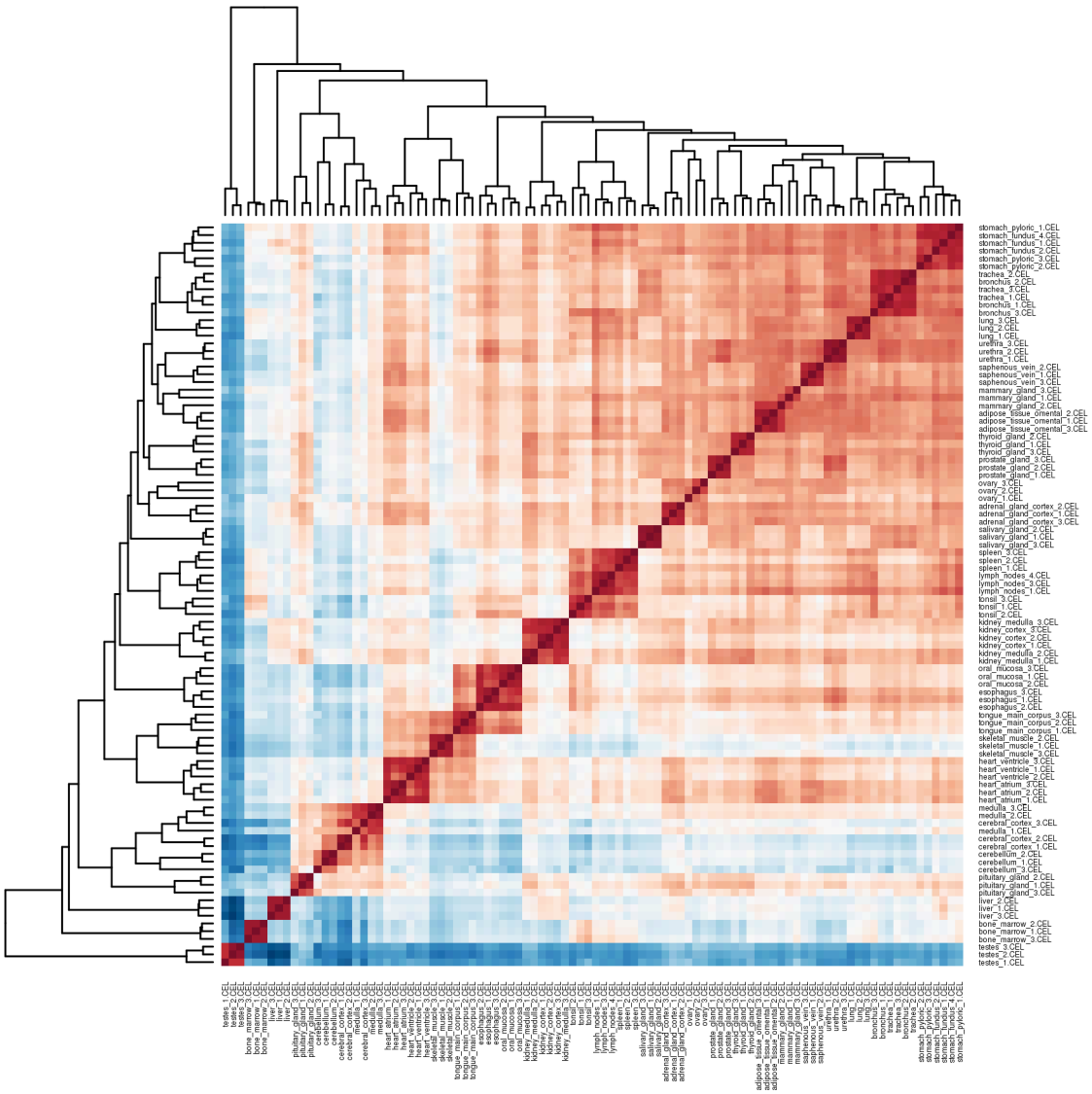


Figura 4.1. Análisis de agrupación (*clustering*) jerárquico no supervisado del conjunto de datos de expresión global –obtenida con microarrays– de 32 tejidos humanos incluyendo 3 réplicas biológicas de cada uno. La figura, de tipo mapa de color (*heatmap*), muestra cercanía y similitud entre las réplicas –todas agrupadas correctamente–, asi como agrupación relativa de tejidos que tienen un origen y función fisiológica común, como por ejemplo los asociados al sistema linfático humano: amígdalas (*tonsils*), bazo (*spleen*) y nodos linfáticos (*lymph nodes*) que aparecen bien agrupados en el centro de la figura.

4.2.2. Métodos de normalización de muestras y medidas de correlación entre genes

A la hora de calcular la expresión de los genes en cada muestra y su correlación se ha elegido una combinación de métodos paramétricos y no paramétricos. Los métodos paramétricos asumen que la población sigue una distribución normal, mientras que los no paramétricos se basan en realizar un ranking de los datos. Según varios autores **(Bolstad et al., 2003; Lim et al., 2007)**, la inclusión de al menos un método paramétrico en cada análisis asegura una mayor robustez en el resultado. En este caso se ha elegido como métodos de normalización RMA (no paramétrico) y MAS5 (paramétrico), y los métodos para calcular las correlaciones fueron la correlación de *Pearson* (paramétrico) y la correlación de *Spearman* (no paramétrico).

Utilizando el *set* de datos descrito en el apartado anterior, se calculó la señal de los microarrays utilizando los algoritmos RMA y MAS5 y realizando ambas normalizaciones por separado con el CDF *GeneMapper* para el array *HG-U133 Plus 2.0* (contenido en ***GATExplorer***). Como se ha explicado en el **capítulo 1**, la utilización de este re-mapeo alternativo al original de *Affymetrix* basado en *probesets*, implica la reducción de ruido posible por hibridación cruzada y proporciona un nivel de expresión para cada gen siguiendo una anotación actualizada en base a *Ensembl*. Después de la normalización, en lugar de realizar los cálculos de correlación con todos los genes detectables por el microarray, se realizó un filtrado previo de los genes "poco informativos" con el objetivo de aumentar la sensibilidad del estudio. Como es sabido, incluso cuando se manejan grandes conjuntos de datos, hay una proporción de genes que no se expresan en ninguna condición, manteniendo su señal en niveles de ruido **(Calza et al., 2007)**. Además, la tecnología de microarrays de expresión de *Affymetrix*, al estar basada en la hibridación de cadenas de oligos de 25 nucleótidos, presenta multitud de sondas que no son reactivas ante la presencia de su RNA complementario o son demasiado inespecíficas permaneciendo su señal saturada y por tanto exhibiendo niveles altos de expresión en todas las muestras. Por ello, se eliminaron de la matriz de expresión los genes que cumplieron las dos siguientes condiciones:

1. **Genes** cuya **diferencia de expresión** (o variabilidad entre dos grupos de muestras cualesquiera) esté por debajo de la mediana de diferencia de todos los genes.
2. **Genes** cuya **expresión media** esté por debajo de la mediana de la expresión de todos los genes.

El filtrado de estos genes "poco informativos" que no cambian suficientemente en ninguna condición solo se realizó para la matriz normalizada con RMA ya que, como demostró un trabajo previo **(Prieto et al., 2008),** el análisis de correlación realizado tras combinación de los algoritmos MAS5+*Spearman* no requiere tal filtrado previo al no cambiar nada los genes antes o después de tal filtrado. La matriz de expresión original contenía 17582 genes de los cuales se filtraron 4979 (28,3%) reduciendo la matriz a 12603 genes.

Como ya se ha indicado, para determinar los pares de genes que tienen una alta similitud en sus perfiles de expresión se utilizaron los métodos de correlación de *Pearson* y de *Spearman*. El umbral considerado para determinar si existe una similitud significativa en la expresión de dos genes se situó en un coeficiente de correlación r >= 0.70 . Finalmente, en el trabajo de Prieto *et al.* **(Prieto et al., 2008)** se revela que las correlaciones negativas son en su mayoría un artefacto y no son consistentes ni reproducibles, por ello no son consideradas en este estudio.

La representación gráfica de la red se realizó mediante la herramienta bioinformática *Cytoscape* **(Shannon et al., 2003; Smoot et al., 2011)**. Este programa de código abierto permite construir y visualizar redes derivadas de experimentos biomoleculares, pudiendo ampliarse también con distintas utilidades para análisis de las redes (ver *plugins* de *Cytoscape*).

4.2.3. Identificación de genes *housekeeping* (HKG)

En los últimos años, se han llevado a cabo experimentos transcriptómicos a gran escala que han aportado gran cantidad de datos sobre los genes humanos. Esta información puede ser útil a la hora de identificar HKG, ya que son los pilares que mantienen los procesos biológicos básicos para la supervivencia de la célula. La comunidad científica asume que los HKG son esenciales para cualquier tipo de célula, sin embargo, no existe aún una lista definitiva de estos genes de referencia para el organismo humano. En este trabajo hemos desarrollado una

estrategia bioinformática para identificar HKG utilizando como fuente de datos la información presente en *GenBank* (**http://www.ncbi.nlm.nih.gov/genbank/**) **(Benson et al., 2010)**. Se utilizaron las librerías de secuencias expresadas (ESTs, versión de 2010) que fueron almacenadas en una base de datos junto con la información individual de cada secuencia, asociándole además el identificador del gen y el tejido en donde se encontró. Esto devolvió un total de 604546 ESTs asociados a 18832 genes, ubicados en 34 tejidos distintos. Debido a que algunos tejidos tenían asociados un número muy bajo de ESTs (p. ej. nodo linfático 2, bronquio 6, timo 7, etc), fue necesaria la aplicación de un filtro de tejidos que se realizó eliminando aquellos con un número arbitrario elegido de ESTs inferior a 1000, lo que resultó en la eliminación de 8 tejidos escasamente representados. Posteriormente se seleccionaron aquellos genes con ESTs presentes en al menos 18 tejidos distintos (≈70%) devolviendo un total de 480 genes. Para completar esta información en términos de cobertura y fiabilidad, se recogieron también 3 listados de HKG previamente realizados por otros autores:

1. **HKG** humanos con expresión altamente correlacionada en perfiles obtenidos a partir de expresión en múltiples tejidos humanos diferentes, obtenidos por nuestro grupo de investigación utilizando de datos de microarrays **(Prieto et al., 2008)**.
2. **HKG** humanos descritos por Hsiao *et al*. **(Hsiao et al., 2001)**.
3. **HKG** humanos descritos por Eisenberg y Levanon **(Eisenberg and Levanon, 2003)**.

Esta aproximación múltiple aporta un alto nivel de fiabilidad en la obtención de un grupo de genes humanos considerados HKG, pudiendo seleccionar aquellos que se encuentran en 2, 3 o 4 de los conjuntos de datos citados. Esto aumenta significativamente el nivel de confianza. El número de HKG detectado por los 4 métodos es tan solo de 36 genes; sin embargo en este grupo están presentes ACTB, ALDOA y GAPDH, que son tres de los genes más conocidos y utilizados como controles en distintas pruebas de validación experimental de expresión como la RT-PCR (ver **tabla 4.2**). El número de HKG detectados por al menos 3 métodos es de 157, lo cual supone un buen balance entre fiabilidad y cobertura. El número de genes detectados por al menos 2 y 1 métodos es de 442 y 1235 genes respectivamente.

ID *Ensembl*	Nombre	Descripción
ENSG00000075624	**ACTB**	actin, beta
ENSG00000149925	**ALDOA**	aldolase A, fructose-bisphosphate
ENSG00000182718	ANXA2	annexin A2 pseudogene 1
ENSG00000134287	ARF3	ADP-ribosylation factor 3
ENSG00000166710	B2M	beta-2-microglobulin
ENSG00000198563	BAT1	small nucleolar RNA, C/D box 84
ENSG00000127022	CANX	calnexin
ENSG00000156508	EEF1A1	eukaryotic translation elongation factor 1 alpha-like 7
ENSG00000175390	EIF3F	eukaryotic translation initiation factor 3, subunit F
ENSG00000110321	EIF4G2	eukaryotic translation initiation factor 4 gamma, 2
ENSG00000111640	**GAPDH**	glyceraldehyde-3-phosphate dehydrogenase
ENSG00000189403	HMGB1	high-mobility group box 1
ENSG00000165119	HNRNPK	microRNA 7-1
ENSG00000185896	LAMP1	lysosomal-associated membrane protein 1
ENSG00000147140	NONO	non-POU domain containing, octamer-binding
ENSG00000070756	PABPC1	poly(A) binding protein, cytoplasmic 1
ENSG00000102144	PGK1	phosphoglycerate kinase 1
ENSG00000159377	PSMB4	proteasome (prosome, macropain) subunit, beta type, 4
ENSG00000187514	PTMA	prothymosin, alpha
ENSG00000067560	RHOA	ras homolog gene family, member A
ENSG00000198755	RPL10A	ribosomal protein L10a

ENSG00000142541	RPL13A	small nucleolar RNA, C/D box 32A
ENSG00000063177	RPL18	ribosomal protein L18
ENSG00000131469	RPL27	ribosomal protein L27
ENSG00000100316	RPL3	small nucleolar RNA, C/D box 43
ENSG00000144713	RPL32	small nucleolar RNA, H/ACA box 7A
ENSG00000137818	RPLP1	ribosomal protein, large, P1
ENSG00000124614	RPS10	ribosomal protein S10
ENSG00000115268	RPS15	ribosomal protein S15
ENSG00000143947	RPS27A	ribosomal protein S27a
ENSG00000083845	RPS5	ribosomal protein S5
ENSG00000168385	SEPT2	septin 2
ENSG00000130985	UBA1	ubiquitin-like modifier activating enzyme 1
ENSG00000109332	UBE2D3	ubiquitin-conjugating enzyme E2D 3 (UBC4/5 homolog, yeast)
ENSG00000134308	YWHAQ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, theta polypeptide
ENSG00000164924	YWHAZ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, zeta polypeptide

Tabla 4.2. Tabla de los 36 genes *housekeeping* humanos detectados por los 4 métodos estudiados: **(i)** Genes expresados en ESTs; **(ii)** genes correlacionados a partir de datos de microarrays **(Prieto et al., 2008)**; **(iii)** genes descritos en **(Hsiao et al., 2001)**; (iv) genes descritos en **(Eisenberg and Levanon, 2003)**.

Para comprobar si los *sets* de genes humanos HKG encontrados es coherente con funciones esenciales de las células, se realizó un análisis de enriquecimiento funcional del conjunto de 157 HKG utilizando la herramienta DAVID **(Huang da et al., 2007)** configurada para analizar únicamente términos de *Gene Ontology* (GO FAT) enriquecidos con p-valores corregidos de FDR < 0.05. El resultado fue la obtención de términos enriquecidos en procesos básicos para el mantenimiento de la actividad celular como: traducción de proteínas y ribosoma, glicólisis, citoesqueleto, etc.

4.2.4. Identificación de genes específicos de tejido (TSG)

La identificación de los TSG se realizó sobre un grupo del *set* GSE3526 de 159 muestras. Este conjunto de muestras es más amplio que el utilizado para el estudio anterior (i.e. las 96 muestras descritas en la **sección 4.2.1**). En este caso la selección se hizo de un modo menos rígido, aumentando el número de tejidos y eliminando la restricción de 3 muestras por tejido. El incluir un mayor número de tejidos permite medir el comportamiento de los genes bajo más condiciones, aumentando por tanto el grado de confianza de que los genes identificados en el análisis posterior sean realmente específicos.

En la **tabla 4.3** se encuentra la relación de tejidos utilizados, detallando las distintas categorías o tipos en tres niveles: desde categorías más específicas en el nivel 1 hasta categorías más genéricas en el nivel 3. La descripción original de la muestra proporcionada por el *set* GSE3526 figura en la columna de nivel 1 en donde se identifican 47 tejidos distintos. En los niveles 2 y 3 se agrupan los tejidos similares bajo categorías más amplias. De esta manera, por ejemplo, los tejidos "córtex cerebral" y "lóbulo frontal" quedarán agrupados bajo la etiqueta "cerebro" en el nivel 2, mientras que "cerebelo" mantendrá la misma etiqueta en el nivel 1 y 2. El nivel 3 agrupa todos estos tejidos bajo la descripción "tejido nervioso". Tejidos con origen y función similar (como p. ej. aquellos que componen el sistema nervioso) puede tener muchos genes en común, pero no serán considerados específicos de tejido si son analizados a nivel 1. Sin embargo, sí pueden ser exclusivos de grupo de tejido cuando se analizan según una categoría más amplia, nivel 3. De esta forma, considerando distintos niveles o categorías fenotípicas, podremos encontrar genes específicos a distintos niveles, aumentando la riqueza de los

análisis y resultados respecto a la identificación de TSG.

Nivel 1	Nivel 2	Nivel 3
adipose_tissue	adipose_tissue	adipose_tissue
adipose_tissue_omental	adipose_tissue	adipose_tissue
adipose_tissue_subcutaneous	adipose_tissue	adipose_tissue
coronary_artery	blood_vessel	endothelial_tissue
saphenous_vein	blood_vessel	endothelial_tissue
bronchus	bronchus	epithelial_tissue
colon_cecum	colon	epithelial_tissue
esophagus	esophagus	epithelial_tissue
lung	lung	epithelial_tissue
mammary_gland	mammary_gland	epithelial_tissue
oral_mucosa	mucosa	epithelial_tissue
pharyngeal_mucosa	mucosa	epithelial_tissue
stomach_fundus	stomach	epithelial_tissue
stomach_pyloric	stomach	epithelial_tissue
tongue_main_corpus	tongue	epithelial_tissue
trachea	trachea	epithelial_tissue
urethra	urethra	epithelial_tissue
adrenal_gland_cortex	adrenal_gland	gland
prostate_gland	prostate	gland
salivary_gland	salivary_gland	gland
thyroid_gland	thyroid_gland	gland
ovary	ovary	gonads
testes	testes	gonads
bone_marrow	bone_marrow	hemato_&_immune system
lymph_nodes	lymph_nodes	hemato_&_immune system
spleen	spleen	hemato_&_immune system
tonsil	tonsil	hemato_&_immune system
kidney_cortex	kidney	kidney
kidney_medulla	kidney	kidney
liver	liver	liver
heart_atrium	heart	muscle_&_heart
heart_ventricle	heart	muscle_&_heart
skeletal_muscle	muscle	muscle_&_heart
cerebral_cortex	brain	nervous_tissue
frontal_lobe	brain	nervous_tissue
hippocampus	brain	nervous_tissue
hypothalamus	brain	nervous_tissue
medulla	brain	nervous_tissue
occipital_lobe	brain	nervous_tissue
parietal_lobe	brain	nervous_tissue
temporal_lobe	brain	nervous_tissue
cerebellum	cerebellum	nervous_tissue
spinal_cord	spinal_cord	nervous_tissue
pituitary_gland	pituitary_gland	pituitary
cervix	uterus	uterus
endometrium	uterus	uterus
myometrium	uterus	uterus

Tabla 4.3. Relación de categorías fenotípicas de tejidos/órganos humanos utilizadas. Cada tejido está detallado en tres niveles distintos de jerarquía, siendo el nivel 3 más genérico y el 1 más específico.

Para encontrar los genes expresados únicamente en cada una de estas categorías considerando los 3 niveles definidos, desarrollamos un algoritmo implementado en R basado en la detección de picos en los perfiles de expresión, el cual se aplicó sobre la matriz normalizada con RMA. Estos picos se detectan mediante la búsqueda de genes cuyo nivel de expresión en el tejido más expresado es significativamente mayor que la expresión en el resto, en los cuales no se debe mostrar modulación o variabilidad significativa. Los pasos concretos del algoritmo son:

1. **Pre-filtrado de genes mediante la identificación de picos de expresión**: Se realizó obteniendo aquellos genes cuya diferencia entre el primer tejido más expresado y el segundo sea mayor que la diferencia entre el segundo y el último.

2. **Significación estadística de la diferencia entre el tejido más expresado con el resto**: Test de tipo *t-Student* de una cola para hallar la significación de la diferencia de expresión entre las muestras pertenecientes al tejido más expresado con respecto al resto. La hipótesis nula de este test es la no existencia de diferencia entre la expresión de ambos grupos, siendo la alternativa que el tejido más expresado muestra una expresión mayor a la del resto. Dado que se realizó un test por cada gen, los p-valores simples calculados fueron corregidos/ajustados para tests múltiples (*multiple testing correction*), utilizando el método de FDR. Posteriormente se seleccionaron los genes con un p-valor corregido FDR ≤ 0.05.

3. **Medición del ruido o cercanía al *background* en los tejidos menos expresados:** Análisis de varianza (ANOVA) **(Scheffe, 1959)** de la expresión de cada gen (con p-valor significativo en el paso anterior) en los diferentes tejidos excluyendo el más expresado. Este test ANOVA pretende identificar los genes que no muestran valores de expresión semejantes entre réplicas biológicas, envidenciando en su lugar valores cambiantes o variables. Esto indica que la expresión detectada sobre esos genes no está midiendo regulación biológica alguna, sino que su señal más bien demuestra cierta cercanía al nivel de ruido o *background* de no señal. Con estos niveles de ruido y, siendo su expresión significativamente más baja que la del tejido más expresado, se infiere que su expresión es "no detectable" por el microarray respecto al *background* y, por lo tanto, muy probablemente el gen no se encuentre expresado en esos tejidos. Igual que en el paso anterior, los p-valores se corrigen por FDR y se eliminan únicamente los genes con un FDR ≤ 0.0001. Este estricto p-valor asegura eliminar únicamente los genes que se encuentran claramente regulados en los diferentes tejidos, alejando su expresión del ruido.

Este algoritmo se aplicó tres veces, una por cada nivel de agrupación de tejidos, lo cual proporcionó tres listas de genes específicos que contienen: 756 genes, 786 genes y 206 genes identificados para los niveles 1, 2 y 3, respectivamente. Para la medición del ruido el test ANOVA (paso 3) se utilizó en todos los casos con las muestras etiquetadas al nivel 1.

4.2.5. Método de análisis de la conservación de los genes

Con el fin de obtener información acerca del grado de conservación de cada uno de los genes humanos desde su punto de vista evolutivo, se comparó su nivel de similitud en otras especies. Se eligieron tres especies a diferentes distancias evolutivas de la humana: *M. musculus* (ratón), *C. elegans* (gusano) y *S. cerevisiae* (levadura) (ver **figura 4.2**).

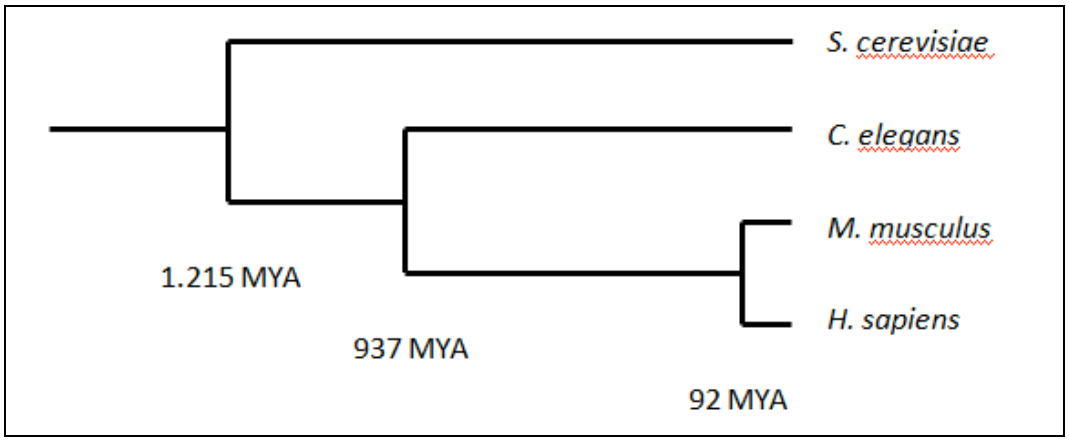


Figura 4.2. Árbol filogenético construido con 4 especies: *S. cerevisiae* (levadura), *C. elegans* (gusano), *M. musculus* (ratón) y *H. sapiens* (humano). Las distancias evolutivas indican hace cuánto tiempo se separaron las distintas especies. La distancia de humano con levadura, gusano y ratón es 1215 millones de años (MYA), 937 y 92, respectivamente **(Hedges et al., 2006)**.

Para obtener esta información se utilizó la base de datos de genes ortólogos de la herramienta *BioMart* **(Haider et al., 2009)** anotados a identificadores de *Ensembl* en su versión 57 en coherencia con *GeneMapper*. De esta manera se recuperó en una lista única la identidad, o grado de conservación, de cada gen para las tres citadas especies en relación a la especie humana. El grado de conservación se mide con un número entre 0 y 100 en función de la identidad de secuencia de aminoácidos con las correspondientes proteínas (i.e. productos génicos) humanas, siempre y cuando exista un gen homólogo. Para obtener un único valor que indique su edad evolutiva se optó por la aproximación más sencilla, promediando las identidades de cada gen a lo largo de las tres especies. Como se puede ver en **tabla 4.4**, la familia de histonas se sitúa a la cabecera de la lista, siendo los genes/proteínas mejor conservados, lo cual es sabido por numerosos estudios de evolución de proteínas. Para este análisis se filtró la lista global de genes manteniendo únicamente aquellos genes humanos con un homólogo en ratón recuperando un total de 36649 pares de genes humanos con ortólogos.

ID *Ensembl*	Nombre Símbolo	Descripción	Identidad *M.mus.* %	Identidad *C.ele.* %	Identidad *S.cer.* %	Media
ENSG00000196176	**HIST1H4A**	histone cluster 1, H4a	100	99	92	97,0
ENSG00000124529	**HIST1H4B**	histone cluster 1, H4b	100	99	92	97,0
ENSG00000197061	**HIST1H4C**	histone cluster 1, H4c	100	99	92	97,0
ENSG00000221983	**UBA52**	ubiquitin A-52 residue ribosomal protein fusion product 1	100	93	90	94,3
ENSG00000159251	**ACTC1**	actin, alpha, cardiac muscle 1	100	91	87	92,7
ENSG00000143632	**ACTA1**	actin, alpha 1, skeletal muscle	100	90	87	92,3
ENSG00000107796	**ACTA2**	actin, alpha 2, smooth muscle, aorta	100	91	86	92,3
ENSG00000163017	**ACTG2**	actin, gamma 2, smooth muscle, enteric	100	91	86	92,3
ENSG00000143761	**ARF1**	ADP-ribosylation factor 1	100	93	77	90,0
ENSG00000213639	**PPP1CB**	protein phosphatase 1, catalytic subunit, beta isozyme	100	89	81	90,0
ENSG00000134287	**ARF3**	ADP-ribosylation factor 3	100	92	76	89,3
ENSG00000169067	**ACTBL2**	actin, beta-like 2	97	87	83	89,0

ENSG00000134287	**ARF3**	ADP-ribosylation factor 3	100	92	75	89,0
ENSG00000132341	**RAN**	RAN, member RAS oncogene family	100	87	80	89,0
ENSG00000070831	**CDC42**	cell division cycle 42 (GTP binding protein, 25kDa)	100	86	80	88,7
ENSG00000075886	**TUBA3D**	tubulin, alpha 3d	100	92	74	88,7
ENSG00000125691	**RPL23**	ribosomal protein L23	100	86	77	87,7
ENSG00000164587	**RPS14**	ribosomal protein S14	97	86	80	87,7
ENSG00000184270	**HIST2H2AB**	histone cluster 2, H2ab	99	87	75	87,0

Tabla 4.4. Listado de genes/proteínas humanas mejor conservados en ratón, gusano y levadura. Se ha elegido el ortólogo más conservado por cada gen humano. Las columnas de identidad muestran el grado de conservación del gen medido como la identidad en secuencia de aminoácidos de la proteína correspondiente respecto a cada especie.

4.2.6. *Sets* de muestras de cáncer e identificación de genes alterados

Para la obtención de un conjunto de genes desregulados en cáncer, se obtuvieron *sets* de datos de distintos tipos de cáncer procedentes de repositorios públicos y de grupos experimentales colaboradores. En todos los casos el tipo de chip hibridado es el *Human Exon 1.0* de *Affymetrix*. Los tipos de cáncer estudiados fueron glioblastoma (GSE9385) **(French et al., 2007)**, cáncer gástrico (en distintos grados de progresión) (GSE27342) **(Cui et al., 2011a; Cui et al., 2011b)**, cáncer de pulmón (GSE12236) **(Xi et al., 2008)**, cáncer de colon **(Gardina et al., 2006)** y dos distintos subtipos de leucemia linfocítica crónica (CLL con pérdida alta y pérdida baja del cromosoma 13q) (ver **capítulo 2**). Estos distintos *sets* de datos fueron normalizados por separado con el algoritmo RMA y posteriormente se buscaron los genes diferencialmente expresados utilizando el algoritmo SAM, situando el punto de corte en FDR < 0,05. El fin de este procedimiento fue obtener los genes sobre e infra-expresados en cada tumor respecto al tejido sano de procedencia que fue usado como control.

4.3. Resultados

4.3.1 Red de coexpresión entre genes humanos

El análisis de correlaciones de expresión entre pares de genes humanos basados en los perfiles de expresión en múltiples tejidos/órganos permite la construcción de una red de coexpresión en donde los nodos son genes y los vínculos, o aristas, corresponden a una correlación significativa entre pares de genes. Los métodos RMA+*Pearson* y MAS5+*Spearman* produjeron su propio *set* de genes correlacionados a partir de los cuales se creó una red de única. Esta red se derivó realizando la unión de las interacciones de ambos métodos, tratando de que ninguno de ellos predominara sobre el otro con el fin de incluir una aportación significativa de cada método ya que son complementarios. Para esto se ajustaron los coeficientes de correlación a 0,75 para MAS5+*Spearman* y 0,95 para RMA+*Pearson*. Con estos umbrales de correlación, MAS5+*Spearman* proporcionó más nodos que RMA+*Pearson* (2098 *versus* 852), pero un número de vínculos entre nodos mucho menor (9355 *versus* 19079). Finalmente la unión proporciona una red con 2875 nodos y 28381 vínculos. Esto indica que el solapamiento entre ambos métodos a estos niveles de significación fue pequeño: sólo de 75 genes y 986 vínculos.

El mapeo de los HKG y de los TSG sobre esta red reveló también diferencias entre ambos métodos. El número de HKG mapeado sobre la red fue de 157 correspondiente a los genes presentes en al menos 3 de los 4 métodos propuestos. MAS5+*Spearman* ubicó en la red 77 de estos genes mientras que RMA+*Pearson*, que es demasiado astringente, tan solo ubicó un 1 HKG. Los TSG ubicados en la red se corresponden con los identificados (de nivel 1) en el **apartado 4.2.4**. En este caso RMA+*Pearson* fue más exitoso en la detección de TSG detectando 351 por ninguno de MAS5+*Spearman*, considerando un total de 756 (ver **tabla 4.5**).

	Umbral de correlación	Número de relaciones	Número de genes	Número de HKG	Número de TSG
MAS5+*Spearman*	0,75	9355	2098	77 / 157	0 / 756
RMA+*Pearson*	0,95	19079	852	1 / 157	351 / 756
Unión		28381	2875	77	352

Tabla 4.5. Datos correspondientes a la red de coexpresión. Se muestran los umbrales seleccionados para los métodos MAS5+*Spearman* y RMA+*Pearson*. También figura el número de nodos/genes y aristas/relaciones que aporta cada método y la capacidad de detección de genes *housekeeping* (HKG) y genes específicos de tejido (TSG).

La **figura 4.3** muestra la red de coexpresión completa representada con *Cytoscape* a partir de ficheros de texto plano generados con R en donde se describen parejas de genes. También se utilizó un fichero de atributos detallando el nombre y tipo de gen (HKG, TSG u otro). Esta red contiene 2875 nodos/genes y 28381 aristas/relaciones en la que los **HKG** han sido coloreados de **azul** y los **TSG** de **rojo**.

Esta red muestra una topología no uniforme, encontrándose distribuida en claros módulos o agrupaciones. A simple vista se puede ver que los nodos **rojos** (**TSG**) están localizados de forma más agrupada, mientras que los **HKG** están en general más dispersos, mostrando agrupación clara únicamente en un caso. Esto puede relacionarse con los datos que indicaban que el método RMA+*Pearson* obtiene más relaciones y detecta más genes de tipo TSG, mientras que MAS5+*Spearman* obtiene más nodos pero menos interconectados entre ellos, detectando también más HKG. Desde el punto de vista biológico, el hecho de que los TSG se agrupen entre ellos formando módulos funcionales localizados, es coherente con el concepto de especificidad de tejido, ya que estos genes actúan en un tipo celular concreto. Por el contrario, el que los HKG muestren una distribución mucho más dispersa, se puede asociar con su presencia ubicua en todas las células del organismo y su esencialidad funcional.

Cada una de las agrupaciones o módulos presentes en la red de coexpresión construida muestra una colección de genes muy relacionados entre sí, por lo que probablemente estén cooperando en una determinada función o proceso biológico. Con el propósito de caracterizar mejor las agrupaciones observadas en la red, se llevó a cabo un análisis de enriquecimiento funcional de cada grupo por separado. Para ello se utilizó la herramienta bioinformática DAVID (**http://david.abcc.ncifcrf.gov/**). El resultado del análisis funcional de varios módulos de genes se presenta en la **figura 4.4**. Cada grupo tiene una función bien definida, con p-valores altamente significativos, indicando que la señal biológica es clara. Dos de los grupos analizados tienen un alto contenido de TSG correspondiendo con funciones muy específicas: "espermatogénesis" y "complemento y cascada de coagulación". También se analizó un grupo de genes con alto contenido de HKG anotado a: "maquinaria molecular del ribosoma".

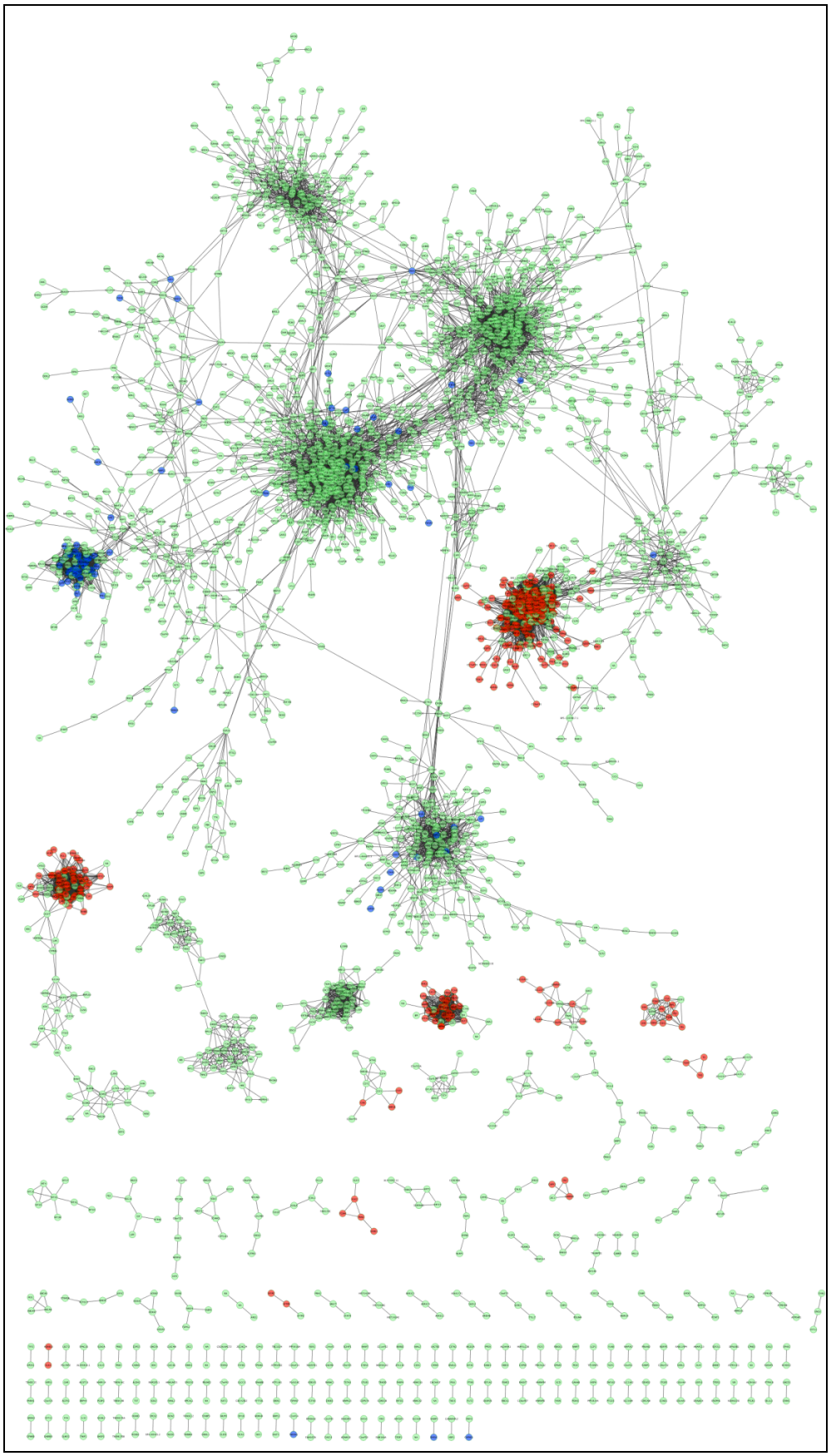


Figura 4.3. Red de coexpresión construida a partir de un *set* de 96 muestras (32 tejidos x 3 réplicas). La tabla tiene 2875 nodos y 28381 vértices. Los genes **HKG** figuran en **azul** y los genes **TSG** en **rojo**.

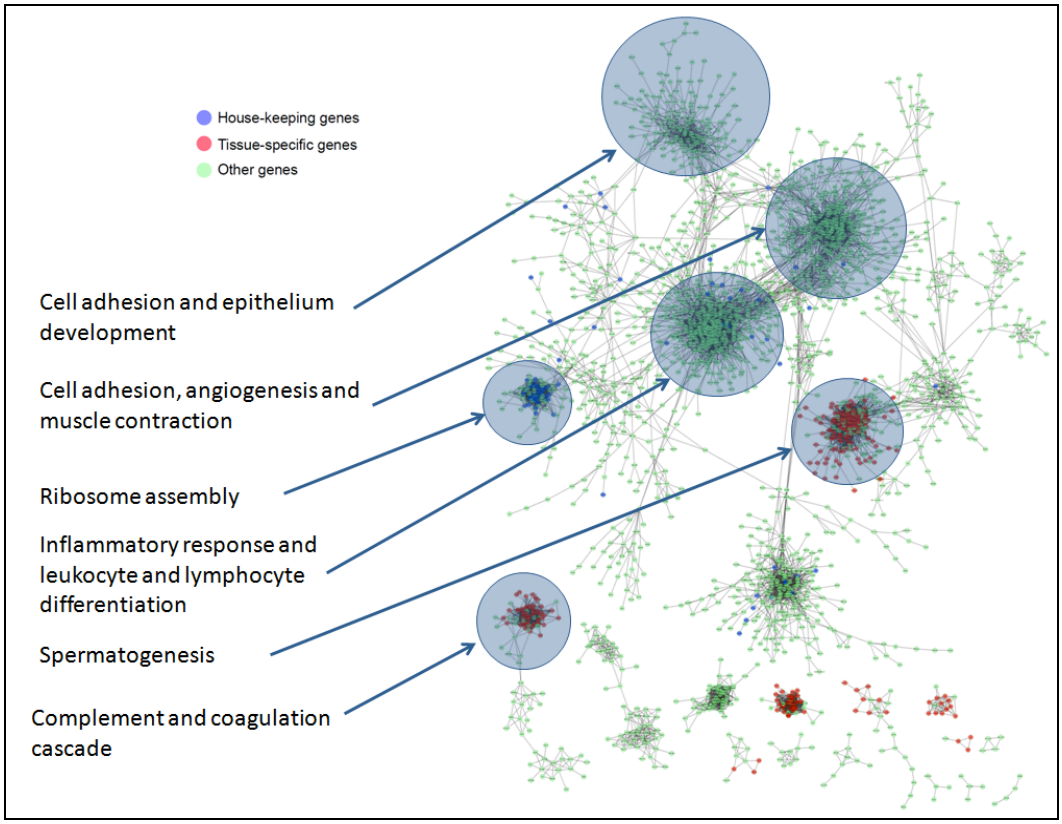


Figura 4.4. Análisis funcional de los grupos o módulos de genes más relevantes de la red de coexpresión. El análisis de enriquecimiento se realizó con la herramienta DAVID.

4.3.2 Diferencias evolutivas entre HKG y TSG

Los HKG y los TSG identificados de acuerdo con los métodos descritos en las **secciones 4.2.3** y **4.2.4** muestran, como cabría esperar, un perfil de expresión muy distinto a lo largo del conjunto de 96 microarrays (ver **figura 4.5**).

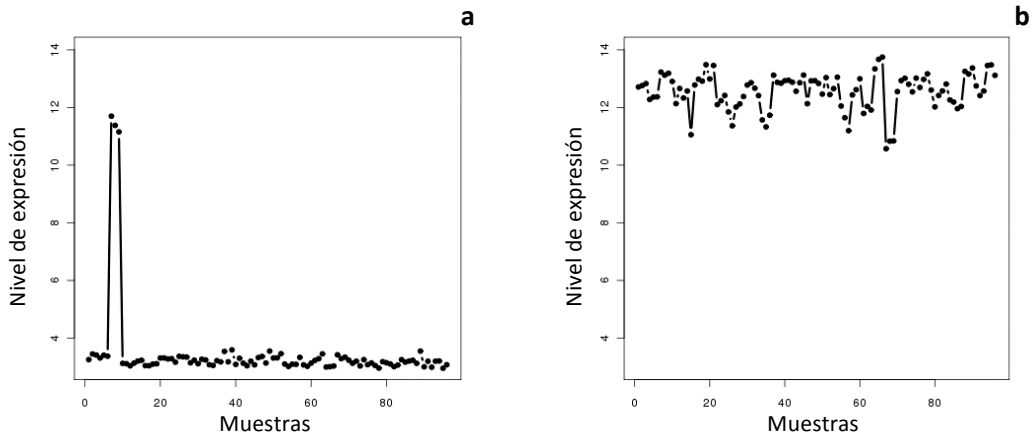


Figura 4.5. Diferencias en el perfil de transcripción entre un gen específico de tejido y un gen *housekeeping*. En la figura **(a)** se muestra el perfil del gen **CEACAM8** expresado únicamente en médula ósea, mientras que en **(b)** se muestra el perfil del gen **ACTB** transcrito en todos los tejidos.

Si analizamos los genes desde el punto de vista evolutivo, se pueden encontrar también marcadas diferencias entre los HKG y los TSG. En la **figura 4.6** se compara el grado de conservación medio en las tres especies entre los HKG y los TSG. Se han identificado los distintos subgrupos de HKG (los presentes en 1, 2, 3 y 4 métodos), y los distintos niveles de TSG (niveles jerárquicos 1, 2 y 3). Los diagramas de cajas muestran una gran variabilidad entre los genes de cada *set* de datos, sin embargo son claras las diferencias entre los distintos grupos. Tanto en las medianas, como en los cuartiles 1 y 3 muestran una mayor conservación de los genes HKG respecto de los TSG. Dentro de cada grupo también se observan diferencias. En los HKG se observa una conservación decreciente desde los presentes en 4 métodos hasta los presentes en uno solo. También entre los tres niveles de TSG existen diferencias, siendo los de nivel 1 y 2 muy similares a nivel evolutivo, mientras que los de nivel 3 muestran una mayor conservación.

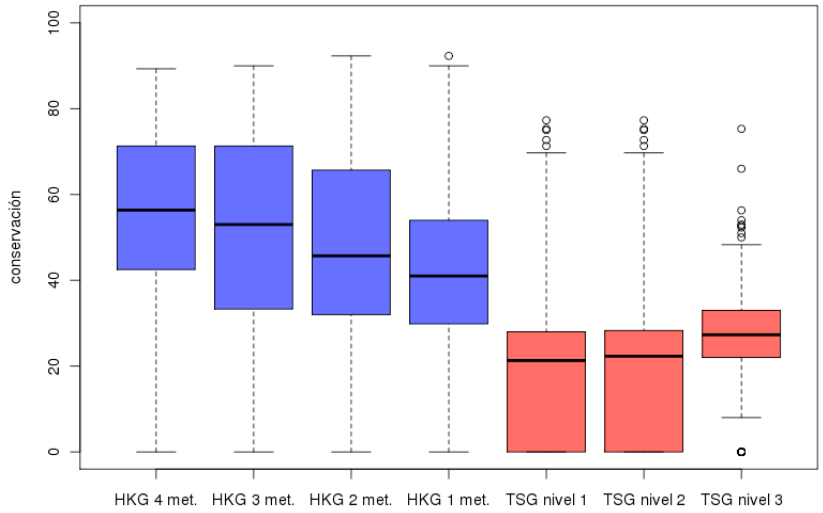


Figura 4.6. Diferencias en el grado de conservación de los **HKG** (en **azul**) y **TSG** (en **rojo**). Los **HKG** presentan mayor conservación cuanto más restrictiva sea su selección. Se aprecia claramente que los **HKG** están más conservados en secuencia a lo largo de la evolución que los **TSG**.

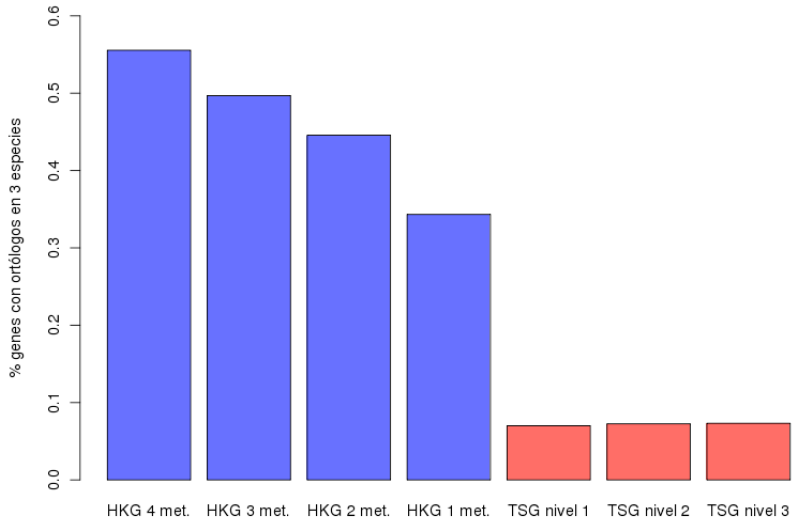


Figura 4.7. Porcentaje de genes presentes en distintos grupos de **HKG** y **TSG** con ortólogos en tres especies: ratón, gusano y levadura. Los **HKG** presentan mayor conservación cuanto más restrictiva sea su selección. Se aprecia de nuevo con este criterio que los **HKG** están más conservados que los **TSG**.

La **figura 4.7** muestra el mismo esquema, pero midiendo el porcentaje de genes que están presentes en las 3 especies: ratón, gusano y levadura. En este caso sólo existe un valor por cada subgrupo de genes marcando aún más las diferencias observadas anteriormente. Estas diferencias ya han sido observadas por otros autores apuntando a una mayor conservación por parte de los HKG **(Zhang and Li, 2004)**. Esto es esperable y se atribuye a la función constitutiva esencial que desempeñan este tipo de genes en la célula, considerando que las necesidades básicas para que una célula funcione no son muy variables a lo largo del tiempo. Sin embargo, los TSG pertenecen a células especializadas altamente diferenciadas que existen de modo más específico en organismos pluricelulares complejos evolutivamente más recientes, debiendo ser unos genes más "jóvenes".

Las diferencias observadas en ambas figuras (**figura 4.6** y **figura 4.7**) entre los 4 subgrupos de HKG se pueden asociar al nivel de evidencia (siendo mayor cuantos más métodos coincidan en catalogar dichos genes como HKG), pero también se pueden asociar a un mayor grado de conservación, lo cual nos permite establecer un rango entre los HKG que indicaría que los de nivel 4 son los más esenciales. En cuanto a las diferencias entre los subgrupos de TSG, en la **figura 4.6** se observa diferencia entre el nivel 3 y el nivel 1 y 2; pero estos entre sí son muy similares. Esto puede deberse a que el nivel 3, agrupa un conjunto más amplio y genérico de tejidos no identificándose con tipos celulares concretos y por lo tanto es menos específico que los de nivel 1 y 2. En la **figura 4.7,** en donde se muestra el porcentaje de genes presente en las 3 especies, se observan sin embargo pocas diferencias entre los 3 grupos de TSG, estando en los 3 casos en valores bajos. Esto confirma que los TSG son genes exclusivos de seres pluricelulares complejos, siendo por lo tanto más "modernos" que los HKG y habrían aparecido más tarde en el proceso evolutivo.

4.3.3 Diferencias evolutivas en genes desregulados en cáncer

Después de obtener los anteriores resultados analizando tejido sano, se trató de entender el comportamiento de los TSG y HKG y la conservación evolutiva de los genes en procesos de cáncer. Para esto, hicimos un análisis evolutivo de los genes perdidos y ganados en varios *sets* de expresión de diferentes tipos de cáncer. Se utilizaron 6 tipos de cáncer: cerebro, colon, pulmón, leucemia linfocítica crónica (con pérdida alta de 13q, 13q-A, y con pérdida baja, 13q-B) y cáncer gástrico. Cada uno de estos *sets* incluye datos de expresión global trascriptómica obtenidos por hibridación con la plataforma de microarrays *Human Exon 1.0*. Los distintos *sets* de datos fueron normalizados por separado con el algoritmo RMA utilizando el paquete *GeneMapper* de ***GATExplorer***. Posteriormente se identificaron los genes sobre-expresados e infra-expresados con el algoritmo SAM para cada tipo de cáncer, utilizando un punto de corte de significación en FDR < 0.05. Debido a que este umbral proporciona un número de genes muy dispar entre los distintos tipos de cáncer, se fijó un límite de 5000 genes en caso de superar esa cifra. El análisis evolutivo trató de identificar las posibles diferencias evolutivas entre los genes activados y suprimidos en cáncer.

La **figura 4.8** representa el porcentaje de genes que tiene ortólogos en las 3 especies para los 5 tipos de cáncer haciendo distinción entre los genes sobre-expresados y los genes reprimidos respecto a un control de tejido sano. Estos resultados muestran un patrón común indicando que los genes que aumentan su expresión en el paso de tejido sano a tumoral, están considerablemente más conservados que los genes que se reprimen en dicho proceso. En la **figura 4.9** se analiza con la misma metodología diferentes grados (G1, G2 y G3) y estadios de cáncer gástrico (S3 *vs* S1B), observando de nuevo que los genes sobre-expresados revelan un mayor grado de conservación que los reprimidos. El análisis de los genes en los grados 1, 2 y 3

se hizo comparando contra el tejido sano, mientras que la comparación "S3 *vs* S1B" compara 2 estadios diferentes. Esta última comparación muestra cómo a medida que avanza la enfermedad existe, en términos darwinianos, una "presión evolutiva" para eliminar los genes poco conservados.

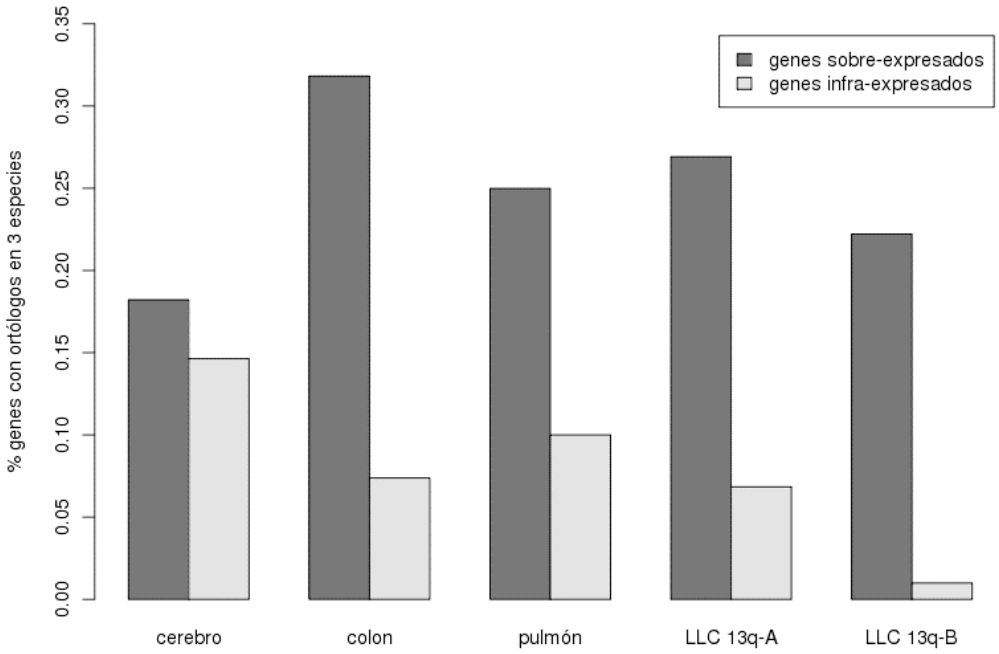


Figura 4.8. Porcentaje de genes diferencialmente expresados en cáncer que presentan ortólogos en 3 especies (ratón, gusano y levadura). Se descubre siempre una mayor conservación en los genes sobre-expresados que en los reprimidos.

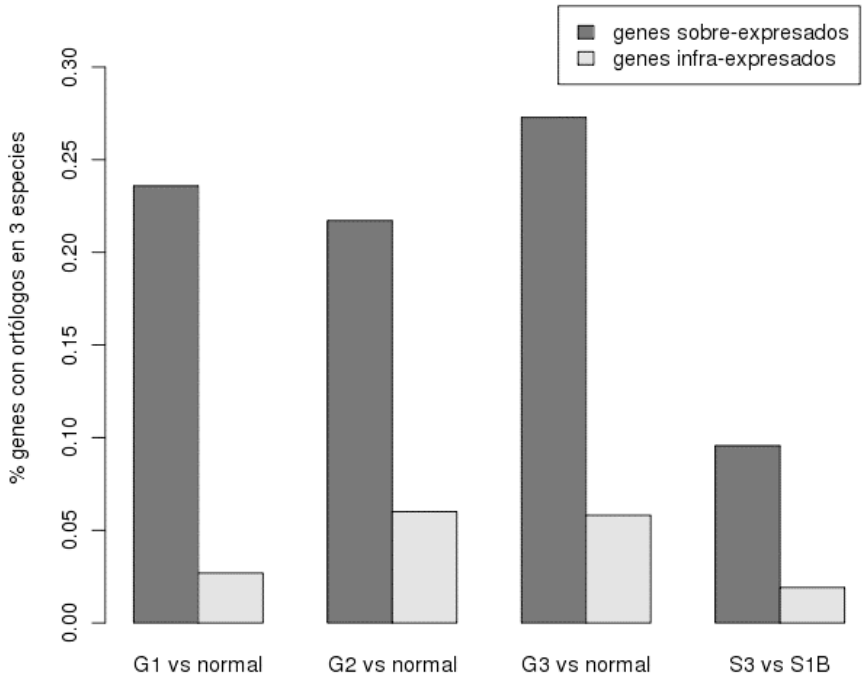


Figura 4.9. Porcentaje de genes diferencialmente expresados en distintos estadios de cáncer gástrico que presentan ortólogos en 3 especies. Se encuentra mayor conservación en los genes sobre-expresados y también una mayor conservación en los genes sobre-expresados en cáncer (tumor *versus* normal) frente a los sobre-expresados comparando estadíos de cáncer (S3 *vs* S1B).

En un organismo pluricelular la división celular está ligada a la diferenciación, proceso que se controla mediante un programa de crecimiento y cambio celular necesario para la especialización. Alteraciones externas pueden inhibir esta diferenciación y especialización, manteniendo la célula en un estado de inmadurez quiescente con una alta capacidad de división **(von Wangenheim and Peterson, 2008)**. Esto significa que las enfermedades neoplásicas tumorales implican un bloqueo en el proceso de diferenciación, que debería suponer una limitación en la expresión de genes específicos de tejido o de tipo celular lo cual es precisamente lo que descubrimos en nuestros datos de la **figura 4.8** y **4.9**. Además, la comparación entre S3 *vs* S1B de la **figura 4.9** sugiere que no solo existe bloqueo en la diferenciación celular durante el proceso de carcinogénesis, sino que también existen ciertas diferencias evolutivas entre los genes sobre e infra-expresados en la progresión tumoral. De este resultado, y con los datos de conservación de los HKG y TSG, se desprende que los genes sobre-expresados en cáncer son genes altamente conservados constitutivos, asociados a actividades esenciales de la célula –como crecimiento y replicación–, y que realizan funciones importantes para su mantenimiento como los procesos energéticos, síntesis de proteínas, ciclo celular, etc. Por contra, los genes reprimidos son genes propios de sistemas pluricelulares especializados, como genes específicos de tejido, que son innecesarios si el proceso de diferenciación se bloquea y que no están presentes en organismos unicelulares más sencillos y arcaicos.

Finalmente, para comprobar si la variación entre la conservación de los genes reprimidos y activados en cáncer se corresponde con su nivel de especificidad en tejidos, se realizó la intersección de los genes específicos de tejido identificados en este trabajo (786 TSG detectados a nivel 2, ver apartado **4.2.4**) con los genes identificados sobre-expresados e infra-expresados en cinco tipos de cáncer estudiados.

La **tabla 4.6** muestra en la segunda columna el número de genes específicos identificados de nivel 2 que fueron a su vez detectados en los 5 diferentes *sets* de cáncer. Ninguno de estos genes específicos aumentó su expresión en algún tipo de cáncer, mientras que varios de ellos la redujeron mostrándose reprimidos respecto a los individuos sanos. En algunos casos, el número de genes reprimidos en cáncer llega casi a la totalidad de los TSG identificados, como en cáncer de pulmón (5 de 6) y en cerebro (8 de 9). Dado que estos análisis se realizaron con distintos *sets* de datos y distintos modelos de microarrays, debe descartarse cualquier hipótesis de artefacto causado por la tecnología utilizada, validando las observaciones respecto al cáncer y las conclusiones desde el punto de vista evolutivo. Respecto a los HKG, debido a que son genes constitutivos y por lo tanto necesarios para la célula, se deduce que no habrá inhibición de ninguno de ellos en ninguna condición.

Procedencia del tejido tumoral	Nº de genes específicos TSG identificados (nivel 2)	Nº de genes específicos TSG detectados sobre-expresados	Nº de genes específicos TSG detectados infra-expresados
colon	5	0	1
pulmón	6	0	5
cerebro	9	0	8
médula ósea (13q- alto)	62	0	13
médula ósea (13q- bajo)	62	0	34

Tabla 4.6. La 1ª columna muestra el número de TSG encontrados alterados en cada tipo de cáncer. En la 2ª y 3ª columnas se muestran el número de esos TSG encontrados como sobre-expresados o infra-expresados, respectivamente, en muestras de cáncer provenientes de cada uno de los tejidos. Se identifican varios TSG infra-expresados en cáncer pero ninguno sobre-expresado.

4.4. Discusión y posible trabajo futuro

El principal objetivo de este trabajo es la identificación de genes co-regulados a nivel de expresión a través de una selección de muestras rigurosamente seleccionada y analizada con unos métodos computacionales robustos. Como resultado se deriva una red de coexpresión con un alto grado de fiabilidad, que muestra diferentes agrupaciones de genes que cooperan para llevar a cabo distintos procesos biológicos. Además en dicha red se observan diferencias entre los HKG y los TSG atendiendo a su ubicación en la topología de la red estando los HKG más distribuidos, mientras que los TSG se concentran en módulos más densos y agrupados. Este trabajo también pone de manifiesto el sesgo que tiene cada metodología hacia uno u otro tipo de genes. Como trabajo futuro se puede plantear la ampliación de la red de coexpresión mediante un meta-análisis que combine múltiples *sets* de datos analizados de forma conjunta y robusta (capaz de eliminar efectos "*batch*"). También se puede considerar analizar *sets* más amplios procedentes de diferentes enfermedades, encontrando mecanismos desregulados en las células patológicas en lugar de analizar genes individuales. Finalmente una identificación de los distintos grupos presentes en la red, y su anotación funcional de forma automatizada, podría revelar nueva información no presente actualmente en las bases de datos biológicas.

Desde un punto de vista evolutivo los HKG muestran un patrón mucho más conservado que los TSG, indicando que su procedencia es más antigua. Por otro lado, los genes sobre-expresados y reprimidos en los procesos cancerosos muestran también asociación con los genes de tipo HKG y TSG. Los genes sobre-expresados muestran un alto grado de conservación, mientras que los reprimidos tienen unos niveles bajos similares a los TSG. Esto da soporte a proponer un modelo evolutivo del proceso canceroso visto como una involución desde células altamente diferenciadas propias de organismos pluricelulares complejos hacia células inmaduras no diferenciadas que se replican rápidamente propias de sistemas unicelulares que pierden los anclajes al tejido circundante y la capacidad de responder a las señales de "control" o "freno". Actualmente existe una gran controversia sobre el origen de la célula tumoral. Los dos modelos principales postulan el bloqueo de la diferenciación celular manteniendo a la célula en un estado inmaduro o la des-diferenciación celular **(Bapat, 2007)**. El proceso inverso a la diferenciación celular, ha sido comprobado experimentalmente por algunos autores mediante la reprogramación células diferenciadas a células pluripotentes **(Takahashi and Yamanaka, 2006)**. También hay evidencias de que dicho proceso de des-diferenciación puede estar implicado en el origen de la célula tumoral **(Farber and Rubin, 1991; Friedmann-Morvinski et al., 2012; Kumar et al., 2012)**. En los resultados derivados del presente trabajo se ha comprobado que las células tumorales se caracterizan por la inhibición de genes específicos, sin embargo esto no aclara por sí mismo la procedencia de la célula tumoral. No obstante la comparación entre distintos estadios de un tumor ha revelado el mismo patrón en la progresión tumoral (ver **figura 4.8**), sugiriendo que en dicha progresión se continua un proceso de des-diferenciación que, podría haberse iniciado en una célula madura y diferenciada del organismo. Como propuesta para el futuro, un análisis de los factores de transcripción asociados a cada grupo de TSGs, podría mejorar nuestro entendimiento sobre los mecanismos de diferenciación celular en tejido sano. De la misma manera se puede tratar de estudiar el proceso inverso enmarcándolo en un contexto de cáncer. Todo ello se debería hacer abordando meta-análisis con múltiples *sets* de datos transcriptómicos integrados. De esta manera se podría contar con un amplio conjunto de genes para analizar los rasgos evolutivos propios de ambos, sobre e infra-expresados, así como su ubicación en la red de coexpresión y sus similitudes con TSG y HKG.

Conclusiones generales

Con la llegada de las técnicas genómicas/transcriptómicas de alto rendimiento, los análisis de expresión génica se han revelado como una herramienta muy eficaz para la investigación biomédica. Estas técnicas permiten obtener perfiles moleculares detallados de las muestras estudiadas, pudiendo comparar posteriormente distintos estados biológicos –como muestras de distintos tipos de tejidos o células sanas o, también, muestras de estados patológicos y enfermedades–, ayudándose de las herramientas bioinformáticas necesarias. La tecnología de expresión génica más utilizada hasta la fecha, ha sido la de microarrays de oligonucleótidos de alta densidad. Aunque en el presente –diciembre de 2012– todo hace pensar que la secuenciación masiva de nueva generación (NGS) está tomando el relevo, la gran cantidad muestras hibridadas con microarrays y almacenadas en los distintos repositorios públicos, hacen provechoso el seguir utilizando y desarrollando métodos computacionales para su análisis. La presente Tesis Doctoral se ha centrado en estudiar y mejorar los análisis computacionales bioinformáticos realizados con esta tecnología, y aplicar esas mejoras al análisis de muestras pertenecientes a series experimentales –principalmente de cáncer– obtenidas con grupos colaboradores, así como para el análisis de muestras procedentes de repositorios públicos.

En el **capítulo 1** se realizó una mejora de la anotación de las sondas de distintos tipos de microarrays de expresión, utilizando para ello una fuente de conocimiento biológico actualizado. Esto ha permitido mejorar los análisis respecto a la anotación original del fabricante apuntando directamente a entidades biológicas como genes, transcritos y exones. Al hacer esto: **(i)** se eliminan muchas sondas que no se pueden asociar con ningún *locus* génico conocido (ya sea genes codificantes o genes ncRNAs); y **(ii)** se eliminan sondas ambiguas que detectan la expresión de más de un gen, las cuales son fuente de ruido a la hora de tratar de identificar la expresión individual de los genes. Además de esto, se desarrolló un portal *web* abierto llamado ***GATExplorer*** en el que, mediante la ayuda de un navegador genómico y otro tipo de gráficos, cualquier investigador puede localizar la región del genoma en la que mapea cada una de las sondas. Esto puede ser particularmente importante cuando se quiere conocer en qué exones se sitúan las diferentes sondas para un gen de interés dado (p. ej. en estudios de *splicing* alternativo), o simplemente cuando se quiere conocer el modelo de microarray más adecuado para su estudio.

Utilizando el trabajo anterior, y en combinación con otras técnicas de aprendizaje automatizado, en el **capítulo 2** se describe cómo se logró identificar y validar genes marcadores para distintos subtipos de cáncer en colaboración con otros grupos de investigación experimental.

Además de esto, en el **capítulo 3** se profundizó en la complejidad del *splicing* alternativo analizando los problemas específicos que presentaban los últimos modelos de microarrays de

exones, y diseñando e implementando un método capaz de minimizarlos. Para su validación se utilizaron fuentes de datos públicas, tanto para la obtención de muestras humanas hibridadas con microarrays, como para la descripción de eventos validados de *splicing* alternativos en genes humanos concretos. Al analizar su comportamiento comparado con otros métodos, nuestro nuevo algoritmo demostró un redimiendo superior en la identificación de exones con regulación alternativa, pudiendo considerarse por lo tanto una aportación de alto valor para la comunidad científica.

Finalmente, en el **capítulo 4**, se realizó un estudio transcriptómico del organismo humano a un nivel más general, incluyendo muestras sanas procedentes de varios tipos tisulares y órganos disponibles públicamente. En este caso no se analizaron los genes de forma individual, sino que se consideraron las relaciones entre ellos mediante cálculo de correlación entre pares. Como resultado se obtuvo una red de coexpresión en la que fueron fácilmente identificables distintos grupos de genes que se correspondieron con funciones biológicas concretas. Para enriquecer los resultados, se analizaron por separado dos grupos de genes con distinto comportamiento transcripcional, confeccionándose así un listado de genes humanos *housekeeping* (HKG) y otro de genes específicos de tejido (TSG). Se encontró que estos dos tipos de genes no se distribuyen de igual forma en una red de coexpresión, mostrando además rasgos distintos al utilizarse información de conservación en otras especies. También se detectaron rasgos evolutivos distintos entre genes sobre-expresados y genes reprimidos en cáncer, ofreciendo una visión nueva interesante sobre la presión evolutiva y conservación existentes sobre los genes implicados en el proceso de carcinogénesis y la progresión tumoral.

De modo global, todo el trabajo descrito en esta Memoria proporciona por lo tanto dos tipos de resultados. El primero es un conjunto de **herramientas bioinformáticas** que están a libre disposición de otros investigadores para mejorar y facilitar el análisis de sus datos genómico / transcriptómicos. El segundo son la **información y nuevo conocimiento biológico** extraídos a partir de datos procedentes de técnicas ómicas de alto rendimiento obtenidos de distintas fuentes. Esta combinación de información y herramientas, es la contribución principal que realiza la presente Tesis Doctoral especialmente dirigida a un mejor conocimiento del transcriptoma humano.

Como **CONCLUSIONES FINALES** concretas resumidas de nuestro trabajo se puede decir:

1ª.– La tecnología de expresión génica de microarrays de oligos de alta densidad enfocada a genes y entidades génicas –por un remapeo a nivel de sonda– resulta eficaz para medir perfiles transcriptómicos globales que, analizados con algoritmos bioinformáticos robustos, permiten obtener firmas génicas muy específicas de cada estado biológico estudiado.

2ª.– La aplicación de la metodología descrita a estudios concretos de series de cáncer con muestras de pacientes resulta eficaz para obtener genes marcadores asociados a subtipos patológicos concretos.

3ª.– Los microarrays de exones permiten medir con precisión eventos de *splicing* alternativo si son analizados con algoritmos robustos que distingan bien la señal global del gen y la señal de cada exón.

4ª.– El análisis de perfil transcriptómico de los genes humanos en numerosos tejidos y órganos sanos permite identificar genes esenciales y genes específicos y construir redes precisas de coexpresión que muestran la agrupación biológica funcional de los genes.

Referencias

Affymetrix (2005a). **Alternative Transcript Analysis Methods for Exon Arrays**. In Affymetrix Technical Note.

Affymetrix (2005b). **Exon Array Design**. In Affymetrix Technical Note.

Affymetrix (2005c). **Exon Probeset Annotations and Transcript Cluster Groupings**. In Affymetrix Technical Note.

Alder, D., and Murdoch, D. (2011). **rgl: 3D visualization device system (OpenGL)**.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). **Basic local alignment search tool**. *J Mol Biol* 215, 403-410.

Alvarez-Garcia, I., and Miska, E.A. (2005). **MicroRNA functions in animal development and human disease**. *Development* 132, 4653-4662.

Amaral, P.P., Dinger, M.E., Mercer, T.R., and Mattick, J.S. (2008). **The eukaryotic genome as an RNA machine**. *Science* 319, 1787-1789.

Anton, M.A., Aramburu, A., and Rubio, A. (2010). **Improvements to previous algorithms to predict gene structure and isoform concentrations using Affymetrix Exon arrays.** *BMC Bioinformatics* 11, 578.

Anton, M.A., Gorostiaga, D., Guruceaga, E., Segura, V., Carmona-Saez, P., Pascual-Montano, A., Pio, R., Montuenga, L.M., and Rubio, A. (2008). **SPACE: an algorithm to predict and quantify alternatively spliced isoforms using microarrays.** *Genome biology* 9, R46.

Auer, H., Lyianarachchi, S., Newsom, D., Klisovic, M.I., Marcucci, u., and Kornacker, K. (2003). **Chipping away at the chip bias: RNA degradation in microarray analysis**. *Nat Genet* 35, 292-293.

Bapat, S.A. (2007). **Evolution of cancer stem cells**. *Semin Cancer Biol* 17, 204-213.

Barash, Y., Dehan, E., Krupsky, M., Franklin, W., Geraci, M., Friedman, N., and Kaminski, N. (2004). **Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays**. *Bioinformatics* 20, 839-846.

Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W., and Edgar, R. (2005). **NCBI GEO: mining millions of expression profiles--database and tools.** *Nucleic Acids Res* 33, D562-566.

Bemmo, A., Benovoy, D., Kwan, T., Gaffney, D.J., Jensen, R.V., and Majewski, J. (2008). **Gene**

expression and isoform variation analysis using *Affymetrix* Exon Arrays. *BMC Genomics* 9, 529.

Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N., and Golani, I. (2001). **Controlling the false discovery rate in behavior genetics research**. *Behav Brain Res* 125, 279-284.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Sayers, E.W. (2010). **GenBank**. *Nucleic Acids Res* 38, D46-51.

Bergsagel, P.L., and Kuehl, W.M. (2005). **Molecular pathogenesis and a consequent classification of multiple myeloma**. *J Clin Oncol* 23, 6333-6338.

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 19, 185-193.

Bushati, N., and Cohen, S.M. (2007). **microRNA Functions**. *Annu Rev Cell Dev Biol* 23, 175-205.

Butte, A.J., Dzau, V.J., and Glueck, S.B. (2001). **Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues"**. *Physiol Genomics* 7, 95-96.

Calin, G.A., and Croce, C.M. (2006). **MicroRNA signatures in human cancers**. *Nat Rev Cancer* 6, 857-866.

Calza, S., Raffelsberger, W., Ploner, A., Sahel, J., Leveillard, T., and Pawitan, Y. (2007). **Filtering genes to improve sensitivity in oligonucleotide microarray data analysis**. *Nucleic Acids Res* 35, e102.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C*., et al.* (2005). **The transcriptional landscape of the mammalian genome**. *Science* 309, 1559-1563.

Catovsky, D., Richards, S., Matutes, E., Oscier, D., Dyer, M.J., Bezares, R.F., Pettitt, A.R., Hamblin, T., Milligan, D.W., Child, J.A*., et al.* (2007). **Assessment of fludarabine plus cyclophosphamide for patients with chronic lymphocytic leukaemia (the LRF CLL4 Trial): a randomised controlled trial**. *Lancet* 370, 230-239.

Clark, T.A., Schweitzer, A.C., Chen, T.X., Staples, M.K., Lu, G., Wang, H., Williams, A., and Blume, J.E. (2007). **Discovery of tissue-specific exons using comprehensive human exon microarrays**. *Genome Biology* 8, R64.

Cline, M.S., Blume, J., Cawley, S., Clark, T.A., Hu, J.S., Lu, G., Salomonis, N., Wang, H., and Williams, A. (2005). **ANOSVA: a statistical method for detecting splice variation from expression data**. *Bioinformatics* 21 Suppl 1, i107-115.

Coomans, D., and Massart, D.L. (1982). **Alternative k-nearest neighbour rules in supervised pattern recognition**. *Analytica Chimica Acta* 136, 15-27.

Cover, T., and Hart, P. (1967). **Nearest neighbor pattern classification**. *IEEE Transactions on Information Theory* 13, 21-27.

Croce, C.M. (2008). **Oncogenes and cancer**. *N Engl J Med* 358, 502-511.

Croce, C.M. (2009). **Causes and consequences of microRNA dysregulation in cancer**. *Nat Rev Genet* 10, 704-714.

Cui, J., Chen, Y., Chou, W.C., Sun, L., Chen, L., Suo, J., Ni, Z., Zhang, M., Kong, X., Hoffman, L.L.*, et al.* (2011a). **An integrated transcriptomic and computational analysis for biomarker identification in gastric cancer**. *Nucleic Acids Res* 39, 1197-1207.

Cui, J., Li, F., Wang, G., Fang, X., Puett, J.D., and Xu, Y. (2011b). **Gene-expression signatures can distinguish gastric cancer grades and stages**. *PLoS One* 6, e17819.

Cheng, A.M., Byrom, M.W., Shelton, J., and Ford, L.P. (2005). **Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis**. *Nucleic Acids Res* 33, 1290-1297.

Chng, W.J., Glebov, O., Bergsagel, P.L., and Kuehl, W.M. (2007a). **Genetic events in the pathogenesis of multiple myeloma**. *Best Pract Res Clin Haematol* 20, 571-596.

Chng, W.J., Kumar, S., Vanwier, S., Ahmann, G., Price-Troska, T., Henderson, K., Chung, T.H., Kim, S., Mulligan, G., Bryant, B.*, et al.* (2007b). **Molecular dissection of hyperdiploid multiple myeloma by gene expression profiling**. *Cancer Res* 67, 2982-2989.

D'Haeseleer, P., Liang, S., and Somogyi, R. (2000). **Genetic network inference: from co-expression clustering to reverse engineering**. *Bioinformatics* 16, 707-726.

Dai, M., Wang, P., Boyd, A.D., Kostov, G., Athey, B., Jones, E.G., Bunney, W.E., Myers, R.M., Speed, T.P., Akil, H.*, et al.* (2005). **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data**. *Nucleic Acids Res* 33, e175.

Dal Bo, M., Rossi, F.M., Rossi, D., Deambrogi, C., Bertoni, F., Del Giudice, I., Palumbo, G., Nanni, M., Rinaldi, A., Kwee, I.*, et al.* (2011). **13q14 deletion size and number of deleted cells both influence prognosis in chronic lymphocytic leukemia**. *Genes Chromosomes Cancer* 50, 633-643.

Draghici, S. (2003). **Data analysis tools for DNA microarrays** (Boca Raton, Fla ; London, Chapman & Hall/CRC).

Durinck, S., Bullard, J., Spellman, P.T., and Dudoit, S. (2009). **GenomeGraphs: integrated genomic data visualization with R**. *BMC Bioinformatics* 10, 2.

Farber, E., and Rubin, H. (1991). **Cellular adaptation in the origin and development of cancer**. *Cancer Res* 51, 2751-2761.

French, P.J., Peeters, J., Horsman, S., Duijm, E., Siccama, I., van den Bent, M.J., Luider, T.M., Kros, J.M., van der Spek, P., and Sillevis Smitt, P.A. (2007). **Identification of differentially regulated splice variants and novel exons in glial brain tumors using exon expression arrays**. *Cancer Res* 67, 5635-5642.

Friedmann-Morvinski, D., Bushong, E.A., Ke, E., Soda, Y., Marumoto, T., Singer, O., Ellisman, M.H., and Verma, I.M. (2012). **Dedifferentiation of neurons and astrocytes by oncogenes can induce gliomas in mice**. *Science* 338, 1080-1084.

Gaidatzis, D., Jacobeit, K., Oakeley, E.J., and Stadler, M.B. (2009). **Overestimation of alternative splicing caused by variable probe characteristics in exon arrays.** *Nucleic Acids Res*

37, e107.

Gardina, P.J., Clark, T.a., Shimada, B., Staples, M.K., Yang, Q., Veitch, J., Schweitzer, A., Awad, T., Sugnet, C., Dee, S.*, et al.* (2006). **Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array.** *BMC Genomics* 7, 325.

Gath, I., and Geva, A.B. (1989). **Unsupervised optimal fuzzy clustering**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11, 773-780.

Gautier, L., Moller, M., Friis-Hansen, L., and Knudsen, S. (2004). **Alternative mapping of probes to genes for *Affymetrix* chips**. *BMC Bioinformatics* 5, 111.

Geng, X., Zhan, D.C., and Zhou, Z.H. (2005). **Supervised nonlinear dimensionality reduction for visualization and classification**. *IEEE Trans Syst Man Cybern B Cybern* 35, 1098-1107.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J.*, et al.* (2004). **Bioconductor: open software development for computational biology and bioinformatics**. *Genome Biology* 5, R80.

Goeman, J., and Oosting, J. (2009). **Testing association of a pathway with a clinical variable**.

Goeman, J.J., van de Geer, S.a., de Kort, F., and van Houwelingen, H.C. (2004). **A global test for groups of genes: testing association with a clinical outcome**. *Bioinformatics* 20, 93-99.

Graveley, B.R. (2001). **Alternative splicing: increasing diversity in the proteomic world**. *Trends Genet* 17, 100-107.

Griffith, O.L., Pleasance, E.D., Fulton, D.L., Oveisi, M., Ester, M., Siddiqui, A.S., and Jones, S.J. (2005). **Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses**. *Genomics* 86, 476-488.

Grosso, A.R., Martins, S., and Carmo-Fonseca, M. (2008). **The emerging role of splicing factors in cancer**. *EMBO Rep* 9, 1087-1093.

Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P., and Kasprzyk, A. (2009). **BioMart Central Portal--unified access to biological data**. *Nucleic Acids Res* 37, W23-27.

Harbig, J., Sprinkle, R., and Enkemann, S.A. (2005). **A sequence-based identification of the genes detected by probesets on the *Affymetrix* U133 plus 2.0 array**. *Nucleic Acids Res* 33, e31.

Hedges, S.B., Dudley, J., and Kumar, S. (2006). **TimeTree: a public knowledge-base of divergence times among organisms**. *Bioinformatics* 22, 2971-2972.

Hernandez, J.A., Rodriguez, A.E., Gonzalez, M., Benito, R., Fontanillo, C., Sandoval, V., Romero, M., Martin-Nunez, G., de Coca, A.G., Fisac, R.*, et al.* (2009). **A high number of losses in 13q14 chromosome band is associated with a worse outcome and biological differences in patients with B-cell chronic lymphoid leukemia**. *Haematologica* 94, 364-371.

Hochberg, Y. (1988). **A sharper Bonferroni procedure for multiple tests of significance**. *Biometrika* 75, 800–803.

Hochreiter, S., Clevert, D.A., and Obermayer, K. (2006). **A new summarization method for**

Affymetrix **probe level data**. *Bioinformatics* 22, 943-949.

Holm, S. (1979). **A simple sequentially rejective multiple test procedure**. *Scandinavian Journal of Statistics* 6, 65-70.

Huang da, W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C*., et al.* (2007). **DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists**. *Nucleic Acids Res* 35, W169-175.

Hubbard, T.J., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L*., et al.* (2009). **Ensembl 2009**. *Nucleic Acids Res* 37, D690-697.

Hubbell, E., Liu, W.M., and Mei, R. (2002). **Robust estimators for expression analysis**. *Bioinformatics* 18, 1585-1592.

Ihaka, R., and Gentleman, R. (1996). **R: A Language for Data Analysis and Graphics**. *Journal of Computational and Graphical Statistics* 5, 299-314.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B., and Speed, T.P. (2003a). **Summaries of *Affymetrix* GeneChip probe level data**. *Nucleic Acids Res* 31, e15.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003b). **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 4, 249-264.

Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., and Shoemaker, D.D. (2003). **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays**. *Science* 302, 2141-2144.

Jolliffe, I. (1986). **Principal Components Analysis**.

Jordan, I.K., Marino-Ramirez, L., Wolf, Y.I., and Koonin, E.V. (2004). **Conservation and coevolution in the scale-free human gene coexpression network**. *Mol Biol Evol* 21, 2058-2070.

Kalnina, Z., Zayakin, P., Silina, K., and Line, A. (2005). **Alterations of pre-mRNA splicing in cancer**. *Genes Chromosomes Cancer* 42, 342-357.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermuller, J., Hofacker, I.L*., et al.* (2007). **RNA maps reveal new RNA classes and a possible function for pervasive transcription**. *Science* 316, 1484-1488.

Kapur, K., Xing, Y., Ouyang, Z., and Wong, W. (2007). **Exon arrays provide accurate assessments of gene expression**. *Genome Biology* 8, R82.

Knudson, A.G. (2001). **Two genetic hits (more or less) to cancer**. *Nat Rev Cancer* 1, 157-162.

Koscielny, G., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Riethoven, J.-J., Nardone, F., Stanley, E., Fallsehr, C., Hofmann, O., Kull, M*., et al.* (2009). **ASTD: The Alternative Splicing and Transcript Diversity database.** *Genomics* 93, 213-220.

Krober, A., Seiler, T., Benner, A., Bullinger, L., Bruckle, E., Lichter, P., Dohner, H., and Stilgenbauer, S. (2002). **V(H) mutation status, CD38 expression level, genomic aberrations,**

and survival in chronic lymphocytic leukemia. *Blood* 100, 1410-1416.

Kumar, S.M., Liu, S., Lu, H., Zhang, H., Zhang, P.J., Gimotty, P.A., Guerra, M., Guo, W., and Xu, X. (2012). **Acquired cancer stem cell phenotypes through Oct4-mediated dedifferentiation**. *Oncogene* 31, 4898-4911.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). **Initial sequencing and analysis of the human genome**. *Nature* 409, 860-921.

Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). **Coexpression analysis of human genes across many microarray data sets**. *Genome Res* 14, 1085-1094.

Lewin, B. (2004). **Genes VIII**, International ed. edn (Upper Saddle River, N.J. ; London, Pearson Prentice Hall).

Lim, W.K., Wang, K., Lefebvre, C., and Califano, A. (2007). **Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks**. *Bioinformatics* 23, i282-288.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., and Lockhart, D.J. (1999). **High density synthetic oligonucleotide arrays**. *Nat Genet* 21, 20-24.

Liu, H., Zeeberg, B.R., Qu, G., Koru, A.G., Ferrucci, A., Kahn, A., Ryan, M.C., Nuhanovic, A., Munson, P.J., Reinhold, W.C.*, et al.* (2007). **AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets**. *Bioinformatics* 23, 2385-2390.

Liu, W.M., Mei, R., Di, X., Ryder, T.B., Hubbell, E., Dee, S., Webster, T.A., Harrington, C.A., Ho, M.H., Baid, J.*, et al.* (2002). **Analysis of high density expression microarrays with signed-rank call algorithms**. *Bioinformatics* 18, 1593-1599.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H.*, et al.* (1996). **Expression monitoring by hybridization to high-density oligonucleotide arrays**. *Nat Biotechnol* 14, 1675-1680.

Lu, J., Getz, G., Miska, E.A., Alvarez-Saavedra, E., Lamb, J., Peck, D., Sweet-Cordero, A., Ebert, B.L., Mak, R.H., Ferrando, A.A.*, et al.* (2005). **MicroRNA expression profiles classify human cancers**. *Nature* 435, 834-838.

MacQueen, J.B. (1967). **Some methods for clustering and analysis of multivariate observations**. *Proc 5th Berkeley Symp on Math Statist Prob*.

Mehes, G. (2005). **Chromosome abnormalities with prognostic impact in B-cell chronic lymphocytic leukemia**. *Pathol Oncol Res* 11, 205-210.

Mercer, T.R., Dinger, M.E., and Mattick, J.S. (2009). **Long non-coding RNAs: insights into functions**. *Nat Rev Genet* 10, 155-159.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat Methods* 5, 621-628.

Muro, E.M., Herrington, R., Janmohamed, S., Frelin, C., Andrade-Navarro, M.A., and Iscove, N.N. (2008). **Identification of gene 3' ends by automated EST cluster analysis**. *Proc Natl Acad*

Sci U S A 105, 20286-20290.

Nakaya, H.I., Amaral, P.P., Louro, R., Lopes, A., Fachel, A.A., Moreira, Y.B., El-Jundi, T.A., da Silva, A.M., Reis, E.M., and Verjovski-Almeida, S. (2007). **Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription**. *Genome Biol* 8, R43.

Nilsen, T.W., and Graveley, B.R. (2010). **Expansion of the eukaryotic proteome by alternative splicing**. *Nature* 463, 457-463.

Nowak, M.A. (2006). **Five rules for the evolution of cooperation**. *Science* 314, 1560-1563.

Osada, H., and Takahashi, T. (2007). **MicroRNAs in biological processes and carcinogenesis**. *Carcinogenesis* 28, 2-12.

Pan, Q., Shai, O., Lee, L.J., Frey, B.J., and Blencowe, B.J. (2008). **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. *Nat Genet* 40, 1413-1415.

Pang, K.C., Stephen, S., Dinger, M.E., Engstrom, P.G., Lenhard, B., and Mattick, J.S. (2007). **RNAdb 2.0--an expanded database of mammalian non-coding RNAs**. *Nucleic Acids Res* 35, D178-182.

Pochet, N., De Smet, F., Suykens, J.A., and De Moor, B.L. (2004). **Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction**. *Bioinformatics* 20, 3185-3195.

Prieto, C., Risueno, A., Fontanillo, C., and De las Rivas, J. (2008). **Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles**. *PLoS One* 3, e3911.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J*., et al.* (2012). **The Pfam protein families database**. *Nucleic Acids Res* 40, D290-301.

Purdom, E., Simpson, K.M., Robinson, M.D., Conboy, J.G., Lapuk, a.V., and Speed, T.P. (2008). **FIRMA: a method for detection of alternative splicing from exon array data.** *Bioinformatics* 24, 1707-1714.

R_Development_Core_Team (2010). **R: A Language and Environment for Statistical Computing** (Vienna, Austria, R Foundation for Statistical Computing).

Rambaldi, D., Felice, B., Praz, V., Bucher, P., Cittaro, D., and Guffanti, A. (2007). **Splicy: a web-based tool for the prediction of possible alternative splicing events from Affymetrix probeset data.** *BMC Bioinformatics* 8 Suppl 1, S17.

Rasche, A., and Herwig, R. (2010). **ARH: predicting splice variants from genome-wide data with modified entropy.** *Bioinformatics* 26, 84-90.

Rodriguez, A.E., Hernandez, J.A., Benito, R., Gutierrez, N.C., Garcia, J.L., Hernandez-Sanchez, M., Risueno, A., Sarasquete, M.E., Ferminan, E., Fisac, R*., et al.* (2012). **Molecular characterization of chronic lymphocytic leukemia patients with a high number of losses in 13q14**. *PLoS One* 7, e48485.

Roth, R.B., Hevezi, P., Lee, J., Willhite, D., Lechner, S.M., Foster, A.C., and Zlotnik, A. (2006). **Gene expression analyses reveal molecular relationships among 20 regions of the human CNS**. *Neurogenetics* 7, 67-80.

Rozman, C., and Montserrat, E. (1995). **Chronic lymphocytic leukemia**. *N Engl J Med* 333, 1052-1057.

Scheffe, H. (1959). **The Analysis of Variance** (pp. xvi. 477. John Wiley & Sons: New York; Chapman & Hall: London).

Schwender, H. (2012). ***siggenes*: Multiple testing using SAM and Efron's empirical Bayes approaches.**

Shai, O., Morris, Q.D., Blencowe, B.J., and Frey, B.J. (2006). **Inferring global levels of alternative splicing isoforms using a generative model of microarray data**. *Bioinformatics* 22, 606-613.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 13, 2498-2504.

Shen, S., Warzecha, C.C., Carstens, R.P., and Xing, Y. (2010). **MADS+: discovery of differential splicing events from Affymetrix exon junction array data**. *Bioinformatics* 26, 268-269.

Sigrist, C.J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A., and Hulo, N. (2010). **PROSITE, a protein domain database for functional characterization and annotation**. *Nucleic Acids Res* 38, D161-166.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). **ROCR: visualizing classifier performance in R**. *Bioinformatics* 21, 3940-3941.

Smith, C.W., and Valcarcel, J. (2000). **Alternative pre-mRNA splicing: the logic of combinatorial control**. *Trends Biochem Sci* 25, 381-388.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.-L., and Ideker, T. (2011). **Cytoscape 2.8: new features for data integration and network visualization.** *Bioinformatics* 27, 431-432.

Smyth, G. (2005). **Bioinformatics and Computational Biology Solutions Using R and Bioconductor**, R. Gentleman, V.J. Carey, W. Huber, R.A. Irizarry, and S. Dudoit, eds. (Springer New York), pp. 397-420.

Smyth, G.K., Ritchie, M., and Thorne, N. (2012). ***limma*: Linear Models for Microarray Data User ' s Guide (Now Including RNA-Seq Data Analysis)**.

Spiegelman, B.M., and Heinrich, R. (2004). **Biological Control through Regulated Transcriptional Coactivators**. *Cell* 119, 157-167.

Stalteri, M.A., and Harrison, A.P. (2007). **Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips**. *BMC Bioinformatics* 8, 13.

Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G*., et al.* (2004). **A gene atlas of the mouse and human protein-encoding transcriptomes**. *Proc Natl Acad Sci U S A* 101, 6062-6067.

Takahashi, K., and Yamanaka, S. (2006). **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors**. *Cell* 126, 663-676.

Thanaraj, T.a., Stamm, S., Clark, F., Riethoven, J.-J., Le Texier, V., and Muilu, J. (2004). **ASD: the Alternative Splicing Database.** *Nucleic Acids Res* 32, D64-69.

Tirosh, I., Weinberger, A., Carmi, M., and Barkai, N. (2006). **A genetic signature of interspecies variations in gene expression**. *Nat Genet* 38, 830-834.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). **Significance analysis of microarrays applied to the ionizing radiation response**. *Proc Natl Acad Sci U S A* 98, 5116-5121.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P.*, et al.* (2000). **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae**. *Nature* 403, 623-627.

Uhlen, M. (2005). **A Human Protein Atlas for Normal and Cancer Tissues Based on Antibody Proteomics**. *Mol Cell Proteomics* 4, 1920-1932.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S.*, et al.* (2010). **Towards a knowledge-based Human Protein Atlas**. *Nat Biotechnol* 28, 1248-1250.

van Noort, V., Snel, B., and Huynen, M.A. (2004). **The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model**. *EMBO Rep* 5, 280-284.

Venables, J.P. (2004). **Aberrant and alternative splicing in cancer**. *Cancer Res* 64, 7647-7654.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A.*, et al.* (2001). **The sequence of the human genome**. *Science* 291, 1304-1351.

von Wangenheim, K.H., and Peterson, H.P. (2008). **The role of cell differentiation in controlling cell multiplication and cancer**. *J Cancer Res Clin Oncol* 134, 725-741.

Wahl, M.C., Will, C.L., and Lührmann, R. (2009). **The spliceosome: design principles of a dynamic RNP machine.** *Cell* 136, 701-718.

Wang, E.T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S.F., Schroth, G.P., and Burge, C.B. (2008). **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 456, 470-476.

Wang, H., Hubbell, E., Hu, J.s., Mei, G., Cline, M., Lu, G., Clark, T., Siani-Rose, M.A., Ares, M., Kulp, D.C.*, et al.* (2003). **Gene structure-based splice variant deconvolution using a microarry platform**. *Bioinformatics* 19, i315-i322.

Wang, Z., Gerstein, M., and Snyder, M. (2009). **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 10, 57-63.

Weinberg, R.A. (2007). **The biology of cancer** (New York ; London, Garland Science).

Wodicka, L., Dong, H., Mittmann, M., Ho, M.H., and Lockhart, D.J. (1997). **Genome-wide expression monitoring in Saccharomyces cerevisiae**. *Nat Biotechnol* 15, 1359-1367.

Xi, L., Feber, A., Gupta, V., Wu, M., Bergemann, A.D., Landreneau, R.J., Litle, V.R., Pennathur, A., Luketich, J.D., and Godfrey, T.E. (2008). **Whole genome exon arrays identify differential expression of alternatively spliced, cancer-related genes in lung cancer**. *Nucleic Acids Res* 36, 6535-6547.

Xie, X.L., and Beni, G. (1991). **A validity measure for fuzzy clustering**. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13, 841-847.

Yates, T., Okoniewski, M.J., and Miller, C.J. (2008). **X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis**. *Nucleic Acids Res* 36, D780-786.

Zhan, F., Hardin, J., Kordsmeier, B., Bumm, K., Zheng, M., Tian, E., Sanderson, R., Yang, Y., Wilson, C., Zangari, M.*, et al.* (2002). **Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells**. *Blood* 99, 1745-1757.

Zhan, F., Huang, Y., Colla, S., Stewart, J.P., Hanamura, I., Gupta, S., Epstein, J., Yaccoby, S., Sawyer, J., Burington, B.*, et al.* (2006). **The molecular classification of multiple myeloma**. *Blood* 108, 2020-2028.

Zhan, F., Tian, E., Bumm, K., Smith, R., Barlogie, B., and Shaughnessy, J., Jr. (2003). **Gene expression profiling of human plasma cell differentiation and classification of multiple myeloma based on similarities to distinct stages of late-stage B-cell development**. *Blood* 101, 1128-1140.

Zhang, L., and Li, W.H. (2004). **Mammalian housekeeping genes evolve more slowly than tissue-specific genes**. *Mol Biol Evol* 21, 236-239.

Apéndice
Publicaciones científicas realizadas durante el desarrollo de la presente Tesis Doctoral

Publicaciones relacionadas con el Capítulo 1

– Risueño, A., Fontanillo, C., Dinger, M.E., and De Las Rivas, J. (2010). **GATExplorer: genomic and transcriptomic explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs.** *BMC Bioinformatics* 11, 221.

Publicaciones relacionadas con el Capítulo 2

– Gutiérrez, N.C., Sarasquete, M.E., Misiewicz-Krzeminska, I., Delgado, M., De Las Rivas, J., Ticona, F.V., Fermiñán, E., Martín-Jiménez, P., Chillón, C., Risueño, a.*, et al.* (2010). **Deregulation of microRNA expression in the different genetic subtypes of multiple myeloma and correlation with gene expression profiling.** *Leukemia* 24, 629-637.

– Rodriguez, A.E., Hernandez, J.A., Benito, R., Gutierrez, N.C., Garcia, J.L., Hernandez-Sanchez, M., Risueno, A., Sarasquete, M.E., Ferminan, E., Fisac, R.*, et al.* (2012). **Molecular characterization of chronic lymphocytic leukemia patients with a high number of losses in 13q14**. *PLoS One* 7, e48485.

Publicaciones relacionadas con el Capítulo 3

– Risueño, A., *et al.* (2013). **Robust estimation of gene and exon expression to identify alternative splicing events: a method applied to human tissue genes.** En preparación

Publicaciones relacionadas con el Capítulo 4

– Prieto, C., Risueño, A., Fontanillo, C., and De Las Rivas, J. (2008). **Human Gene Coexpression Landscape: Confident Network Derived from Tissue Transcriptomic Profiles**. *PLoS One* 3, e3911.

Otras publicaciones

– Rodríguez, A.E., Robledo, C., García, J.L., González, M., Gutiérrez, N.C., Hernández, J.A.M., Sandoval, V., García de Coca, A., Recio, I., Risueño, A.*, et al.* (2012). **Identification of a novel recurrent gain on 20q13 in chronic lymphocytic leukemia by array CGH and gene expression profiling.** *Annals of Oncology* 23, 2138-2146.

BMC Bioinformatics

DATABASE Open Access

GATExplorer: Genomic and Transcriptomic Explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs

Alberto Risueño[1], Celia Fontanillo[1], Marcel E Dinger[2] and Javier De Las Rivas*[1]

Abstract

Background: Genome-wide expression studies have developed exponentially in recent years as a result of extensive use of microarray technology. However, expression signals are typically calculated using the assignment of "probesets" to genes, without addressing the problem of "gene" definition or proper consideration of the location of the measuring probes in the context of the currently known genomes and transcriptomes. Moreover, as our knowledge of metazoan genomes improves, the number of both protein-coding and noncoding genes, as well as their associated isoforms, continues to increase. Consequently, there is a need for new databases that combine genomic and transcriptomic information and provide updated mapping of expression probes to current genomic annotations.

Results: GATExplorer (Genomic and Transcriptomic Explorer) is a database and web platform that integrates a gene loci browser with nucleotide level mappings of oligo probes from expression microarrays. It allows interactive exploration of gene loci, transcripts and exons of human, mouse and rat genomes, and shows the specific location of all mappable *Affymetrix* microarray probes and their respective expression levels in a broad set of biological samples. The web site allows visualization of probes in their genomic context together with any associated protein-coding or noncoding transcripts. In the case of all-exon arrays, this provides a means by which the expression of the individual exons within a gene can be compared, thereby facilitating the identification and analysis of alternatively spliced exons. The application integrates data from four major source databases: *Ensembl*, *RNAdb, Affymetrix* and *GeneAtlas*; and it provides the users with a series of files and packages (R CDFs) to analyze particular query expression datasets. The maps cover both the widely used *Affymetrix GeneChip* microarrays based on 3' expression (e.g. human HG U133 series) and the all-exon expression microarrays (Gene 1.0 and Exon 1.0).

Conclusions: GATExplorer is an integrated database that combines genomic/transcriptomic visualization with nucleotide-level probe mapping. By considering expression at the nucleotide level rather than the gene level, it shows that the arrays detect expression signals from entities that most researchers do not contemplate or discriminate. This approach provides the means to undertake a higher resolution analysis of microarray data and potentially extract considerably more detailed and biologically accurate information from existing and future microarray experiments.

Background

As our knowledge of metazoan genomes and transcriptomes improves, the number of both protein-coding and noncoding transcripts continues to increase [1,2]. To take account of the increasing emphasis on transcriptomics, genomic databases need to be adapted to better accommodate this type of data. Consideration of transcriptomic data necessitates improvements in both the correct mapping of all actively transcribed units and the accurate determination of their expression levels. *Ensembl* maintains and provides visualization of a comprehensive database of all publicly available eukaryotic genome sequences and contains all major biomolecular entities (such as RNAs and proteins) with extensive additional information including mapping of microarray probes [3]. Other databases, such as *RNAdb*, complement *Ensembl* by providing details and annotations of larger collections of non-coding RNAs (ncRNAs) [4]. However, these bio-

* Correspondence: jrivas@usal.es
[1] Bioinformatics and Functional Genomics Research Group, Cancer Research Center (CiC-IBMCC, CSIC/USAL), Salamanca, Spain
Full list of author information is available at the end of the article

logical databases do not integrate expression signal data and they do not provide tools to use up-to-date probe mapping with query expression datasets. Finally, databases such as *GEO* include large collections of expression datasets with powerful analysis tools, but they lack microarray probe mapping at nucleotide level and presentation in a genomic context, and instead consider "probesets" as genes [5]. Several recent transcriptomic studies are showing that gene loci are considerably more complex than previously thought, often with networks of overlapping transcripts on both strands [1], emphasizing the importance to examine expression data at the nucleotide level. Nucleotide level mapping facilitates the identification of particular probes that uniquely represent the expression of specific transcripts. It also provides the possibility to discriminate between alternate isoforms of the same gene. Such analyses require unambiguous assignment of the array probes to the functional entities defined in current transcriptomes (i.e. gene loci, transcripts, exons, ncRNAs), including their specific genomic location. The huge number of transcriptomic studies conducted in recent years illustrates the potential demand for improved analytical approaches of microarray data as well as the opportunity to reinterpret existing datasets. To provide a means to analyze microarray data at the nucleotide level in a genomic context we have developed a database and web platform called GATExplorer. The application integrates information from multiple biological sources and includes several bioinformatic tools to allow a novel perspective and interpretation of microarray expression data.

Construction and content

Database integrating genomes, transcriptomic entities and expression

To analyze transcriptomic data in a genomic context, GATExplorer integrates five **datasets**: **(i)** the human, mouse and rat genomes (derived from *Ensembl* http://www.ensembl.org); **(ii)** the sequences and IDs of all oligonucleotide probes (perfect match only) from all *Affymetrix* expression microarrays http://www.affymetrix.com for these species; **(iii)** *de novo* mapping data of each array probe to the transcriptome of the corresponding organism, with the genomic coordinates for each locus (including locations on exons, introns and across exon-exon junctions) and identification of any intersecting genes, transcripts and exons; **(iv)** mapping data of unmapped probes to transcripts in *RNAdb* (research.imb.uq.edu.au/RNAdb), a database of ncRNAs of human and mouse; and **(v)** detailed expression data derived from a set of microarrays from different cell types, tissues or organs (*GeneAtlas* GEO ID GSE1133 [6]) calculated at probe- and probeset-level using complete *de novo* mapping.

BLASTN sequence **alignment** was used to map the 25-mer oligo probes of the main *Affymetrix* expression microarrays to the RNA sequences of human, mouse and rat, selecting only complete perfect match alignments. The mapped probes were then placed in the corresponding genome based on the coordinates of the main genomic entities defined by *Ensembl*. The versions of the *genomes assemblies* and the *source databases* in current use are indicated on the website (PROBE MAPPING section, "Genomes ASSEMBLY and Databases VERSION").

Each of the source and newly derived datasets are structured and integrated in a relational SQL database (MySQL), which can be queried and viewed via the website. For a specified gene locus, the web interface presents a hierarchical display of the corresponding genomic entities (chromosome, locus, exons, transcripts and protein domains), together with detailed mapping of array probes and probesets and their signal in a set of sample arrays. This data is presented as follows: **(i)** Description; **(ii)** Chromosome global view (chr [chr number]); **(iii)** Chromosomal regional view (indicating the *specie*); **(iv)** Gene locus and transcripts view; **(v)** Expression view (profile in different tissues); **(vi)** Probesets table: *Affymetrix* Probesets which map on [gene locus name]; **(vii)** Probes table: *Affymetrix* Probes which map on [gene locus name].

An illustrative **workflow** for the general use of GATExplorer is included in the front page of the website to facilitate a practical guide of the application. The transcript and protein sequences within each gene locus are also provided in a link called "Show SEQUENCES (cDNA)" included within the "Description" of each gene. Some other useful links and tools are included in the "Description" box: one external link to the corresponding gene in *Ensembl* (indicating the ENSG ID); another external link to the corresponding proteins associated to this gene in the *Protein Atlas* database http://www.proteinatlas.org; a tool called "Bookmark GENE" that builds a new box inside the web with direct links to genes selected by the user: bookmarked genes.

The server can be queried using five access-boxes located on the left side which receive the following types of queries: a **keyword** related to any gene locus; a **probe** or **probeset** ID from *Affymetrix*; a **list of probesets** to find corresponding genes; a **sequence** (nucleotide or amino acid) via BLAST; or a range of **chromosomal coordinates**. The usage of each of these access tools is explained in detail in the "Help" section (link on the top right side of the main page).

Accurate graphical representations of genes, transcripts and probes in the "Chromosomal regional view" and the "Gene locus and transcripts view" are achieved using MING (a library for generating *Flash* files), which produces vector drawings in SWF format maintaining the

scale of each exon and intron in proportion to their sequence length. Interactive links, gene descriptions, exonic structure and probe positions are included in the drawings. Each microarray probe is identified by its sequence, which is included in the "Probes table" together with its GC content (%). The "Probesets table" and "Probes table" include links ("Download PROBESETS (.txt)", "Download PROBES (.txt)") on the top right to download text files containing the probesets or the probes that map to the selected gene locus. The complete mappings for each microarray platform are included in the PROBE MAPPING section, which is opened in another browser window (link on the top right side of the front page of the website).

The PROBE MAPPING section includes several pages divided into two parts: **(i)** pages providing the complete collection of files that can be downloaded by the user to facilitate the application of the mappings to any particular microarray dataset that is to be analyzed ("Text Files", "R Packages" and "Annotation Files"); **(ii)** pages to explain how the mapping has been performed and provide data to compare the results with other methods previously reported ("Methods", "Statistics", "Comparative Analysis" and "Genomes & Databases VERSION"). Detailed descriptions of the downloadable files included in this section of the database are as follows:

• text files (.txt) with complete unambiguous mapping of the microarray probesets to genes (*probesets2 genes*); • text files (.txt) with complete unambiguous mapping of array probes to genes (*probes2 genes*); • text files (.txt) with complete unambiguous mapping of array probes to transcripts (*probes2transcripts*); • text files (.txt) with complete mapping of microarray probes that are ambiguous because they map on more than one gene locus (*ambigprobes2 genes*).

• R chip definition files (CDFs) with complete unambiguous mapping of microarray probes to genes (*GeneMapper*); • R chip definition files (CDFs) with complete unambiguous mapping of array probes to transcripts (*TranscriptMapper*); • R chip definition files (CDFs) with complete unambiguous mapping of array probes to exons (*ExonMapper*); • R chip definition files (CDFs) with mapping to ncRNAs of the probes that did not map any known protein-coding exon (*ncRNA Mapper*).

• annotation files with information about the mapped entities derived from the *Ensembl* database: genes (ENSGs), transcripts (ENSTs) and exons (ENSEs); • annotation files that include only the selected subset of the protein-coding genes (i.e. the *Ensembl* gene loci, ENSGs, that correspond to mRNAs) or the selected subsets of known microRNAs (i.e. the *Ensembl* gene loci that have been assigned to microRNAs).

The PROBE MAPPING section also presents details regarding the specific "Methods" used, the "Statistics" regarding the mapping to different transcribed entities and a "Comparative Analysis" with other related applications. The "Methods" page provides descriptions and links to the main data sources used in GATExplorer and a graphical schematic view of the pipeline followed to build the web platform, presenting the main steps and procedures applied and the files and packages provided by the server. The "Statistics" page provides the data derived from the sequence mapping of all the oligonucleotide probes from *Affymetrix* expression microarrays into different types of RNAs. Probes are classified as mapping to: protein-coding RNAs (mature mRNAs), non protein-coding RNAs (ncRNAs) or unassigned to any known RNA (NA). Probes that only map to introns were classified as mapping to putative ncRNAs. The percentage of probes mapping to each class is provided for four types of widely used human expression microarrays platforms. The page also provides statistics on the number and percentages of transcripts and genes mapped by the probes in each *Affymetrix* expression microarray (for human *Homo sapiens*, mouse *Mus musculus* and rat *Rattus norvegicus*); and the number and percentages of probes that map to transcripts and genes with respect to the total in each array. The "Comparative Analysis" page includes a comparison of GATExplorer with other related applications that have been previously published. The page examines four studies that have undertaken an alternative mapping of probes to genes for *Affymetrix* microarrays. Some of these re-mapping approaches and tools are limited to a subset of microarray platforms or do not apply to whole-transcript expression microarrays (i.e. Gene 1.0 and Exon 1.0). Among the previous studies, none present mapping to intronic regions or ncRNAs.

Figure 1 presents a graphical view of the main data sources and methods included in GATExplorer described above. The graph shows the pipeline followed to build the database and web platform. The probe mapping files (Text files) and packages (R CDFs) are freely provided as part of the repository to allow researchers to use the microarray probe remapping data for their own expression analyses.

Each gene can be queried to find detailed information on the mapping of probes to their corresponding locus, transcripts and exons. When a gene loci is shown in the associated GATExplorer web page, the "Gene locus and transcripts view" presents all the probes that map to such loci for each of the *Affymetrix* microarrays. As mentioned above, the information about such probes indicating whether they are ambiguous (i.e. multi-mapping) or not is included in the table called "Probes table". After mapping, each probe is designated with a COLOR CODE (green, yellow, red and black) to indicate whether it is ambiguous or not (see HELP section). As a result, all ambiguous probes that can cross-hybridize with several

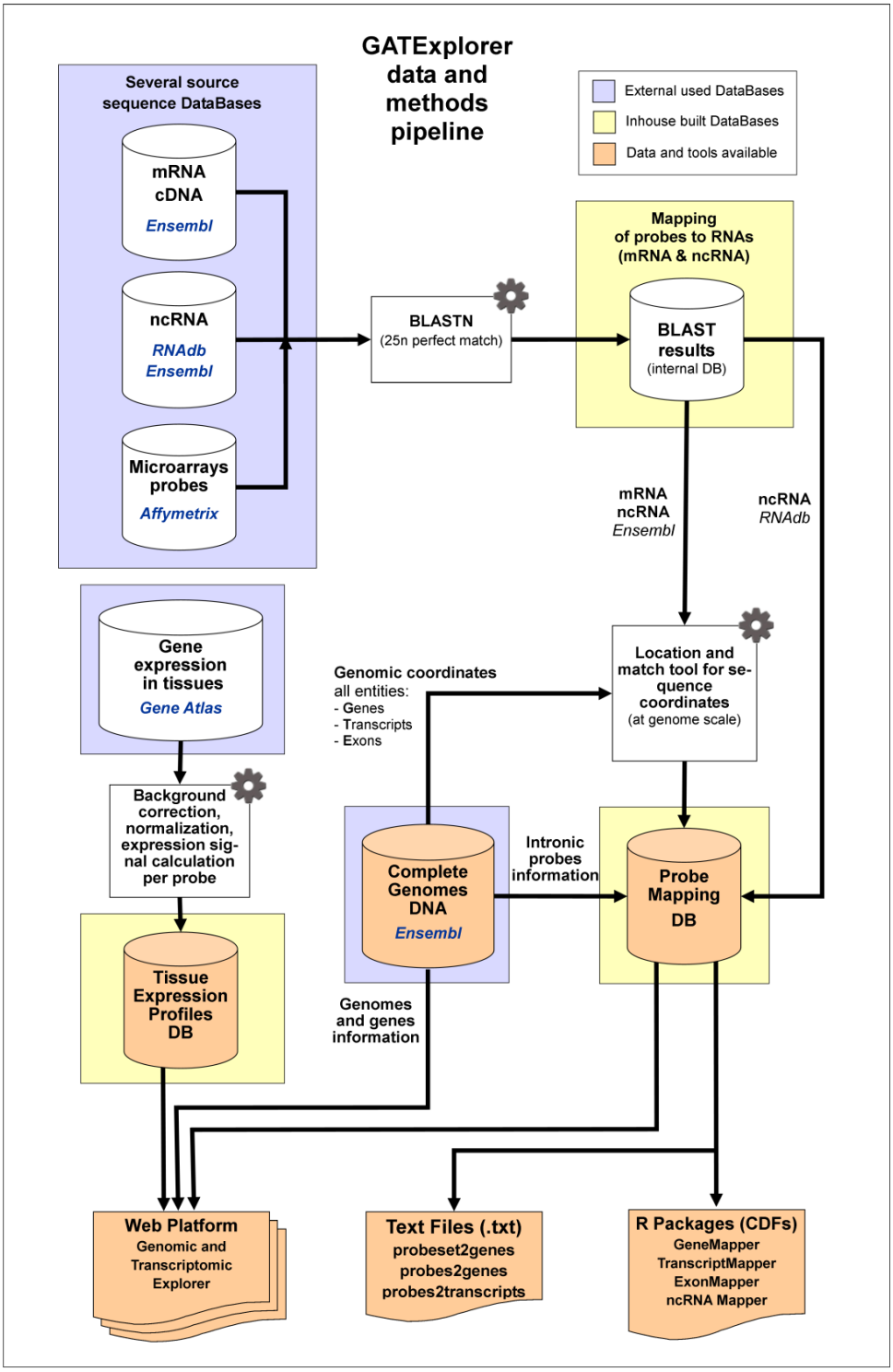


Figure 1 Graphical view of the pipeline followed to build the GATExplorer application, showing the data sources integrated (*Ensembl*, *RNAdb*, *Affymetrix* and *GeneAtlas* -i.e. the four external databases marked in blue square frames-) and a schematic view of the main methods applied at each step -marked with grey gears. As a result, the outputs from the pipeline provide the web platform, which encloses several inhouse built databases -marked in **yellow** square frames- and several data and tools -marked in **orange**-, including the files (Text files) and packages (R CDFs) with the *de novo* mapping of the array probes to the transcriptome of the corresponding organism (human, mouse or rat). The application also provides the genomic coordinates to each locus (including mapping on exons, introns and along exon-exon junctions) and identification of any intersecting genes, transcripts, exons and ncRNAs.

biological entities (i.e. >1 gene or transcript or exon) are identified. The probes that are transcript-specific or exon-specific are provided in a link ("Probes ... specific" link) to another page that includes the list of corresponding *Ensembl* ENST or ENSE IDs. The "Expression view" provides the expression profile of the queried gene in a set of different organs, tissues or cell types (Su *et al*. 2004 dataset [6]) and shows the expression signal per probe for each of the probes assigned to this gene locus or the global expression signal as log2 of the mean of all probes.

We provide an **example** to facilitate the use of the described views and tools included in GATExplorer: human gene MEST (mesoderm-specific transcript homolog genes, *Ensembl* ENSG00000106484). This gene is located on chromosome 7 and its locus is 20.08 Kbp long. It has 4 transcripts and 16 exons. It is mapped by 175 distinct *Affymetrix* probes, which are included amonst 9 different microarray platforms. In the case of array *HGU133 plus 2*, 11 probes map to this gene, which correspond to *Affymetrix* probeset 202016_at. This probeset does not include any "transcript-specific probe" because all probes map to the 4 known transcripts. For this gene the highest expression measured corresponds to bone marrow samples.

Utility and discussion

Mapping genes, transcripts and exons: coverage and efficiency

Accurate expression determination requires that microarray probes have minimal cross-hybridization with other genes or other transcribed entities. The GATExplorer database includes detailed information regarding the *coverage* and *efficiency* of the probe mapping (Tables 1, 2). *Coverage* is defined as the proportion (i.e. %) of gene loci or transcripts from the total genes/transcripts of the *Ensembl* genomes (human, mouse or rat) that are mapped by the probes of a given microarray. *Efficiency* is defined as the proportion (%) of probes from the total probes of a given microarray that map to *Ensembl* genes or transcripts. The term "unique mapped" refers to those gene loci or transcripts that are targeted by a set of probes of a given microarray that do not cross-hybridize (i.e. map unambiguously) with any other known gene loci or transcript.

The quantity and percentage of human gene loci and transcripts targeted by the most widely used human *Affymetrix* expression microarrays based on 3' expression (U133A and U133 Plus 2.0) and the new all-exon arrays (Gene 1.0 and Exon 1.0) is summarized in Table 1. The data shows that the Gene 1.0 and Exon 1.0 arrays achieve the highest coverage over gene loci: 82.57% and 95.82%, respectively (mapping to a total of 27184 human genes, obtained from genome assembly *Ensembl v53 NCBI36*). Such coverage (which depends on the quality of the genome annotation) has improved with respect to previous array models; for example, U133A shows 55.36% coverage of the current *Ensembl* genes. The transcript coverage also improves in the newer models (mapping to a total of 53024 human transcripts, obtained from the same genome assembly). However, the coverage decreases when "unique mapped" genes or transcripts are considered. For example, in the case of the human Gene 1.0 array, 73.29% of the genes are mapped by unique sets of probes. In any case, the overall coverage to measure expression from most human gene loci has improved by 27%, from U133A (55%) to Gene 1.0 (82%).

With respect to the efficiency of the probe mapping, Table 2 presents the number and percentage of distinct probes in each microarray (i.e. probes of distinct sequence) that map to one or more transcripts or gene loci, for the most commonly used human microarray models. Therefore, the columns with >1 include the probes that map to more than one transcript or gene locus. Those probes that map to several transcripts or loci, can be considered "ambiguous" probes. The figures show that the best mapping efficiency (88.41%) is obtained with the Gene 1.0 array. For the U133A array, 78.86% of the probes map to known gene loci of the current human genome version (*Ensembl v53 NCBI36*). The mapping efficiency decreases even further, to 74.5% (180188/241898), when only probes that hybridize to one gene locus are considered (e.g. 180188 probes for U133A). Therefore, probe mappings to human cDNA show that a significant portion (5.54% for U133A and 7.44% for Gene 1.0) hybridize "ambiguously" to more than one gene locus. A larger percentage of probes (55.40% for U133A and 45.55% for Gene 1.0) can hybridize to more than one transcript. Therefore, only a certain percentage of probes can be regarded as gene-specific or transcript-specific. As a general conclusion, these calculations indicate that a significant proportion of probes (about 20 to 25% when mapping genes with U133A) can produce noise using standard expression signal calculations based on the probesets assigned by *Affymetrix*.

The described problem is also present in the new Exon 1.0 arrays, which show the lowest efficiency with only 25.5% of the probes mapping to known genes. This apparently low efficiency is not contradictory with a newly manufactured array, because the Exon 1.0 arrays have been designed with a different goal to previous gene expression arrays, which was not just to cover known genes, but to be able to distinguish the expression of each exon in a given locus. To achieve this goal, the array includes a complex collection of exon probes that correspond to five types of probesets based on different degrees of evidence: core, extended, full, ambiguous and free. Descriptions of each of these probesets can be found in the *Affymetrix* white paper on the Exon arrays (called

Table 1: *Coverage* of the probe mapping.

	Transcripts				Gene Loci				TOTAL N of Transcripts	TOTAL N of Gene Loci
	Unique mapped		All mapped		Unique mapped		All mapped			
	N transcripts	%	N transcripts	%	N gene loci	%	N gene loci	%		
Microarray										
HG U133A	7198	**13,57%**	31219	**58,88%**	12218	**44,95%**	15048	**55,36%**	53024	27184
HG U133 Plus 2.0	11755	**22,17%**	42819	**80,75%**	17710	**65,15%**	20764	**76,38%**	53024	27184
Human Gene 1.0	19947	**37,62%**	48169	**90,84%**	19923	**73,29%**	22446	**82,57%**	53024	27184
Human Exon 1.0	28024	**52,85%**	51851	**97,79%**	23967	**88,17%**	26047	**95,82%**	53024	27184

Coverage of the probe mapping for four human *Affymetrix* microarray models: U133A, U133 Plus 2.0, Gene 1.0 and Exon 1.0. *Coverage* is defined as the proportion of gene loci or transcripts from the total genes/transcripts of the *Ensembl* human genome that are mapped by the probes of a given microarray. The term "unique mapped" indicates the gene loci or transcripts mapped by a unique set of probes of a given array that do not cross-hybridize with any other known gene loci or transcript (i.e. such probes are not ambiguous). N corresponds to number of gene loci or transcripts.

exon_array_design_technote.pdf), which is available from the *Affymetrix* website http://www.affymetrix.com/support/help/exon_glossary. The most important probesets correspond to the "core" type, which are the ones supported by the most reliable evidence from *RefSeq* and full-length mRNA *GenBank* records containing complete CDS information (see exon_array_design_technote.pdf). Recent analytical tools for the Exon 1.0 arrays recommend use of just the "core" set [7]. In the case of human Exon 1.0, the "core" set is composed of 1,082,385 probes http://www.aroma-project.org/chipTypes/ and these probes are mostly included in the set of 1,252,500 probes that GATExplorer assigns to mRNAs exons for this array (see Figure 2). These numbers show that the probe remapping data used in GATExplorer allows the use of a larger set of probes than the "core" set described by *Affymetrix*.

The analysis of coverage and efficiency presented in Tables 1 and 2 should be considered together with the analysis presented in Figure 2, which includes information about the mapping and assignment of all probes from different arrays to protein-coding genes and to ncRNAs (derived from *Ensembl* and from *RNAdb*). The first part, corresponding to the assignment to mRNAs, is marked in green in the pie graphs in Figure 2. These green sectors indicate the proportion of probes that map to known genes, which are large in the case of expression arrays specially designed to measure genes, as human U133A and Gene 1.0 with green sectors of 78.0% and 71.8%, respectively. The blue sector corresponds to ncRNAs, which is exclusively provided by the GATExplorer database. In this sector, we have included those probes that map within "introns" because these probes may measure signal from putative exons included in alternative mRNAs. Indeed, the proportion of probes mapping in such putative or hypothetical exons is comparatively large in the exon arrays (1,352,630 probes of human Exon 1.0), showing that this array is designed to measure many alternative mRNAs. As indicated above, the exon arrays are not only designed to detect mRNAs and for this reason they include many other probes apparently not assigned to any RNA (NA, red sector in pie graphs). Some of these probes correspond to oligos designed for the exon boundaries or for the UTR 3' and 5' borders. Many of these probes map to non-genic regions with little evidence to support their transcription. However, it has been reported that the UTR regions of many human and mouse genes are not well defined, so expression of some of these probes would be anticipated [8]. Nevertheless most of the probes on the human Exon 1.0 array are outside of the "core" reliable set. Complete information and statistics regarding the coverage and efficiency of each microarray platform from human, mouse and rat are included in the GATExplorer database in the PROBE MAPPING section.

Comparison of the use of genes versus probesets in expression calculations

In the expression profiles provided by the *GEO* database and in the many associated publications http://www.ncbi.nlm.nih.gov/geo, the most common approach to calculate gene expression signals is using the *Affyme-*

Table 2: *Efficiency* of the probe mapping.

	Transcripts				Gene Loci				TOTAL N of probes mapped	TOTAL N of probes in the microarray	Mapping efficiency
	1		>1		1		>1				
	N probes	%	N probes	%	N probes	%	N probes	%			
Microarray											
HG U133A	85075	**44,60%**	105677	**55,40%**	180188	**94,46%**	10564	**5,54%**	190752	241898	**78,86%**
HG U133 Plus 2.0	150060	**47,74%**	164260	**52,26%**	299482	**95,28%**	14838	**4,72%**	314320	594532	**52,87%**
Human Gene 1.0	387229	**54,45%**	323912	**45,55%**	658258	**92,56%**	52883	**7,44%**	711141	804372	**88,41%**
Human Exon 1.0	614549	**45,75%**	728669	**54,25%**	1263553	**94,07%**	79665	**5,93%**	1343218	5270588	**25,49%**

Efficiency of the probe mapping for four human *Affymetrix* microarray models: U133A, U133 Plus 2.0, Gene 1.0 and Exon 1.0. *Efficiency* is defined as the proportion of probes from the total probes of a given microarray that map to *Ensembl* human genes or transcripts. N corresponds to number of probes.

trix probesets as direct synonyms of genes. The underlying assumptions carry considerable risks and there are few comparative expression studies that investigate the value of using up-to-date mapping of probes to genes, although it has been reported that this approach improves the precision and accuracy of microarrays [9]. Therefore, to investigate how the application of the remapping may affect the expression data, we present in Figure 3 the results of a comparative study of several microarray datasets that were analyzed either using the standard Chip Definition Files (CDFs) to "probeset" or using the new Chip Definition Files (CDFs) that include the "gene-specific" remapping and assignment, provided by GATExplorer. These analyses are performed using first three different expression signal calculation algorithms (**MAS5.0**, **FARMS** and **RMA**) with CDFs to "probesets" and then using **RMA** with CDFs to "genes" (i.e. using the *GeneMapper* packages) [10-12]. These three algorithms are well-known (*Affycomp* website affycomp.biostat.jhsph.edu) and **RMA** is nowadays the most widely used to calculate microarray gene expression signals [13,14]. Following the application of the expression calculation algorithms with different CDFs, a common robust algorithm for differential expression called **SAM** was applied to all the data [15]. All the analyses were performed using R and the *BioConductor* packages (see website: http://www.bioconductor.org/).

The datasets are a collection of mouse microarray experiments (including four different *Affymetrix* platforms) corresponding to five sets of six samples. Each set includes three biological replicates of knock-out (KO) mice for a specific gene that are compared to three biological replicates of the corresponding wild-type (WT) mice. The five gene KOs are: APOE-/-, IRS2-/-, NRAS-/-, SCD1-/- and ENG+/-. The full name of these genes, the *Ensembl* ID number (ENSG) and the probesets assigned by *Affymetrix* are indicated in Figure 3. Three genes have a unique *Affymetrix* probeset (APOE, IRS2 and ENG) and two genes have two probesets (NRAS and SCD1).

The main feature to be evaluated in the comparison is: how the mapping with the CDFs to "probesets" and the mapping with the CDFs to "genes" affect the detection of the KO genes. In optimum conditions, the gene that is not present in the KO mice should suffer one of the most dramatic differences when compared with the WT and show a significant "repression" or "down-regulation". *A priori* we do not know how many other genes can be affected by the KO gene and we do not know the overall biological/functional signature associated to each KO gene. For this reason, we do not assume that the KO gene will always be the most repressed.

The data and statistical parameters calculated in the comparison, shown in Figure 3, are: **(1)** full name of the gene, corresponding probesets assigned by *Affymetrix* and *Ensembl* ENSG ID number; **(2)** rank of the KO gene across down-regulated genes; **(3)** rank of the KO gene across all genes; **(4)** p-value from **SAM** for the KO gene; **(5)** d-value from **SAM** for the KO gene; **(6)** number of significant genes with q-value < 0.10 (using the assignment of probesets to genes provided either by *Affymetrix*

Mapping of microarray probes to transcribed entities:	HG U133A		HG U133 Plus 2.0		Human Gene 1.0		Human Exon 1.0	
mRNAs (EXONs encoding for proteins)	**188720**	78.0%	**311022**	52.3%	**577259**	71.8%	**1252500**	23.8%
ncRNAs	**24013**	9.9%	**157445**	26.5%	**176525**	21.9%	**1564362**	29.7%
INTRONs (as putative RNAs)	14916		91306		29527		1352630	
ncRNAs (*Ensembl*)	2032		3298		133882		90718	
ncRNAs (*RNAdb*)	11860		92094		27555		190188	
Not Assigned to RNAs (NA)	**29165**	12.1%	**126065**	21.2%	**50588**	6.3%	**2453843**	46.6%
TOTAL number of probes	**241898**		**594532**		**804372**		**5270588**	

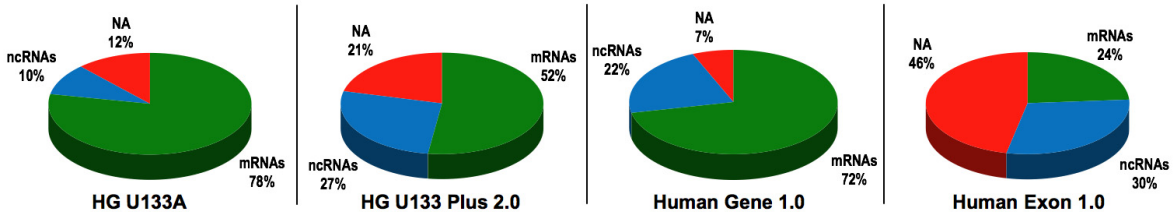


Figure 2 Table and pie graphs summarizing the probe mapping from four widely used human *Affymetrix* microarray models (U133A, U133 Plus 2.0, Gene 1.0 and Exon 1.0) to different transcribed entities. The figure includes information about the mapping and assignment of probes considering tree types: **(i) green**, probes mapped to translated genes (mRNAs, i.e. exons encoding for proteins); **(ii) blue**, probes mapped to ncRNAs (derived from *RNAdb* and from genes labeled as non-protein-coding in *Ensembl*); **(iii) red**, probes not mapped to any known RNA (NA, not assigned).

or by *GeneMapper*); **(7)** total number of mouse genes assigned within the microarray; **(8)** percentage of significant genes with respect to the total.

The highest ranked statistical value among the four comparisons is highlighted in yellow (although, it is important to note that the highest statistical rank does not imply the most biologically relevant change). In four out of five cases (IRS2, NRAS, SCD1 and ENG) the newly calculated gene mapping provides a better rank than the standard mapping, according to the statistical significance of the differential expression of the KO gene. The number of genes with *q-value* < 0.10 (which indicates the extension of the significant change) was the largest with the newly calculated gene mapping for two KO genes: APOE, 9.29% changed genes; NRAS, 0.28% changed genes. Finally, the *p-value* of the statistical test was lowest with the new mapping for KO genes ENG and IRS2. The results consistently indicate that the method using CDFs with the new remapping to "genes" provides at least as significant changes as the best of the three methods based on *Affymetrix* "probesets" CDFs.

We emphasize that the purpose of these analyses is not to propose a new algorithm, but rather to determine in a comparative approach whether the array probe remappings provide results that are at least of equal quality to the original probesets. A complete evaluation of the methods will need a deep biological and functional analysis of the results that goes beyond the scope of this paper. To facilitate further analysis and independent comparison, we provide in the website the raw datasets (CEL files) corresponding to the results presented in Figure 3. Moreover, APOE, NRAS and SCD1 microarray samples can be downloaded in GEO database: GSE2372, GSE14829 and GSE2926, respectively http://www.ncbi.nlm.nih.gov/geo.

Remapping expression probes to ncRNAs

As indicated above, the proportion of probes on the human arrays that map to genes was 78.9% for U133A, 52.9% for U133 Plus 2.0 and 25.5% for Exon 1.0. This efficiency is fractionally lower when only "protein-coding gene loci" (i.e. loci that encode mRNAs translated to proteins) are considered: 78.0% for U133A, 52.3% for U133 Plus 2.0 and 23.8% for Exon 1.0 arrays. This shows that a large fraction of probes within these microarrays do not map to any known protein-coding RNA (i.e. mRNA). Therefore, we performed a remapping of those probes not assigned to mRNAs to a database of ncRNA sequences (*RNAdb v.1* from 2009) [4]. These ncRNAs belong to the mRNA-like class of long ncRNAs. These were predominantly identified in cDNA libraries, such as those used in the *FANTOM3* and *H-Invitational* datasets [1,16]. Because cDNA library generation typically involves poly-dT priming, such cDNA sequences largely arise from polyadenylated transcripts. However, due to the possibility of internal priming from polyA-rich tracts, some non-polyadenylated transcripts may also be present. Nevertheless, such transcripts are represented in any gene expression study that employs a microarray protocol that selectively amplifies polyadenylated transcripts by poly-dT priming. The new generation *Affymetrix* microarrays Gene 1.0 and Exon 1.0 use WT random primed amplification, which does not necessitate the

		MAS5 with CDF to *Affymetrix* probesets	FARMS with CDF to *Affymetrix* probesets	RMA with CDF to *Affymetrix* probesets	RMA with CDF to genes (using *GeneMapper*)
APOE -/-	**Entity (1)**	**apolipoprotein e**		1432466_a_at [1]	ENSMUSG00000002985
3 WT vs 3 KO	Rank (in DOWN) (2)	38	17	2	17
mouse4302	Rank (in ALL) (3)	250	72	22	47
45101 gp	p-value (4)	0.00128600	0.00085227	0.00005321	0.00043362
	d-value (5)	-4.76	-9.21	-8.54	-9.31
	nº gn loci q-value<0.10 (6)	2	208	1350	1564
	nº gn loci total (7)	24100	24100	24100	16835
	% (nº glc sig / nº glc total) (8)	0.01	0.86	5.60	9.29
IRS2 -/-	**Entity (1)**	**insulin receptor substrate 2**		1443969_at [1]	ENSMUSG00000038894
3 WT vs 3 KO	Rank (in DOWN) (2)	45	82	10	4
moe430a	Rank (in ALL) (3)	99	137	19	8
22690 gp	p-value (4)	0.00374174	0.01178727	0.00011018	0.00009661
	d-value (5)	-3.41	-4.40	-3.82	-3.94
	nº gn loci q-value<0.10 (6)	2	0	12	2
	nº gn loci total (7)	13702	13702	13702	12421
	% (nº glc sig / nº glc total) (8)	0.01	0.00	0.09	0.02
NRAS -/-	**Entity (1)**	**neuroblastoma ras oncogene**		94362_at [1] & 160925_at [2]	ENSMUSG00000027852 *
3 WT vs 3 KO	Rank (in DOWN) (2)	1 & >200	1 & 22	1 & 17	1
mgu74av2	Rank (in ALL) (3)	4 & >200	6 & 48	1 & 53	2
12488 gp	p-value (4)	0.00006406	0.00016576	0.00000801	0.00002552
	d-value (5)	-6.97	-15.53	-15.48	-10.83
	nº gn loci q-value<0.10 (6)	2	0	10	22
	nº gn loci total (7)	9557	9557	9557	7837
	% (nº glc sig / nº glc total) (8)	0.02	0.00	0.10	0.28
SCD1 -/-	**Entity (1)**	**stearoyl-Coenzyme A desaturase 1**		94056_at [1] & 94057_g_at [2]	ENSMUSG00000037071
3 WT vs 3 KO	Rank (in DOWN) (2)	2 & 1	18 & 1	2 & 1	1
mgu74a	Rank (in ALL) (3)	2 & 1	22 & 1	2 & 1	1
12654 gp	p-value (4)	0.00000790	0.00001486	0.00000791	0.00001407
	d-value (5)	-15.61	-33.45	-16.60	-11.32
	nº gn loci q-value<0.10 (6)	2	2414	2589	2049
	nº gn loci total (7)	9662	9662	9662	7122
	% (nº glc sig / nº glc total) (8)	0.02	24.98	26.80	28.77
ENG +/-	**Entity (1)**	**endoglin**		1417271_a_at [1]	ENSMUSG00000026814
3 WT vs 3 KO	Rank (in DOWN) (2)	32	28	2	2
moe430a	Rank (in ALL) (3)	40	32	2	3
22690 gp	p-value (4)	0.00174967	0.00694714	0.00004847	0.00004025
	d-value (5)	-5.34	-4.74	-3.21	-3.62
	nº gn loci q-value<0.10 (6)	0	0	1	0
	nº gn loci total (7)	13702	13702	13702	12421
	% (nº glc sig / nº glc total) (8)	0.00	0.00	0.01	0.00

Entity (1)	name of the KO gene; probesets for this gene in the microarray; ENSEMBL gene ID
Rank (in DOWN) (2)	rank of the KO gene in the list of DOWN-regulated significant genes ordered by p-value
Rank (in ALL) (3)	rank of the KO gene in the list of ALL significant genes ordered by p-value
p-value (4)	p-value of the KO gene in the analysis with SAM
d-value (5)	d-value of the KO gene in the analysis with SAM
nº gn loci q-value<0.10 (6)	number of significant gene loci with a q-value lower than 0.10 in the analysis with SAM
nº gn loci total (7)	total number of mouse gene loci assigned within the microarray
% (nº glc sig / nº glc total) (8)	percentage of significant gene loci with respect to the total
*	this gene (NRAS) correspond to ENSEMBL v47, all the rest to v53

Figure 3 Comparison of the differential expression calculated by the SAM algorithm for a series of data of mouse microarrays (five sets of six samples) analyzed using three different expression signal calculation algorithms (MAS5.0, FARMS and RMA) with standard CDFs to "probesets" or using RMA with CDFs to "genes" (*GeneMapper* CDFs). Each set includes three biological replicates of knock-out (KO) mice for a specific gene compared to three replicates of the corresponding wild-type (WT) mice. The gene KOs are: APOE-/-, IRS2-/-, NRAS-/-, SCD1-/- and ENG+/-. The full name of these genes, the *Ensembl* ID number (ENSG) and the probesets assigned by *Affymetrix* are indicated in the top line of each set, labelled Entry **(1)**. The table shows the numbers for the statistical parameters calculated in the comparison, which are: **(2)** rank of the KO gene across down-regulated genes; **(3)** rank of the KO gene across all genes; **(4)** p-value from SAM for the KO gene; **(5)** d-value from SAM for the KO gene; **(6)** number of significant gene loci with q-value < 0.10 (this calculation was performed such that all probesets were assigned to specific genes following the *Affymetrix* assignment or the *GeneMapper* assignment; therefore the methods are comparable since the number of gene loci indicated are the fraction of total mouse genes assigned); **(7)** total number of mouse gene loci assigned within the microarray; **(8)** percentage of significant gene loci with respect to the total. Yellow background indicates the top values for each statistical parameter calculated with each of the four procedures used. The comparison that includes the identical methods for expression calculation (**RMA**) and for differential expression (**SAM**) changing only the CDFs is presented in the last two columns, framed with a black line.

presence of a poly-A tail. This type of microarrays can detect many more ncRNAs.

The results of the remapping of array probes to ncRNAs showed that 29.7% of the probes from human Exon 1.0 and 26.5% of the probes from U133 Plus 2.0 map to ncRNAs. A summary of this remapping is presented in Figure 2, which shows the percentages of probes that map to ncRNAs, combining both the information from *RNAdb* and from genes labeled as non-protein-coding in

Ensembl. These data also include the probes that only map within introns.

Expression of ncRNAs is becoming of increasing interest due to the accumulating evidence showing that ncRNAs are biologically relevant. A key question in investigating ncRNA function is determining whether ncRNAs produce distinct expression signals or just reflect background "noise". To check the variability and detectability of changes in expression provided by the ncRNAs, we selected the 92,094 probes from the U133 Plus 2.0 array that map to ncRNAs according to our mapping to *RNAdb* (see Figure 2). This set of probes is provided in the CDF package "ncRNA Mapper" (file ncrnamapperhgu133plus2cdf_1.0). This CDF file was applied to a microarray dataset obtained from GEO (GSE3526), which includes 353 arrays corresponding to samples from 65 different normal human tissues. The expression of the ncRNAs assigned by the CDF package was calculated using the **RMA** algorithm. Following the calculation of the expression signals, we determined the number of ncRNAs showing significant differential expression by performing a statistical analysis of variance using an *anova* test (function *aov* from *stats* R package). This analysis indicated that 70.5% of the assigned ncRNAs showed differential expression with p-values < 0.01 (p-values corrected by *Bonferroni* method). This means that 4,274 ncRNAs (out of 6,062) changed in all replicates in at least one set of tissues. These results reveal the importance of considering expression signals coming from ncRNAs in transcriptomic studies. Moreover, there is an increasing number of reports showing the biological importance of new transcribed entities that do not encode for proteins, and demonstrate that many play important and diverse roles in cellular function [17,18].

Conclusions

Beyond the "gene" in microarray studies

Genome-wide expression studies have developed exponentially in recent years due to the use of microarray technology [5]. Presently, the most reproducible and widely used microarrays are high-density oligonucleotide microarrays, which feature synthetic oligos based on cDNA and EST sequences. New high-throughput RNA-sequencing will become an excellent complement to microarray datasets, providing highly detailed information about all transcribed entities [19]. However, due to the large number of studies performed with expression microarrays (both past and present) it remains worthwhile to improve the manner by which these data are analyzed. The specific assignment of array probes to current gene annotations, transcripts and exons and the provision of tools to visualize array expression signals in an updated genomic context represent a significant enhancement to currently available methods. With the continued erosion of traditional definitions of "gene" being exposed through transcriptomic sequence data [20,21], as well as the increasing importance of ncRNAs in understanding disease and development [17], the data and analytical techniques enabled by GATExplorer comprise an important aspect for the meaningful interpretation of microarray expression information and its integration within the transcriptome.

The first attempt to provide alternative mapping of *Affymetrix* microarray probes to the latest versions of human genes was reported by Gautier *et al*. in 2004 [22]. Since this report, several studies have been published providing redefinitions of *Affymetrix* microarray probe and probesets to genes and transcripts, including tools to use such redefinitions [22-29]. Dai *et al*. developed probably the most comprehensive mapping of microarray probes from several species [23]. Despite the reannotation of *Affymetrix* microarray probes and probesets to genes and transcripts having been reported previously, GATExplorer is the first system that integrates mapping of probes (including maps to ncRNAs) with simple genomic contextual views, as well as expression signals at probe level. A study and comparison of the characteristics of five major applications that have undertaken an alternative mapping of probes to genes for *Affymetrix* microarrays can be seen in the "Comparative Analysis" page of the PROBE MAPPING section of GATExplorer (the comparison corresponds to references [22-24,28] and this work).

Regarding the visualization of the probes in a genomic context, current genome browsers (such as the *UCSC* browser: http://genome.ucsc.edu/, and *Ensembl* browser: http://www.ensembl.org) incorporate large amounts of data with complex genome-wide information, including location of the probesets from microarrays. Other web sites, like *X:map*, provide specific annotation and visualization of *Affymetrix* exon arrays probesets and probes within the genome structure [30]. *Exon Array Analyzer* is a web tool that allows analysis of exon arrays to detect differentially expressed exons and places the probes within the corresponding genes [31]. The open-source software *BioConductor* (http://www.bioconductor.org/), includes a package called *GenomeGraphs* to plot genomic information from *Ensembl*, which uses *biomaRt* to query the genomes database and transform gene/transcript structures to graphical views, with the possibility of including probes from exon arrays. *Affymetrix* has also developed an application for visualization and exploration of genes, genomes and genome-scale data sets, called *Integrated Genome Browser* (IGB) that uses the *GenoViz* and *Genometry* software (http://genoviz.sourceforge.net/). Although these applications are useful, they fulfil demands that differ to that of GATExplorer. The importance of the work presented here lies in the demonstra-

tion of the large proportion of probe targets that microarrays detect which most researchers do not consider in an expression experiment, and to allow them to use the expression signals arising from non-coding RNAs and hypothetical exons.

In conclusion, GATExplorer is an integrated database and web platform that is useful to visualize, analyze and explore the increasing complexity of eukaryotic transcriptomes (human, mouse and rat), which includes microarray probes mapping to gene loci, transcripts and exons (even exon-exon junctions), as well as introns and ncRNAs.

Availability and requirements

The database is available at http://bioinfow.dep.usal.es/xgate/. GATExplorer is open access and the website makes available all the files and packages described here. The website also includes a "Help" section to facilitate the use of the application.

Authors' contributions

AR is the main developer and current system manager of the database, who has carried out most of the programming code including several interactive tools and the construction of the relational database. CF has helped in the software development and in the improvement of the web pages. She has also contributed in the active discussion and correction of the manuscript. MED has provided updated ncRNAs databases and has actively contributed to the improvement of the web pages of the database and to the writing of the manuscript. JR has made a major contribution to conception and design of the database and several of the tools included, he is the corresponding author who carried out the writing of the manuscript and revising its intellectual content. All authors read and approved the final manuscript.

Acknowledgements

We thank the financial support provided by the spanish Ministry of Science and Innovation MICINN (ISCiii, ref. projects PI061153 and PS09/00843) and the local government Junta de Castilla y León (ref. projects CSI03A06 and CSI07A09).

Author Details

[1]Bioinformatics and Functional Genomics Research Group, Cancer Research Center (CiC-IBMCC, CSIC/USAL), Salamanca, Spain and [2]Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia

Received: 23 October 2009 Accepted: 29 April 2010
Published: 29 April 2010

References

1. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, *et al.*: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309:**1559-1663.
2. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316:**1484-1488.
3. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, *et al.*: **Ensembl 2009.** *Nucleic Acids Res* 2009, **37:**D690-D697.
4. Pang KC, Stephen S, Dinger ME, Engström PG, Lenhard B, Mattick JS: **RNAdb 2.0: an expanded database of mammalian non-coding RNAs.** *Nucleic Acids Res* 2007, **35:**D178-D182.
5. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R: **NCBI GEO: archive for high-throughput functional genomic data.** *Nucleic Acids Res* 2009, **37:**D885-D890.
6. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101:**6062-6067.
7. Gellert P, Uchida S, Braun T: **Exon Array Analyzer: a web interface for Affymetrix exon array analysis.** *Bioinformatics* 2009, **25:**3323-3324.
8. Muro EM, Herrington R, Janmohamed S, Frelin C, Andrade-Navarro MA, Iscove NN: **Identification of gene 3' ends by automated EST cluster analysis.** *Proc Natl Acad Sci USA* 2008, **105:**20286-20290.
9. Sandberg R, Larsson O: **Improved precision and accuracy for microarrays using updated probe set definitions.** *BMC Bioinformatics* 2007, **8:**48.
10. Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho MH, Baid J, Smeekens SP: **Analysis of high density expression microarrays with signed-rank call algorithms.** *Bioinformatics* 2002, **18:**1593-1599.
11. Hochreiter S, Clevert DA, Obermayer K: **A new summarization method for *Affymetrix* probe level data.** *Bioinformatics* 2006, **22(8):**943-949.
12. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of *Affymetrix* GeneChip probe level data.** *Nucleic Acids Res* 2003, **31(4):**e15.
13. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19:**185-193.
14. Barash Y, Dehan E, Krupsky M, Franklin W, Geraci M, Friedman N, Kaminski N: **Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays.** *Bioinformatics* 2004, **20:**839-846.
15. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98:**5116-5121.
16. Genome Information Integration Project And H-Invitational 2, Yamasaki C, Murakami K, Fujii Y, Sato Y, Harada E, Takeda J, Taniya T, Sakate R, Kikugawa S, Shimada M, Tanino M, Koyanagi KO, Barrero RA, Gough C, Chun HW, Habara T, Hanaoka H, Hayakawa Y, Hilton PB, Kaneko Y, Kanno M, Kawahara Y, Kawamura T, Matsuya A, Nagata N, Nishikata K, Noda AO, Nurimoto S, Saichi N, Sakai H, *et al.*: **The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts.** *Nucleic Acids Res* 2008, **36:**D793-799.
17. Mercer TR, Dinger ME, Mattick JS: **Long noncoding RNAs: insights into functions.** *Nat Rev Genet* 2009, **10:**155-159.
18. Dinger ME, Amaral PP, Mercer TR, Mattick JS: **Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications.** *Brief Funct Genomic Proteomic* 2009, **8:**407-423.
19. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Res* 2008, **18:**1509-1517.
20. Brent MR: **Genome annotation past, present, and future: How to define an ORF at each locus.** *Genome Res* 2005, **15:**1777-1786.
21. Mattick JS, Taft RJ, Faulkner GJ: **A global view of genomic information - moving beyond the gene and the master regulator.** *Trends Genet* 2009, **26:**21-28.
22. Gautier L, Møller M, Friis-Hansen L, Knudsen S: **Alternative mapping of probes to genes for Affymetrix chips.** *BMC Bioinformatics* 2004, **5:**111.
23. Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F: **Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.** *Nucleic Acids Res* 2005, **33:**e175.
24. Harbig J, Sprinkle R, Enkemann SA: **A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array.** *Nucleic Acids Res* 2005, **33:**e31.
25. Carter SL, Eklund AC, Mecham BH, Kohane IS, Szallasi Z: **Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray**

probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics* 2005, **6:**107.
26. Leong HS, Yates T, Wilson C, Miller CJ: **ADAPT: a database of affymetrix probesets and transcripts.** *Bioinformatics* 2005, **21:**2552-2553.
27. Lu J, Lee JC, Salit ML, Cam MC: **Transcript-based redefinition of grouped oligonucleotide probe sets using AceView: high-resolution annotation for microarrays.** *BMC Bioinformatics* 2007, **8:**108.
28. Liu H, Zeeberg BR, Qu G, Koru AG, Ferrucci A, Kahn A, Ryan MC, Nuhanovic A, Munson PJ, Reinhold WC, Kane DW, Weinstein JN: **AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets.** *Bioinformatics* 2007, **23:**2385-2390.
29. Ferrari F, Bortoluzzi S, Coppe A, Sirota A, Safran M, Shmoish M, Ferrari S, Lancet D, Danieli GA, Bicciato S: **Novel definition files for human GeneChips based on GeneAnnot.** *BMC Bioinformatics* 2007, **8:**446.
30. Yates T, Okoniewski MJ, Miller CJ: **X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis.** *Nucleic Acids Res* 2008, **36:**D780-786.
31. Gellert P, Uchida S, Braun T: **Exon Array Analyzer: a web interface for Affymetrix exon array analysis.** *Bioinformatics* 2009, **25:**3323-3324.

doi: 10.1186/1471-2105-11-221
Cite this article as: Risueño *et al.*, GATExplorer: Genomic and Transcriptomic Explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs *BMC Bioinformatics* 2010, **11**:221

Leukemia (2010) 24, 629–637

www.nature.com/leu

ORIGINAL ARTICLE

Deregulation of microRNA expression in the different genetic subtypes of multiple myeloma and correlation with gene expression profiling

NC Gutiérrez[1,4], ME Sarasquete[1,4], I Misiewicz-Krzeminska[1], M Delgado[1], J De Las Rivas[2], FV Ticona[1], E Fermiñán[3], P Martín-Jiménez[1], C Chillón[1], A Risueño[2], JM Hernández[1], R García-Sanz[1], M González[1] and JF San Miguel[1]

[1]*Servicio de Hematología, Hospital Universitario, Centro de Investigación del Cáncer-IBMCC (USAL-CSIC), Salamanca, Spain;* [2]*Grupo de Bioinformática y Genómica Funcional, Centro de Investigación del Cáncer-IBMCC (USAL-CSIC), Salamanca, Spain and* [3]*Unidad de Genómica, Centro de Investigación del Cáncer-IBMCC (USAL-CSIC), Salamanca, Spain*

Specific microRNA (miRNA) signatures have been associated with different cytogenetic subtypes in acute leukemias. This finding prompted us to investigate potential associations between genetic abnormalities in multiple myeloma (MM) and singular miRNA expression profiles. Moreover, global gene expression profiling was also analyzed to find correlated miRNA gene expression and select miRNA target genes that show such correlation. For this purpose, we analyzed the expression level of 365 miRNAs and the gene expression profiling in 60 newly diagnosed MM patients, selected to represent the most relevant recurrent genetic abnormalities. Supervised analysis showed significantly deregulated miRNAs in the different cytogenetic subtypes as compared with normal PC. It is interesting to note that miR-1 and miR-133a clustered on the same chromosomal loci, were specifically over-expressed in the cases with t(14;16). The analysis of the relationship between miRNA expression and their respective target genes showed a conserved inverse correlation between several miRNAs deregulated in MM cells and *CCND2* expression level. These results illustrate, for the first time, that miRNA expression pattern in MM is associated with genetic abnormalities, and that the correlation of the expression profile of miRNA and their putative mRNA targets is useful to find statistically significant protein-coding genes in MM pathogenesis associated with changes in specific miRNAs.
Leukemia (2010) **24,** 629–637; doi:10.1038/leu.2009.274; published online 7 January 2010
Keywords: microRNA; myeloma; gene expression

Introduction

The genetics of multiple myeloma (MM) has been increasingly investigated in recent years.[1,2] Genomic data generated by high-throughput technologies in the last decade, particularly by gene expression profiling analysis, has contributed to demonstrate the enormous genetic diversity exhibited by MM[3–6] and genetic classifications which incorporate genomic signatures, cyclin D expression, ploidy status and translocations of the immunoglobulin heavy-chain gene (*IGH*) have been proposed. Their final goal is to identify a connection between clinical behaviour of MM patients and biological features of myeloma cells to eventually individualize treatment.[5,6] However, all these advances in the understanding of MM biology are not sufficient to explain the genesis and evolution of this malignancy. The discovery of small non-coding RNAs called microRNAs (miRNA), which control gene expression at post-transcriptional level, by degrading or repressing target mRNAs, revealed a new mechanism of gene regulation.[7,8] It is well-known that miRNAs are involved in critical biological processes, including cellular growth and differentiation.[9] miRNA expression patterns have been explored in several hematological malignancies, such as chronic lymphocytic leukemia[10,11] and acute myeloid leukemia;[12,13] however, the available information in MM is limited.[14,15] Pichiorri *et al*,[14] have recently investigated the possible role of miRNA in the malignant transformation of plasma cells (PCs) using 49 MM-derived cell lines and a small number of MM (16) and monoclonal gammopathies of undetermined significance (6) patients. In acute leukemias and recently in chronic lymphocytic leukemia, specific miRNA signatures have been associated with different genetic subtypes.[12,13,16–18] Following this approach, we wanted to search for such types of associations in MM patients. For this purpose, we have investigated the expression level of 365 miRNAs by quantitative PCR in 60 primary MM patient samples, specifically selected according to their cytogenetic features, in order to include the most relevant genetic abnormalities in MM. Although animal miRNAs were initially reported to function as translational repressors without mRNA cleavage, it is now evident that miRNAs can also induce mRNA degradation, even if the target sites do not have complete complementarity to the miRNA.[19,20] These findings prompted us to investigate the global gene expression profiling in the same 60 patients to look for candidate mRNAs that were susceptible to miRNA induced knockdown. Our results show the presence of deregulation in miRNA expression of MM cells, which seems to be associated with cytogenetic abnormalities and correlated with gene-expression changes characteristic of MM genetic subtypes.

Materials and methods

Patients

In all, 60 patients with symptomatic newly diagnosed MM were included in the study. Five healthy controls of bone marrow (BM) samples were obtained from subjects undergoing BM harvest for allogeneic transplantation. In all the BM samples a CD138 positive PC isolation using the AutoMACs automated separation system (Miltenyi-Biotec, Auburn, CA, USA) was carried out (purity was above 90%). All patients as well as healthy donors provided written informed consent in accordance with the Helsinki Declaration, and the research ethics committee of the University Hospital of Salamanca approved the study. The total 65 samples were analyzed from both miRNAs and mRNA gene expression profiling.

Correspondence: Professor JF San Miguel, Hospital Universitario de Salamanca, Paseo de San Vicente, 58-182, Salamanca 37007, Spain. E-mail: sanmigiz@usal.es
[4]These authors contributed equally to this work.
Received 9 June 2009; revised 27 October 2009; accepted 12 November 2009; published online 7 January 2010

Cytogenetic analysis

The selection of patients was based on cytogenetic features in order include a representative number of samples with the most relevant and recurrent genetic abnormalities. The systematic screening for genomic aberrations in our institution includes interphase fluorescence *in situ* hybridization studies for the detection of *IGH* rearrangements, *RB1* and *P53* deletions (Abbott Molecular/Vysis, Des Plaines, IL, USA) as previously described,[21] and 1q gains (ON 1q21/SRD 1p36, Kreatech Diagnostics, Amsterdam). Furthermore, only patients with more than 80% of PCs exhibiting genetic abnormalities were considered for the analysis, with the exception of gains on 1q and *P53* deletions, which were present in a median of 77% (range, 28–100%) and 72% (range, 46–88%) of the PCs, respectively. The distribution of cytogenetic abnormalities in the 60 MM patients is summarized in Table 1.

RNA extraction

Total RNA was extracted from normal and tumor plasma cells using miRNEasy Mini Kit (Qiagen, Valencia, CA, USA) following manufacturer's protocol. The RNA integrity was assessed using Agilent 2100 Bioanalyzer (Agilent Tech, Palo Alto, CA, USA).

miRNA profiling

Complementary DNA was synthesized from total RNA using the so-called hairpin RT primer according to the TaqMan miRNA Reverse Transcription Kit (PE Applied Biosystems, Foster City, CA, USA). Reverse transcriptase reactions contained the following: 20 ng of RNA, 1.5 μl 10X RT buffer, 0.15 μl dNTP mix (100 mM total), 1 μl MultiScribe reverse transcriptase (50 U/μl), 0.19 μl AB RNase inhibitor (20 U/μl), 1 μl multiplex RT primers and 4.16 μl H_2O (final volume 10 μl). Reactions were incubated in an Applied Biosystems GeneAmp PCR System 9700 for 30 min at 16 °C, 30 min at 42 °C, 5 min at 85 °C and then held at 4 °C. A total of eight independent RT reactions must be run per sample. Diluted RT reaction product is mixed with TaqMan® Universal PCR Mastermix (no AmpErase UNG) and loaded into the corresponding TaqMan low-density arrays fill ports (Applied Biosystems, part number: 4384792). This panel contains 368 TaqMan miRNA assays enabling accurate quantification of 365 human miRNAs and three endogenous controls (RNU48, RNU48 and RNU6B) to aid in data normalization. Real-time PCR was carried out using an Applied Biosystems 7900 HT Fast Real-time PCR sequence detection system. The reactions were incubated at 94.5 °C for 10 min, followed by 50 cycles of 97 °C for 30 s and 59.7 °C for 1 min. The threshold cycle (Ct) data was determined using 0.3 as a threshold. The Ct is defined as the fractional cycle number at which the fluorescence passes the fixed threshold. Full miRNA data are available at the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/, accession number GSE16558).

Table 1 Cytogenetic characteristics of the 60 patients with newly diagnosed MM

Cytogenetic abnormalities and their combinations[a]	*No. of cases*
t(4;14)	17
With *RB* deletion	14
With *P53* deletion	5
With 1q gains	11
t(11;14)	11
With *RB* deletion	1
t(14;16)	4
With *RB* deletion	2
With 1q gains	2
RB deletion as a unique abnormality	14
With 1q gains	8
RB deletion + *P53* deletion	1
Normal FISH	13

Abbreviations: FISH, fluorescence *in situ* hybridization; MM, multiple myeloma.
[a]Only patients with more than 80% of PC exhibiting genetic abnormalities were considered for the analysis.

miRNAs with Ct values higher than 35 were excluded from the analysis, leaving a set of 192 miRNAs. Normalization was carried out with the mean of RNU44 and RNU48, as they were uniformly expressed across the patient dataset. Relative quantification of miRNA expression was calculated with the $2^{-\Delta Ct}$ and $2^{-\Delta\Delta Ct}$ methods, where $\Delta Ct = Ct_{(miRNA)} - Ct_{(control\ miRNA)}$ and $\Delta\Delta Ct = \Delta Ct_{(MM)} - \Delta Ct_{(average\ normal\ PC)}$, depending upon whether comparisons were made between MM samples and normal PC or between MM samples, respectively.[22] The data was presented as $\log_{10}$ of the relative quantity of each miRNA.[23]

We used hierarchical clustering (Cluster and TreeView software) based on the average-linkage method with the centered correlation metric for unsupervised analysis.[24] Differentially expressed miRNAs were identified using Significant Analysis of Microarrays (SAM) algorithm, by using the two-class (unpaired) format, not considering equal variances.[25] Significant genes were selected based on false discovery ratio and controlling the q-value for the gene list. To estimate whether a given group of MM patients were significantly associated with a set of miRNAs, a Global Test algorithm was used.[26] This method is based on a prediction model for predicting a response variable from the expression measurements of a set of bioentities. The null hypothesis to be tested is that the expression profile of the miRNAs set selected is not associated with the samples. If the null hypothesis were true, then the expected influence parameter should be 0. The Global Test algorithm used in this study corresponds to its version in R, a package called globaltest.

mRNA gene expression profiling

RNA labeling and microarray hybridization have been previously reported.[27] Briefly, 100–300 ng of total RNA were amplified and labeled using the WT Sense Target labelling and control reagents kit (Affymetrix, Santa Clara, CA, USA), and then hybridized to Human Gene 1.0 ST Array (Affymetrix). Washing and scanning were carried out using *GeneChip* System of Affymetrix (GeneChip Hybridization Oven 640, GeneChip Fluidics Station 450 and GeneChip Scanner 7G). Full microarray data compliant with the MIAME guidelines (http://www.mged.org/) are available at the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo/, accession number GSE16558). Expression value for each probe set was calculated using RMAExpress program that uses RMA (Robust Multi-Array Average) algorithm.[28] A differential expression analysis was carried out on the data using SAM algorithm to identify genes with statistically significant changes in expression between different classes.

miRNA target prediction and correlation with gene expression

Target prediction was carried out using miRecords, an integrated resource for animal miRNA-target interactions.[29] The Predicted

Targets component of miRecords integrates the predicted targets of the following miRNA target prediction tools: DIANA-microT, MicroInspector, miRanda, MirTarget2, miTarget, NBmiRTar, PicTar, PITA, RNA22, RNAhybrid and TargetScan.

To identify the putative mRNA targets of the significant miRNAs, which at the same time were selected as significantly deregulated genes in the class comparison analysis of gene-expression data, as well as to look for associations between intronic miRNAs and host genes, a Pearson correlation analysis was carried out using a cutoff *p*-value <0.05 (two-tailed). Calculation of the *p*-values assigned to the Pearson's correlation coefficients was done assuming that the variables follow a bivariate normal distribution.

Even after extensive studies, it is not well known what makes that an mRNA follows the cleavage or translational repression pathway. There are experimental analyses which support the presence of specific molecular requirements that can contribute to a preferential degradation of the messages after introducing a miRNA.[30] On the basis of that we have used four programs (TargetScan 5.1; miRDB; miRanda; and Pictar) which, in addition to consider the perfect complementarity between miRNA and mRNA in the 'seed' region, also use for predictions the following requirements: 3′ region pairing, conservation of the site, untranslated region (UTR) context, number of sites and free energy of the complex. To eliminate the probable false-positive predictions, we have recognized, as probably leading to mRNA degradation, only those miRNA–mRNA pairs that were predicted by at least three programs. The mRNA databases or the versions of the same mRNA database may be different in each program. Occasionally, it can then occur that the sequences of some genes were not present in a database, so the prediction of miRNA target could not be carried out. Only in these cases we considered those genes as probable targets that were predicted by two programs. To further eliminate the possibility of false-positive predictions we have chosen only these predictions that had simultaneously prediction score higher than: −0.3 for TargetScan; 60 for miRDB; 2 for Pictar and 150 for miRanda. Detailed information about prediction criteria of every program is included in the Supplementary Methods.

The functional analysis to identify the most relevant functional categories in the datasets of miRNA target genes selected by statistical analysis was generated through the use of Ingenuity Pathways Analysis (Ingenuity Systems, Mountain View, CA, USA). To analyze the effect of target genes on biology of multiple myeloma we have analyzed the molecular function of each gene. To do this we have searched subsequent databases: Entrez, GeneCard, UniProtKB/SwissProt and KEGG.

Results

Unsupervised analysis of miRNA expression

Initially, we investigated the miRNA expression data set using unsupervised analysis. A two-dimensional clustering analysis carried out on the filtered data revealed that the MM samples were not distributed into clearly separated clusters according to chromosomal abnormalities. Nevertheless, the four samples with t(14;16) were tightly clustered, and MM with t(4;14) and those with *RB* deletion showed a trend to be segregated into the same subclusters (Supplementary Figure 1).

miRNAs differentially expressed in the cytogenetic subtypes of myeloma

It is well-known that subtle differences in gene expression among closely related neoplastic subtypes may escape detection using unsupervised clustering analysis, whereas supervised methods can detect them. Therefore, in a supervised approach, we looked for miRNAs significantly deregulated in the different cytogenetic subtypes of MM compared with normal PC. For this purpose, we used the SAM algorithm with an false discovery rate (FDR) <0.001 in all class comparisons. As compared with normal PC, we identified 11, 8, 7, 37 and 18 miRNAs differentially expressed in MM cells from patients with t(4;14), t(14;16), t(11;14), *RB* deletion as a unique abnormality and cytogenetically normal MM samples, respectively. Significantly deregulated miRNAs with their chromosomal band location, in the different cytogenetic subsets are presented in Table 2. Only miR-214 and miR-375 were commonly deregulated in these five comparisons (both being underexpressed). A previous report has described that the inhibition of mir-214 was associated with decreased apoptosis in HeLa cells.[31] The 11 miRNAs significantly deregulated in MM cells with t(4;14) were downregulated compared with normal PC. miRNA-203 and miRNA-342 were located at 14q32, but they were not clustered together. Similarly, using the same restrictive FDR, the seven miRNAs deregulated in MM patients with t(11;14) were also underexpressed. Those patients with t(14;16) showed a singular signature characterized by the underexpression of five miRNAs and the overexpression of three miRNAs. It is noteworthy to mention, that miR-1 and miR-133a, both upregulated in MM with t(14;16), were clustered on the same chromosomal locus at 18q11 (Figure 1). When we compared the expression of these two miRNAs in MM with t(14;16) with their expression in the other cytogenetics groups we observed that these miRNA were overexpressed only in the four cases with t(14;16) (Figure 2).

All the miRNAs deregulated in patients with *RB* deletion compared with normal PC samples were infra-expressed and interestingly, the miR-20a, miR-19b, miR-15a and miR-19a, were located in 13q. Furthermore, these miRNAs were not significantly downregulated in the other genetic subtypes. We confirmed this finding when we compared the miRNA expression between MM patients with only *RB* deletion and MM patients without this abnormality. Among the set of 8 miRNAs downregulated in the samples with *RB* deletion, miR-20a, miR-18a and miR-19b, located at 13q31, were included. The reason why the other miRNAs localized in 13q, such as miR-16-1 at 13q14, as well as miR-622 at 13q31, were not infra-expressed in our study is unknown, but differences in PCR efficiency due to highly multiplexed pools in the reverse transcription reaction cannot be excluded from consideration.

In this series we also included 13 patients with normal fluorescence *in situ* hybridization (FISH) analysis, which displayed a particular miRNAs signature characterized by the downregulation of 18 miRNAs. Mir-362 and miR-501 were clustered together at Xp11. We failed to find a specific miRNA expression profile for MM cases with 1q gains or *P53* deletion. Probably, their association with other abnormalities, such as t(4;14) and *RB* deletion, prevents the identification of miRNA exclusively deregulated in MM with these abnormalities. Nevertheless, clustering analysis of MM with t(4;14) and MM with *RB* deletion did not segregate samples according to 1q status.

When we applied the Global Test algorithm to all the MM samples (the only sample with monosomy 13 and *P53* deletion was not considered) using the 49 miRNAs found significative (miRNAs duplicated in the different MM subgroups were eliminated), the different groups of samples agreed with the miRNAs signature and they provided a positive 'influence' in such signature. The MM subtype that had higher average influence in the signature was that with t(14,16) and the one that

Table 2 Deregulated microRNAs (miRNAs) in each of the cytogenetic groups in comparison with normal plasma cell (NPC); all the miRNAs were downregulated except for miR-1, miR-449 and miR-133a, which were upregulated in multiple myeloma (MM) with t(14;16). miRNAs are listed according to the statistical significance as determined by Significant Analysis of Microarrays (SAM) analysis (false discovery rate (FDR) <0.001)

Cytogenetic group	*Deregulated miRNA*	*Chromosomal region*
t(4;14) (N = 17)		
	hsa-miR-203	14q32
	hsa-miR-155	21q21
	hsa-miR-650	22q11
	hsa-miR-375	2q35
	hsa-miR-196b	7p15
	hsa-miR-342	14q32
	hsa-miR-214	1q24
	hsa-miR-193a	17q11
	hsa-miR-135b	1q32
	hsa-miR-146a	5q33
	hsa-miR-133b	6p12
t(11;14) (N = 11)		
	hsa-miR-650	22q11
	hsa-miR-125a	19q13
	hsa-miR-375	2q35
	hsa-miR-184	15q25
	hsa-miR-214	1q24
	hsa-miR-95	4p16
	hsa-miR-199a	19p13
t(14;16) (N = 4)		
	hsa-miR-1	18q11
	hsa-miR-449	5q11
	hsa-miR-133a	18q11
	hsa-miR-196b	7p15
	hsa-miR-135b	1q32
	hsa-miR-214	1q24
	hsa-miR-375	2q35
	hsa-miR-642	19q13
RB deletion as a unique abnormality (N = 14)		
	hsa-miR-196a	17q21
	hsa-miR-486	8p11
	hsa-miR-375	2q35
	hsa-miR-501	Xp11
	hsa-miR-320	8p21
	hsa-miR-20a	13q31
	hsa-miR-133b	6p12
	hsa-miR-135b	1q32
	hsa-miR-126	9q34
	hsa-miR-650	22q11
	hsa-miR-214	1q24
	hsa-miR-19b	13q31
	hsa-miR-10a	17q21
	hsa-miR-15a	13q14
	hsa-miR-133a	18q11
	hsa-miR-139	11q13
	hsa-miR-197	1p13
	hsa-miR-10b	2q31
	hsa-miR-95	4p16
	hsa-miR-126	9q34
	hsa-miR-186	1p31
	hsa-miR-19a	13q31
	hsa-miR-451	17q11
	hsa-let-7b	22q13
	hsa-miR-140	16q22
	hsa-miR-125a	19q13
	hsa-miR-362	Xp11
	hsa-miR-33	22q13
	hsa-miR-223	Xq12
	hsa-miR-224	Xq28
	hsa-miR-221	Xp11
	hsa-miR-30e	1p34
	hsa-miR-374	Xq13
	hsa-let-7c	21q21
	hsa-miR-99b	19q13
	hsa-miR-130a	11q12
	hsa-miR-193a	17q11
Normal fluorescence in situ *hybridization (FISH) (N = 13)*	hsa-miR-135b	1q32
	hsa-miR-375	2q35
	hsa-miR-155	21q21
	hsa-miR-650	22q11
	hsa-miR-572	4p15
	hsa-miR-152	17q21
	hsa-miR-362	Xp11
	hsa-miR-486	8p11
	hsa-miR-95	4p16
	hsa-miR-214	1q24
	hsa-miR-501	Xp11
	hsa-miR-196a	17q21
	hsa-miR-642	19q13
	hsa-miR-10a	17q21
	hsa-miR-452	Xq28
	hsa-miR-342	14q32
	hsa-let-7c	21q21
	hsa-miR-203	14q32

had lower influence was the normal FISH subgroup (Figure 3).

We also compared miRNA expression in the total 60 MM samples with that of normal PC using SAM analysis, and we found 11 miRNAs downregulated in MM patients: miR-375, miR-650, miR-214, miR-135b, miR-196a, miR-155, miR-203, miR-95, miR-486, miR-10a and miR-196b.

Relationship between miRNA and mRNA expression profiling

In addition to the miRNA expression analysis, protein-coding gene expression profiling was carried out for all the MM samples using high-density oligonucleotide microarrays. miRNAs were believed to act mainly by interfering translation rather than reducing the expression levels of the target mRNAs. However, there are studies indicating that the expression regulation at the mRNA level may be a common mechanism for miRNA regulation.[32] Moreover, there is increasing evidence supporting that the simultaneous profiling of mRNA, by using microarrays, and miRNA is an effective strategy for aiding in the detection of functional miRNA targets.[33–35] Accordingly, we combined our miRNA and mRNA expression data sets in order to find deregulated genes than can be specific targets of the miRNAs forming the signature of each MM cytogenetic group. First, we identified those genes with statistically significant changes in the expression between different cytogenetic subclasses of MM and normal PC using SAM algorithm with a cutoff FDR $\leqslant 0.03$. Second, we searched for putative target genes for each significant miRNA using miRecords database, selecting only those target gene transcripts proposed by at least 5 of 11 established miRNA target prediction programs. In a further step, we selected the intersection between miRNA target genes

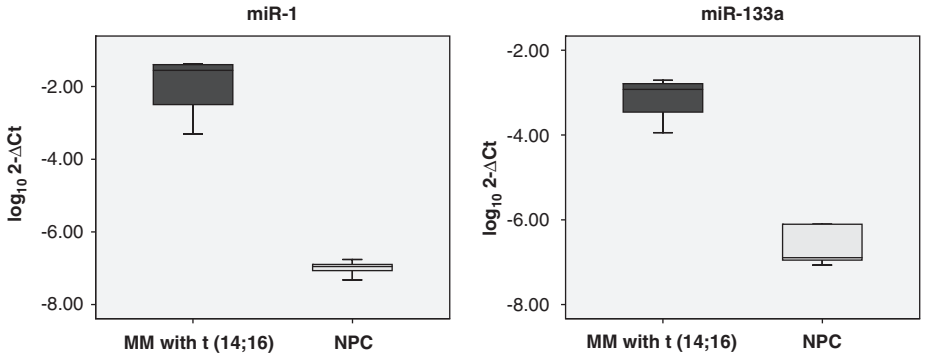


Figure 1 Box plot representation of the normalized expression of miR-1 and miR-133a in MM with t(14;16) and normal PC ($q<0.0001$).

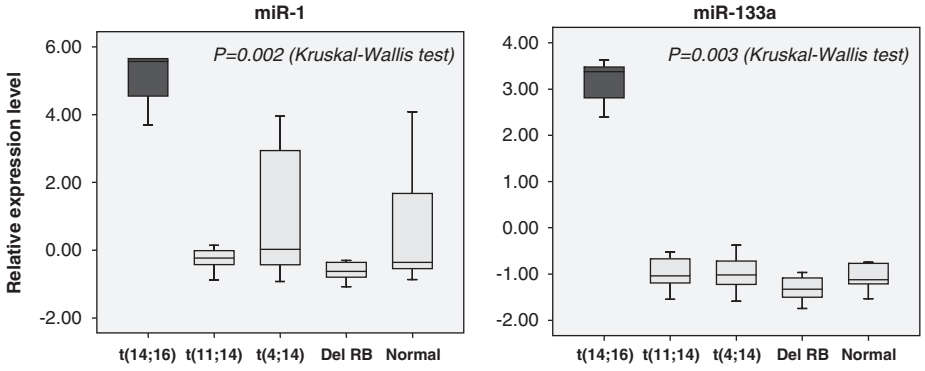


Figure 2 Box plot representation of the normalized expression of miR-1 and miR-133a in MM with t(14;16) compared with the other cytogenetic groups.

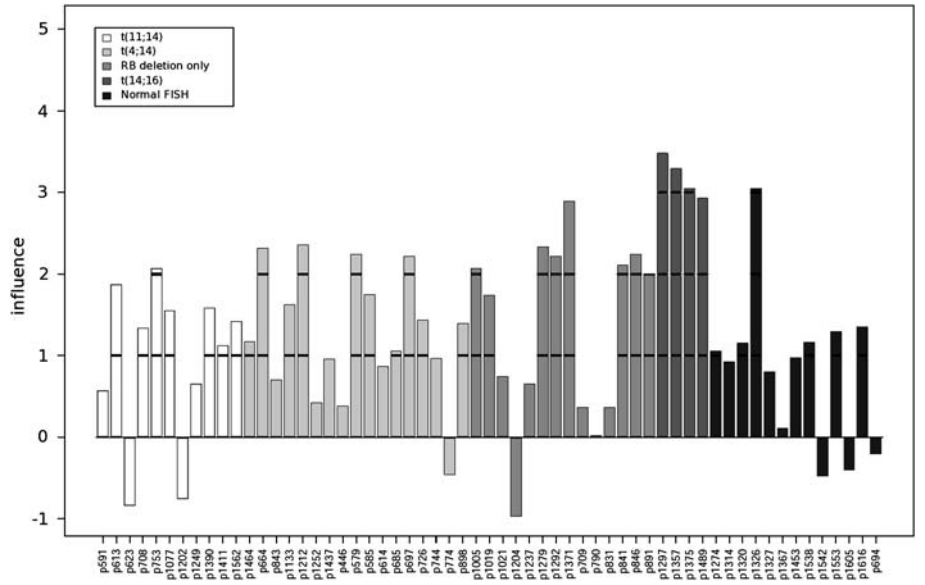


Figure 3 Result of the Global Test algorithm applied to all the MM samples (the sample with monosomy 13 and P53 deletion was not considered). This algorithm shows how each individual sample agrees with the miRNA signature found for MM and constituted by the 49 significant miRNAs. The units in the 'influence plot' mark the s.d. of the influence in the samples with respect to the null hypothesis ($P<0.001$).

and differentially expressed genes to discriminate only significant deregulated genes that are putative targets of the miRNAs associated to MM. Finally, we carried out a statistical correlation analysis between the raw expression levels of the MM-deregulated miRNAs and the MM-deregulated target genes selected. Because of miRNAs function as negative regulators, we selected the significant inverse correlations. The target genes for the miRNAs specifically deregulated in MM with t(4;14), t(14;16), t(11;14) and with monosomy 13 as a unique abnormality are shown in Supplementary Tables 1, 2, 3 and 4,

respectively. Most of the predicted miRNA target genes for the miRNA characterizing MM with t(4;14) and MM with t(14;16) were assigned to the functional categories of cellular growth and proliferation: *JUND, MYBL1, SOCS3, CCND2, SLC7A5* and *IRAK1* upregulated in MM with t(4;14), and *AKAP12, CREG1, EEF2K, IGFBP5, RASSF5, S1PR1, TFAP2C, WNT5A* and *CCND2* upregulated in MM with t(14;16). This observation suggests a possible link between the miRNA downregulation in these MM subsets and an advantage in myeloma cell growth. The majority of candidate miRNA target genes inversely correlated with miRNA signature of MM with monosomy 13 were involved in cell death category in the functional analysis.

Although the analysis strategy up to this point was based on target gene selection from at least five miRNA-target prediction programs and on a further inverse correlation, filters for discriminating those miRNA–mRNA interactions more susceptible to a degradation process were not applied. Therefore, we carried out a final step in the bioinformatic analysis selecting stringent criteria filters from the combination of four programs, including criteria of seed region complementarity and UTR context, which greatly reduces false-positive predictions and provide miRNA–mRNA interactions potentially more sensitive to a cleavage pathway.[20,30,34] The miRNA target genes selected using this more accurate analysis approach are shown in Tables 3 and 4.

A significant number of miRNAs are located within the intronic regions of either coding or non-coding transcription units. The expression of intronic miRNA is believed to be regulated by the expression of the host mRNA.[36] In this study, we selected the intronic miRNAs among the deregulated miRNAs, in order to investigate whether their host genes were also deregulated in the MM gene expression profile. However, we did not find any significant correlation between the intronic miRNAs and their host transcripts.

Discussion

In this study, a genome-wide miRNA expression profiling on MM samples with different cytogenetic abnormalities was carried out. Our results show that the genetic subgroups in MM are associated with singular miRNA signatures. Supervised approaches identified miRNA sets differentially expressed in the MM genetic groups, although not clearly separated clusters containing specific cytogenetic abnormalities were observed. The fact that genetic subgroups were not distinguishable in the miRNA clustering is in contrast to the studies in acute lymphoid and myeloid leukemias, wherein miRNA expression grouped samples according to the genetic rearrangements. It is possible that the complexity of the genetic portrait of myeloma cells with multiple associations of genetic abnormalities within the same cell as compared with that of myeloid and lymphoid blasts, may explain the inability to segregate MM cytogenetic subgroups by miRNA unsupervised clustering.[12,37,38] In fact, there is a strong association between the different genetic abnormalities in MM, particularly between t(4;14) and *RB* deletion and 1q gains.[1]

One of the most significant findings was the upregulation of miR-1 and miR-133a, belonging to a cluster at 18q11, in MM with t(14;16). These two miRNAs have distinct roles in modulating skeletal and cardiac muscle proliferation and differentiation, and their potential function in the MM with t(14;16) deserves further investigation.[39] With respect to the MM samples harboring *RB* deletion, the underexpression of miRNAs located in 13q would reflect a reduction in dosage of these miRNAs as a consequence of monosomy 13, since in most MM cases *RB* deletion detected by FISH represents whole chromosome monosomy. This observation is in agreement with the haploinsufficiency of many of the genes mapped at chromosome 13, demonstrated by gene expression analysis.[40]

Table 3 Potential microRNA (miRNA)–mRNA interactions deregulated in the multiple myeloma (MM) with immunoglobulin heavy-chain gene (*IGH*) translocations

Deregulated target genes[a]	*GenBank ID*	*TargetScan*[b]	*miRBD*[b]	*Pictar*[b]	*miRanda*[b]	*No. of sites (predicted by TargetScan)*
t(4;14)						
hsa-miR-135b (MIMAT0000758)						
PELI2	NM_021255	−0.48	64	7.54	468	2
hsa-miR-146a (MIMAT0000449)						
IRAK1	NM_001569	−0.76	87	6.94	156	2
hsa-miR-133b (MIMAT0000770)						
TAGLN2	NM_003564	−0.43	64	—	289	2
t(11;14)						
hsa-miR-125a-5p (MIMAT0000443)						
ARID3B	NM_006465	−0.49	82	20.42	291	3
MAP3K11	NM_002419	−0.3	55	3.24	150	1
t(14;16)						
hsa-miR-214 (MIMAT0000271)						
TFAP2C	NM_003222	−0.33	65	3.08	149	1
RASSF5	NM_182663	−0.44	61	4.98	291	3
hsa-miR-1 (MIMAT0000416)						
GPD2	NM_001083112	−0.62	84	3.48	—	2
GLCCI1	NM_138426	−0.7	85	0	303	2
FNDC3B	NM_022763	−0.42	66	1.52	293	1
ASH2L	NM_004674	−0.44	65	2.93	153	1
hsa-miR-449 (MIMAT0001541)						
FUT8	NM_178155	−0.45	74	0	158	1

[a]All the target genes were upregulated in the gene expression analysis, except miR-1 and miR-449 target genes which were downregulated.
[b]Prediction score.

Table 4 Potential microRNA (miRNA)–mRNA interactions deregulated in the multiple myeloma (MM) with monosomy 13 as a unique abnormality

Deregulated target genes[a]	*GenBank ID*	*TargetScan*[b]	*miRBD*[b]	*Pictar*[b]	*miRanda*[b]	*No. of sites (predicted by TargetScan)*
hsa-miR-486-5p (MIMAT0002177)						
PIM1	NM_002648	−0.33	74	-	146	2
hsa-miR-320a (MIMAT0000510)						
MLLT3	NM_004529	−0.49	87	5.56	295	2
hsa-miR-20a(MIMAT0000075)						
CDKN1A	NM_078467	−0.3	78	5.27	144	2
ADAM9	NM_003816	−0.38	67	1.68	159	1
FURIN	NM_002569	−0.33	50	6.23	153	2
YES1	NM_005433	−0.32	0	4.28	152	2
hsa-miR-214 (MIMAT0000271)						
RASSF5	NM_182663	−0.44	61	4.98	291	3
PLAGL2	NM_002657	−0.32	60	6.28	443	3
hsa-miR-19b (MIMAT0000074)						
SGK1	NM_005627	−0.56	79	6.83	153	2
IGF1	NM_001111283	−0.48	71	-	314	2
hsa-miR-15a (MIMAT0000068)						
CCND2	NM_001759	−0.33	-	3.43	155	3
hsa-miR-10b (MIMAT0000254)						
KLF11	NM_003597	−0.4	75	2.21	140	1
hsa-miR-186 (MIAMT0000456)						
BTBD3	NM_014962	−0.78	87	0	582	3
UBE2B	NM_003337	−0.45	71	0.31	292	1
FBXO33	NM_203301	−0.51	88	1.28	289	2
hsa-miR-19a (MIMAT0000073)						
KRAS	NM_0033360	−0.35	0	3.72	152	2
IGF1	NM_001111283	−0.48	71	-	314	2
CCND2	NM_001759	−0.46	70	2.99	151	2
MAPK6	NM_002748	−0.43	77	-	296	1
hsa-miR-let-7b (MIMAT0000063)						
MAPK6	NM_002748	−0.4	72	-	158	1
LRIG3	NM_153377	−0.51	93	4.4	166	1
hsa-miR-125a-5p (MIMAT0000443)						
ARID3B	NM_006465	−0.49	82	20.42	291	3
hsa-miR-223 (MIMAT0000280)						
DUSP10	NM_007207	−0.57	80	6.99	308	2
STIM1	NM_003156	−0.3	52	2.57	159	1
hsa-miR-374a (MIMAT0000727)						
GADD45A	NM_001924	−0.54	78	0.92	162	1
PELI1	NM_020651	−0.86	99	3.18	462	3
MAPK6	NM_002748	−0.68	92	-	286	2
hsa-miR-let-7c (MIMAT0000064)						
LRIG3	NM_153377	−0.51	92	4.4	163	1
hsa-miR-130a (MIMAT0000425)						
CFL2	NM_138638	−0.46	77	9.9	281	2

[a]All the target genes were upregulated in the gene expression analysis.
[b]Prediction score.

None of the 11 miRNAs downregulated in the total 60 MM samples has been related to lymphoid cells except for miR-155, which participates in B-cell differentiation and could act as a tumor suppressor by reducing potentially oncogenic translocations generated by *AID* gene.[41] It has also been demonstrated that the failure of miR-155-deficient B cells to generate high-affinity switched antibodies appears not because of a defect in somatic hypermutation or class-switch recombination, but more likely to a defect in the differentiation or survival of plasmablasts.[42] The predominance of infra-expressed miRNAs in myeloma samples is consistent with the global decrease in miRNA levels observed in other human cancers.[37,43,44] Nevertheless, Pichiorri *et al*.[14] have recently reported distinctive miRNA signatures in MM and MGUS characterized mainly by the overexpression of miRNAs. Moreover, the deregulated miRNAs are not the same in the two studies. There are some plausible explanations for these discrepancies. First, the platforms used for miRNA analysis were different and there are around 110 miRNAs only explored by one of the two approaches; second, the number of primary myelomas analysed in this study is higher and the samples are selected according to the genetic abnormalities, whereas the genetic status of MM samples from Pichiorri's study is unknown; and at last, the statistical design in terms of class comparisons and significance level is also different.

miRNA target prediction remains a major bioinformatic challenge, particularly because of false-positive predictions. In this study, we combined several strategies to identify miRNA–mRNA interactions more susceptible to a degradation process. Using these stringent criteria, we found that the miR-135b and miR-146a (Table 3), both downregulated in MM with t(4;14), targeted two genes (*PELI2* and *IRAK1*) involved in IL-1 signalling

pathway, which is a potent inducer of IL-6. IRAK1 (IL-1 receptor-associated kinase) activates and recruits TRAF6 to the IL-1 receptor complex,[45,46] and PELI2 is a scaffold protein probably involved in the IL-1 signalling through its interaction with the complex containing IRAK kinases and TRAF6.[47] Both *IRAK1* and *PELI2* display two matches to the miR-146a and miR-135b respectively, which increases the prediction specificity. In addition, *IRAK1* is a validate target of miR-146a.[48,49] Concerning MM with t(14;16) four genes (*GPD2, GLCCI1, FNDC3B* and *ASH2L*) were targeted by miR-1 (Table 3). A role in cancer has been reported for all of them, although an involvement in myeloma biology has not been found. Among the potential miRNA–mRNA interactions detected in MM patients with monosomy 13, it should be noted that miR-19a and miR-19b, belonging the miR-17∼92 cluster, target the overexpressed *IGF1* transcript (Table 4). Likewise, *MAPK6* is targeted by three of the miRNAs downregulated in MM with monosomy 13 (miR-19a, miR-let-7b and miR-374a). As *MAPK* route and *IGFR1* pathway have important roles in the control of MM cell biology, the investigation of these miRNAs interactions would be interesting.[50,51] When we applied this stringent approach to the 11 miRNAs deregulated in all MM comparing with normal PC no predicted miRNA–mRNA interactions fulfilled the filters, but miR-10a and its predicted target *BACH2*, and miR-135b and its putative target *CD47* were very close to the established thresholds. The transcription factor *BACH2*, together with other factors such as *BLIMP1* appears to constitute a main transcriptional regulatory network for the terminal differentiation of B cells to PC.[52] *CD47* is an integrin-associated protein which modulates cell activation and adhesion.[53,54] Our group has reported that *CD47* upregulation is only present in PC from MM, but not in PC from Waldenstrom's macroglobulinemia.[55] The potential role of miR-10a and miR-135b in the *CD47* and *BACH2* expression deserves to be explored.

It is interesting to note that *CCND2*, which was upregulated both in MM subtypes t(4;14), t(14;16) and monosomy 13, have target sites for several miRNA significantly deregulated in these cytogenetic subtypes. It has been proposed that dysregulation of a *CYCLIN D* gene is a unifying oncogenic event in MM.[5] About 45% of MM express *CCND2*, particularly those MM with t(14;16) and t(4;14). However, the mechanism responsible for the increased expression of *CCND2* in MM often remains unknown. Therefore, it could be speculated that the underexpression of miR-196b, miR-135b, miR-320, miR-20a, miR-19b, miR-19a and miR-15a found in many MM (see Supplementary Tables) could contribute to the overexpression of their predicted target *CCND2*. Noteworthy, miR-15a and miRNA-19a were the only ones selected using the stringent analysis and displayed three and two sites in the 3′UTR (Table 4). These results are supported by a recent study, which demonstrates that the transfection of myeloma cells with the pre-miRNA-15a results in *CCND2* inhibition.[15]

Unfortunately, although the presence of perfect seed pairing and other bioinformatic criteria predict miRNA regulation more likely led to mRNA degradation, false positives cannot be ruled out. On the other hand, the possibility that there are additional redundant indirect pathways also contributing to their degradation should be kept in mind. Hence the functional characterization of the results obtained by using bioinformatic analysis is of paramount importance.

In summary, these results indicate that miRNA expression is deregulated in myeloma cells, and what is more important, that the miRNA pattern of expression in MM seems to be associated with specific genetic abnormalities. Of particular interest is our finding of the upregulation of the cluster miR-1/miR-133a in MM samples with t(14;16), which supports a role for miRNAs in the pathogenic pathways of MM putatively delineated by chromosomal aberrations. In addition, the strategy proposed of finding potential miRNA–mRNA interactions could be useful to identify relevant protein-coding genes in MM pathogenesis that can be regulated by miRNAs.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This study was partially supported by Spanish FIS (PI080568) and 'Gerencia Regional de Salud, Junta de Castilla y León' (GRS202/A08) grants, and the Spanish Myeloma Network Program (RD06/0020/0006). M.E.S. was supported by the 'Ministerio de Sanidad y Consumo (Contrato de Técnicos de Apoyo a la Investigación, CA08/00212). J.D.L.R. was supported by Junta de Castilla y León grant (CSI07A09).

We are grateful to 'Grupo Español de Mieloma' clinicians for providing MM samples; to JA Pérez-Simón and F Sánchez-Guijo for providing healthy bone marrow samples; to I. Isidro, T Prieto, A Antón and M Hernández for technical assistance; and to E Bandrés and MD Odero for their support in the data analysis.

References

1 Chng WJ, Glebov O, Bergsagel PL, Kuehl WM. Genetic events in the pathogenesis of multiple myeloma. *Best Pract Res Clin Haematol* 2007; **20**: 571–596.
2 Hideshima T, Bergsagel PL, Kuehl WM, Anderson KC. Advances in biology of multiple myeloma: clinical applications. *Blood* 2004; **104**: 607–618.
3 Davies FE, Dring AM, Li C, Rawstron AC, Shammas MA, O'Connor SM *et al*. Insights into the multistep transformation of MGUS to myeloma using microarray expression analysis. *Blood* 2003; **102**: 4504–4511.
4 Zhan F, Hardin J, Kordsmeier B, Bumm K, Zheng M, Tian E *et al*. Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells. *Blood* 2002; **99**: 1745–1757.
5 Bergsagel PL, Kuehl WM, Zhan F, Sawyer J, Barlogie B, Shaughnessy Jr J. Cyclin D dysregulation: an early and unifying pathogenic event in multiple myeloma. *Blood* 2005; **106**: 296–303.
6 Zhan F, Huang Y, Colla S, Stewart JP, Hanamura I, Gupta S *et al*. The molecular classification of multiple myeloma. *Blood* 2006; **108**: 2020–2028.
7 Calin GA, Croce CM. MicroRNA signatures in human cancers. *Nat Rev Cancer* 2006; **6**: 857–866.
8 Esquela-Kerscher A, Slack FJ. Oncomirs—microRNAs with a role in cancer. *Nat Rev Cancer* 2006; **6**: 259–269.
9 Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004; **116**: 281–297.
10 Calin GA, Liu CG, Sevignani C, Ferracin M, Felli N, Dumitru CD *et al*. MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias. *Proc Natl Acad Sci USA* 2004; **101**: 11755–11760.
11 Cimmino A, Calin GA, Fabbri M, Iorio MV, Ferracin M, Shimizu M *et al*. miR-15 and miR-16 induce apoptosis by targeting BCL2. *Proc Natl Acad Sci USA* 2005; **102**: 13944–13949.
12 Jongen-Lavrencic M, Sun SM, Dijkstra MK, Valk PJ, Lowenberg B. MicroRNA expression profiling in relation to the genetic heterogeneity of acute myeloid leukemia. *Blood* 2008; **111**: 5078–5085.
13 Li Z, Lu J, Sun M, Mi S, Zhang H, Luo RT *et al*. Distinct microRNA expression profiles in acute myeloid leukemia with common translocations. *Proc Natl Acad Sci USA* 2008; **105**: 15535–15540.
14 Pichiorri F, Suh SS, Ladetto M, Kuehl M, Palumbo T, Drandi D *et al*. MicroRNAs regulate critical genes associated with multiple

myeloma pathogenesis. *Proc Natl Acad Sci USA* 2008; **105**: 12885–12890.
15 Roccaro AM, Sacco A, Thompson B, Leleu X, Azab AK, Azab F *et al.* microRNAs 15a and 16 regulate tumor proliferation in multiple myeloma. *Blood* 2009; **113**: 6669–6680.
16 Garzon R, Garofalo M, Martelli MP, Briesewitz R, Wang L, Fernandez-Cymering C *et al.* Distinctive microRNA signature of acute myeloid leukemia bearing cytoplasmic mutated nucleophosmin. *Proc Natl Acad Sci USA* 2008; **105**: 3945–3950.
17 Marcucci G, Radmacher MD, Maharry K, Mrozek K, Ruppert AS, Paschka P *et al.* MicroRNA expression in cytogenetically normal acute myeloid leukemia. *N Engl J Med* 2008; **358**: 1919–1928.
18 Visone R, Rassenti LZ, Veronese A, Taccioli C, Costinean S, Aguda BD *et al.* Karyotype specific microRNA signature in chronic lymphocytic leukemia. *Blood* 2009; **114**: 3872–3879.
19 Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. *Nature* 2008; **455**: 64–71.
20 Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 2009; **136**: 215–233.
21 Gutierrez NC, Castellanos MV, Martin ML, Mateos MV, Hernandez JM, Fernandez M *et al.* Prognostic and biological implications of genetic abnormalities in multiple myeloma undergoing autologous stem cell transplantation: t(4;14) is the most relevant adverse prognostic factor, whereas RB deletion as a unique abnormality is not associated with adverse prognosis. *Leukemia* 2007; **21**: 143–150.
22 Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* 2001; **25**: 402–408.
23 Bandres E, Cubedo E, Agirre X, Malumbres R, Zarate R, Ramirez N *et al.* Identification by real-time PCR of 13 mature microRNAs differentially expressed in colorectal cancer and non-tumoral tissues. *Mol Cancer* 2006; **5**: 29.
24 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.
25 Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001; **98**: 5116–5121.
26 Goeman JJ, van de Geer SA, de KF, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 2004; **20**: 93–99.
27 Gutierrez NC, Lopez-Perez R, Hernandez JM, Isidro I, Gonzalez B, Delgado M *et al.* Gene expression profile reveals deregulation of genes with relevant functions in the different subclasses of acute myeloid leukemia. *Leukemia* 2005; **19**: 402–409.
28 Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* 2003; **31**: e15.
29 Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res* 2009; **37**: D105–D110.
30 Carthew RW, Sontheimer EJ. Origins and mechanisms of miRNAs and siRNAs. *Cell* 2009; **136**: 642–655.
31 Cheng AM, Byrom MW, Shelton J, Ford LP. Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic Acids Res* 2005; **33**: 1290–1297.
32 Bagga S, Bracht J, Hunter S, Massirer K, Holtz J, Eachus R *et al.* Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* 2005; **122**: 553–563.
33 Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* 2005; **433**: 769–773.
34 Wang X, Wang X. Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic Acids Res* 2006; **34**: 1646–1652.
35 Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL *et al.* Using expression profiling data to identify human microRNA targets. *Nat Methods* 2007; **4**: 1045–1049.
36 Li SC, Tang P, Lin WC. Intronic microRNA: discovery and biological implications. *DNA Cell Biol* 2007; **26**: 195–207.
37 Lu J, Getz G, Miska EA, varez-Saavedra E, Lamb J, Peck D *et al.* MicroRNA expression profiles classify human cancers. *Nature* 2005; **435**: 834–838.
38 Mi S, Lu J, Sun M, Li Z, Zhang H, Neilly MB *et al.* MicroRNA expression signatures accurately discriminate acute lymphoblastic leukemia from acute myeloid leukemia. *Proc Natl Acad Sci USA* 2007; **104**: 19971–19976.
39 Xu C, Lu Y, Pan Z, Chu W, Luo X, Lin H *et al.* The muscle-specific microRNAs miR-1 and miR-133 produce opposing effects on apoptosis by targeting HSP60, HSP70 and caspase-9 in cardiomyocytes. *J Cell Sci* 2007; **120**: 3045–3052.
40 Zhan F, Barlogie B, Shaughnessy Jr J. Toward the identification of distinct molecular and clinical entities of multiple myeloma using global gene expression profiling. *Semin Hematol* 2003; **40**: 308–320.
41 Dorsett Y, McBride KM, Jankovic M, Gazumyan A, Thai TH, Robbiani DF *et al.* MicroRNA-155 suppresses activation-induced cytidine deaminase-mediated Myc-Igh translocation. *Immunity* 2008; **28**: 630–638.
42 Vigorito E, Perks KL, breu-Goodger C, Bunting S, Xiang Z, Kohlhaas S *et al.* microRNA-155 regulates the generation of immunoglobulin class-switched plasma cells. *Immunity* 2007; **27**: 847–859.
43 Gaur A, Jewell DA, Liang Y, Ridzon D, Moore JH, Chen C *et al.* Characterization of microRNA expression levels and their biological correlates in human cancer cell lines. *Cancer Res* 2007; **67**: 2456–2468.
44 Kumar MS, Lu J, Mercer KL, Golub TR, Jacks T. Impaired microRNA processing enhances cellular transformation and tumorigenesis. *Nat Genet* 2007; **39**: 673–677.
45 Chen H, Li M, Campbell RA, Burkhardt K, Zhu D, Li SG *et al.* Interference with nuclear factor kappa B and c-Jun NH2-terminal kinase signaling by TRAF6C small interfering RNA inhibits myeloma cell proliferation and enhances apoptosis. *Oncogene* 2006; **25**: 6520–6527.
46 Song KW, Talamas FX, Suttmann RT, Olson PS, Barnett JW, Lee SW *et al.* The kinase activities of interleukin-1 receptor associated kinase (IRAK)-1 and 4 are redundant in the control of inflammatory cytokine expression in human cells. *Mol Immunol* 2009; **46**: 1458–1466.
47 Moynagh PN. The Pellino family: IRAK E3 ligases with emerging roles in innate immune signalling. *Trends Immunol* 2009; **30**: 33–42.
48 Taganov KD, Boldin MP, Chang KJ, Baltimore D. NF-kappaB-dependent induction of microRNA miR-146, an inhibitor targeted to signaling proteins of innate immune responses. *Proc Natl Acad Sci USA* 2006; **103**: 12481–12486.
49 Pauley KM, Satoh M, Chan AL, Bubb MR, Reeves WH, Chan EK. Upregulated miR-146a expression in peripheral blood mononuclear cells from rheumatoid arthritis patients. *Arthritis Res Ther* 2008; **10**: R101.
50 Maiso P, Ocio EM, Garayoa M, Montero JC, Hofmann F, Garcia-Echeverria C *et al.* The insulin-like growth factor-I receptor inhibitor NVP-AEW541 provokes cell cycle arrest and apoptosis in multiple myeloma cells. *Br J Haematol* 2008; **141**: 470–482.
51 Hideshima T, Anderson KC. Molecular mechanisms of novel therapeutic approaches for multiple myeloma. *Nat Rev Cancer* 2002; **2**: 927–937.
52 Igarashi K, Ochiai K, Muto A. Architecture and dynamics of the transcription factor network that regulates B-to-plasma cell differentiation. *J Biochem* 2007; **141**: 783–789.
53 Cooper D, Lindberg FP, Gamble JR, Brown EJ, Vadas MA. Transendothelial migration of neutrophils involves integrin-associated protein (CD47). *Proc Natl Acad Sci USA* 1995; **92**: 3978–3982.
54 Mateo V, Brown EJ, Biron G, Rubio M, Fischer A, Deist FL *et al.* Mechanisms of CD47-induced caspase-independent cell death in normal and leukemic cells: link between phosphatidylserine exposure and cytoskeleton organization. *Blood* 2002; **100**: 2882–2890.
55 Gutierrez NC, Ocio EM, Maiso P, Ferminan E, Delgado M, Lopez-Perez R *et al.* Gene expression profiling of B-lymphocyte and plasma cell populations from Waldenström's macroglobulinemia comparison with expression patterns of the same cell-counterparts from other B-cell neoplasms. *Leukemia* 2007; **21**: 541–549.

Supplementary Information accompanies the paper on the Leukemia website (http://www.nature.com/leu)

OPEN ACCESS Freely available online

PLOS ONE

Molecular Characterization of Chronic Lymphocytic Leukemia Patients with a High Number of Losses in 13q14

Ana Eugenia Rodríguez[1], Jose Ángel Hernández[2], Rocío Benito[1], Norma C. Gutiérrez[3], Juan Luis García[4], María Hernández-Sánchez[1], Alberto Risueño[5,6], M. Eugenia Sarasquete[3], Encarna Fermiñán[7], Rosa Fisac[8], Alfonso García de Coca[9], Guillermo Martín-Núñez[10], Natalia de las Heras[11], Isabel Recio[12], Oliver Gutiérrez[13], Javier De Las Rivas[5], Marcos Gonźalez[1,3], Jesús M. Hernández-Rivas[1,3]*

1 IBSAL,IBMCC, Centro de Investigación del Cáncer, Universidad de Salamanca-CSIC, Salamanca, Spain, **2** Servicio de Hematología, Hospital Universitario Infanta Leonor, Madrid, Spain, **3** Servicio de Hematología, Hospital Clínico Universitario de Salamanca, Salamanca, Spain, **4** Instituto de Estudios de Ciencias de la Salud de Castilla y León, (IECSCYL)–HUSAL, Castilla y León, Spain, **5** Grupo de Bioinformática y Genòmica Funcional, IBMCC, Centro de Investigación del Cáncer, Universidad de Salamanca-CSIC, Salamanca, Spain, **6** Celgene Institute for Translational Research Europe (CITRE), Sevilla, Spain, **7** Unidad de Genómica, IBMCC, Centro de Investigación del Cáncer, Universidad de Salamanca-CSIC, Salamanca, Spain, **8** Servicio de Hematología, Hospital General de Segovia, Segovia, Spain, **9** Servicio de Hematología, Hospital Clínico Universitario, Valladolid, Spain, **10** Servicio de Hematología, Hospital Virgen del Puerto, Plasencia, Spain, **11** Servicio de Hematología, Hospital Virgen Blanca, León, Spain, **12** Servicio de Hematología, Hospital Nuestra Señora de Sonsoles, Ávila, Spain, **13** Servicio de Hematología, Hospital del Río Hortega, Valladolid, Spain

Abstract

Background: Patients with chronic lymphocytic leukemia and 13q deletion as their only FISH abnormality could have a different outcome depending on the number of cells displaying this aberration. Thus, cases with a high number of 13q- cells (13q-H) had both shorter overall survival and time to first therapy. The goal of the study was to analyze the genetic profile of 13q-H patients.

Design and Methods:: A total of 102 samples were studied, 32 of which served as a validation cohort and five were healthy donors.

Results: Chronic lymphocytic leukemia patients with higher percentages of 13q- cells (>80%) showed a different level of gene expression as compared to patients with lower percentages (<80%, 13q-L). This deregulation affected genes involved in apoptosis and proliferation (BCR and NFkB signaling), leading to increased proliferation and decreased apoptosis in 13q-H patients. Deregulation of several microRNAs, such as miR-15a, miR-155, miR-29a and miR-223, was also observed in these patients. In addition, our study also suggests that the gene expression pattern of 13q-H cases could be similar to the patients with 11q- or 17p-.

Conclusions: This study provides new evidence regarding the heterogeneity of 13q deletion in chronic lymphocytic leukemia patients, showing that apoptosis, proliferation as well as miRNA regulation are involved in cases with higher percentages of 13q- cells.

Citation: Rodríguez AE, Hernández JÁ, Benito R, Gutiérrez NC, García JL, et al. (2012) Molecular Characterization of Chronic Lymphocytic Leukemia Patients with a High Number of Losses in 13q14. PLoS ONE 7(11): e48485. doi:10.1371/journal.pone.0048485

Editor: Javier S. Castresana, University of Navarra, Spain

Received July 11, 2012; **Accepted** October 2, 2012; **Published** November 13, 2012

Funding: The study was partially supported by grants from the Spanish Fondo de Investigaciones Sanitarias 02/1041 and FIS 09/01543; Caja de Burgos-Banca Cívica, Proyectos de Investigación del SACYL 106/A/06 and by the Acción Transversal del Cáncer project, through an agreement between the Instituto de Salud Carlos III (ISCIII), the Spanish Ministry of Science and Innovation, the Cancer Research Foundation of Salamanca University and the Redes de Investigación RTIIC (FIS). AR is fully supported by an Ayuda Predoctoral FIS de Formación en Investigación by the Spanish Fondo de Investigaciones Sanitarias. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jmhr@usal.es

Introduction

Chronic lymphocytic leukemia (CLL) is characterized by the progressive accumulation of mature, monoclonal B lymphocytes in the blood, bone marrow (BM) and secondary lymphoid tissues [1]. The clinical course ranges from an indolent disorder with a normal lifespan to a rapidly progressive disease leading to death [2,3]. The variable clinical course of CLL is driven, at least in part, by the immunogenetic and molecular heterogeneity of the disease [4,5]. The genomic aberrations and the immunoglobulin (Ig) VH mutation status provide us with two separate genetic parameters of prognostic relevance. Thus, patients whose leukemic cells express unmutated *IgVH* regions (Ig-unmutated CLL) often have progressive disease, whereas those whose leukemic cells express mutated *IgVH* regions (Ig-mutated CLL) more often have an indolent disease [4,6]. Fluorescent *in situ* hybridization (FISH) can detect genomic abnormalities in more than 80% of CLL cases and the

genetic subtypes of CLL show different biological and clinical features [5]. Although unfavorable aberrations (losses on 17p and 11q) are more frequent in the Ig-unmutated subgroup [7–9], and favorable aberrations (loss on 13q as a single abnormality) are more frequent in the Ig-mutated subgroup, they have independent value in predicting outcome in CLL [8,9].

Deletion at 13q14 (13q-) is the most common genomic aberration in CLL. It is present in more than 50% of cases, and is the sole documented cytogenetic abnormality in 36% of the patients. These latter cases are known to have a more favorable clinical course [5,10]. However, recent data from our group and others, suggest that patients with CLL and 13q deletion as the only FISH abnormality could have a different outcome depending on the number of cells displaying this aberration [11–13]. Moreover, previous studies had demonstrated that the percentage of cells displaying a particular cytogenetic abnormality (e.g. loss of *P53)* [14] or antigenic markers (e.g. CD38 or ZAP-70) [7] can be related to prognosis. We have demonstrated that cases with a high number of 13q- cells (13q-H) usually had both shorter overall survival and time to first therapy. However, to the best of our knowledge the molecular characteristics of 13-H CLLs have not been so far analyzed in detail in order to better understand why these patients have a poor outcome.

The value of gene expression profiling (GEP) in the study of CLL is widely accepted. Such studies have identified new prognosis markers such as ZAP-70, LPL, PEG10 and CLLU1. Some of these are already well-established factors used in clinical practice, while the application of others is under study.

As a next step toward elucidation of biological differences within 13q- subgroup, the current study used the Affymetrix Human Exon arrays 1.0 ST, which offer a more fine-grained view of gene expression than the former generation of chips. Thus, the data obtained provide great insights into the biological mechanisms underlying the clinical differences observed in this CLL subgroup [11–13].

Materials and Methods

Patients

A total of 102 samples were selected for the study, 32 of which served as a validation cohort and five were healthy donors. CLL diagnosis was performed according to the World Health Organization (WHO) classification [15] and the Working Group of National Cancer Institute (NCI) criteria [16]. A complete immunophenotypic analysis by flow cytometry [17] and FISH studies were carried out in all cases. The median age at the time of study was 68 years (range, 35 to 90 years). Most patients were male (66%) and were in Binet clinical stage A (69%), while 26% were in stage B, and the remaining 5% were in stage C. The clinical and biological features of the CLL patients included in the study are shown in Table S1. The study was approved by the local ethical committees ''Comité Ético de Investigación Clínica, Hospital Universitario de Salamanca''. Written informed consent was obtained from each patient before they entered the study.

Methods

B cell isolation. Peripheral blood mononuclear cells (PBMCs) were isolated from fresh peripheral blood samples using Ficoll gradient, snap-frozen and stored at –80°C.

For the validation cohort, CD19-positive B cells were purified by magnetically activated cell sorting (MACS) CD19 MicroBeads (Miltenyi Biotec, Bergisch Gladbach, Germany) resulting in a >98% purity, as analyzed by flow cytometry. CD19-positive normal B cells from peripheral blood of five healthy donors served as controls.

Genomic aberrations. For the purpose of the study, only samples with one cytogenetic abnormality were included. For the gene expression profile analysis, according to our previous results [11], two groups of patients with 13q- were compared: those in whom 80% or more of cells showed 13q- (13q-H) and those in whom fewer than 80% of cells showed 13q losses (13q-L). The distribution of cases in the study cohort was: 13q-H (n = 25; 36%), 13q-L (n = 27; 39%), normal FISH (nCLL, n = 8; 11%) and 17p−/11q- (n = 10; 14%).

In the validation cohort, the distribution of samples was similar: 13q-H (n = 7; 22%), 13q-L (n = 11; 34%) and nCLL (n = 9; 28%). The remaining five cases were healthy donors.

Mutation status of IGVH genes. IGVH genes were amplified and sequenced according to the ERIC recommendations on IGHV gene mutational status analysis in CLL [18].

Global gene expression using high density microarrays. Genome-wide expression analysis of the isolated samples was performed using Human Exon 1.0 ST microarrays (Affymetrix). RNA isolation, labeling and microarray hybridization were carried out following the manufacturer's protocols for the GeneChip platform by *Affymetrix*. Methods included synthesis of first- and second-strand cDNAs, the purification of double-stranded cDNA, synthesis of cRNA by in vitro transcription, recovery and quantization of biotin-labeled cRNA, fragmentation of this cRNA and subsequent hybridization to the microarray slide, post-hybridization washings, and detection of the hybridized cRNAs using a streptavidin-coupled fluorescent dye. Hybridized *Affymetrix* arrays were scanned with an *Affymetrix* Gene-Chip 3000 scanner. Images were generated and features extracted using Affymetrix GCOS Software.

Table 1. miRNAs significantly deregulated between 13q- CLL subgroups (patients with 80% or more of cells with 13q deletion and patients with less than 80% 13q cells).

miRNA	Map	q-value	R fold
Down-regulated			
hsa-mir-1-1*	20q13.33	0.0125	0.7027
hsa-mir-7-1	9q21.32	0.0397	0.5453
hsa-mir-15a	13q14.3	0.0329	0.4917
hsa-mir-29a	7q32.3	0.0354	0.5101
hsa-mir-34a*	1p36.23	0.0366	0.6874
hsa-mir-106b*	7q22.1	0.0280	0.5190
hsa-mir-181b	1q31.3	0.0256	0.6775
hsa-mir-204	9q21.11	0.0294	0.5693
hsa-mir-206	6p12.2	0.0476	0.7077
hsa-mir-221*	Xp11.3	0.0133	0.4622
hsa-mir-223*	Xq12	0.0017	0.1016
Up-regulated			
hsa-mir-134	14q32.31	0.0095	1.8096
hsa-mir-105-2	Xq28	0.0182	1.4040
hsa-mir-155*	21q21.3	0.0046	3.7013
hsa-mir-205	1q32.2	0.0161	1.3830

Upregulation or downregulation refers to 13q-H relative to13q-L CLL patients. miRNA: microRNA.
*deregulation shared with 17p/11q CLL patients.
doi:10.1371/journal.pone.0048485.t001

Table 2. Enriched functional analysis of the 3450 genes differentially expressed between the two 13q- patient subgroups: 1244 genes were upregulated (i) and 2206 genes were downregulated (ii) in CLL patients with ≥80% cells displaying 13q deletion.

	Up-regulated				Down-regulated		
i.	Ingenuity Canonical Pathway	p-value	Up-regulated genes	ii.	Ingenuity Canonical Pathway	p-value	Down-regulated genes
	EIF2 Signaling	1,70E-07	RPL24,RPL27A,RPL26,RPS11,RPS27,RPS3A,SOS1,RPL35, RPL19,RPL13,RPL39L,RPL34,RPL27,RPL21,RPS19,RPL23A, RPS29,RPL36,RRAS2,RPS13,RPL26L1,RPL32,RPS25,RPS15A, RPL13A,RPS27A,RPL41,RPS14,RPSA		Mitotic Roles of Polo-Like Kinase	1,35E-05	KIF23,CDC25C,ESPL1,CDC20,PPP2CA,PRC1,CDC7 (includes EG:12545), CCNB2,CDC23,PLK1,PPP2R5A,CDK1,CCNB1,SLK,HSP90B1,PLK4, PKMYT1,PPP2R1B,KIF11,CDC27,CDC25A
	B Cell Receptor Signaling	1,95E-05	MAP2K6,BLNK,MAP3K14,MAP3K9,CD19,CD79B,BAD, POU2F2,IKBKE,NFATC1,FCGR2B,PTEN,MAP3K12,RRAS2, CAMK2D,SYK,SOS1,CD22,NFATC2,PIK3AP1,PPP3CA,PRKCB		Cell Cycle Control of Chromosomal Replication	2,75E-05	MCM6,CDC45,CDT1,CDC6,CDC7 (includes EG:12545),CDK6,ORC6, MCM4,MCM3,MCM2,CDK2,MCM7,ORC1
	PI3K Signaling in B Lymphocytes	6,92E-05	BLNK,CD19,CD79B,IKBKE,NFATC1,FCGR2B,PRKCZ, PTEN,BLK,CAMK2D,RRAS2,CD180,SYK,IRS1,SH2B2, NFATC2,PIK3AP1,PPP3CA,PRKCB		Caveolar-mediated Endocytosis Signaling	6,17E-04	FYN,ITGA2B,ITSN1,RAB5A,ACTB,COPA,ITGA6,ITGA5,COPB1,ACTG1, COPG,COPB2,DYRK3,ITGB2,ITGAE,ITGAM,ITGA9,ITGAV,HLA-C,ITGA4
	CD27 Signaling in Lymphocytes	2,75E-03	MAP2K6,MAP3K12,MAP3K9,MAP3K14,CD70,IKBKE, TRAF5,CD27,MAP2K5		Glycolysis/ Gluconeogenesis	7,94E-04	PGK1,ALDH4A1,PGM2,PKLR,GAPDH,PGM1,BPGM,PDHA1,HK1,ALDH2, GPI,HK2,ALDH1A1,DHRS9,ENO1,DLAT,DLD,FBP1,ALDH3B1,LDHA,ACSL1
	mTOR Signaling	5,37E-03	VEGFB,RHOC,RPS19,RPS11,PRKCZ,RPS29,RPS27,RPS3A, RRAS2,RPS13,IRS1,GPLD1,RPS15A,RPS25,GNB1L, RPS27A,RPS14,RPSA,PRKCB		Integrin Signaling	1,12E-04	RAP2B,RAF1,FYN,ITGA2B,ARHGAP26,TSPAN7,PIK3R1,PIK3R5,PPP1CB,NCK1, SHC1,ITGAE,PARVB,ARF6,WASL,RHOG,ITGA9,ARF4,PIK3CG,RHOU, ITGAV,VCL,MAP2K1,ACTN1,ITGA4,PXN,NRAS,ASAP1,CRKL,ACTB,ITGA6, TSPAN2,ITGA5,ACTG1,ITGB2,ARF1,ITGAM,TLN2,ZYX,PIK3CB,ACTN4,CTTN
	Role of JAK1 and JAK3 in γc Cytokine Signaling	6,76E-03	BLNK,IL2RG,RRAS2,IRS1,SYK,SH2B2,JAK2,STAT1,IL7		Cyclins and Cell Cycle Regulation	1,17E-03	RAF1,E2F4,CCNE2,TFDP1,HDAC2,PPP2CA,SUV39H1,CDK6,CDKN2C, CCNB2,E2F3,PPP2R5A,CDK1,CCNB1,CCNA2,CCNE1,E2F1,PPP2R1B,E2F2, CDK2,CDC25A
	Nucleotide Excision Repair Pathway	1,20E-02	ERCC4,ERCC1,GTF2H1,ERCC2,MNAT1,XPA		Role of CHK Proteins in Cell Cycle Checkpoint Control	1,32E-03	CDC25C,E2F4,E2F1,RFC2,E2F3,BRCA1,CDK1,E2F2,CDK2,CDC25A, CHEK1,RFC3
	Regulation of eIF4 and p70S6K Signaling	1,66E-02	RPS19,RPS11,PRKCZ,RPS29,RPS27,RRAS2,RPS3A,RPS13, IRS1,SOS1,RPS25,RPS15A,RPS27A,RPS14,RPSA		Nicotinate and Nicotinamide Metabolism	1,66E-03	DAPK1,PRKCQ,SGK1,MAPK6,CSNK1A1,CDK6,CSNK1D,PLK1,TTK,CDK1, SACM1L,VNN1,NEK2,ARAF,GRK6,PRKAA1,PNP,CD38,HIPK1,MAP2K1, NMNAT3,CDK2,BST1,DUSP16
	Phospholipase C Signaling	1,86E-02	BLNK,PEBP1,CD79B,RHOC,MEF2A,HDAC9,NFATC1,FCGR2B, MYL6B,PRKCZ,RRAS2,SYK,SOS1,GPLD1,NFATC2,MEF2C, GNB1L,ARHGEF9,PPP3CA,PRKCB		Inositol Phosphate Metabolism	1,78E-03	MINPP1,SGK1,PIK3R1,PIK3R5,CSNK1A1,TTK,OCRL,NEK2,PIK3CG,PRKAA1, PLCB1,IMPA2,PI4K2B,HIPK1,MAP2K1,PMPCA,MTMR3,DAPK1,IMPA1, PRKCQ,MTMR14,MAPK6,CDK6,CSNK1D,PLK1,CDK1,ARAF,SYNJ1, GRK6,PIK3CB,CDK2
	PKCθ Signaling in T Lymphocytes	1,86E-02	MAP3K12,MAP3K9,MAP3K14,POU2F1,RRAS2,CAMK2D, SOS1,NFATC2,IKBKE,NFATC1,CARD11,PPP3CA		Cell Cycle Regulation by BTG Family Proteins	4,57E-03	CCNE2,E2F4,CCNE1,PPP2CA,E2F1,E2F3,PPP2R1B,CCRN4L, E2F2,CDK2,PPP2R5A
	April Mediated Signaling	2,34E-02	MAP3K14,NFATC2,IKBKE,NFATC1,TRAF5,TNFRSF17		Clathrin-mediated Endocytosis Signaling	6,61E-03	AP2A1,STON2,PIK3R1,PIK3R5,PDGFC,VEGFA,ARF6,ARRB1,WASL,SNX9, PIK3CG,DAB2,CSNK2B,AAK1,AP2M1,RAB5A,ACTB,CHP,CLTC,RAB7A, ITGA5,ACTG1,TSG101,ITGB2,ARRB2,LDLR,SYNJ1,TFRC,PIK3CB,DNM1L,CTTN

Table 2. Cont.

i.	Up-regulated			ii.	Down-regulated		
	Ingenuity Canonical Pathway	p-value	Up-regulated genes		Ingenuity Canonical Pathway	p-value	Down-regulated genes
	Interferon Signaling	2,63E-02	OAS1,IFI35,JAK2,STAT1,BCL2		Sphingolipid Metabolism	6,76E-03	LASS6,GLA,GALC,SGMS2,ASAH1,SACM1L,LASS2,VNN1,LPIN1,GBA, SMPD4,GLB1,PPAP2B,SPHK1,ARSB,FUT4,KDSR,DUSP16
	IL-4 Signaling	2,69E-02	IL2RG,RRAS2,IRS1,SOS1,NFATC2,NFATC1,JAK2,FCER2		Role of BRCA1 in DNA Damage Response	7,08E-03	E2F4,BARD1,RBBP8,PLK1,E2F3,CHEK1,RAD51,GADD45A,E2F1, RFC2,BRIP1,BRCA1,HLTF,E2F2,RFC3
	B Cell Activating Factor Signaling	2,95E-02	MAP3K14,NFATC2,IKBKE,NFATC1,TRAF5,TNFRSF17		Protein Ubiquitination Pathway	7,94E-03	USP24,USP14,USP12,UBE2H,PSMD7,CDC20,USP20,DNAJC3,CDC23, HSPA5,USP39,SMURF1,USP3,HSP90B1,USP42,USP47,NEDD4L,BRCA1, PSMC2,HLA-C,DNAJB12,USP15,MED20,USP36,USP38,HSPA9,USP19,PSMD6, PSMD5,HSPD1,PSMD3,USP1,UBE2D1,NEDD4,TRAF6,PSMD11,DNAJC5, USP4,PSMD2,DNAJB11,PSMD12,HSPA13,BAP1,PSMD1,DNAJB6,PSMC3, UBE2C,BIRC2
	NF-κB Signaling	6,46E-02	MAP2K6,MAP3K14,FLT1,BMPR2,PRKCZ,TNFRSF17,TLR10, RRAS2,BMPR1A,TLR6,TLR7,TRAF5,CARD11,PRKCB				

doi:10.1371/journal.pone.0048485.t002

Bioinformatic analysis: normalization, signal calculation, significant differential expression, and sample/gene profile clustering. The Robust Microarray Analysis (RMA) algorithm was used for background correction, intra- and inter-microarray normalization, and expression signal calculation [19]. The absolute expression signal for each gene was calculated for each microarray. For the expression signal calculation of the Human Exon arrays we used a new CDF package, called GeneMapper (from GATExplorer) [20], instead of the *Affymetrix* original probe-set definition. This mapping represents an improvement thanks to the reannotation of updated Ensembl gene loci and removal of cross-hybridization noise [20]. It also allows operations to be carried out from the outset using gene identifications (Ensembl IDs) instead of probe-sets (*Affymetrix* IDs). Mapping to genome version Ensembl v53 (assembly NCBI36) was done for these analyses.

Significance Analysis of Microarray (SAM) [21] was used to calculate significant differential expression and to identify the gene probe sets that characterize the samples of each compared state. In this method, permutations provide robust statistical inference of the most significant genes and, by using a false discovery rate (FDR) [22], adjust the raw p-values to take multiple testing into account. An FDR cut-off of <0.05 was used for all the differential expression calculations.

Finally, the Global Test [23] algorithm was used to test the resulting lists of candidate genes associated with 13q-H subgroup. The Global Test allows us to identify the genes that have the global expression pattern most significantly related to the clinical feature studied.

All the bioinformatic analyses were performed with the statistical program R, using the custom packages Bioconductor [24] and GATExplorer [20].

Principal component analysis. To explore and represent the differences among the different categories studied (13q-HCLLs, 13q-L CLLs, nCLLs and healthy controls), we applied Principal Component Analysis (PCA) to the expression data sets, using the normalized gene expression matrices of all samples of the validation cohort as the input. The expression matrices were filtered beforehand removing 25% of the least variable genes to avoid noise produced by non-expressed genes (i.e. the remaining 28 806 genes). For each of these genes, the median expression value across samples within each category was calculated. Next, the following formula was designed to calculate the expression values per gene and sample considering their variability within each category:

$$Y_{ij} = \frac{X_{ij(k)} - median(ik)}{sd(ik) + \beta} + median(ik) \qquad (1)$$

where Y_{ij} is the PCA input matrix, X_{ij} is the original expression matrix, i is the gene, j the sample, k the category and $\beta = 2$ is a small positive constant added to the denominator to ensure that the variance of Y_{ij} is independent of the genes [21]. This formula represents a way of calculating the dispersion of the biological replicates plus its median in each category. In this way, the clustering derived from the principal components includes a small amount of variation between individual samples, highlighting the differences between the categories.

Functional analysis and gene annotation. The functional assignment of the genes included in the expression signature of CLL cytogenetic subgroups was carried out by the Database for Annotation, Visualization and Integrated Discovery (DAVID) and the GeneCodis program [25], which identifies concurrent

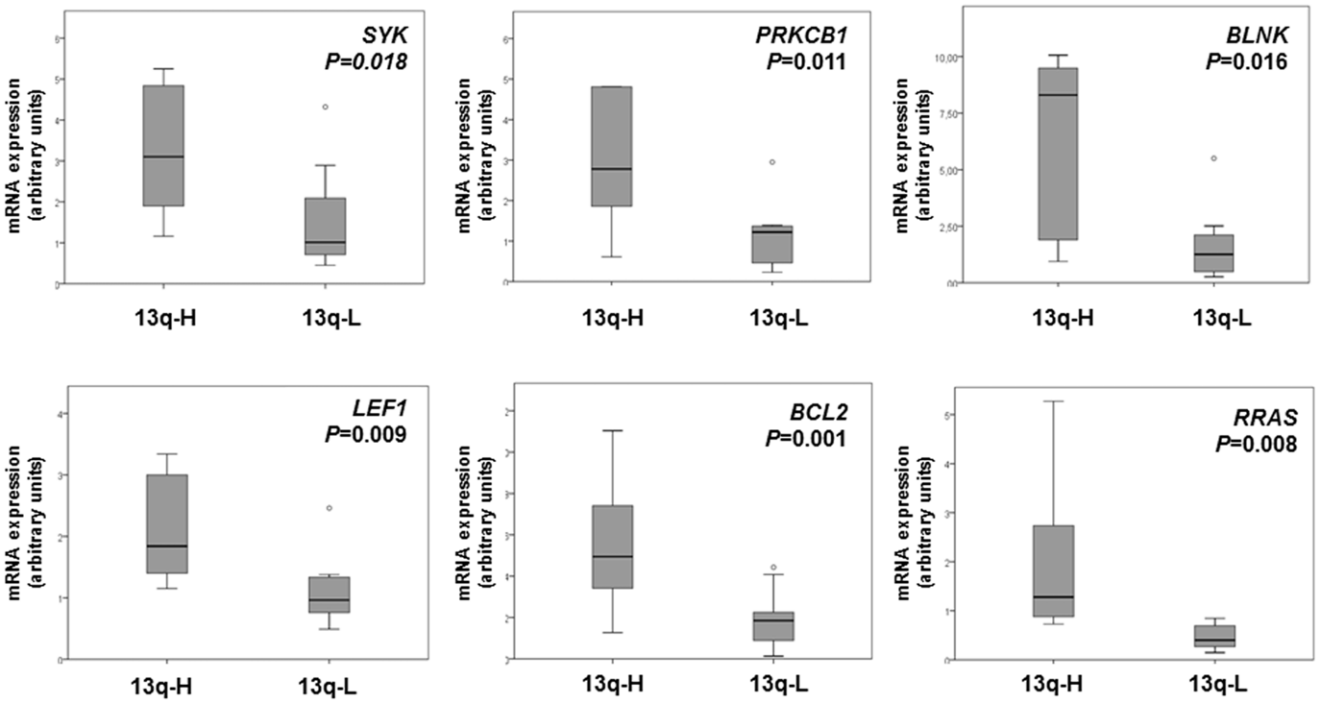


Figure 1. Gene expression levels of genes significantly upregulated in 13q-H CLL patients. Box plot of the expression levels [represented as arbitrary units (a.u.)] of six genes with significantly different expression between 13q-H and 13q-L patients, assessed by semi-quantitative PCR analysis. Box plots show the relative upregulation of BCR (SYK, PRKCB1 and BLNK), proliferation (LEF1 and RRAS2) and antiapoptotic (BCL2) related genes in patients with a high number of 13q- cells compared with CLL patients with lower percentages of losses in 13q. The thick line inside the box plot indicates the median expression levels and the box shows the 25th and 75th percentiles, while the whiskers show the maximum and minimum values. Outliers (extreme values falling out of the main distribution) are represented by open circles. Statistical significance was determined using the Mann-Whitney U test ($P<0.05$).
doi:10.1371/journal.pone.0048485.g001

annotations in GO and KEGG, and thereby constructs several groups of genes of functional significance. The most significant biological mechanisms, pathways and functional categories in the data sets of genes selected by statistical analysis were identified through the use of Ingenuity Pathways Analysis Sep2011 (Ingenuity Systems, Mountain View, CA, USA).

Gene-specific semi-quantitative PCR. Semi-quantitative SYBRgreen PCR was done in triplicate with iQ^{TM} SYBR® Green Supermix kit (BioRad) using the IQ5 Multicolor Real-Time PCR Detection System (Bio-Rad). Expression data for selected genes were validated in a subset of CLL patients (n = 40). Sense and antisense primers were designed based on the probe-sets used by Affymetrix to synthesize the GeneChip Primer sequences (Table S2). The *ABL1* gene was used as the internal control and the quantification of relative expression [reported as arbitrary units (a.u.)] were performed using the comparative Ct method. The data were not normally distributed, so non-parametric tests were used. Expression levels of the selected genes in both groups (13q-H and 13q-L) were analyzed using the Mann-Whitney U test with a two-tailed value of $P<0.05$ for statistical significance. All tests were performed using SPSS v19.0.

Quantification of miRNA expression levels. The expression of selected mature miRNAs was assayed using the Taqman MicroRNA Assays (Applied Biosystems) specific to hsa-mir-15a, hsa-mir-29a, hsa-mir-155 and hsa-mir-223 in 24 CLL patients displaying 13q- according to the manufacturer's recommendations. The Taqman MicroRNA Assays for U43 RNA (RNU43, Applied Biosystems) was used to normalize the relative abundance of miRNA using the $2^{-\Delta Ct}$ method. All experiments were performed in duplicate. Expression levels [reported as arbitrary units (a.u.)] of the selected miRNAs in both groups (13q-H and 13q-L) were analyzed using the Mann-Whitney U test in SPSS v19.0. Values of $P<0.05$ were considered statistically significant.

Integrative analysis of miRNA and gene expression profile. A summary of the miRNA analysis performed in the study is shown in the Figure S1. miRNAs with significantly different expression (FDR<0.05) between 13q-H and 13q-L were further analyzed to identify the networks and pathway targets. For this purpose, IPA's microRNA Target Filter, which enables prioritization of experimentally validated and predicted mRNA targets from TargetScan, TarBase, miRecords and the Ingenuity Knowledge Base was used. This tool identified the putative targets for the input miRNAs and then developed the networks among the targets and identified the known and most relevant biological functions, pathways and annotations in this enriched set of target genes. By applying the expression pairing tool, the analysis was focused on targets exhibiting altered expression in our analysis, finding miRNAs and their target genes with opposite or same expression.

Results

13q-H CLLs are Characterized by a Specific Genetic Signature and miRNA Expression

A total of 3 450 genes significantly distinguished 13q-H from 13q-L patients. These comprised 1 244 overexpressed genes and 2

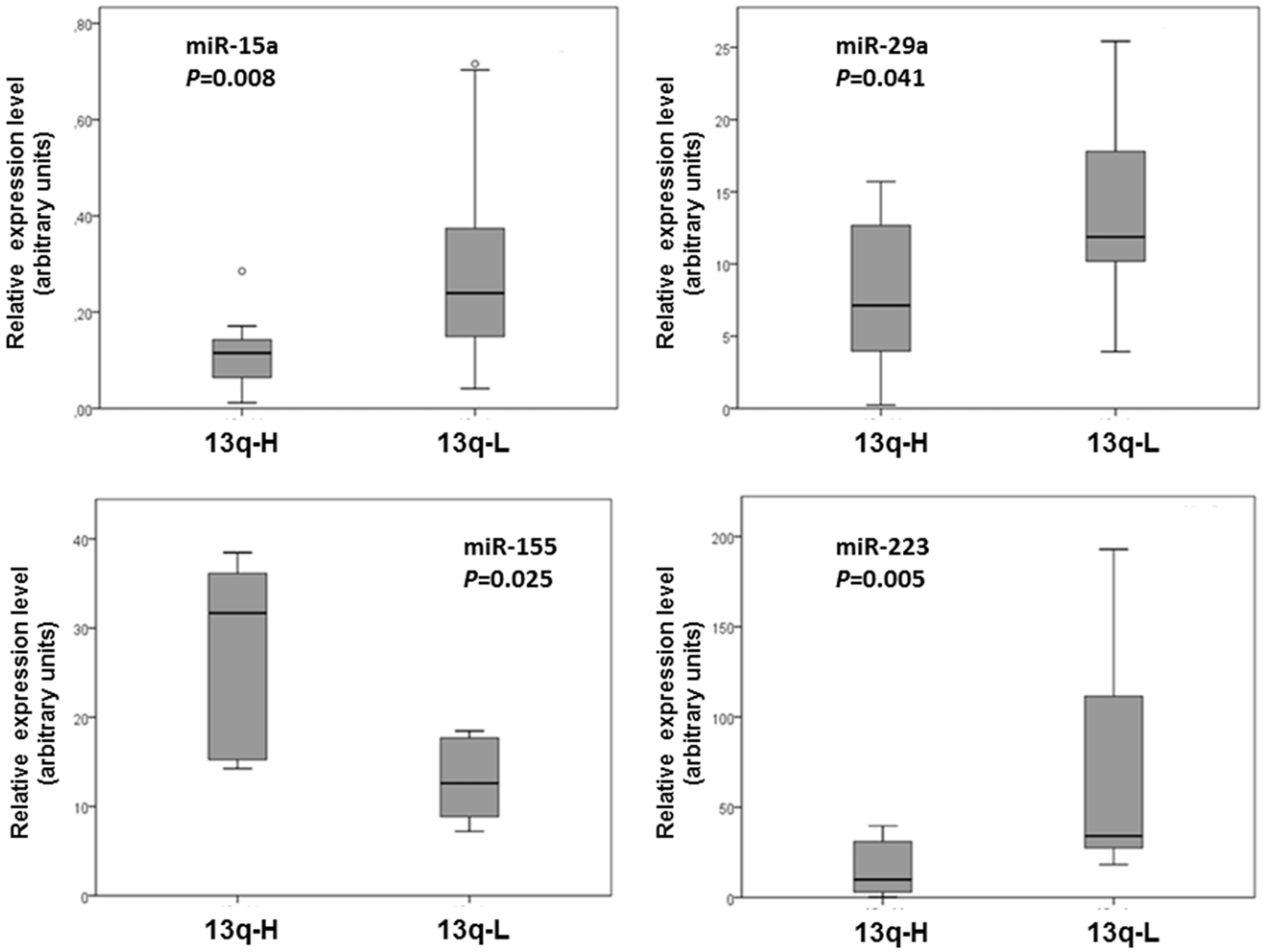


Figure 2. Quantitative RT-PCR validation for miR-15a, miR-29a, miR-155 and miR-223 in independent CLL patients. Relative expression of miR-15a, miR-29a, miR-155 and miR-223 [represented as arbitrary units (a.u.)] was evaluated by individual TaqMan miRNA assays performed in duplicate and normalized to RNU43 (2^{-dCt}). Box plots indicate the median value (horizontal line) and the 25^{th}–75^{th} percentile range (box) while whiskers showing the maximum and minimum values. Values outside this range are shown as outliers (open circles). *P*-values were determined by the Mann-Whitney U test. In every case, miRNAs downregulated in 13q-H CLL patients relative to 13q-L patients were also found to be downregulated by quantitative RT-PCR. Similar observations were made for miR-155, which was upregulated in 13q-H patients. All comparisons were statistically significant ($P<0.05$).
doi:10.1371/journal.pone.0048485.g002

206 underexpressed in the 13q-H group, defining the 13q-H signature. The deregulated genes of the 13q-H signature were annotated and analyzed for the presence of overrepresented "Gene Ontology categories" (Table S3). The most significant overrepresented GO biological processes in 13q-H were related to cell cycle ($P<0.0001$), ribosome ($P<0.0001$) and regulation of transcription ($P<0.0001$). Moreover, 13q-H CLLs had higher levels of expression of *LEF1*, *BCL2*, *CARD11*, *HDAC9*, *NAFTC1*, *NFATC2*, *PAX5*, *FCRL2* and *SOS1*, while we identified several other genes downregulated in 13q-H, such as *GAS7*, *E2F1*, *RRM1*, *KIT*, *NP* and *EPOR*. Many of these genes have been reported to be deregulated in CLL, as we confirmed in our analyses that showed overexpression of *LEF1*, *NFATC1*, *NFATC2* and *PAX5* in B lymphocytes from CLL patients compared with B lymphocytes from healthy controls (data not shown). PCR results confirmed the microarray data in the analyzed genes such as *GAS7*, *E2F1* and *FCRL2* (Figure S2).

Moreover, 13q-H CLL patients were also characterized by a striking overrepresentation of deregulated miRNAs. A total of 15 miRNAs were deregulated in 13q-H relative to 13q-L patients. Most of them (eleven) were downregulated while four were upregulated in 13q-H CLL (Table 1).

Signaling Pathways and Functional Ontology Analyses of Genes Differentially Expressed in 13q-H Patients

To determine the biological significance of the deregulated genes, a further analysis of the 3 450 deregulated genes characterizing the 13q-H CLL was carried out, revealing in this group of patients the involvement of several pathways (Table 2). These pathways are primarily related to cell proliferation, apoptosis and cell signaling. Thus the BCR pathway was upregulated in 13q-H CLL patients. In fact, 21 genes from this pathway were overexpressed in 13q-H CLLs, some of which, such as *SYK*, *BLNK* and *PRKCB1*, were previously related to CLL pathogenesis (Figure S3). We also observed an imbalance in proliferation and apoptosis in 13q-H patients, due to upregulation of antiapoptotic genes (*BCL2*) and decreased expression of proapoptotic genes (*RASSF5*, *BAD*, *CASP8*, *CASP10*, *FAS*) in 13q-

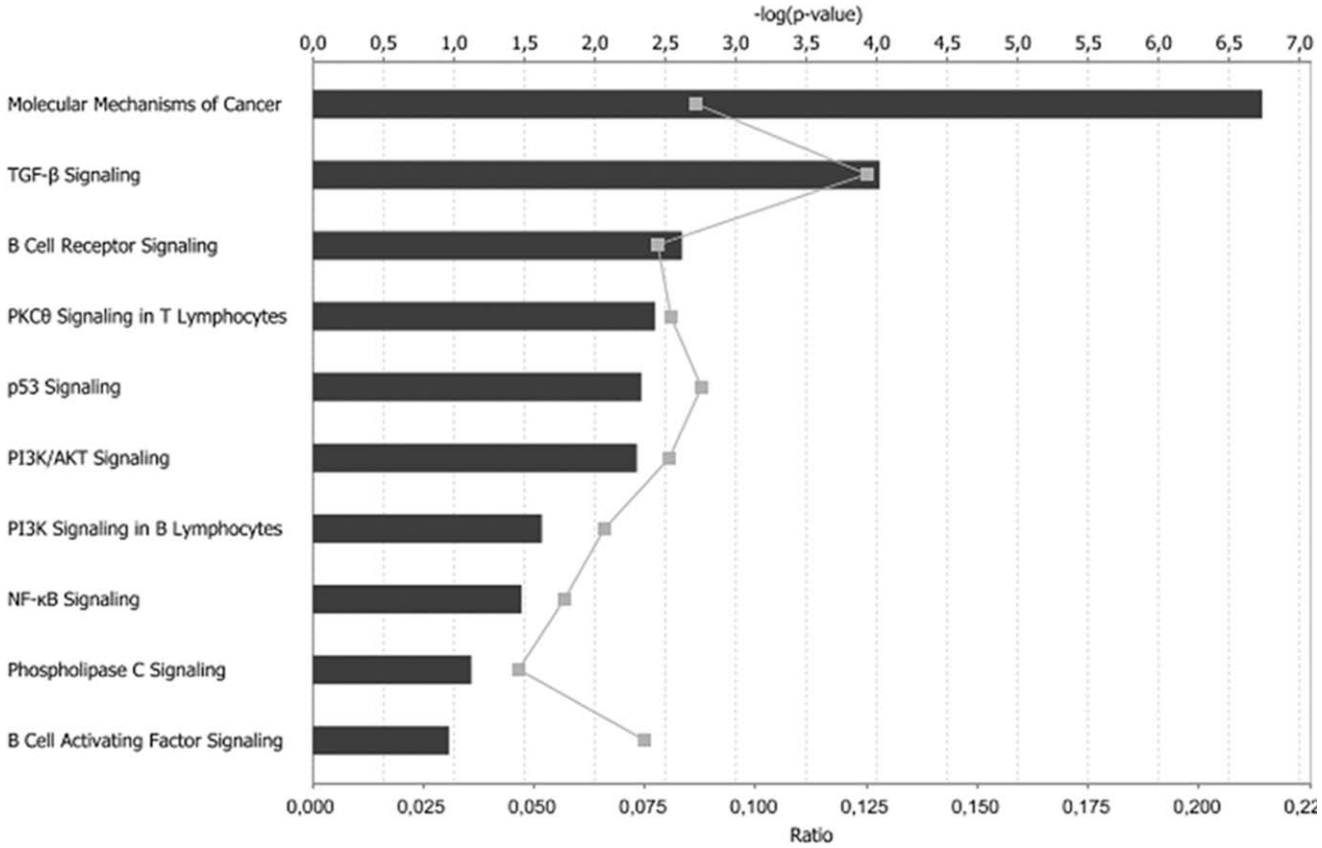


Figure 3. Most significant cellular functions affected by the deregulation of miRNAs in 13q-H CLL patients. 432 out of the 1027 predicted mRNA target genes of the deregulated miRNAs in 13q-H CLL patients appeared also deregulated in our analysis. A functional enrichement analysis was performed in this dataset. Category names are presented on the vertical axis. Of note, B cell receptor signaling and NF-kB signaling were among the most significant cellular functions affected. The significance of the association between the dataset and the canonical pathway was measured in two ways: (1) the ratio of the number of genes from the dataset that met the expression value cut-off that map onto the pathway divided by the total number of molecules that exist in the canonical pathway, represented by grey squares in the graph and (2) the *P*-value determining the probability of the association between the genes in the dataset and the canonical pathway, calculated by Fisher's exact test. The horizontal axis on the top indicates the −log (*P* value) and the horizontal axis at the bottom, the ratio. In both cases, the higher value indicates the higher significance.
doi:10.1371/journal.pone.0048485.g003

H patients. Moreover, our analysis showed an overexpression of genes promoting proliferation, such as *LEF1*, *E2F5* and *RRAS2*. To ensure that the gene expression profiles accurately reflected the upregulation of BCR signaling pathway and the deregulation of apoptosis-related genes, representative genes that were differentially expressed in 13q-H patients were assessed by semi-quantitative SYBRgreen PCR analysis. These included *SYK*, *BLNK* and *PRKCB1* (BCR signaling pathway), *BCL2* (apoptosis) and *LEF1* and *RRAS2* (proliferation). The semi-quantitative PCR results were in close agreement with the microarray data (Figure 1) confirming the overexpression of these genes in 13q-H CLLs compared with 13q-L. Western blot analysis should be made for a more concluding validation after mRNA screening. Unfortunately, due to the lack of material, this was not possible in this study.

miRNA Deregulation in 13q-H CLL Patients

The analysis of miRNA expression in 13q-H and 13q-L CLL patients revealed that fifteen miRNAs were deregulated in 13q-H CLL patients: hsa-miR-155 was the most highly upregulated miRNA (Rfold = 3.70), while hsa-miR-223 was the most significantly downregulated (Rfold = 0.10). Four of the deregulated miRNAs (miR-15a, miR-29a, miR-155 and miR- 223) were further assayed by quantitative RT-PCR for validation purposes in 24 CLL samples displaying 13q-. Results confirmed the upregulation of miR-155 and the downregulation of miR-15a, miR-29a and miR- 223 in 13q-H samples relative to 13q-L (Figure 2).

The influence of these deregulated miRNAs on 13q- patients was assessed (Figure S2). Specifically, we investigated whether observed changes in miRNAs were correlated with changes in the expression of genes. Therefore the post-transcriptional regulatory network of miRNA and genes in CLL patients with more than 80% of 13q- cells was carried out by analyzing the miRNA-mRNA relationships. A total of 1 027 mRNA putative targets with altered expression in 13q-H CLL patients were found (Table S4). Indeed, because miRNAs tend to downregulate the target genes, we focused our study on the subset of 11 miRNAs selected for analysis in IPA and the 432 genes predicted to be regulated by them and characterized by expression profiles stringly anticorrelated. Functional analysis revealed that transcription was the cell function most strongly affected by these miRNAs, with a total of 97 genes affected by the 11 selected miRNAs. Modification of proteins (n = 41), proliferation of immune cells (n = 34), and activation of protein binding sites (n = 32) were other important functions affected by these miRNAs (Table S5). Finally we performed a functional analysis of the 11 miRNAs and their 432 putative targets. The pathway analysis demonstrated that, again, B

Table 3. Most significant target genes affected by deregulation in miRNA in 13q-H CLL patients.

Target			miRNA	
Symbol	**Fold Change**	**B-cells related pathways**	**ID**	**Fold Change**
			hsa-mir-206	0.708
			hsa-mir-15a	0.492
BCL2	2.132	Apoptosis	hsa-mir-106b	0.519
			hsa-mir-204	0.569
			hsa-mir-34a	0.687
			hsa-mir-206	0.708
E2F5	2.624	DNA Damage Response	hsa-mir-106b	0.519
			hsa-mir-34a	0.687
FOS	0.447	B Cell Activating Factor,CD27	hsa-mir-155	3.701
LEF1	2.835	ILK, Wnt	hsa-mir-34a	0.687
MAP2K6	3.558	BCR,CD27	hsa-mir-29a	0.510
MAP3K12	1.254	BCR,CD27	hsa-mir-106b	0.519
MAP3K14	1.348	Apoptosis,B Cell Activating Factor,BCR,CD27	hsa-mir-106b	0.519
MAP3K9	1.400	BCR,CD27	hsa-mir-106b	0.519
MYD88	0.752	NF-κB,Toll-like Receptor	hsa-mir-155	3.701
PLCB1	0.773	PI3K	hsa-mir-205	1.383
RRAS2	1.931	Apoptosis, BCR	hsa-mir-223	0.102
			hsa-mir-15a	0.492
SOS1	2.352	BCR	hsa-mir-106b	0.519
			hsa-mir-204	0.569
TCL1A	7.848	Akt	hsa-mir-29a	0.510

doi:10.1371/journal.pone.0048485.t003

cell receptor signaling, PI3K signaling and NFkB signaling were among the most strongly affected pathways in 13q-H patients (Figure 3), highlighting the importance of miRNA regulation in CLL. MiR-155, the most overexpressed miRNA in 13q-H, was negatively correlated with the expression of 90 of the 182 expected genes (49%), demonstrating a relationship between miRNA and gene deregulation. Interestingly, most of these putative targets were assigned to the functional categories of transcription regulation ($P = 0.002$). Moreover, we found several miRNAs whose targets that were experimentally observed or predicted with high confidence were strongly related to CLLs such as *BCL2* (miR-15, miR-206, miR-106b and miR-34a), *TCL1A* (miR-29a) and *LEF1* (miR-34a) (Table 3).

The GEP of 13q-H CLL Patients is Similar to that in CLL Patients with 11q or 17p Losses

We also analyzed the gene signature of CLL high risk cytogenetic subgroups in comparison with 13q- patients. Surprisingly, a significant number of deregulated genes were found to be shared between the genes that differentiate 13q subgroups and 13q-L and high risk subgroup of patients. That is, the GEP of 13q-H CLL patients resembled the gene expression pattern of patients with 17p- or 11q- abnormalities (Figure 4A). In fact, both subgroups of CLL patients (13q-H and the 17p- and 11q- subgroup) shared 1 325 genes (46%) of the deregulated genes in the global analysis including all CLL subtypes. By contrast, the comparison between the GEP of 13q-H patients and those with losses in either 17p or 11q showed fewer differences in expression (Figure S4).

To evaluate the biological significance of the observed similarity between the 13q-H and the 17p−/11q- signatures, we used the Ingenuity Pathway Analysis comparative tool, which facilitates the functional comparison of several panels of differentially expressed genes. Thus, we identified several commonly deregulated biological functions in both gene signatures (Figure 4B), such as cell cycle, cell death, cellular growth and proliferation. Finally, pathway analysis was performed on those genes commonly upregulated or downregulated in 13q-H, 17p- and 11q- patients in comparison with the 13q-L subgroup (Table S6). In accordance with the comparative analysis results, several commonly deregulated pathways of relevance in CLL pathogenesis were observed. The most significant of these were the B cell receptor signaling pathway for commonly upregulated genes, and the cell cycle control of chromosomal replication pathway for commonly downregulated genes in patients showing 13q-H, 17p- or 11q- (Table S6). The expression of the *TCL1* gene had one of the lowest q-values (0.002) with higher expression levels in patients with 13q-H, 17p- and 11q-. Of note, 13q-H, 17p- and 11q- patients also shared the deregulation of several miRNAs (Table 1).

Genome-wide Expression Differentiates 13q-H CLLs from 13q-L CLLs and Controls

To validate the differences observed between the subgroups of 13q- CLL patients and get a visualization of these, we applied the Principal Component Analysis (PCA) in an independent series of patients. The clustering algorithm of PCA reduces complex multidimensional data to a few specified dimensions so that it can be visualized effectively. For a better characterization of the

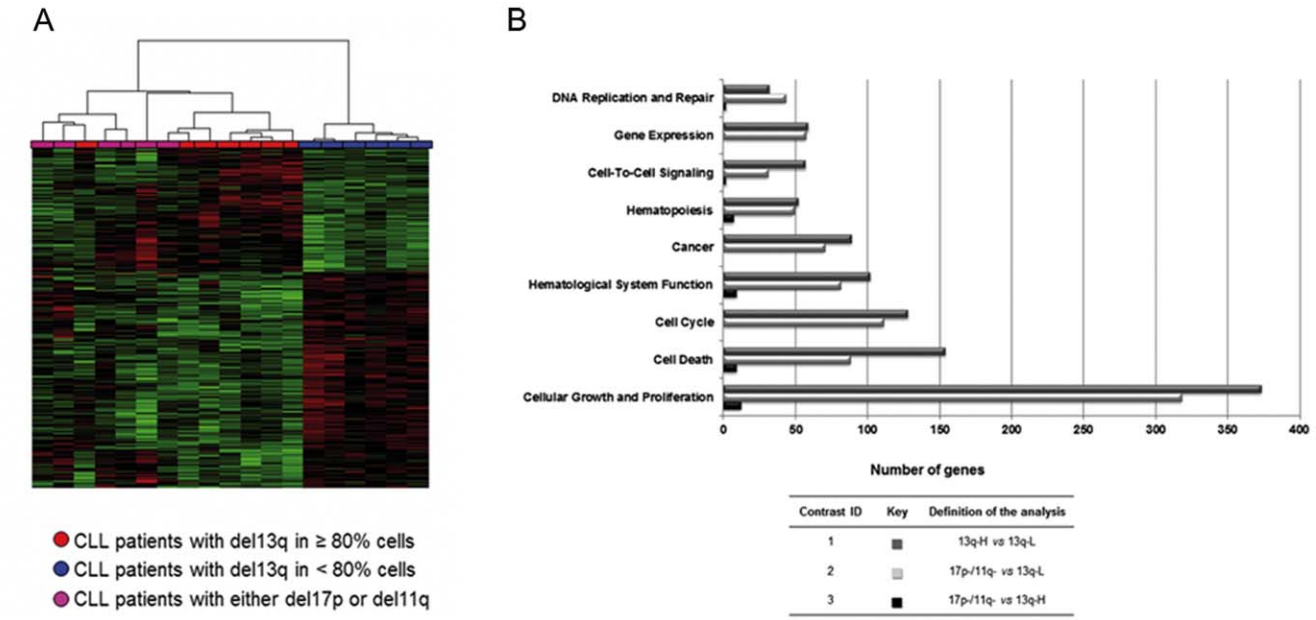


Figure 4. Differential expression analysis followed by pathway analysis revealed commonly deregulated biological processes in CLL patients with a high load of 13q- cells, 17p- and 11q-. A. Heatmap of 3450 differentially expressed genes in CLL patients with a high number of losses in 13q (red), losses in 17p or 11q (magenta) and a low number of losses in 13q (blue). Differentially regulated genes were identified using Significance Analysis of Microarray (SAM), with a false discovery rate 5%, followed by the Global Test algorithm to test the candidate genes associated with the group of patients with a high number of losses. Individual patients are arranged in columns with the expression level for each gene across rows. Normalized gene expression values are color-coded (standard deviation from mean): red and green indicate high and low expression, respectively. All patients with 13q-L were clustered on the right side of the map in a homogeneous manner and separately from 13q-H and 17p−/11q-, which clustered together, showing that the gene expression profile (GEP) of CLL cases with higher percentages of 13q- cells is similar to that of 17p- and 11q-, while CLL patients with lower percentages of 13q- cells had a different gene profile. **B. Commonly deregulated biological functions in 13q-H and 17p−/11q- CLL patients compared with 13q-L CLL subgroup.** Biological function names are presented on the vertical axis and the number of deregulated genes involved in each function, in the horizontal one. Fisher's exact test was used to examine the probability of the association between the genes in the dataset and the functional category. The color-coded bar plot (dark grey, light grey and black bars) depicts the analysis results. 13q-H patients showed marked differences in the expression of genes related to several cellular functions compared with 13q-L CLL patients (comparison 1, dark grey bars). In addition, most of these cellular functions were also deregulated in comparison with high-risk cytogenetic subgroups (17p- and 11q-) and 13q-L CLL patients (comparison 2, light grey bars). Thus, 13q-H, 17p- and 11q- patients share the deregulation of several important functions relative to 13q-L patients. Furthermore, a small number of genes related to cell cycle, cell growth and DNA repair (comparison 3, black bars) were found to be differentially expressed in the 13q-H group in a comparison of this subgroup of patients and high-risk cytogenetic subgroups.
doi:10.1371/journal.pone.0048485.g004

differences, we included in this cohort patients with normal FISH (nCLL) and healthy donors as two different types of controls.

Overall, the expression pattern of B lymphocytes from 13q-H and 13q-L CLL patients and nCLLs was notably different from the gene expression profile of B lymphocytes from healthy donors, as expected (Figure 5). PCA revealed a cumulative variance between groups of 48.3%, 60.9% and 68.3% corresponding to one, two and three of the initial components, respectively. Since the first three principal components explained a considerable proportion of the overall variance (68%), the 3D representation was able to show the main similarities and differences between categories. Notably, the 13q-H samples were largely separated from the others. Thus, 13q-H patients had a distinctive GEP that was different not only from healthy donors but also from all other CLLs, including 13q-L patients. By contrast, the gene expression of B lymphocytes from 13q-L CLL and nCLL was similar (Figure 5). SAM analysis revealed differences in the expression of 15 332 and 16 754 genes between CD19+cells from 13q-L or nCLL compared with B lymphocytes from healthy donors, respectively, while both subgroups (13q-L and nCLL patients) shared the deregulation of 13 749 genes (data not shown). Moreover, the analysis failed to demonstrate differences between nCLL and 13q-L patients, while 131 genes were differentially expressed in comparison with 13q-H (data not shown).

Thus, both qualitatively (PCA) and quantitative (SAM) analysis showed that the gene expression profile of 13q- CLLs is different depending on the percentage of cells displaying this aberration.

IgVH Mutational Status and Mono/biallelic 13q14 Deletion in 13q-patients

Given that the prognostic significance of IgVH mutations is independent from that of cytogenetic abnormalities, we also analyzed the IgVH mutational status in the 13q- subgroups. There was no significant difference between both 13q- subgroups ($P = 0.664$).

Regarding the distribution of biallelic 13q14 deletion in both 13q- subgroups, no correlation between the presence of biallelic 13q14 deletion and the percentage of 13q- cells was observed. Thus in the group of patients with 3 of the 32 cases (9%) had a biallelic loss of 13q, while in the group of 13q-L 5 of 38 patients (13%) showed a biallelic loss on 13q. ($P = $N.S.).

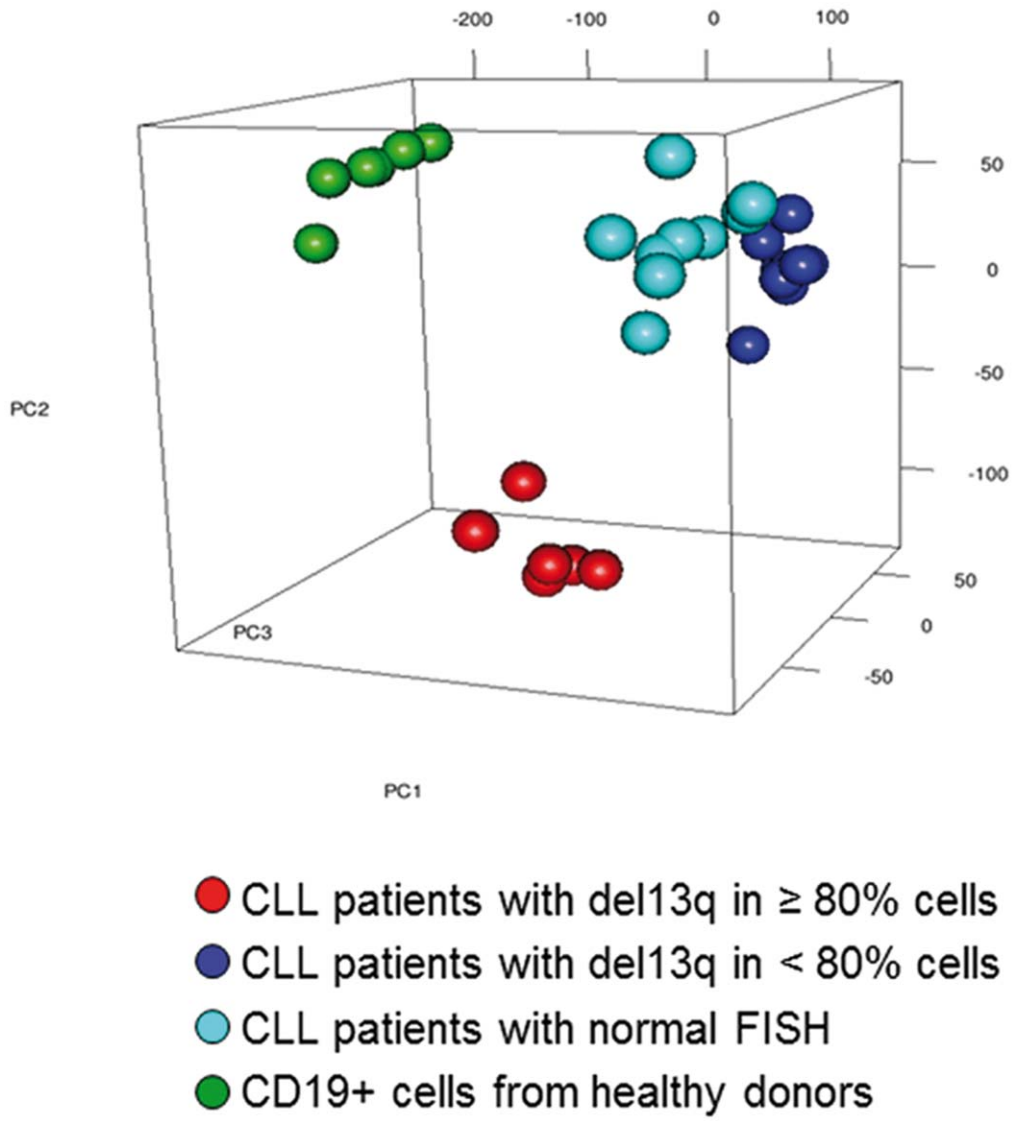


Figure 5. CLL patients with a high number of 13q- cells can be differenciated based on their expression profile. Principal component analysis (PCA) plot of CD19+cells from healthy controls (green), CLL with normal FISH (sky blue), 13q-H CLL (red) and 13q-L CLL (dark blue) was carried out using the 28,806 remaining genes after filtering the normalized gene expression matrices to remove the least variable genes (25%). Each sphere represents a single GEP. The result of the PCA shows a cumulative variance of 48.3%, 60.9% and 68.3% corresponding to one, two or three of the initial components, respectively. The expression pattern of CD19+cells from CLL patients is notably different from the gene expression profile of CD19+cells from healthy donors. Of note, the PCA analysis shows that 13q-H CLL patients have a distinctive gene expression profile. By contrast, the gene expression of B lymphocytes from 13q-L CLL and nCLL was similar.
doi:10.1371/journal.pone.0048485.g005

Discussion

13q deletion (13q-) is the most common cytogenetic aberration in CLL and it is usually associated with the most favourable prognosis as a sole abnormality [5]. However, recent studies have shown that CLL patients carrying higher percentages of 13q- cells have more aggressive clinical courses [11–13]. By combining gene expression profile and miRNA analysis, we have shown that 13q- patients are also a biologically heterogeneous group, in which a higher number of 13q- cells (13q-H) could involve the deregulation of relevant cellular pathways. Thus, several pathways are involved in 13q-H patients (Table 2 and Table S3), BCR signaling, NFkB signaling and antiapoptotic pathways being of special interest in CLL. Deregulation of several miRNAs (Table 1) was also observed. The influence of other factors with prognostic relevance in CLL, such as IGVH mutational status, was discarded.

The BCR is an essential signal transduction pathway for the survival and proliferation of mature B lymphocytes. In the present study, monoclonal B-cells in 13q-H CLL patients exhibit a molecular signature characterized by the overexpression of genes mainly involved in BCR signaling (Figure 1). There is now strong evidence that signaling via the B cell receptor plays a major role in the development of CLL, and it could be related to the different clinical outcomes of CLL [26]. Thus, the BCR pathway is activated in poor prognosis CLL patients (IGHV unmutated), and the overexpression of several molecules involved in this pathway has been reported in advanced stages of the disease [27,28]. In addition, SYK expression is enhanced in CLL relative to healthy B cells and also in unmutated compared with mutated CLL, possibly

reflecting the increased BCR signaling in these patients [29]. In our study 13q-H CLL also overexpressed *SYK* (Figure 1), providing new evidence of the involvement of the BCR pathway in this group of CLLs. In addition, this group of patients also showed upregulation of *CD79b*. Chronic active BCR signaling due to point mutations in *CD79b* has recently been identified as a key pathogenic mechanism in aggressive B-cell lymphoma, and results in constitutive nuclear factor-kB (NF-kB) activation [30]. Interestingly, CLL patients with deletions on 17p or 11q or those with losses in 13q in a high percentage of cells had an increased expression of a cluster of genes comprising several PKCs, such as *PRKCB1* and *PRKCZ*. Previous studies have shown an overexpression of *PKC* in human CLLs, which is part of a poor-prognosis gene cluster in CLL linked to the transmission of BCR signals such as calcineurin-NFAT and NF-kB, which our analysis also revealed to be deregulated (Table 2) [31,32]. Furthermore, the overexpression of calcium metabolism-related genes as well as several MAPK in 13q-H patients was also observed in the present study, which would be consistent with these previous studies (Table 2).

One of the hallmarks of this clinically heterogeneous disease is defective apoptosis, which is considered to contribute not only to cell accumulation but also to disease progression and resistance to therapy [26]. In this study we report the overexpression of genes involved in promoting cell survival and antiapoptotic pathways, as well as the downregulation of several proapoptotic genes in 13q-H CLL patients (Table 2 and Table S3). We confirm the overexpression of *LEF-1* in CLL B cells compared with B cells from healthy donors (data not shown), as previously reported [33], but we also observed upregulation of *LEF-1* and other genes involved in the Wnt signaling pathway in 17p-, 11q- and 13q-H patients in comparison with 13q-L cases. Wnt pathway gene expression is widely known to be deregulated in CLL [34,35]. Alterations of RAS signaling are associated with potent oncogenic effects, which keep the cell in a proliferative state and block apoptosis, thereby paving the way for cancer formation. Overexpression of *RRAS* and other molecules involved in this signaling cascade, such as SOS1, RHOC and several MAP kinases, was also observed. In addition, apoptosis was also deregulated in 13q-H patients by the involvement of both mitochondrial (*BCL2* and several caspases) and extrinsic (*FAS*) pathways. Interestingly, the apoptotic signature of 13q-H patients showed a similar pattern of deregulation to that of high-risk cytogenetic groups (Figure 4B), since they both featured the alteration of several genes involved in the classic apoptotic pathway (mitochondrial). Sustained BCR signaling has also been reported to have an antiapoptotic effect [36]. Thus, in 13q-H CLL patients, our study shows an imbalance between the proliferative and apoptotic signals, which could explain the higher level of lymphocytosis and the poor outcome previously described in these patients [11].

An aberrant cellular miRNA expression profile in CLL cells has already been described and the changes correlate well with prognostic factors, including ZAP-70 expression status and *IgVH* mutations in CLL patients [37]. A recent study evaluating microRNAs as a signature for CLL patients with specific chromosomal abnormalities found nine miRNAs whose expression values were correlated with a specific karyotype [38]. In our study we found that several miRNAs were deregulated in 13q-H patients (Table 1), some of which had been previously reported in CLL (Table 3). Several important miRNAs, such as miR223, miR-29a and miR-181, were downregulated in 13q-H and high-risk cytogenetic subgroups, which could be related to the worse outcome in these groups of patients [39,40]. By contrast, overexpression of miR-155 was observed, which could be related to enhanced BCR-activation, as previously reported [41]. The pathogenic role of deletion 13q in CLL has been related to the lack of B-cell proliferation control allegedly determined by the deletion of the *DLEU2/MIR15A/MIR16-1* locus [42]. Interestingly, miR-15a was downregulated in 13q-H CLL patients and it has been reported to induce apoptosis through the negative regulation of *BCL2,* overexpressed in the 13q-H group of patients. It should be noted that a third of deregulated genes in 13q-H compared with 13q-L were putative targets of miRNAs also altered in this analysis, supporting the presence of a specific relationship between miRNA and gene expression in 13q-H CLL patients. Most of these genes were related to TGF or BCR signaling and confirmed these pathways to be those most commonly affected by miRNA deregulation in 13q-H patients. Among the putative target mRNAs we found many genes, such as *TCL1A, BCL2, LEF1* [33,43,44], to be closely involved in CLLs (Table 3). These results suggest that miRNAs have a key role in the reported heterogeneity of 13q- patients. Surprisingly, our results suggest that some of the biological characteristics of 13q-H CLL patients are similar to those of high-risk cytogenetic subgroups, since they share the deregulation of several key signaling pathways (Figure 4B; Table S6). However, 13q-L patients had similar gene expression to that of CLL with normal FISH (Figure 5).

Therefore, this study provides new evidence regarding the heterogeneity of 13q deletion in CLL patients, showing that apoptosis, BCR and NF-kB signaling as well as miRNA regulation are the most significant affected pathways in 13q-H CLL patients. The identification of the mechanisms responsible for the clinical heterogeneity of CLL, including the mutations recently described [45,46] and the critical signaling pathways affected can lead to a better understanding of the molecular pathogenesis of the disease.

Supporting Information

Figure S1 Summary of the miRNA analysis performed in the study. The chart explains the steps involved in the identification and validation of the miRNAs and their deregulated targets in 13q- CLL patients.
(TIF)

Figure S2 Box plot of the expression levels of three genes with significant differences between 13q-H and 13q-L patients, assessed by semi-quantitative PCR. Box plots show the values for *GAS7*, *E2F1* and *FCRL2* relative expression [represented as arbitrary units (a.u.)], showing a significant difference in the level of expression between 13q-H and 13q-L CLL patients. The thick line inside the box plot indicates median expression levels, the limits of the box represent the 25th and 75th percentiles, and the whiskers show the maximum and minimum values. Outliers (extreme values falling outside the main distribution) are represented by open circles. Statistical significance was determined using the Mann-Whitney U test ($P<0.05$).
(TIF)

Figure S3 BCR signaling pathway identified as the top canonical pathway altered in CLL patients with higher percentages of 13q- losses according to the Ingenuity Pathway Analysis knowledge base. Genes significantly differentially expressed between CLL with 80% or more of cells with loss of 13q (13q-H) and CLL with losses in 13q in fewer than 80% of cells (13q-L) were mapped to the pathway and colored in red if the expression levels were higher, or in green if they were lower in 13q-H than in 13q-L cases. Significant positions of the pathway are occupied by genes deregulated in our analysis, indicating that this pathway is affected in 13q-H patients. CLL

patients with 17p and 11q deletions showed similar deregulation in this pathway.
(TIF)

Figure S4 Overlap of differentially expressed genes as analyzed by SAM. Venn diagram illustrating the number of significantly affected genes in common and distinct for the contrasts (1) and (2). 13q-H and 17p−/11q- shared the deregulation of 46% of genes (n = 1325) relative to 13q-L.
(TIF)

Table S1 Clinical and biological features of CLL patients included in the study.
(XLS)

Table S2 Sequences of primers used for SYBR Green detection.
(XLS)

Table S3 Enriched functional annotations terms associated with the 3450 differentially expressed genes in 13q-H CLL patients. Genes were clustered into functional categories using the DAVID Bioinformatics Database Gene Functional Classification Tool (NIAID/NIH). The P-value is provided by DAVID bioinformatics resources.
(XLS)

Table S4 miRNAs and their predicted targets (n = 1027) that are significantly deregulated in 13q-H CLL patients. By applying an integrated miRNA-mRNA analysis, mRNA targets were identified for the list of miRNAs deregulated in 13q-H CLL patients. The P-value for each predicted target gene refers to the contrast between 13q-H and 13q-L CLL patients.
(XLS)

Table S5 Functional analysis of the potential target genes of the deregulated miRNAs in CLL patients with a high number of 13q- cells (13q-H). The 432 mRNA target genes that showed an inverse relationship with miRNA expression level were input into Ingenuity (Ingenuity Systems, Inc.) and core analysis was then performed to retrieve the target genes' association with biological functions of relevance in CLL.
(XLS)

Table S6 Most significant differentially expressed genes in patients with 80% or more cells showing 13q- (13q-H) and 17p/11q deletion compared with 13q-L patients. (Upper: overexpressed; Lower: underexpressed in 13q-H, 17p- and 11q- patients with respect to the 13q-L CLL patients).
(XLS)

Acknowledgments

We thank Irene Rodríguez, Sara González, Teresa Prieto, M[a] Ángeles Ramos, Almudena Martín, Ana Díaz, Ana Simón, María del Pozo and Vanesa Gutiérrez of the Centro de Investigación del Cáncer, Salamanca, Spain, for their technical assistance and Jesús F. San Miguel for his critical review of the manuscript.

Author Contributions

Conceived and designed the experiments: AER JAE JMHR. Performed the experiments: AER MHS EF MES MG. Analyzed the data: RB NCG AR JdR. Contributed reagents/materials/analysis tools: JLG RF AGC GMN IR NDLH OG. Wrote the paper: AER JAH JMHR.

References

1. Chiorazzi N, Rai KR, Ferrarini M (2005) Chronic lymphocytic leukemia. N Engl J Med 352: 804–815.
2. Keating MJ (1999) Chronic lymphocytic leukemia. Semin Oncol 26: 107–114.
3. Dighiero G (2003) Unsolved issues in CLL biology and management. Leukemia 17: 2385–2391.
4. Hamblin TJ, Davis Z, Gardiner A, Oscier DG, Stevenson FK (1999) Unmutated Ig V(H) genes are associated with a more aggressive form of chronic lymphocytic leukemia. Blood 94: 1848–1854.
5. Dohner H, Stilgenbauer S, Benner A, Leupolt E, Krober A, et al. (2000) Genomic aberrations and survival in chronic lymphocytic leukemia. N Engl J Med 343: 1910–1916.
6. Damle RN, Wasil T, Fais F, Ghiotto F, Valetto A, et al. (1999) Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. Blood 94: 1840–1847.
7. Krober A, Seiler T, Benner A, Bullinger L, Bruckle E, et al. (2002) V(H) mutation status, CD38 expression level, genomic aberrations, and survival in chronic lymphocytic leukemia. Blood 100: 1410–1416.
8. Lin K, Sherrington PD, Dennis M, Matrai Z, Cawley JC, et al. (2002) Relationship between p53 dysfunction, CD38 expression, and IgV(H) mutation in chronic lymphocytic leukemia. Blood 100: 1404–1409.
9. Oscier DG, Gardiner AC, Mould SJ, Glide S, Davis ZA, et al. (2002) Multivariate analysis of prognostic factors in CLL: clinical stage, IGVH gene mutational status, and loss or mutation of the p53 gene are independent prognostic factors. Blood 100: 1177–1184.
10. Mehes G (2005) Chromosome abnormalities with prognostic impact in B-cell chronic lymphocytic leukemia. Pathol Oncol Res 11: 205–210.
11. Hernandez JA, Rodriguez AE, Gonzalez M, Benito R, Fontanillo C, et al. (2009) A high number of losses in 13q14 chromosome band is associated with a worse outcome and biological differences in patients with B-cell chronic lymphoid leukemia. Haematologica 94: 364–371.
12. Van Dyke DL, Shanafelt TD, Call TG, Zent CS, Smoley SA, et al. (2010) A comprehensive evaluation of the prognostic significance of 13q deletions in patients with B-chronic lymphocytic leukaemia. Br J Haematol 148: 544–550.
13. Dal BM, Rossi FM, Rossi D, Deambrogi C, Bertoni F, et al. (2011) 13q14 Deletion size and number of deleted cells both influence prognosis in chronic lymphocytic leukemia. Genes Chromosomes Cancer 50: 633–643.
14. Catovsky D, Richards S, Matutes E, Oscier D, Dyer MJ, et al. (2007) Assessment of fludarabine plus cyclophosphamide for patients with chronic lymphocytic leukaemia (the LRF CLL4 Trial): a randomised controlled trial. Lancet 370: 230–239.
15. Harris NL, Jaffe ES, Diebold J, Flandrin G, Muller-Hermelink HK, et al. (1999) World Health Organization classification of neoplastic diseases of the hematopoietic and lymphoid tissues: report of the Clinical Advisory Committee meeting-Airlie House, Virginia, November 1997. J Clin Oncol 17: 3835–3849.
16. Binet JL, Caligaris-Cappio F, Catovsky D, Cheson B, Davis T, et al. (2006) Perspectives on the use of new diagnostic tools in the treatment of chronic lymphocytic leukemia. Blood 107: 859–861.
17. Sanchez ML, Almeida J, Gonzalez D, Gonzalez M, Garcia-Marcos MA, et al. (2003) Incidence and clinicobiologic characteristics of leukemic B-cell chronic lymphoproliferative disorders with more than one B-cell clone. Blood 102: 2994–3002.
18. Ghia P, Stamatopoulos K, Belessi C, Moreno C, Stilgenbauer S, et al. (2007) ERIC recommendations on IGHV gene mutational status analysis in chronic lymphocytic leukemia. Leukemia 21: 1–3.
19. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185–193.
20. Risueno A, Fontanillo C, Dinger ME, de las RJ (2010) GATExplorer: genomic and transcriptomic explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs. BMC Bioinformatics 11: 221-.
21. Tusher VG, Tibshirani R, Chu G (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A 98: 5116–5121.
22. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I (2001) Controlling the false discovery rate in behavior genetics research. Behav Brain Res 125: 279–284.
23. Goeman JJ, van de Geer SA, de KF, van Houwelingen HC (2004) A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 20: 93–99.
24. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80-.
25. Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. Genome Biol 8: R3-.
26. Chiorazzi N (2007) Cell proliferation and death: forgotten features of chronic lymphocytic leukemia B cells. Best Pract Res Clin Haematol 20: 399–413.
27. Guarini A, Chiaretti S, Tavolaro S, Maggio R, Peragine N, et al. (2008) BCR ligation induced by IgM stimulation results in gene expression and functional changes only in IgV H unmutated chronic lymphocytic leukemia (CLL) cells. Blood 112: 782–792.

28. Rodriguez A, Villuendas R, Yanez L, Gomez ME, Diaz R, et al. (2007) Molecular heterogeneity in chronic lymphocytic leukemia is dependent on BCR signaling: clinical correlation. Leukemia 21: 1984–1991.
29. Buchner M, Fuchs S, Prinz G, Pfeifer D, Bartholome K, et al. (2009) Spleen tyrosine kinase is overexpressed and represents a potential therapeutic target in chronic lymphocytic leukemia. Cancer Res 69: 5424–5432.
30. Davis RE, Ngo VN, Lenz G, Tolar P, Young RM, et al. (2010) Chronic active B-cell-receptor signalling in diffuse large B-cell lymphoma. Nature 463: 88–92.
31. Su TT, Guo B, Kawakami Y, Sommer K, Chae K, et al. (2002) PKC-beta controls I kappa B kinase lipid raft recruitment and activation in response to BCR signaling. Nat Immunol 3: 780–786.
32. Bernal A, Pastore RD, Asgary Z, Keller SA, Cesarman E, et al. (2001) Survival of leukemic B cells promoted by engagement of the antigen receptor. Blood 98: 3050–3057.
33. Gutierrez A Jr, Tschumper RC, Wu X, Shanafelt TD, Eckel-Passow J, et al. (2010) LEF-1 is a prosurvival factor in chronic lymphocytic leukemia and is expressed in the preleukemic state of monoclonal B-cell lymphocytosis. Blood 116: 2975–2983.
34. Lu D, Zhao Y, Tawatao R, Cottam HB, Sen M, et al. (2004) Activation of the Wnt signaling pathway in chronic lymphocytic leukemia. Proc Natl Acad Sci U S A 101: 3118–3123.
35. Reya T, O'Riordan M, Okamura R, Devaney E, Willert K, et al. (2000) Wnt signaling regulates B lymphocyte proliferation through a LEF-1 dependent mechanism. Immunity 13: 15–24.
36. Longo PG, Laurenti L, Gobessi S, Sica S, Leone G, et al. (2008) The Akt/Mcl-1 pathway plays a prominent role in mediating antiapoptotic signals downstream of the B-cell receptor in chronic lymphocytic leukemia B cells. Blood 111: 846–855.
37. Calin GA, Ferracin M, Cimmino A, Di LG, Shimizu M, et al. (2005) A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. N Engl J Med 353: 1793–1801.
38. Visone R, Rassenti LZ, Veronese A, Taccioli C, Costinean S, et al. (2009) Karyotype-specific microRNA signature in chronic lymphocytic leukemia. Blood 114: 3872–3879.
39. Stamatopoulos B, Meuleman N, Haibe-Kains B, Saussoy P, Van Den NE, et al. (2009) microRNA-29c and microRNA-223 down-regulation has in vivo significance in chronic lymphocytic leukemia and improves disease risk stratification. Blood 113: 5237–5245.
40. Li S, Moffett HF, Lu J, Werner L, Zhang H, et al. (2011) MicroRNA expression profiling identifies activated B cell status in chronic lymphocytic leukemia cells. PLoS One 6: e16956-.
41. Yin Q, Wang X, McBride J, Fewell C, Flemington E (2008) B-cell receptor activation induces BIC/miR-155 expression through a conserved AP-1 element. J Biol Chem 283: 2654–2662.
42. Cimmino A, Calin GA, Fabbri M, Iorio MV, Ferracin M, et al. (2005) miR-15 and miR-16 induce apoptosis by targeting BCL2. Proc Natl Acad Sci U S A 102: 13944–13949.
43. Pekarsky Y, Santanam U, Cimmino A, Palamarchuk A, Efanov A, et al. (2006) Tcl1 expression in chronic lymphocytic leukemia is regulated by miR-29 and miR-181. Cancer Res 66: 11590–11593.
44. Calin GA, Cimmino A, Fabbri M, Ferracin M, Wojcik SE, et al. (2008) MiR-15a and miR-16–1 cluster functions in human leukemia. Proc Natl Acad Sci U S A 105: 5166–5171.
45. Puente XS, Pinyol M, Quesada V, Conde L, Ordonez GR, et al. (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature 475: 101–105.
46. Quesada V, Conde L, Villamor N, Ordonez GR, Jares P, et al. (2011) Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. Nat Genet 44: 47–52.

OPEN ACCESS Freely available online

PLoS one

Human Gene Coexpression Landscape: Confident Network Derived from Tissue Transcriptomic Profiles

Carlos Prieto, Alberto Risueño, Celia Fontanillo, Javier De Las Rivas*

Bioinformatics and Functional Genomics Research Group, Cancer Research Center (CIC-IBMCC, CSIC/USAL), Salamanca, Spain

Abstract

Background: Analysis of gene expression data using genome-wide microarrays is a technique often used in genomic studies to find coexpression patterns and locate groups of co-transcribed genes. However, most studies done at global "omic" scale are not focused on human samples and when they correspond to human very often include heterogeneous datasets, mixing normal with disease-altered samples. Moreover, the technical noise present in genome-wide expression microarrays is another well reported problem that many times is not addressed with robust statistical methods, and the estimation of errors in the data is not provided.

Methodology/Principal Findings: Human genome-wide expression data from a controlled set of normal-healthy tissues is used to build a confident human gene coexpression network avoiding both pathological and technical noise. To achieve this we describe a new method that combines several statistical and computational strategies: robust normalization and expression signal calculation; correlation coefficients obtained by parametric and non-parametric methods; random cross-validations; and estimation of the statistical accuracy and coverage of the data. All these methods provide a series of coexpression datasets where the level of error is measured and can be tuned. To define the errors, the rates of true positives are calculated by assignment to biological pathways. The results provide a confident human gene coexpression network that includes 3327 gene-nodes and 15841 coexpression-links and a comparative analysis shows good improvement over previously published datasets. Further functional analysis of a subset core network, validated by two independent methods, shows coherent biological modules that share common transcription factors. The network reveals a map of coexpression clusters organized in well defined functional constellations. Two major regions in this network correspond to genes involved in nuclear and mitochondrial metabolism and investigations on their functional assignment indicate that more than 60% are house-keeping and essential genes. The network displays new non-described gene associations and it allows the placement in a functional context of some unknown non-assigned genes based on their interactions with known gene families.

Conclusions/Significance: The identification of stable and reliable human gene to gene coexpression networks is essential to unravel the interactions and functional correlations between human genes at an omic scale. This work contributes to this aim, and we are making available for the scientific community the validated human gene coexpression networks obtained, to allow further analyses on the network or on some specific gene associations. The data are available free online at http://bioinfow.dep.usal.es/coexpression/.

Citation: Prieto C, Risueño A, Fontanillo C, De Las Rivas J (2008) Human Gene Coexpression Landscape: Confident Network Derived from Tissue Transcriptomic Profiles. PLoS ONE 3(12): e3911. doi:10.1371/journal.pone.0003911

Editor: Nicholas James Provart, University of Toronto, Canada

Received July 3, 2008; **Accepted** November 5, 2008; **Published** December 15, 2008

Funding: Funding and grant support was provided by the Ministery of Health, Spanish Government (ISCiii-FIS, MSyC; Project reference PI061153) and by the Ministery of Education, Castilla-Leon Local Government (JCyL; Project reference CSI03A06). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jrivas@usal.es

Introduction

Exploration and analysis of gene expression data using genome-wide microarrays is a technique often used in genomic studies to find coexpression patterns and locate groups of co-transcribed genes. This kind of studies has been used in model organisms, like yeast [1], to discover gene functions, to define biological processes and to find related transcription factors and their products. The main features of expression patterns that give a wide utility in bioinformatic studies are: the functional information associated [2], the high conservation of gene coexpression groups along evolution [3] and the high correlation of these groups with biomolecular pathways or reactions [4]. All these features leverage genome-wide expression profiling, and convert this topic in a hot research area.

Despite the described interest, coexpression studies done at global "omic" scale are not focused in many cases on human samples [5], and, when they correspond to human, very often they include heterogeneous datasets, mixing "normal" samples with "disease altered" samples from patients suffering from some kind of pathological state. This is the case, for example, in several human gene expression large studies [2,6]. The inclusion of many disease datasets (mainly from cancer) in such meta-analyses may introduce strong bias and produce a lot of biological noise in the results. In fact, it is well known that cancer cells have altered genomes. Therefore, these kind of studies cannot be used to clarify

how a normal-healthy human cellular system works, and they cannot be used to draw a reliable map of the human gene coexpression landscape.

The technical noise in the genome-wide expression microarray studies is another well reported problem that can not be ignored when gene coexpression studies at "omic" scale are undertaken. Considering all these problems and knowing the interest of having a reliable normal human gene coexpression network, we have undertaken this task selecting human genome-wide expression microarrays from a controlled set of different normal tissues to build a confident human transcriptomic network using several statistical and computational methods. These methods (which include robust data normalization and signal calculation, combined parametric and non-parametric correlation and random cross-validation) help to avoid both biological and technical noise and provide a human gene coexpression network that shows good accuracy and coverage. Moreover, the network reveals well defined biological functions and pathways that map to specific coexpression clusters.

Results and Discussion

Genome-wide expression profiles from a broad set of human samples

An expression matrix was calculated for a dataset of human genome-wide microarrays hybridized with mRNA samples coming from different human tissues, glands and organs from healthy normal individuals. As indicated in **Materials and Methods** the dataset included two biological replicates of samples from 24 parts of the body: *adrenal gland*, *appendix*, *blood*, *bone marrow*, *brain*, *kidney*, *liver*, *lung*, *lymph node*, *muscle heart*, *ovary*, *pancreas*, *pituitary gland*, *prostate gland*, *salivary gland*, *skin*, *spinal cord*, *testis*, *thymus gland*, *thyroid gland*, *tongue*, *tonsil gland*, *trachea* and *uterus*. **Figure 1** presents the heatmaps and clustering of the 48 samples analyzed by two different methods following the strategy and steps described in **Methods**: (**1st**) "MAS5-Spearman" method, that applies MAS5 algorithm for signal calculation and Spearman correlation coefficient (**r**) for distance calculation (based on the sample expression profiles and displayed in the heatmap as $\mathbf{1-r}$); (**2nd**) "RMA-Pearson" method, that applies RMA algorithm for signal calculation and Pearson correlation coefficient (**r**) for distance calculation (also based on the sample expression profiles and displayed as $\mathbf{1-r}$). We use "Spearman with MAS5" and "Pearson with RMA" because it has been shown that the inclusion of at least one non-parametric step based on ranks in the analyses of microarray data offers statistically more robust and more accurate estimation of expression values [7] and expression correlations [8]. The two methods proposed provide such non-parametric transformation (i.e. change to ranks), because Spearman is a rank correlation coefficient and RMA includes a quantile normalization.

The heatmaps (**Figures 1A and 1B**) show a clear and coherent clustering of each pair of biological replicates. A color bar with scales for each heatmap is included in the figure, indicating that **dark-red** corresponds to minimum distance (i.e. maximum correlation) and **dark-blue** to maximum distance (i.e. minimum correlation). White color corresponds to medium values and the distributions inside the color bars show that the two methods are similar but not identical: MAS5-Spearman provides larger distances between samples (more blue values in the heatmap) than RMA-Pearson (more red values in the heatmap).

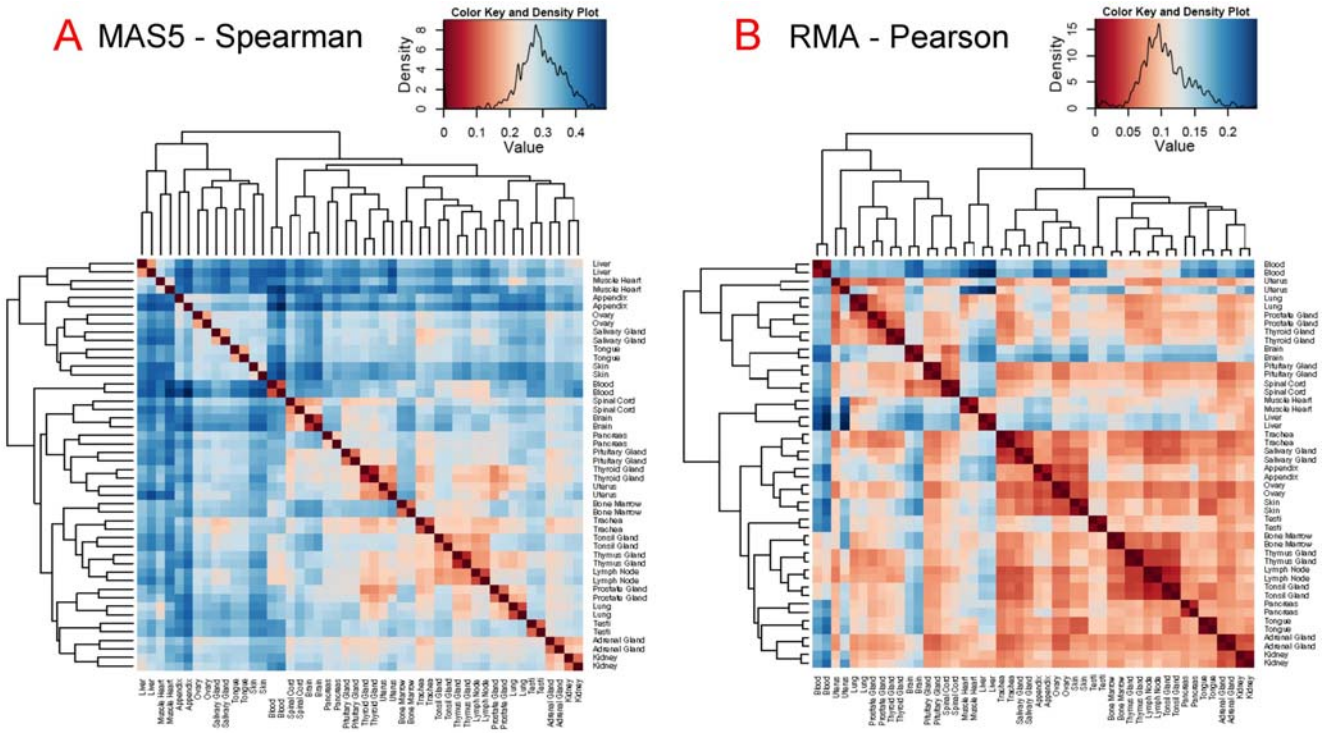


Figure 1. Clustering of human tissue expression profiles. Heatmaps and clustering of the 48 human genome-wide expression microarray samples from 24 different tissues and organs analyzed by two different methods: **(A)** MAS5-Spearman: MAS5 for signal calculation and Spearman for distance calculation based on the sample expression profiles; and **(B)** RMA-Pearson: RMA for signal calculation and Pearson for distance calculation based on the sample expression profiles. A color bar with scales for each heatmap is included, indicating that **dark-red** corresponds to minimum distance and **dark-blue** to maximum distance. The color distributions observed in the heatmaps are also included inside the bars.
doi:10.1371/journal.pone.0003911.g001

The similarity and proximity of the replicates is closer in the case of the second method, but in both cases there is not confusion or separation of any pair of replicates. By contrast to this clear clustering, the ordering and clustering of the different tissues, glands and organs is not fixed in the heatmaps, changing quite a lot from **1A** to **1B**. This observation was confirmed by bootstrap analysis done with *pvclust* [9] which allows the assessment of the uncertainty in hierarchical clusters (see **Methods**). The results of *pvclust* showed that the biological "replicate pairs" gave in all cases stable groups with optimum probability values (AU and BP = 100%). However, within the tissues and organs only two stable groups were found with both methods: the group that includes *lymph node*, *thymus gland* and *tonsil gland* (that gave a AU value of 0.98); and the group that includes *kidney* and *adrenal gland* (with AU value 0.97). These groups have clear biological meaning since they correspond to physiologically and functionally related organs (i.e. *lymph node*, *thymus* and *tonsil* are related to the lymphatic and immune systems). Thus the functional relationship between samples is captured by the gene expression profiles. However, all the other tree branches produced low AU values, therefore the overall sample clustering observed in the heatmaps indicates a lack of well defined and stable groups. In conclusion, these results show neat separation of most of the sample expression profiles, which is an adequate condition for the exploration of a global broad human gene expression landscape.

In order to consider if these observations are reliable enough, we explored the data changing some conditions following another two different strategies (data not shown). **First** strategy, the same analyses with 48 microarrays were done again twice: one not using the total number of genes (i.e. 22 283 gene probesets) but only the 25% of the genes that showed the largest variance; and another using only the 25% of the genes that showed the highest signal. In both cases, the heatmap and trees obtained were very similar to the ones presented in **Figure 1**, and the bootstrap gave similar results. **Second** strategy, we included in the data set two new groups of microarrays corresponding to samples from specific organs: 16 microarrays from different parts of the brain and 10 microarrays from different hematologic cell types. In this case (data not shown) the analyses provided larger trees, where two main clusters were segregated from other branches: one corresponding to brain related samples (i.e. nervous system) including the two whole-brain samples; and another cluster corresponding to the hematologic related samples including the two whole-blood samples. These results indicate again that any functional relation between samples is well captured by the global gene expression profiles, and provide validity to the genome-wide expression profiles of human normal tissues obtained, allowing us to proceed to the next step of the study.

From sample expression profiles to gene expression signatures

The main data presented so far correspond to the analysis of the genome-wide expression profiles of samples from different human normal tissues, organs or glands. These genome-wide "sample profiles" are numerical vectors including the expression values of each one of the gene probesets present in the microarray (i.e. each one of the detectable genes of the human genome). As shown above, the "sample profiles" can resemble the physiological relationships expected between the samples (tissues, glands and organs). However, in order to achieve a mapping of the human gene coexpression landscape, we needed to move from the analysis of the "sample expression profiles" based on the genes, to the analysis of each "gene expression signature" based on the sample set.

It is difficult to achieve a proper gene coexpression study due to several obstacles that have to be taken in consideration: **(i)** the technical noise present in the microarrays at genomic scale [10], despite the fact that the *Affymetrix* high density oligonucleotide genechips have been reported quite reliable and reproducible [11,12]; **(ii)** the small number of samples used to define each gene expression signature (specially in comparison to the large number of genes); **(iii)** the strong heterogeneity of the data sets frequently studied, that include in many cases samples from pathological or altered states [2,13] which are not adequate samples to find "normal" gene expression behavior.

The approach and strategies taken in this study to solve or minimize these problems were the following: **(a)** careful selection of expression samples from different parts of the human body (tissues, whole glands and whole organs) from normal healthy individuals; **(b)** calculation of expression signals and correlations using two different independent methods: MAS5-Spearman, RMA-Pearson; **(c)** use of a robust random cross-validation strategy to find the most stable correlation pairs and distinguish the consistent biological-signal from the noise-signal; **(d)** statistical estimation of the accuracy and the coverage for each coexpression dataset obtained. All the details and description of these strategies are presented in **Materials and Methods**. The results associated with them have been partially described above and are explained in the following paragraphs.

Gene pairs coexpression analyzed with cross-validated correlations

The complete expression data matrix analyzed had, as indicated, 48 samples (24 duplicates) and 22,283 gene probesets (which correspond to 13,068 distinct known human genes according to *Affymetrix* annotation). Therefore the global pair-wise gene coexpression matrix including all possible pairs had 248,254,903 data points and was calculated twice, once for each independent method used (MAS5-Spearman and RMA-Pearson). These huge data matrices have many pairs that are false coexpression pairs and to detect those positive gene pairs that had stable and significant correlation we use cross-validation. The results corresponding to the gene pairs correlation obtained with the cross-validation method (described in **Methods**) are presented in **Figure 2**, that shows what we called "**rN-plots**". The rN-plots are graphics representing: **r** at *y* axis, that is, for each gene probeset pair, the "correlation coefficient" of their expression signatures along the complete dataset of 48 samples, calculated as Spearman or Pearson distance (for MAS5 or RMA data, respectively) (with values from 0 to 1 for positive correlations and from 0 to −1 for negative correlations); **N** at *x* axis, that is the "cross-validation coefficient" defined as the number of times that a given gene pair has a significant correlation (i.e. $r \geq |0.70|$) out of the 1000 times random selection (as explained in **Methods**). This graphical analysis presents the positive and negative correlations well segregated and it allows to identify those gene pairs that have a significant "cross-validated correlation", discriminated from those false gene-pairs that have low **r** or low **N** values. Such false gene-pairs do not correlate in a stable and consistent way, being undistinguishable from noise.

To demonstrate how the **rN-plots** represent stable and consistent correlations, we selected in the case of the **red** circles or dots only the gene probeset pairs that correspond to probesets assigned to "the same gene". For example, pairs between the 4 probesets that correspond to gene ALDOB, *fructose bisphosphate aldolase B* (204704_s_at, 204705_x_at, 211357_s_at, and 217238_s_at in microarray HGU133A); or pairs between the 3 probesets that correspond to gene CDK10, *cell division protein kinase 10* (203468_at, 203469_s_at and 210622_x_at in HGU133A). When correlation is found between these kind of "common gene probesets" they are drawn as **red** circles in **Figure 2**. The analysis indicates that the red circles accumulate at high **r** correlations and

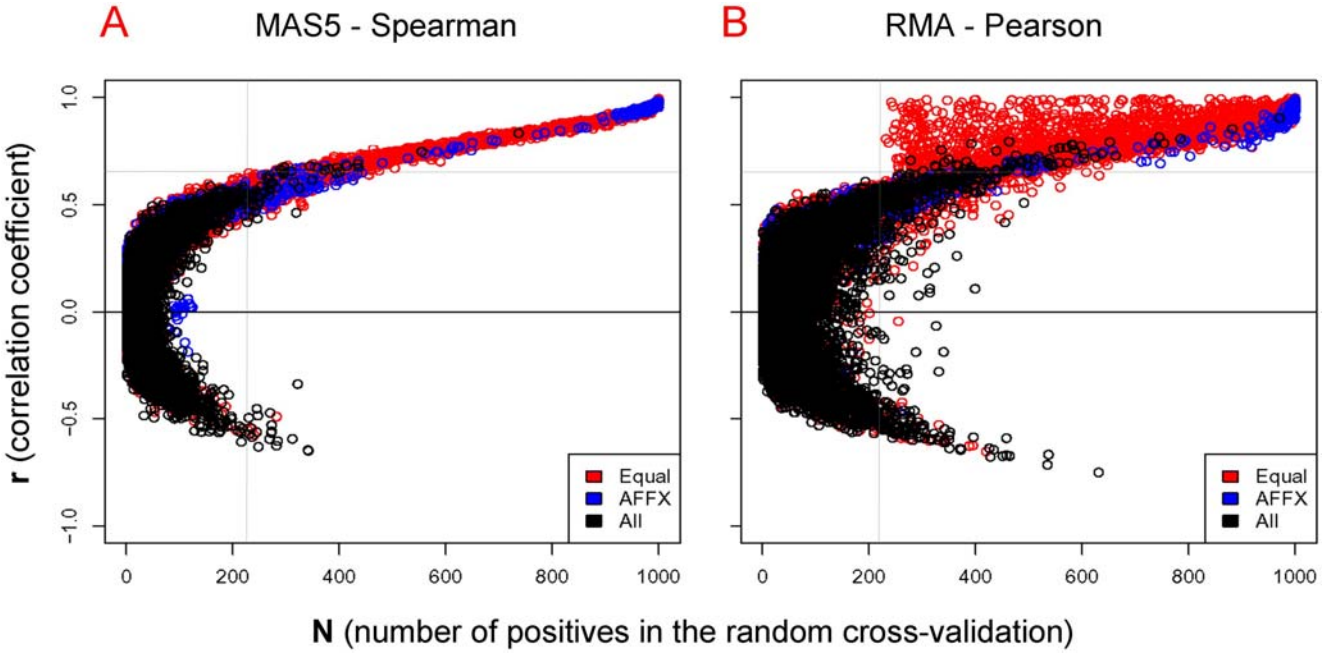


Figure 2. Plot of r and N coefficients calculated for each gene coexpression pair. rN-plots that represent the correlation coefficient (from 0 to 1) versus the cross-validation coefficient (from 0 to 1000) of each gene pair by two different methods: **(A)** MAS5-Spearman and **(B)** RMA-Pearson. The cross-validation is considered positive for a given gene pair when it gives **r**>|0.7| in each sampling. As indicated in **Methods** 1000 samplings are run for each gene-probeset pair. The gene probeset pairs that correspond to the same gene are drawn as **red** circles. The probeset pairs of *Affymetrix* controls are drawn as **blue** circles. A random selection of 10,000 coexpressed gene probeset pairs are drawn as **black** circles. Two dotted lines are drawn to indicate an approximate threshold that can be considered the border of noisy data. These lines are drawn just to show the minimal **r** and **N** values bellow which the coexpressed gene pairs are mainly noise; therefore the coexpression signal appears mostly at **r**>0.65 and **N**>220.
doi:10.1371/journal.pone.0003911.g002

high **N** values. This is the result that should be expected considering that these groups of probesets are measuring the same gene; and, despite the fact that this is not always true, it is a good way to evaluate the meaning of the **rN-plot**. A more stringent evaluation was to find out the correlation between probesets that correspond to "control RNAs" that are added in each microarray assay in the hybridization process. Such controls, named with prefix AFFX in the chip, are spike controls (i.e. series of mRNAs added during hybridization protocol that correspond to different concentrations of non-human genes like AFFX-BIO) and human house-keeping controls (like AFFX-HUMGAPDH). These controls should have strong correlation since they have been added to the microarrays in known concentrations. We draw such correlations in the **rN-plots** as **blue** circles (**Fig. 2**); and it could be seen that the distribution of these true positive gene correlated pairs was very much accumulated at high **N** values and high **r** correlations. This observation again shows that the **rN-plots** are very useful and valuable to separate noisy false correlations from stable true correlations.

The differences observed between **Fig. 2A and 2B** are due to the differences in the methods and to the characteristics of the cross-validation (described in **Methods**). Some **red** circles with high-**r** and low-**N** appear only in the RMA-Pearson method because the correlations derived from this method give in some instances high correlation values to gene pairs that are correlated just in only one tissue (shown in **Fig. 2B**). The cross-validation values of these gene pairs are low because they only appear when such one tissue samples are selected. The probability to select one sample pair out of 24 is: $1-(23/24)^6 = 0.225$; and this is why the red circles with high-**r** and low-**N** only appear for values **N**>225. By contrast, the MAS5-Spearman method does not find any **red** circle in the high-**r** and low-**N** region, because Spearman is a "rank correlation coefficient" which does not produce high correlation values for gene pairs that correlated in only one tissue (just once out of 6). The **r** value obtained with the Spearman method is proportional to the number of tissues or samples that co-express and so it is quite proportional to **N**.

Data filtering to clear genes with low information content

The calculations and analysis presented in **Figure 2**, were done without using any previous filter of gene probesets. No filtering means using the complete gene expression matrices with all the human gene probesets present in the microarrays. It is known that in most samples and conditions genome-wide microarrays include a large proportion of the genes that are not expressed and therefore they give signal close to the background or noise. This situation is not very likely to occur all along the complete sample set of 24 different tissues and organs studied here. However, out of the 22,283 gene probesets some may have no significant change, and therefore, it is important to find out the possible presence and effect of these "non-changing genes" (that we also called "flat-genes") [14]. The most adequate filter to be used in most of the expression analyses is a variance-filtering between samples (i. e. between-array variability), because this approach filters out elements of low information content within the sample set and covers the complete signal range (from low to high expression), therefore, it does not bias the data by signal intensity or signal ratios [14,15]. However some genes with high signal may be significant despite showing relative low variance, and for these reasons it is better to apply combined filters that explore the variance, but also consider the intensity of the probes [15].

As described in **Methods** we use a combined filter based on between-sample variability and gene minimal signal, that is designed to get rid of genes with low information content. The use of this filter with the 48 microarrays sample set gave different results for the data expression matrix obtained with RMA method and the expression matrix obtained with MAS5 method. In the first case the filter leaves out 6,893 gene probesets (leaving 69.06%) and in the second 3,682 (leaving 83.48%) from 22,283 total gene probesets. The difference in these numbers shows that these two methods do not provide an equal calculation of expression signal and variance and therefore, as explained bellow, both methods can be considered complementary.

Analysis of accuracy and coverage along gene coexpression data

Using the filtered data sets we follow a more thorough analysis of the coexpression distributions with respect to the parameters **r** and **N**. In the rN-plots (**Fig. 2**) two dotted lines were drawn to indicate an approximate threshold for coexpressed gene pairs that could be considered the border of noisy data. These lines are tentatively drawn just to show the minimal **r** and **N** values bellow which the coexpression pairs are mainly noise; therefore, the coexpression signal appears mostly at **r**>0.65 and **N**>220. However, this estimation is not robust enough and a proper calculation of the statistical "accuracy" and "coverage" along all the gene coexpression data matrices was done. The details about the calculation of these parameters are described in **Materials and Methods**. KEGG pathway database was used to estimate the true positives. After these calculations, for all data presented (i.e. all next **Figures**) the nodes correspond to genes and not any more to "gene probesets" from the microarrays. This change was done taking the correspondence of the probesets to the specific genes according to the *Affymetrix* annotation files for HG-U133A from 31.May.2007 (that can be found in URL: http://www.affymetrix.com/support/technical/byproduct.affx?product = hgu133). In this conversion all probesets of the microarray were used. Previously, we calculated the coexpression values for each gene pair considering each probeset independently. When multiple probesets map to one gene, we merged the multiple probesets to the corresponding gene and we only take the gene coexpression pairs with maximum values of correlation (**r**) and cross-validation (**N**) in which its probesets participate.

In **Figure 3** the positive predictive value (PPV) was computed for each coexpression data set obtained at a given correlation factor **r** (**Fig. 3** top graphs) or at a given the number of cross-validations **N** (**Fig. 3** bottom graphs). The change or evolution of the accumulated PPV is drawn as a curve (solid **red** and **blue** circles) for both methods (**Fig. 3A**: MAS5-Spearman; **B**: RMA-Pearson). The graphs show that the rate of true positives increases with higher expression correlation and with higher number of cross-validation. The increase is more significant for the MAS5-Spearman method that achieves PPV about 80% for **r**≥0.8 and

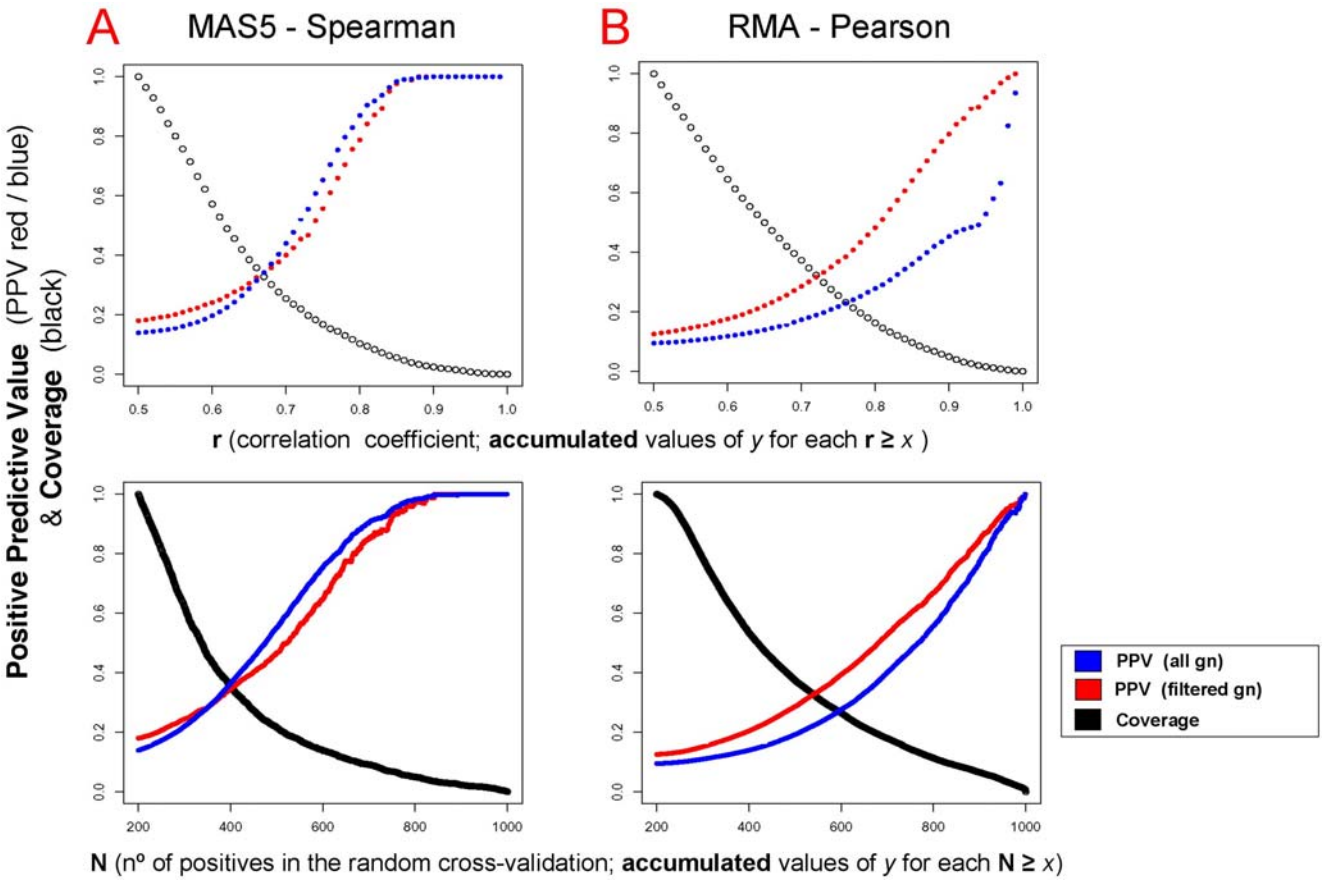


Figure 3. Accuracy and coverage of the coexpression data. Accuracy measured as Positive Predictive Value PPV (for all genes in **blue** and filtered genes in **red**) and coverage as True Positive Rate TPR (in **black**) computed for each coexpression dataset obtained at a given correlation coefficient **r** (top figures) or at a given number of cross-validations **N** (bottom figures) for both methods: **(A)** MAS5-Spearman and **(B)** RMA-Pearson. The accuracy and coverage (in *y* axis) correspond to accumulated values for each **r**≥*x* or for each **N**≥*x*.
doi:10.1371/journal.pone.0003911.g003

for **N**≥700. However, RMA-Pearson provides higher coverage since the amount of positive gene coexpression pairs annotated to common KEGGs for **r** and **N** values is quite different in both methods (larger for RMA-Pearson). The results for the coverage calculated for each method are shown by the curves in black in **Figure 3** (**black** circles), presenting the amount of gene pairs annotated to common KEGGs that remain at each **r**≥*x* or **N**≥*x*. This is calculated considering as "total amount of positive pairs" (value 1.0 at the beginning of the curve, 100%): the number of gene coexpressing pairs annotated to common KEGGs at **r**≥0.5 and **N**≥200. This coverage parameter indicates, as it should be expected, that the number of gene coexpressing pairs decreases when the conditions (**r** and **N**) are more stringent. The decrease is steeper for the MAS5-Spearman method since for **r**≥0.75 it retains about 16.7% of the positive data points, but RMA-Pearson retains 25.4%. Equally for **N**≥600 the MAS5-Spearman method retains 13.9% of the positive data points and RMA-Pearson retains 26.4%. The total amount of positive pairs, which corresponds to value 100% at the beginning of the curve, was: 15,657 for RMA-Pearson and only 2,198 for MAS5-Spearman. These numbers seem small but they only correspond to the "positive pairs", and so, if we take the total number of gene probeset coexpression pairs of the study (i.e. not including only the genes annotated to KEGGs but the complete coexpression data sets) the figures are much larger: 1,340,472 for RMA-Pearson and 180,305 for MAS5-Spearman. These results also indicate that the coverage is larger with the RMA-Pearson method.

In conclusion, the study shows that the RMA-Pearson method has better coverage of the coexpression landscape and the MAS5-Spearman is more accurate to find coexpression pairs. These results support the use of both methods in order to find a confident human coexpression network, since they do not find exactly the same expression signal and both provide important and complementary data allowing a progressive improvement of the significance and confidence of the coexpression set. Moreover, a better knowledge of the strength of each method is a discovery that complements previous comparative studies about RMA [7] and MAS5 [8].

Effects of gene filtering

The original coexpression data used in **Figure 2** are obtained without any gene filtering, however for the analyses in **Figure 3** it was convenient to study the effect of gene filtering upon the accuracy and coverage of the methods. The evolution of the coverage did not show any significant change (data not shown). The evolution of the accuracy was studied by plotting the relative changes of the positive predictive values (PPV) of the coexpressing data with **r** (**Fig. 3** top graphs) and **N** (**Fig. 3** bottom graphs) for each method. In these graphs the **blue** circles correspond to non-filtered data and **red** circles to filtered data. This analysis indicates that for the case of RMA-Pearson method (**Fig. 3B**) a significant improvement was obtained with the gene filtered versus non-filtered. However, in the case of MAS5-Spearman there was not any relative improvement, as it can be seem in **Fig. 3A** both for **r** and **N**. This means that **r** and **N** are already very stringent in MAS5-Spearman dataset and the filter takes out approximately the same amount of estimated true positives and false positives within the data, and so it does not improve the coexpression accuracy (i.e. PPV). This observation, together with the fact that filtered data with the MAS5-Spearman method gives low coverage (as indicated above the total amount of positive pairs was only 2,198), brings us to the resolution of not using the filter for MAS5 dataset. By doing this, the MAS5-Spearman non-filtered dataset at **r** = 0.5 and **N** = 200 included 15,623 positive coexpression pairs; and this number was very similar to the 15,657 pairs found for RMA-Pearson filtered.

Integration of correlation, cross-validation and PPV for datasets obtained with two balanced methods

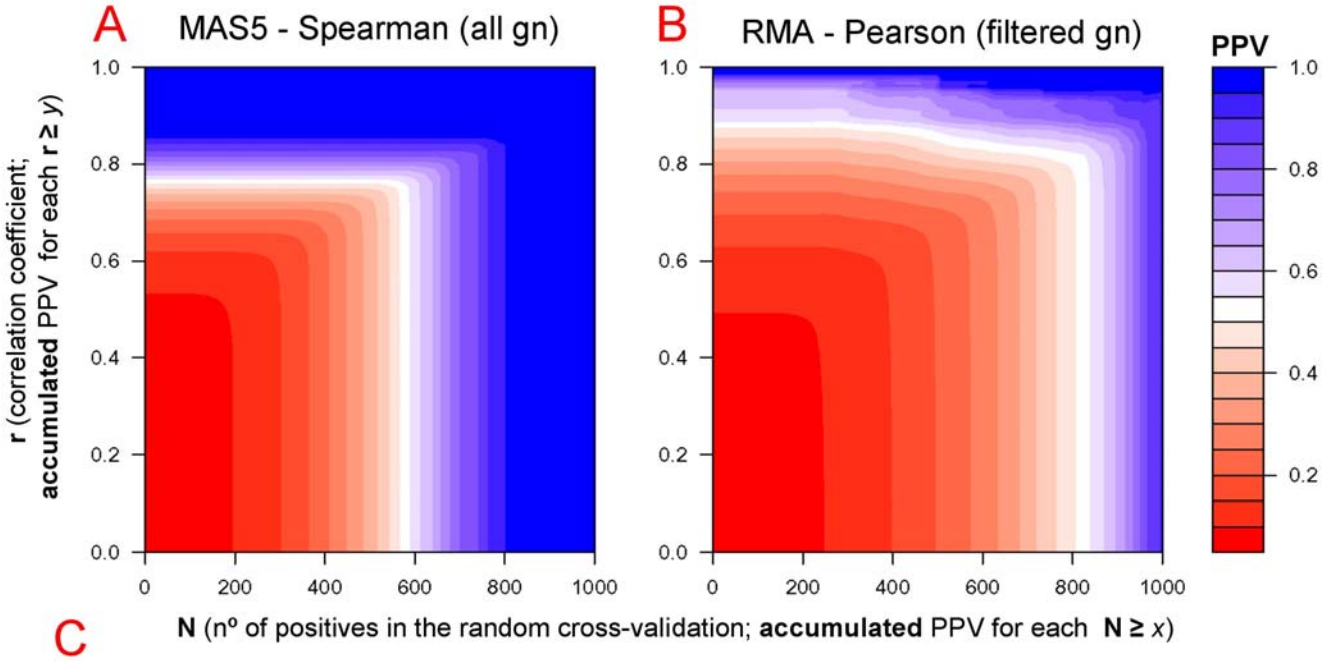
Following the observations and arguments described above we proceed to integrate in "three-dimensions color plots" the data corresponding to the values of correlation (**r**), cross-validation (**N**) and PPV obtained with each method. The results are shown in **Figure 4**. The graphic considers all the calculated subsets of coexpression gene pairs and represents, for each one, the numerical relationship between the accumulated values of the estimated accuracy (**PPV**) corresponding to the correlation coefficients (**r** in *y* axis) and to the cross-validation coefficients (**N** in *x* axis). PPV ranges from 0.05 to 1.0 as indicated in the color scale of **Fig. 4**: **red** low and **blue** high. The graph are calculated for the data corresponding to two methods: MAS5-Spearman without gene filtering (all gn) (**Fig. 4A**) and RMA-Pearson with gene filtering (filtered gn) (**Fig. 4B**). As indicated above, in these conditions both methods include a similar number of coexpression pairs and so they are "balanced" with respect to the coverage.

The three-dimensions color plots allow to assess in a graphic way the level of confidence for a given coexpression data subset. We use them to select three data subsets derived from each method at three specific PPV values: ≥0.60, ≥0.70 and ≥0.80. The values of the correlation and cross-validation coefficients that correspond to these data subsets are indicated in the table enclosed as **Fig. 4C**. The figures show that the second method (RMA-Pearson) is more stringent, since the same given PPVs correspond to higher values of **N** and **r**. The size of the gene coexpression networks that correspond to the three selected accuracy values are also presented in **Fig. 4C**, including for each network the number of nodes (i.e. number of genes) and the number of links (i.e. number of coexpression pairwise relations). The selection and combination of these subsets at well defined and precise accuracy allows the identification of stable and confident human coexpression networks. This was done in the table enclosed as **Fig. 4D**, where the results of the union and the intersection of the datasets provided by the two methods at each PPV are presented. The union with accuracy ≥0.60 provides a full confident and cross-validated human gene coexpression network that includes 3327 genes and 15841 coexpression links. As indicated bellow, we have analyzed in detail a core transcriptomic network that corresponds to the intersection of both methods with accuracy ≥0.60 and includes 731 gene nodes and 2249 coexpression links.

Biological significance of the coexpression datasets: house-keeping gene pairs and tissue-specific gene pairs

Once significant human gene coexpression datasets have been found and evaluated using statistical parameters, we started exploring the biological meaning and functional consistency of these datasets.

In a first approach, we investigate the location of house-keeping gene pairs in the coexpression datasets, taking two different published compendiums of human house-keeping genes [16,17]. *Hsiao et al.* identified 451 genes that are expressed in all 19 different human tissue types. *Eisenberg et al.* identified 575 human genes that show constitutive expression in all conditions tested in several publicly available databases. Mapping these genes in the general distribution of coexpression data shows that the ratio of house-keeping genes increases at high **N** and **r** coefficient values (**Fig. 5A,B**). The top panels in **Fig. 5A and B** present the density distributions of coexpression data for **N**>220 corresponding to all gene pairs (in **black**), to *Eisenberg's* house-keeping gene pairs (in **green**) or to *Hsiao's* house-keeping gene pairs (in **red**). Bottom panels in **Fig. 5A and B** show the same information including now



C

PPV [1]	Coefficients		Nº of Nodes [2]	Nº of Links [2]
	N	r		
MAS5-Spearman (all gn)				
≥ 0.60	605	0.77	3052	**12669**
≥ 0.70	645	0.79	2295	7874
≥ 0.80	695	0.81	1762	4910
RMA-Pearson (filtered gn)				
≥ 0.60	765	0.85	1672	**5945**
≥ 0.70	835	0.87	1215	3273
≥ 0.80	925	0.84	983	2423

1 Accuracy as PPVs that correspond to the networks derived for KEGG annotated genes
2 Values that corresponds to the full networks including all genes

D

	Coexpression networks			
PPV [1]	Union		Intersection	
	Nº Nodes	Nº Links	Nº Nodes	Nº Links
≥ 0.60	3327	**15841**	731	**2249**
≥ 0.70	2411	9264	542	1447
≥ 0.80	1863	5935	387	1008

Figure 4. Coexpression networks obtained at different levels of accuracy. Color plots **(A** and **B)** that represent the Positive Predictive Value (**PPV**) calculated for each set of gene coexpression data for different values of correlation coefficient (**r**) and cross-validation coefficient (**N**). The PPV corresponds to accumulated values for **N**≥*x* and **r**≥*y*. Calculations are done for data derived from two methods: **(A)** MAS5-Spearman without gene filtering (all gn) and **(B)** RMA-Pearson with gene filtering (filtered gn). Table **(C)** shows the specific values of correlation and cross-validation for three coexpression datasets derived from each method at 3 specific PPVs: ≥0.60, ≥0.70 and ≥0.80. This table also shows the number of nodes and links included in each coexpression dataset. Table **(D)** shows the number of gene-nodes and interaction-links that are included in the combined coexpression networks at 3 specific PPVs.
doi:10.1371/journal.pone.0003911.g004

all data points of coexpression pairs with **N**>220 and **r**>0.65 for either all gene pairs (in **black**) or only *Hsiao's* house-keeping gene pairs (in **red**). Panels **A** correspond to coexpression data obtained with method MAS5-Spearman and **B** to RMA-Pearson. The results reveal that house-keeping genes have a clear tendency to coexpress in many different tissues. This can be expected from the mere definition of house-keeping; however, since the result is obtained by mapping external datasets [16,17] on our human gene coexpression data, it provides functional validity to our coexpression study. The analysis also reveals a clear difference between the data obtained with different methods. Meanwhile MAS5-Spearman method finds mainly house-keeping gene coexpression, the RMA-Pearson method finds many gene pairs that are not in the major house-keeping region, but rather they show high levels of **r** correlation with lower levels of **N** cross-validation (**N**>220 and **N**<600).

We further investigate this observation by selecting subsets of the coexpression data for genes included in specific KEGG pathways. Examples of this subsetting are presented in **Fig. 5C**, that includes 6 panels with the coexpression data obtained with the RMA-Pearson method for the human genes included in 6 different pathways: **(1)** ribosome (KEGG ID = hsa03010), **(2)** oxidative phosphorylation (hsa00190), **(3)** proteasome (hsa03050), **(4)** cytokine-cytokine receptor interaction (hsa04060), **(5)** neuroactive ligand-receptor interaction (hsa04080), and **(6)** complement and coagulation cascades (hsa04610). First three pathways can be considered as general constitutive, present in all tissues and cellular types. The other three pathways are tissue-specific, only present in some cell types, like: nervous system cells in the case of the neuroactive ligand-receptor interaction pathway or blood cells in the case of the complement and coagulation cascades pathway. These differences in functional specificity are reflected in the coexpression distributions: only the three panels on the right (**Fig. 5C 4,5,6**) present data points with high **r** values but relatively lower **N** values (220<**N**<600). In conclusion, this analysis reveals that such coexpression pairs correspond to genes expressed in specific cells or specific tissue types, and so they are tissue-specific genes.

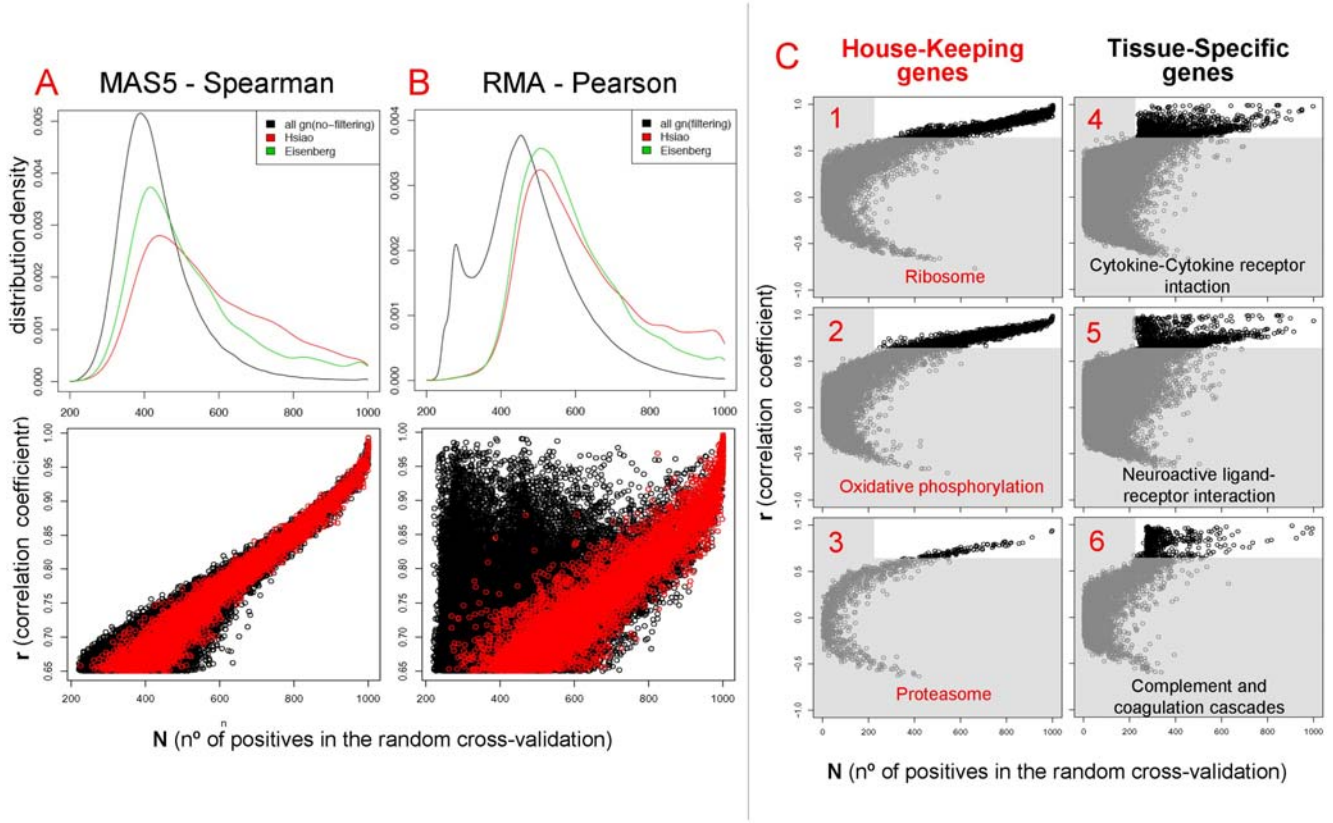


Figure 5. Coexpression of house-keeping and tissue-specific genes. Top panels A and B: Density distributions of coexpression data for N>220 corresponding to all gene pairs (in **black**), to Eisenberg's house-keeping gene pairs (in **green**) or to Hsiao's house-keeping gene pairs (in **red**). **Bottom panels A and B:** rN-plots with all data points of coexpression pairs with N>220 and r>0.65 for either all gene pairs (in **black**) or only Hsiao's house-keeping gene pairs (in **red**). In these panels **(A)** correspond to data from MAS5-Spearman method and **(B)** from RMA-Pearson method. **Panels (C)** 6 rN-plots that present the coexpression data obtained with the RMA-Pearson method corresponding to the human genes included in 6 different pathways: **(1)** ribosome (KEGG ID = hsa03010), **(2)** oxidative phosphorylation (hsa00190), **(3)** proteasome (hsa03050), **(4)** cytokine-cytokine receptor interaction (hsa04060), **(5)** neuroactive ligand-receptor interaction (hsa04080), and **(6)** complement and coagulation cascades (hsa04610).
doi:10.1371/journal.pone.0003911.g005

Comparison of human coexpression datasets: molecular machines and pathways consistently co-regulated

In a second approach, we investigate the functional assignment of the gene coexpression data following the strategy taken by *Stuart et al.* [5], who explored functional coverage on a coexpression network obtained for four organisms looking at the percentage of genes that are connected to at least one other gene in the same "functional category". We proceed to the same percentage calculation using the KEGG pathways as "functional categories". The analysis was done for the coexpression dataset derived from RMA-Pearson method with **r**>0.63 and **N**>500. The same functional analysis was also done using two other external human coexpression datasets previously published by *Lee et al.* [2] and *Griffith et al.* [6].

The results are presented in **Table 1**, that includes the ten-top pathways found with best percentage of genes coexpressing within the gene groups assigned to KEGG pathways for 3 different human coexpression datasets (this work, *Lee et al.* and *Griffith et al.*). This comparative analysis of functional coverage shows some interesting results: **(i)** All coexpression datasets find the most significant coexpression for 3 key molecular machines: ribosome, proteasome and oxidative phosphorylation. **(ii)** Genes involved in cell scaffolding and cell to cell interaction or anchoring are also found to coexpress quite often, as indicated by the presence of pathways like focal adhesion, extracellular matrix (ECM) interaction and cytoskeleton regulation. **(iii)** Genes involved in cell cycle pathway are also common to the three datasets, indicating that cells keep a tight regulation of the genes involved in essential living functions (maintenance, proliferation, survival). **(iv)** An important difference between our coexpression dataset and *Lee et al.* or *Griffith et al.* datasets is that this work only includes samples coming from normal non-pathological tissues, but the others include quite heterogeneous samples mixing normal and disease altered samples (for example, *Lee et al.* includes many human cancer samples). The inclusion of pathological samples can bias the results and this may be the reason of the appearance of "pathogenic infection pathways" in *Lee et al.* data. **(v)** Finally, the data obtained in this work also includes many coexpressing pairs involved in cell-cell communication like cytokine-receptor and ligand-receptor interactions.

As a general conclusion of this analysis, we can say that KEGG pathways is revealed as a good database to investigate the biological functions of human genes, because it includes groups of genes that really work together in well defined biomolecular processes.

The comparative calculation of the coverage for the three human coexpression datasets included in **Table 1** indicates that the data obtained in this work present a higher level of functional coherence than previously published datasets [2,6]. This comparison was also done taking coexpression networks of similar sizes (including in each case around 12,000 best coexpression relations) and calculating the statistical accuracy for all of them. The result presented in **Table 2** shows that the accuracy estimated as PPV

Table 1.

This work (2008)

Pathway Name (KEGG ID number)	n° gn [1]	gn coexp/gn [2]	% gn coexp	mean r [3]
Proteasome (3050)	31	28/28	**100.0%**	0.69
Ribosome (3010)	120	52/55	**94.5%**	0.75
Oxidative phosphorylation (190)	129	88/95	**92.6%**	0.73
Focal adhesion (4510)	194	154/168	**91.7%**	0.68
Antigen processing and presentation (4612)	86	71/78	**91.0%**	0.75
Glycan structures - degradation (1032)	30	20/22	**90.9%**	0.65
Neuroactive ligand-receptor interact. (4080)	299	227/255	**89.0%**	0.68
Cell cycle (4110)	114	90/102	**88.2%**	0.66
Regulation of actin cytoskeleton (4810)	208	141/161	**88.2%**	0.66
Cytokine-cytokine receptor interact. (4060)	256	196/223	**87.9%**	0.69

Lee et al. (2004)

Pathway Name (KEGG ID number)	n° gn [1]	gn coexp/gn [2]	% gn coexp
Ribosome (3010)	120	43/44	**97.7%**
Proteasome (3050)	31	19/22	**86.4%**
Oxidative phosphorylation (190)	129	31/44	**70.5%**
Cell cycle (4110)	114	33/47	**70.2%**
ECM-receptor interaction (4512)	87	16/23	**69.6%**
Gap junction (4540)	92	9/13	**69.2%**
Pathogenic Escherichia coli infection (5130)	49	11/16	**68.8%**
Pathogenic Escherichia coli infection (5131)	49	11/16	**68.8%**
T cell receptor signaling pathway (4660)	93	15/22	**68.2%**
Metabolism of xenobiotics by cytP450 (980)	70	7/11	**63.6%**

Griffith et al. (2005)

Pathway Name (KEGG ID number)	n° gn [1]	gn coexp/gn [2]	% gn coexp
Ribosome (3010)	120	36/38	**94.7%**
Proteasome (3050)	31	20/24	**83.3%**
Oxidative phosphorylation (190)	129	55/67	**82.1%**
Val, Leu and isoleucine degradation (280)	50	15/19	**78.9%**
ECM-receptor interaction (4512)	87	16/22	**72.7%**
Cell cycle (4110)	114	36/51	**70.6%**
Propanoate metabolism (640)	34	9/14	**64.3%**
Butanoate metabolism (650)	44	9/14	**64.3%**
Hematopoietic cell lineage (4640)	88	18/28	**64.3%**
beta-Alanine metabolism (410)	24	7/11	**63.6%**

[1]n° gn = whole number of genes included in this KEGG pathway.
[2]gn coexp/gn = genes that coexpress within the genes included for this pathway in the network.
[3]mean value of the correlation factor (r) for the coexpressing gene pairs included in this pathway.
doi:10.1371/journal.pone.0003911.t001

was 0.61 for our dataset obtained with MAS5-Spearman, 0.56 for *Lee et al.* and 0.49 for *Griffith et al.* As a whole these numbers indicate that the human coexpression network derived from this work includes very consistent co-regulation of genes many times involved in common pathways.

A high confidence human coexpression network reveals a map of ubiquitous biological functions

As far as we know, none of the previously published human coexpression networks [2,5,6] has a comprehensive calculation of the estimated statistical error in the datasets at different levels of coverage. However, following the analysis and data presented in **Figure 4** we can select coexpression datasets at specific thresholds of PPV accuracy. In order to gain in reliability, we can also combine the data obtained with 2 methods: MAS5-Spearman and RMA-Pearson. This was done taking the datasets of both methods with PPV≥0.60 (3052 and 1672 genes) to produce an intersect coexpression network that includes 731 genes and 2249 coexpression interactions (see **Fig. 4D**). We also restrict the network including only coexpressing groups including at least three genes. In this way, a high confidence core subset of 615 gene nodes and 2190 coexpression links was obtained.

Table 2.

	Nodes [1]	Links [2]	TP [3]	All [4]	PPV [5]
This work (2008)	3052	12669	729	1189	**0.613**
Lee et al. (2004)	1751	12187	1275	2265	**0.563**
Griffith et al. (2005)	2922	12686	1265	2588	**0.489**

[1]N° of genes as nodes in the network (the values correspond to the full networks including all genes).
[2]N° of coexpression links (the values correspond to the full networks including all links).
[3]True Positives = gene-pairs that coexpress and are annotated to the same KEGG.
[4]All the genes that coexpress and are annotated to KEGG.
[5]Accuracy as PPVs that correspond to the networks derived for KEGG annotated genes.
doi:10.1371/journal.pone.0003911.t002

Figure 6 presents a graphical view of this coexpression network where the nodes correspond to genes and the edges to coexpression. The network was produced introducing the coexpression dataset of 615 genes and 2190 pairwise interactions in *Cytoscape* (a bioinformatics software platform for visualizing molecular interaction networks, [18]. In the graphical view the most significant regions of this human gene coexpression network have been marked with background colors to enhance them as constellations within the coexpression landscape. Labels have been placed to each colored region to describe the main biological processes that are common to most of the genes in each region. The map shows that the larger subnetwork corresponds to genes involved in nuclear activity and nuclear-driven metabolism (region in **blue**), with a side part (in dark **blue**) that includes most of the ribosomal proteins and proteins involved in ribosomal function. The second major constellation (region in **green**) includes many genes involved in mitochondrial metabolism and redox homeostasis (like genes of the COX family, the NDUF family and the UQCR family). The third main region (in **red**) corresponds to genes involved in the immune response, genes of the major histocompatibility complex (MHC), genes that produce the cell surface clusters of differentiation (CD) and genes that encode antigen-specific molecules. Finally some smaller regions include: genes involved in metal ion homeostasis (in **grey**); genes related to the extracellular matrix and cell adhesion (in **orange**); genes related to the cytoskeleton (in **yellow**).

As a whole the network is quite stringent but it is functionally very coherent. Moreover, coming from the intersection of two methods it will be expected to include mainly essential human genes. To prove if this network is enriched in house-keeping and essential genes we identified the nodes of the network that are included in the *Hsiao* human house-keeping gene set [16] and we also identified the nodes that correspond to genes that are orthologous to known essential yeast genes (taken from SGD database). In this way, we found that the two major constellations of the network, including mainly genes involved in nuclear related and mitochondrial related metabolism, show respectively 63% and 58% of genes assigned to be house-keeping. This result reveals that the coexpression network is enriched in essential genes.

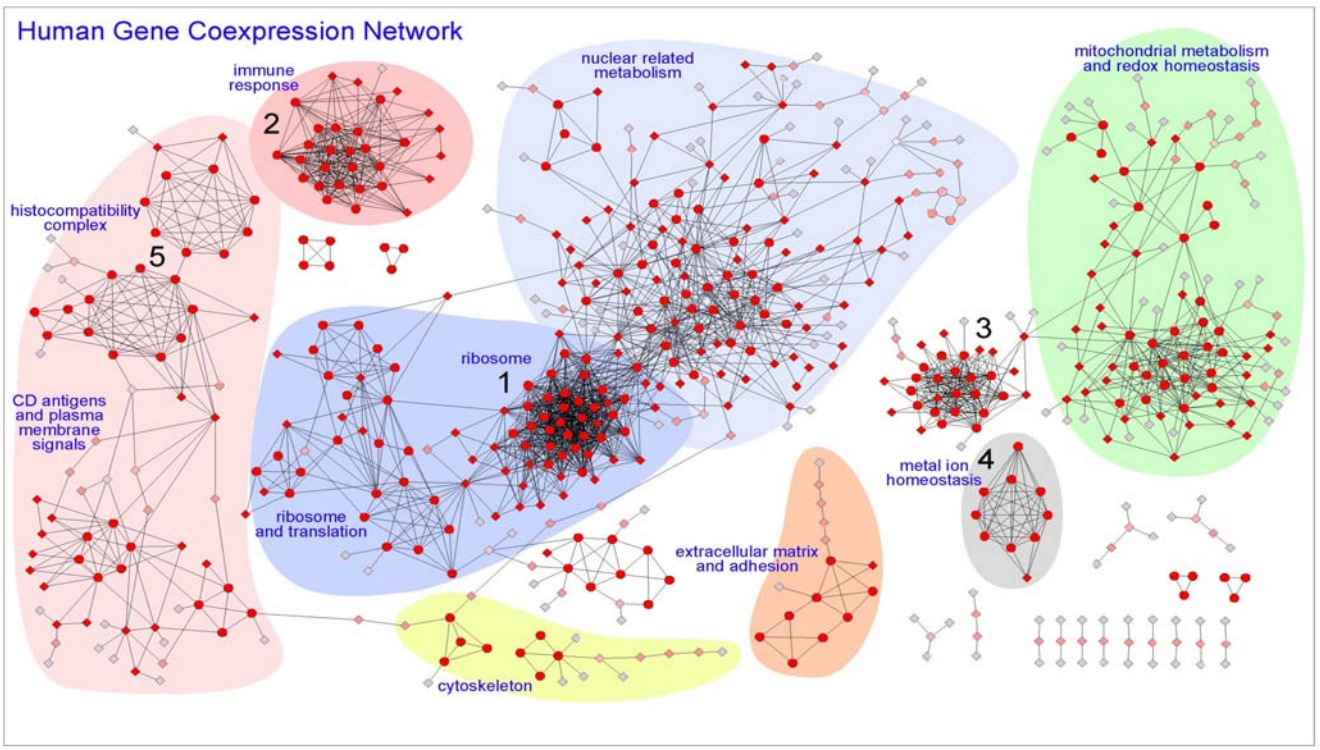


Figure 6. Human Gene Coexpression Network. Graphical view of the human gene coexpression network where the nodes correspond to genes and the edges to coexpression links. The network was produced as the intersection of two datasets (MAS5-Spearman and RMA-Pearson datasets with PPV≥0.60) to provide a confident coexpression network that includes 615 genes and 2190 pairwise coexpression interactions. The network includes only groups of coexpressing genes with at least three nodes. The most significant regions have been marked with background colors and labels describe main functions assigned. For each node the color (from **red** to **grey**) and shape (circles or diamonds) were obtained with MCODE algorithm. The circular nodes are the ones found with high cluster coefficient and the diamond nodes are the ones with lower cluster coefficient. The intensity of the **red** color in the nodes also indicates the degree of clustering, changing till pale **grey** for the most peripheral nodes that only have one link.
doi:10.1371/journal.pone.0003911.g006

In conclusion, the functional consistency observed in the constellations and regions defined by the coexpression network and the enrichment on house-keeping genes place the genes in a new integrative relational context that has strong biological coherence and, in many cases, can reveal essential or ubiquitous biological processes. The network also unravels new non-described human gene associations.

All the details about this coexpression network are provided in a supplementary file for *Cytoscape* (Supporting Information File S1: **S1_HumanCoexpNtw_615g_cys.zip**; that can be downloaded and used as a *.cys* file to be explored interactively using *Cytoscape*). This file also includes information about each node with GO and KEGG functional annotations.

Analysis of the network with clustering algorithms

The network described above was analyzed using a graph theoretic clustering algorithm called MCODE [19] as indicated in **Materials and Methods**. The result of this analysis is presented in **Figure 6**, where the circular nodes are the ones with high "cluster coefficient" and the diamond nodes are the ones with lower "cluster coefficient". The intensity of the **red** color of each node indicates the degree of clustering; changing up to pale **grey** for the most peripheral nodes (that only have one link). MCODE found 5 major gene coexpressing clusters marked with numbers in **Figure 6**: **(cluster 1)** corresponds to ribosomal genes, it includes 29 nodes and 366 links and many of the genes are RPL or RPS; **(cluster 2)** corresponds to immunoglobulins and immune response related genes (many belong to families IGH, IGK and IGL) and it includes 19 nodes and 151 interactions; **(cluster 3)** includes 19 nodes and 140 interactions and corresponds to an heterogeneous group of genes strongly clustered with no apparent common functional theme; **(cluster 4)** includes 9 nodes and 36 interactions and corresponds to genes related to metal ion homeostasis (several MT1 and MT2); and **(cluster 5)** corresponds to genes related to the major histocompatibility complex (MHC), it includes 17 nodes split in two clusters with 63 interactions, where most of the genes are HLA. There are other less dense clusters also found by MCODE that have lower score and significance for this algorithm.

We also applied another cluster algorithm for graphs called MCL [20] (see **Methods**). The analysis with MCL provided similar results to MCODE for the large clusters mentioned, although it splits the network in more clusters being the smaller ones more coherent in functional terms that the ones found by MCODE. For example, MCL algorithm finds another cluster form by 15 genes, with 7 assigned to RNA binding gene products, 3 to DNA binding gene products (all included in region **blue** in **Figure 6**), other 3 genes members of the gene family HNRP (heterogeneous nuclear ribonucleoproteins: HNRPA2B1, HNRPR, HNRPU) and 2 genes translation initiation factors (EIF3M, EIF4G2).

These results show that the gene clusters obtained with the graph algorithms from the coexpression network can help to understand the function of many human genes and the active relations between them. As expected, we find that stable and consistent coexpression clusters of genes are involved in specific functions, at cellular or systemic level. A complete analysis of all clusters is not possible in just one article but, as indicated above, the coexpression datasets of this study are open to new studies.

Functional coherence of the coexpressing modules: finding coregulation and new biological assignments

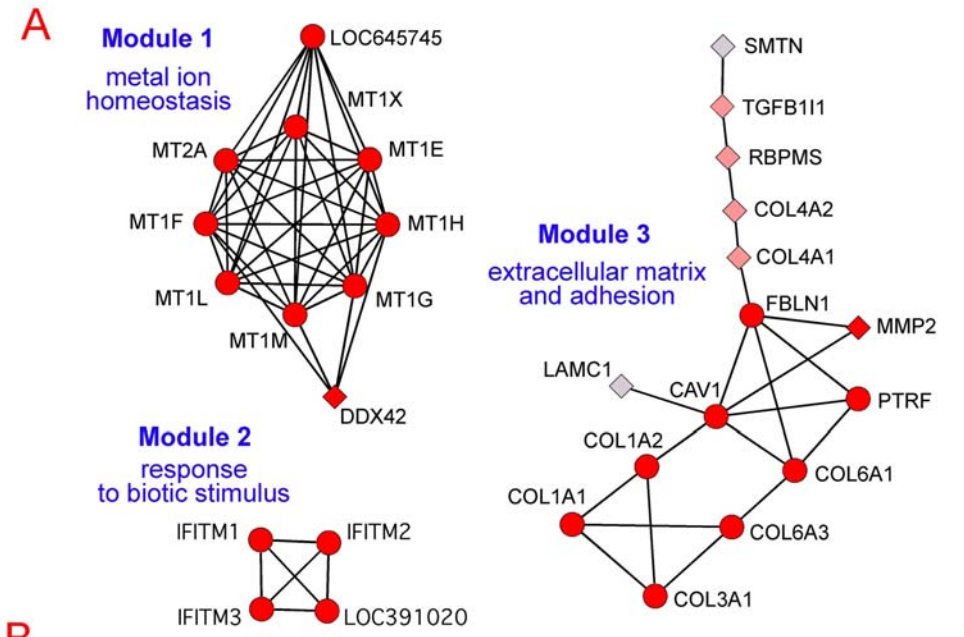
To show some specific examples about the functional coherence of the gene coexpressing modules and the adequate correlation of the genes with common regulatory elements (i.e. transcription factors, TFs, and corresponding promoters) we analyzed three specific clusters or modules found in the core coexpression network.

The first module includes 10 genes: 8 forming a full cross-related octogonal structure plus 2 nodes linked to them. The 8 genes are all metallothioneins: MT1E, MT1F, MT1G, MT1H, MT1L, MT1M, MT1X, MT2A. The other 2 genes are not well annotated: DDX42 (that encodes a member of the DEAD box protein family with unclear function) and LOC645745 (that has been recently and provisionally identified as a putative MT1, metallothionein 1 pseudogene 2). The coexpression of these two genes with a well defined and stable cluster of metallothioneins allows to infer that they will be genes also involved in metal ion homeostasis. This module can be seen in **Figure 7**.

A further analysis was done to find if these coexpressing genes have any common transcription factor (**TF**) that can act on the promoters and regulation regions of these genes. Two bioinformatic tools were used to find out **TFs** associated in a significant way to the coexpressing genes: PAP [21] and FactorY (see **Methods**). Using PAP we found that the 10 coexpressing genes of module 1 are regulated in common by the transcription factor MTF1 (found with p-value = 0.001). This result could be expected since MTF1 is a metal-regulatory transcription factor that induces expression of metallothioneins and other genes involved in metal homeostasis (such as zinc and copper). In any case, the association of MTF1 to module 1 provides strong coherence to the data, showing that this coexpression network is correlated with an underlying transcription regulatory entity.

The second module shown in **Figure 7** includes 4 genes: 3 correspond to interferon-induced transmembrane proteins (IFITM1, IFITM2, IFITM3) and the fourth is an unknown gene LOC391020 recently annotated by inference as similar to interferon-induced transmembrane protein 3. The coexpression of these four genes in a full related cluster gives support to the indication that all produce IFITM proteins. The analysis of transcription factors done with PAP and FactorY (**Figure 7B**) indicated that these 4 genes can be significantly correlated with the transcription factor CRE-BP1 (also called ATF2, activating transcription factor 2), that is a protein which binds to the cAMP-responsive element promoter (CRE, an octameric palindrome) and forms a homodimer or heterodimer with JUN. The deduction that IFITM genes can be coregulated by ATF2 makes biological sense because it has been observed that transcriptional activation of interferon related genes requires assembly of an enhanceosome containing the transcription factors ATF2 and JUN [22,23].

Finally, the third module shown in **Figure 7** includes 15 genes: 6 encode for collagen proteins (COL1A1, COL1A2, COL3A1, COL4A1, COL4A2, COL6A1) that are fibrillar proteins found in most connective tissues, related to the extracellular matrix. Other proteins within this module are also related to cell adhesion and extracellular matrix, like: Fibulin 1 (FBLN1), a secreted glycoprotein that becomes incorporated into the fibrillar extracellular matrix; Laminin gamma 1 (LAMC1), another extracellular matrix glycoprotein which is part of the major noncollagenous constituent of basement membranes; and matrix metalloproteinase 2 (MMP2), that belongs to a family of proteins involved in the breakdown of extracellular matrix in normal physiological processes and in altered disease processes. In fact MMP2 gene encodes an enzyme which degrades type IV collagen. All these data indicate functional consistency and proximity for the genes included in this coexpression module. The analysis, using PAP



B

Coexp Modules	Search in	TF found	p-value	TransFac_db	TF Gene Name
Module 1	*PAP*	MTF-1	**0.001**	T02354	**MTF1** metal-regulatory transcription factor 1
10 genes	*Factory*	–	–		
Module 2	*PAP*	CRE-BP1	**0.0172**	T00167	**ATF2** activating transcription factor 2
4 genes	*Factory*	CRE-BP1	**0.0033**		
Module 3	*PAP*	Sp1	**0.13**	T00759	**SP1** Sp1 transcription factor
15 genes	*Factory*	Sp1	**0.017**		

Figure 7. Coexpressed gene modules regulated by specific transcription factors. (A) Graphical enlarged view of three coexpressing modules selected from the network presented in Figure 6, indicating the name of each gene corresponding to each node and the functional labels: **(Module 1)** metal ion homeostasis; **(Module 2)** response to biotic stimulus; **(Module 3)** extracellular matrix and adhesion. **(B)** Table showing the results of the search for common transcription factors (**TFs**) most significantly associated to the genes included in each of the three modules described above. The search was done using the bioinformatic tools PAP and FactorY.
doi:10.1371/journal.pone.0003911.g007

and FactorY, of the regulatory promoters of this 15 genes shows a significant association with SP1 transcription factor, and recent experimental data have reported that in fact SP1 transcription factor is involved in the regulation of the collagen promoters [24–26].

The results presented for three coexpression modules can be extended to most of the clusters present in the network, and they indicate that the coexpression network can be correlated with an underlying regulatory network driven by specific transcription factors. This observation provides biological and functional coherence to the human gene pairwise coexpression network presented in this paper deduced from the analysis of normal-healthy human samples (whole tissues, glands or organs).

Finally, it is clear that a complete pairwise coexpression network of human genes will be only obtained using a comprehensive and systematic set of samples including all different human cell types. This achievement is at present quite far and difficult, since there are more than two hundred different cell types in the human body and that each cell type can be at different development or differentiation stages. Meanwhile, however, we think that the present study reports a reliable gene-gene coexpression network that includes very valuable information about many human genes, placing them in an integrated transcriptomic context. These coexpression networks selected at specific levels of confidence include a lot of information to better understand the complexity of the human expressing genome.

Materials and Methods

Sample selection: dataset of genome-wide expression microarrays from human normal whole tissues/glands/organs

The data used in this work corresponds to a set of human genome-wide expression microarrays hybridized with mRNA samples coming from different human tissues, glands or organs from healthy normal individuals. The complete list of tissues, glands and organs is: *adrenal gland*, *appendix*, *blood*, *bone marrow*, *brain*, *kidney*, *liver*, *lung*, *lymph node*, *muscle heart*, *ovary*, *pancreas*, *pituitary gland*, *prostate gland*, *salivary gland*, *skin*, *spinal cord*, *testis*, *thymus gland*, *thyroid gland*, *tongue*, *tonsil gland*, *trachea* and *uterus*. These 24 samples where selected from a larger set of 68 human samples (GEO GSE1133; Su et al. 2004) that also included some cell specific sources, like: lung bronchial epithelial cells HBEC, blood B-cells CD19 and T-cells CD4. The samples selection done was driven under the criteria of including mRNA samples from whole organs, glands or tissues covering the main parts of the human body and avoiding samples of very specific cell types within a tissue. This selection was validated performing global expression analyses of the samples, using a series of algorithms described bellow. The total mRNA from these 24 different samples came form a mix of 3 different individuals, that were: two men and one woman or one man and two women for the samples non sex-associated; three men for *testis* and *prostate* samples and three women for *ovary* and *uterus* samples. Moreover two biological replicates were used in each case, producing

a total set of 48 microarrays. The microarrays used were high density oligonucleotide microarrays HGU133A GeneChips from *Affymetrix*, that include 22,283 probesets (corresponding to 13,068 human genes according to *Affymetrix* annotation).

Genome-wide sample expression profiles and gene expression signatures

The global expression matrix including the genome-wide expression profiles of each sample and the expression signature of each gene-probeset was calculated and evaluated using a set of algorithms and methods in four consecutive steps: **(1st)** use of two different background correction, normalization and signal calculation methods: MAS5 [8,27] and RMA [28]; **(2nd)** use of two distance measuring methods based in the global gene expression profile of each sample: first, distance based on Spearman correlation coefficient applied to MAS5 data; second, distance based on Pearson correlation coefficient applied to RMA data (both methods provided robust non-parametric distance distributions); **(3rd)** analysis by hierarchical clustering with complete linkage of the samples using the tool *hclust* from **R** (http://www.r-project.org/), taking as distance (**1**−**r**), where **r** is the correlation coefficient between sample expression profiles [29]; **(4th)** analysis by bootstrapping of the sample hierarchical trees to assay the stability of the associations, using the tool *pvclust* from **R**. The *pvclust* algorithm allows to assess the uncertainty in hierarchical cluster analysis via multiscale bootstrap resampling. This assessment is provided by two parameters: the *approximately unbiased p-value* (AU) and the *bootstrap probability value* (BP). The maximum and optimum values of AU and BP are 1 (or 100 in %).

Gene pairs coexpression and cross-validation

As indicated above the global gene to gene (i.e. pair-wise) coexpression matrix was calculated using two different and independent methods: MAS5-Spearman and RMA-Pearson. Furtherly, cross-validation was used to discriminate stable and significant correlations. The cross-validation strategy applied was a 1000 times random selection of a 25% subset sampling (that are 12 samples, corresponding to 6 duplicates out of 24 duplicated samples) and calculation of the **r** correlation coefficient for each gene-probeset pair in such 1000 samplings. Only when the **r** correlation coefficient for a given time was higher than |0.70|, such was considered a positive event (positive cross-validation) and counted for the corresponding gene-probeset pair. In this way, for example, a given gene pair with **N** = 620 means that it gave 620 positive times out of the 1000 samplings. Therefore **N** can be considered a cross-validation coefficient or cross-validation factor (**N** = 620 is equivalent to 620/1000 = 0.62).

Gene filtering method

In order to get rid of genes with low information content a combined filter based on between-sample variability and gene minimal signal was used. The filter leaves out only those gene probesets that fulfilled both of the two following conditions: **1st**.- Genes which have an expression difference or variability between samples ($\Delta Exp^{gi}_{highest\text{-}lowest}$) lower than the median of all the expression differences calculated for each gene ($\Delta Exp^{gi}_{highest\text{-}lowest} < median\ \Delta Exp_{highest\text{-}lowest}$); **2nd**.- Genes which have a mean expression signal between samples ($meanExp_{samples}$) lower than the median of all the expression signals calculated for each gene.

Statistical estimation of accuracy and coverage of the coexpression datasets

The ***accuracy*** measured as "**Positive Predictive Value**" (**PPV**) in statistical terms is defined as the ratio TP/(TP+FP), where TP is the number of true positives and FP is the number of false positives [30,31]. This parameter is related to "error type I", and it is the inverse to the ratio of "false positives" (i.e. FP/(TP+FP), percentage of false positives within all the positives). The ***coverage*** (sometimes also named recall) can be measured as the proportion of true positives that remain in a given subset selected, with respect to an initial reference set of positives. We consider that both the accuracy and coverage are critical statistical parameters to evaluate the error and validity of a method. They are directly related to ***specificity*** = TN/(TN+FP), −where (TN+FP) are all the "false"−, and ***sensitivity*** = TP/(TP+FN) −where (TP+FN) are all the "true"− [30], though these can only be applied when the real true and real false data of a test are known; while the accuracy defined as "positive predictive value" and the defined coverage can be applied when it is only possible to know or estimate the "positive data".

Therefore, in this study if the true data are not known (i.e. if we do not know *a priori* which are true gene coexpressing pairs) a proper calculation of the sensitivity and specificity is not possible. This is the most common situation in many biological and biomolecular studies where many of the true occurring relations between molecules are not yet known. Therefore, we need to design a way to at least estimate the percentage or ratio of "true positives" of the method, and so estimate the accuracy and coverage. These parameters will provide a good indication of how valuable is the method that we have applied to find human coexpressing gene pairs. The estimation was done considering the idea that genes that work together in the same biological pathway are much more likely to coexpress than genes that are not involved in a common biological reaction or pathway. This biomolecular axioma in our case was tested annotating all the genes of the microarrays to the KEGG pathway database (www.genome.jp/kegg/), that is one of the most complete and expert curated repository of human genes involved in biological reactions or pathways [32]. Therefore, selecting only the subset of the genes annotated to KEGGs, a gene coexpression pair was considered a "true positive" when both genes of the pair were included in a common KEGG human pathway. This strategy allows to calculate the statistical parameters ***accuracy*** and ***coverage*** defined above, and therefore to explore how the values of the **r** and **N** coefficients change such parameters.

Analytic algorithms to find groups and modules in the coexpression networks

The gene to gene coexpression networks obtained were analyzed using a graph theoretic clustering algorithm called MCODE (Molecular Complex Detection) [19] that allows to detect densely connected regions in large interaction networks which may represent molecular associations. This algorithm follows a vertex weighting by local neighbourhood density and outward traversal from locally dense seed nodes to isolate the dense regions. Furthermore, the networks were also analyzed using another cluster algorithm for graphs called MCL (Markov Cluster algorithm, http://micans.org/mcl/) [20] that finds cluster structure in graphs by a mathematical bootstrapping procedure. MCL has been shown very robust to find relevant modules in protein interaction networks [33].

Mapping transcription factors associated to gene coexpressing modules

Two bioinformatic tools were used to find out transcription factors that can be associated in a significant way to groups or modules of coexpressing genes: Promoter Analysis Pipeline (PAP) and Transcription Factor Enrichment Analysis (FactorY).

PAP is based in a systematic, statistical model of mammalian transcriptional regulatory sequence analysis and it is suitable for the identification of the potential transcriptional regulators of co-expressed genes and the identification of the potential regulatory targets of transcription factors. A typical PAP analysis includes input of a co-expressed gene cluster, identification of several high scoring transcription factors and visualization of the predicted transcription factor binding sites [21]. The bioinformatic tool is at: http://bioinformatics.wustl.edu/webTools/portalModule/PromoterSearch.do.

FactorY is another bioinformatic tool that explores the 1000 bp upstream sequence signature of co-expressed genes to find homology with transcription factor binding sites (TFBs) based on JASPAR and TRANSFAC databases. The tool calculates the significant enrichment in known given TFBs for a group of genes and it was used at the web site: http://www.garban.org/factory/.

Supporting Information

File S1 Human Gene Coexpression Network. Network that corresponds to the core with the most confident human gene pairwise coexpression data and includes 615 gene-nodes and 2190 coexpression-links. This network is provided in Cytoscape format (.cys file compressed as .zip) with full annotations about the genes. The file to be run in Cytoscape should have .cys extension: S1_HumanCoexpNtw_615g.cys
Found at: doi:10.1371/journal.pone.0003911.s001 (0.30 MB ZIP)

Acknowledgments

We thank the support provided by the *Instituto de Salud Carlos Tercero*, Ministery of Health, Spanish Government (ISCiii-FIS, MSyC) and by the *Consejería de Educación*, Castilla-Leon Local Government (JCyL).

Author Contributions

Conceived and designed the experiments: CP JDLR. Performed the experiments: CP AR CF. Analyzed the data: CP AR CF. Contributed reagents/materials/analysis tools: CP AR CF. Wrote the paper: JDLR.

References

1. van Noort V, Snel B, Huynen MA (2004) The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Rep 5: 280–284.
2. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res 14: 1085–1094.
3. Tirosh I, Weinberger A, Carmi M, Barkai N (2006) A genetic signature of interspecies variations in gene expression. Nat Genet 38: 830–834.
4. Magwene PM, Kim J (2004) Estimating genomic coexpression networks using first-order conditional independence. Genome Biol 5: R100.
5. Stuart JM, Segal E, Koller D, Kim SK (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science 302: 249–255.
6. Griffith OL, Pleasance ED, Fulton DL, Oveisi M, Ester M, et al. (2005) Assessment and integration of publicly available SAGE, cDNA microarray, and oligonucleotide microarray expression data for global coexpression analyses. Genomics 86: 476–488.
7. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185–193.
8. Lim WK, Wang K, Lefebvre C, Califano A (2007) Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics 23: i282–288.
9. Suzuki R, Shimodaira H (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22: 1540–1542.
10. Wang Y, Miao ZH, Pommier Y, Kawasaki ES, Player A (2007) Characterization of mismatch and high-signal intensity probes associated with Affymetrix genechips. Bioinformatics 23: 2088–2095.
11. Barnes M, Freudenberg J, Thompson S, Aronow B, Pavlidis P (2005) Experimental comparison and cross-validation of the Affymetrix and Illumina gene expression analysis platforms. Nucleic Acids Res 33: 5914–5923.
12. Dallas PB, Gottardo NG, Firth MJ, Beesley AH, Hoffmann K, et al. (2005) Gene expression levels assessed by oligonucleotide microarray analysis and quantitative real-time RT-PCR – how well do they correlate? BMC Genomics 6: 59.
13. Choi JK, Yu U, Yoo OJ, Kim S (2005) Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics 21: 4348–4355.
14. Prieto C, Rivas MJ, Sanchez JM, Lopez-Fidalgo J, De Las Rivas J (2006) Algorithm to find gene expression profiles of deregulation and identify families of disease-altered genes. Bioinformatics 22: 1103–1110.
15. Calza S, Raffelsberger W, Ploner A, Sahel J, Leveillard T, et al. (2007) Filtering genes to improve sensitivity in oligonucleotide microarray data analysis. Nucleic Acids Res 35: e102.
16. Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, et al. (2001) A compendium of gene expression in normal human tissues. Physiol Genomics 7: 97–104.
17. Eisenberg E, Levanon EY (2003) Human housekeeping genes are compact. Trends Genet 19: 362–365.
18. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498–2504.
19. Bader GD, Hogue CW (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4: 2.
20. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 30: 1575–1584.
21. Chang LW, Fontaine BR, Stormo GD, Nagarajan R (2007) PAP: a comprehensive workbench for mammalian transcriptional regulatory sequence analysis. Nucleic Acids Res 35: W238–244.
22. Falvo JV, Parekh BS, Lin CH, Fraenkel E, Maniatis T (2000) Assembly of a functional beta interferon enhanceosome is dependent on ATF-2-c-jun heterodimer orientation. Mol Cell Biol 20: 4814–4825.
23. Panne D, Maniatis T, Harrison SC (2004) Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-beta enhancer. Embo J 23: 4384–4393.
24. Kypriotou M, Beauchef G, Chadjichristos C, Widom R, Renard E, et al. (2007) Human collagen Krox up-regulates type I collagen expression in normal and scleroderma fibroblasts through interaction with Sp1 and Sp3 transcription factors. J Biol Chem 282: 32000–32014.
25. Magee C, Nurminskaya M, Faverman L, Galera P, Linsenmayer TF (2005) SP3/SP1 transcription activity regulates specific expression of collagen type X in hypertrophic chondrocytes. J Biol Chem 280: 25331–25338.
26. Poree B, Kypriotou M, Chadjichristos C, Beauchef G, Renard E, et al. (2008) Interleukin-6 (IL-6) and/or Soluble IL-6 Receptor Down-regulation of Human Type II Collagen Gene Expression in Articular Chondrocytes Requires a Decrease of Sp1{middle dot}Sp3 Ratio and of the Binding Activity of Both Factors to the COL2A1 Promoter. J Biol Chem 283: 4850–4865.
27. Liu WM, Mei R, Di X, Ryder TB, Hubbell E, et al. (2002) Analysis of high density expression microarrays with signed-rank call algorithms. Bioinformatics 18: 1593–1599.
28. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31: e15.
29. Murtagh F (1985) Multidimensional Clustering Algorithms. COMPSTAT Lectures. Wuerzburg: Physica-Verlag.
30. Loong TW (2003) Understanding sensitivity and specificity with the right side of the brain. Bmj 327: 716–719.
31. Suojanen JN (1999) False false positive rates. N Engl J Med 341: 131.
32. Aoki-Kinoshita KF, Kanehisa M (2007) Gene annotation and pathway mapping in KEGG. Methods Mol Biol 396: 71–91.
33. Brohee S, van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics 7: 488.

Annals of Oncology Advance Access published January 6, 2012

original article

Annals of Oncology
doi:10.1093/annonc/mdr579

Identification of a novel recurrent gain on 20q13 in chronic lymphocytic leukemia by array CGH and gene expression profiling

A. E. Rodríguez[1], C. Robledo[1], J. L. García[2], M. González[3], N. C. Gutiérrez[3], J. A. Hernández[4], V. Sandoval[5], A. García de Coca[6], I. Recio[7], A. Risueño[8], G. Martín-Núñez[9], E. García[10], R. Fisac[11], J. Conde[12], J. de las Rivas[8] & J. M. Hernández[1,3]*

[1]IBMCC, Centro de Investigación del Cáncer, Universidad de Salamanca-CSIC, Salamanca; [2]Instituto de Estudios de Ciencias de la Salud de Castilla y León (IECSCYL)–HUSAL, Castill y León; [3]Department of Hematology, Hospital Clínico Universitario de Salamanca, Salamanca; [4]Department of Hematology, Hospital Infanta Leonor, Madrid; [5]Department of Hematology, Hospital Virgen Blanca, León; [6]Department of Hematology, Hospital Clínico Universitario, Valladolid; [7]Department of Hematology, Hospital Nuestra Señora de Sonsoles, Ávila; [8]Bioinformatics and Functional Genomics, Centro de Investigación del Cáncer, Universidad de Salamanca-CSIC, Salamanca; [9]Department of Hematology, Hospital Virgen del Puerto, Plasencia; [10]Genomics and Proteomics Unit, Centro de Investigación del Cáncer, Universidad de Salamanca-CSIC, Salamanca; [11]Department of Hematology, Hospital General de Segovia, Segovia; [12]Department of Hematology, Hospital del Río Hortega, Valladolid, Spain

Received 3 August 2011; revised 11 November 2011; accepted 16 November 2011

Background: The presence of genetic changes is a hallmark of chronic lymphocytic leukemia (CLL). The most common cytogenetic abnormalities with independent prognostic significance in CLL are 13q14, *ATM* and *TP53* deletions and trisomy 12. However, CLL displays a great genetic and biological heterogeneity. The aim of this study was to analyze the genomic imbalances in CLL cytogenetic subsets from both genomic and gene expression perspectives to identify new recurrent alterations.

Patients and methods: The genomic imbalances and expression levels of 67 patients were analyzed. The novel recurrent abnormalities detected with bacterial artificial chromosome array were confirmed by FISH and oligonucleotide microarrays. In all cases, gene expression profiling was assessed.

Results: Copy number alterations were identified in 75% of cases. Overall, the results confirmed FISH studies for the regions frequently involved in CLL and also defined a new recurrent gain on chromosome 20q13.12, in 19% (13/67) of the CLL patients. Oligonucleotide expression correlated with the regions of loss or gain of genomic material, suggesting that the changes in gene expression are related to alterations in copy number.

Conclusion: Our study demonstrates the presence of a recurrent gain in 20q13.12 associated with overexpression of the genes located in this region, in CLL cytogenetic subgroups.

Key words: CLL, cytogenetic aberrations, gene expression profile, genomic arrays

introduction

Chronic lymphocytic leukemia (CLL) is the most common leukemia in the western world and is characterized by a highly variable clinical course with survival times ranging from months to decades despite a remarkable phenotypic homogeneity [1, 2]. This clinical heterogeneity reflects its biological diversity [3]. Our understanding of the biology of CLL has helped to identify several markers of prognostic significance, delineating CLL into several distinct diseases. These markers include cytogenetic abnormalities, the mutational status of the immunoglobulin heavy chain variable (*IGHV*) and ZAP-70, CD38 and CD49d expression [2, 4–6]. Conventional cytogenetic analyses have revealed chromosomal aberrations in 40%–50% of patients, but detection of abnormalities is limited by the low mitotic activity of CLL cells. By contrast, interphase FISH (iFISH) has identified chromosomal changes in ~80% of patients with CLL, the presence of specific chromosomal abnormalities being a prognostic indicator of disease progression and survival [2, 7]. Thus, half of the CLL patients carry deletions of 13q, which is correlated with an indolent disease course in patients with this abnormality as their sole aberration. In contrast, deletions of 11q and 17p (which cover the *ATM* and *TP53* genes, respectively) have a poorer outcome. Furthermore, trisomy 12 is related to an intermediate prognosis, whereas deletion of 6q has been identified as a recurrent CLL progression marker [8]. In addition, great genomic complexity has been associated with worse survival and is also closely related to markers of poor prognosis [9–11].

*Correspondence to: Prof. J. M. Hernández, Hematology Unit, Department of Medicine, Hospital Universitario de Salamanca, Paseo San Vicente 58, 37007 Salamanca, Spain. Tel: +34-923-291-100; Fax: +34-923-294-624; E-mail: jmhr@usal.es

Considering the great heterogeneity of CLL from both genetic and prognostic points of view, microarray technology is a powerful tool for the analysis of genetic alterations in CLL. Thus, comparative genomic hybridization using high-density arrays, array comparative genomic hybridization (aCGH), allows high-resolution genome-wide scan for detection of copy number alterations in a single hybridization and aCGH using bacterial artificial chromosome (BAC) clones has been widely applied in the analysis of hematological malignancies [12–15]. Regarding oligonucleotide microarrays, the study of the gene expression profile (GEP) in CLL has given us insights into the molecular mechanisms involved in its pathogenesis by analyzing the impact of genomic aberrations on the expression of genes located on the corresponding loci [16–18].

Although the application of microarray technology in CLL has provided additional knowledge of the known recurrent aberrations as well as enabling novel aberrations, such as gain of 2p and deletion of 22q to be identified [19–23], to date, few studies have investigated genomic aberrations specifically in relation to CLL cytogenetic subsets. Therefore, the aim of this study was not only to screen and identify new genomic events in CLL patients but also to compare the prevalence of these genomic aberrations in cytogenetic CLL subsets. Furthermore, our data revealed an association between altered transcription levels and genomic imbalances in the genetic subsets of CLL, indicating that gene dosage might have pathogenic effects in CLL and delineate a new gained region, on 20q13, in CLL patients.

methods

patients

Peripheral blood samples from 67 patients with CLL were analyzed. The diagnoses were confirmed by standardized clinical, morphological and immunological data according to the World Health Organization classification and the criteria of the Working Group of the National Cancer Institute [24]. FISH studies and *IGH* mutational status were determined in all patients. The study protocol was approved by the local ethical committees and prior written informed consent was obtained from the patients. All patients were untreated and most of them were studied at the moment of diagnosis (Table 1). The main characteristics of the 67 CLL patients included in the study are reported in Table 2.

FISH analysis

Interphase FISH was carried on all the samples using commercially available probes for the following regions: 13q14, 12q13, 11q22/*ATM*, 17p13/*TP53* and 14q32/*IGH* (Abbott Co., Downers Grove, IL) using the previously described methods [25].

To confirm the gains and losses assessed by aCGH, FISH analysis was done using Vysis LSI ZNF217, the commercially available probe for 20q13.2 (Abbott Co.) and the BAC clones dJ1028D15–dJ781B1, mapping to 20q13.12, as previously described [25]. The clones were located in the same region of gain as detected by aCGH and were selected from the aCGH BAC clone library (Wellcome Trust Sanger Institute), whereas the commercial probe was located in 20q13.2 (breast tumor amplicon). DNA from the BAC clones was isolated, labeled and hybridized, as previously described [26]. The changes were validated in fixed cells from the same diagnostic samples as used for aCGH ($n = 20$).

FISH analysis was carried out on 400 interphase cells using standard fluorescent microscopy.

mutation status of IGVH genes

IGVH genes were amplified and sequenced according to the ERIC recommendations on *IGHV* gene mutational status analysis in CLL [27].

array comparative genomic hybridization

BAC arrays. DNA samples were analyzed using a BAC array containing 3523 sequence-validated BACs covering the genome with a mean resolution of 1 Mb, as previously described [26].

Table 1. Status of disease in the total series ($n = 67$) and in +20q CLL patients ($n = 13$)

	CLL patients ($n = 67$) n (%)	+20q CLL patients ($n = 13$) n/CLL (%)
At diagnosis	50 (75)	13/50 (26)
Progressive	17 (25)	4/17 (24)

CLL, chronic lymphocytic leukemia.

Table 2. Clinical and molecular characteristics of the CLL patients

Characteristics	(%)
Median age in years (range)	68 (35–90)
Male/Female (ratio)	73/23 (2.7)
White blood cells, range/ml	39 000 (7600–175 000)
Lymphocytes/ml (range)	32 000 (5000–160 000)
Hemoglobin, g/dl (range)	13.6 (7.1–16.3)
Platelet count/ml (range)	167 000 (59 000–306 000)
LDH	
Normal	82
High	12
β_2-microglobulin	
Normal	52
High	48
Status of the disease	
At diagnosis	71.6
Progressive	28.4
Binet stage	
A	66
B	26
C	9
ZAP-70 expression	
Positive	44
Negative	56
CD38 expression	
Positive	26
Negative	74
IgVH mutational status	
Mutated	41
Unmutated	59
Interphase FISH analysis	
Normal karyotype	22
13q deletion	37
Trisomy 12	16
11q deletion	7
17p deletion	9
IGH translocation	9
20q13.12 gain	19

CLL, chronic lymphocytic leukemia; LDH, lactate dehydrogenase.

oligonucleotide microarrays. In order to confirm the results of the BAC aCGH analysis, a subset of 35 patients were analyzed using a NimbleGen Human CGH 4×72K Whole Genome v2.0 array (Roche Diagnostics, Mannheim, Germany).

The complete description of BAC and oligonucleotide microarrays experiments is available as supplementary Material (available at *Annals of Oncology* online).

GEP analysis

RNA isolation, labeling and microarray hybridization were carried out, as previously reported [28]. The GEP was analyzed in all cases with Human Genome U133A microarray (Affymetrix, Santa Clara, CA). Data analysis is available as supplementary material (available at *Annals of Oncology* online).

comparative analysis of CGH and expression arrays

In order to achieve a comparative analysis of the copy number changes, from the CGH arrays, and the gene changes, from the expression arrays, for the same patients, we select the patients who showed significant gains in the aCGH data and their corresponding expression data. We normalized the expression dataset using the R package GeneMapper [29] that allows an accurate assignment to ENSEMBL genes (instead to Affymetrix probesets) including their location in the genome. Following this, we selected the three regions in chromosome 20q where the gains detected by aCGH were significative. For such regions, we calculate, in the corresponding samples, the mean and median expression signal based on the genes included. On these expression numbers, we carried out a statistical one-tail *t*-test (using R) to check if there was a significant correlation between the aCGH gain observed in the 20q regions and the overexpression of the genes included in such regions.

statistical analysis

Two-tailed Chi-square and Fisher's exact tests were used to analyze the associations between variables. For all tests, values of $P < 0.05$ were considered to indicate statistical significance. The calculations were carried out using SPSS 17.0 for Windows (SPSS Inc., Chicago, IL).

results

FISH and mutational status

FISH analyses revealed that 25 of the 67 cases analyzed (37%) carried the 13q14 deletion and that this was the only abnormality in 20 patients (30%). Overall, the 11q22.3 and 17p13.1 deletions were present in 5 (7%) and 6 (9%) patients, respectively, while trisomy 12 was present in 11 (16%) and t(14q32) in 9 (13%) patients. The remaining 21 (31%) patients did not show aberrations by FISH. To better characterize the 20q13.12 gain, FISH analysis was also carried out in a validation series of 58 patients: 17% patients showed this alteration in ≥4.5% cells (ranging from 4.5% to 12%). In relation to mutational status, 49% of cases had unmutated *IgVH* gene.

aCGH showed recurrent genomic imbalances in CLL

Fifty of 67 patients (75%) displayed genomic changes with aCGH. In addition to the regions detected by FISH abnormalities, aCGH enabled the presence of novel recurrent genomic imbalances to be demonstrated. In order to rule out previously described single nucleotide polymorphisms, the minimal regions of overlap for all the recurrent lesions were compared with the frequencies of known copy number variations. A total of 443 altered chromosomal regions were found, of which 237 (53%) were deletions. The median number of changes per patient was five (range 0–14). The most commonly recurring alterations (observed in >5% of cases), their boundaries and frequencies are shown in Table 3. Losses in 13q14.2–q14.3 (21%), 11q13.3 (16%), 17p13.2–p13.1 (10%), 11q22.3–q23.1 (9%) and 5q13.3–q14.1, 5q31.1 and 7q22 (6% each) as well as gains in 1q21.3–q22 (22%), 11q13.3 (21%), 16q23.2–q24.2 (21%), 6p21.31–p21.1 (19%) and 10q22.3 (7.5%)

Table 3. Recurrent aberrations identified by aCGH in CLL

Chromosome	Cytoband	Start position	End position	Size	Frequency of gains	Frequency of losses	Number of genes covered by the aCGH
1	q21.3–q22	151208615	155645185	4.44	22		31
1	q31.1–q31.2	189804292	192389810	2.59	13		1
5	q13.3–q14.1	74766678	77235911	2.47		6	18
5	q31.1	131568885	132631297	1.06		6	21
6	p21.31–p21.1	33722842	43897461	10.17	19		146
7	q22.1	98853924	101547956	2.69		6	78
10	q22.3	80565766	81502183	0.94	7.5		2
11	q13.1	64268965	65470124	1.2	22		51
11	q13.3–q13.4	69974549	73001497	3.03	21	16	37
11	q22.3–q23.1	108518932	112964950	3.73		9	24
12	p13.33–q24.33	1	133851895	133.9	15		
13	q14.2–q14.3	48863579	54941189	6.08		21	27
16	q23.2–q24.2	80972437	88692209	7.72	21		56
17	p13.2–p13.1	4571828	7483888	2.91		10	104
17	q25.3	75993754	80470659	4.48	19		105
18	q21.2	49844734	50270501	0.43	9	7.5	5
20	q13.12	42188467	44495323	2.31	19		52

Positions and sizes are expressed in base pairs. Bp locations according to GRCh37, February 2009 (hg 19).
aCGH, array comparative genomic hybridization; CLL, chronic lymphocytic leukemia.

were the most frequent changes revealed by aCGH (Table 3). Interestingly, a critical segment of gain was delineated on chromosome 20q in 13 patients (19%). The analysis identified a minimal region of gain on 20q13.12 of ∼2.31 Mb involving three clones at linear positions (42 188 467–44 495 323), as shown in Figure 1. Most of these cases (75%) were studied at the time of diagnosis (Table 1). Changes detected in diagnostic and progression groups are shown in Table 4.

genomic abnormalities in the cytogenetic subgroups of CLL

Overall, correlation between FISH and aCGH was observed for +12, 11q- or 17p- cases (100%, 91% and 83%, respectively). However, this was not the case for 13q- subgroup. Interestingly, most of these cases displayed <30% of 13q deletions and in the 14 cases with a deletion of 13q revealed by aCGH, losses were located in 13q14.2–q21.1, with heterogeneous breakpoints. It should be noted that the 21 CLL samples showing no aberrations with FISH also appeared to be normal for the CLL FISH regions when analyzed by aCGH. However, novel recurrent alterations by aCGH were detected in this group of CLL patients (Table 5).

The analysis of the relationship between the recurrent abnormalities revealed by FISH and the presence of novel chromosomal imbalances detected by aCGH showed a significant association between the loss of 13q and the loss of 5q13.3–q14.1 and 5q31.1 ($P < 0.05$). No other additional abnormalities were observed in any of the cytogenetic subgroups.

In order to assess the genomic complexity in the cytogenetic CLL subsets, a comparison between the number of genetic changes ascertained by aCGH and the FISH CLL subgroup was carried out. As the median of changes per patient was five (range 0–14), we defined two groups to analyze the number of changes with respect to the FISH categories: ≤5 (low genomic complexity) and >5 (high genomic complexity) (Table 6). Interestingly, an association between the presence of a large number of changes detected by aCGH and ATM deletion ($P = 0.026$) by FISH was observed. The presence of gains on 20q12.13 ($P = 0.002$) was also associated with a high frequency of changes as revealed by aCGH (Table 6).

oligonucleotide and FISH studies validated the changes observed by aCGH

Oligonucleotide aCGH was carried out in 35 cases to confirm the BAC array results. Genomic patterns of gains and losses representative of the probe sizes (∼150 kb) were compared with those obtained by BAC aCGH and found to be 100% concordant.

FISH experiments were carried out on 20 patients to confirm the gains on 20q13.12 observed with aCGH (supplemental Figure S1, available at *Annals of Oncology* online). All but one of the cases (95%) was concordant with the aCGH results. The median of cells showing this aberration was 20% (range 16%–25%). In addition, the cases were analyzed with a probe covering 51 992 266–52 410 801 bp (Vysis LSI ZNF217, breast tumor amplicon at 20q13.2). The results failed to show any

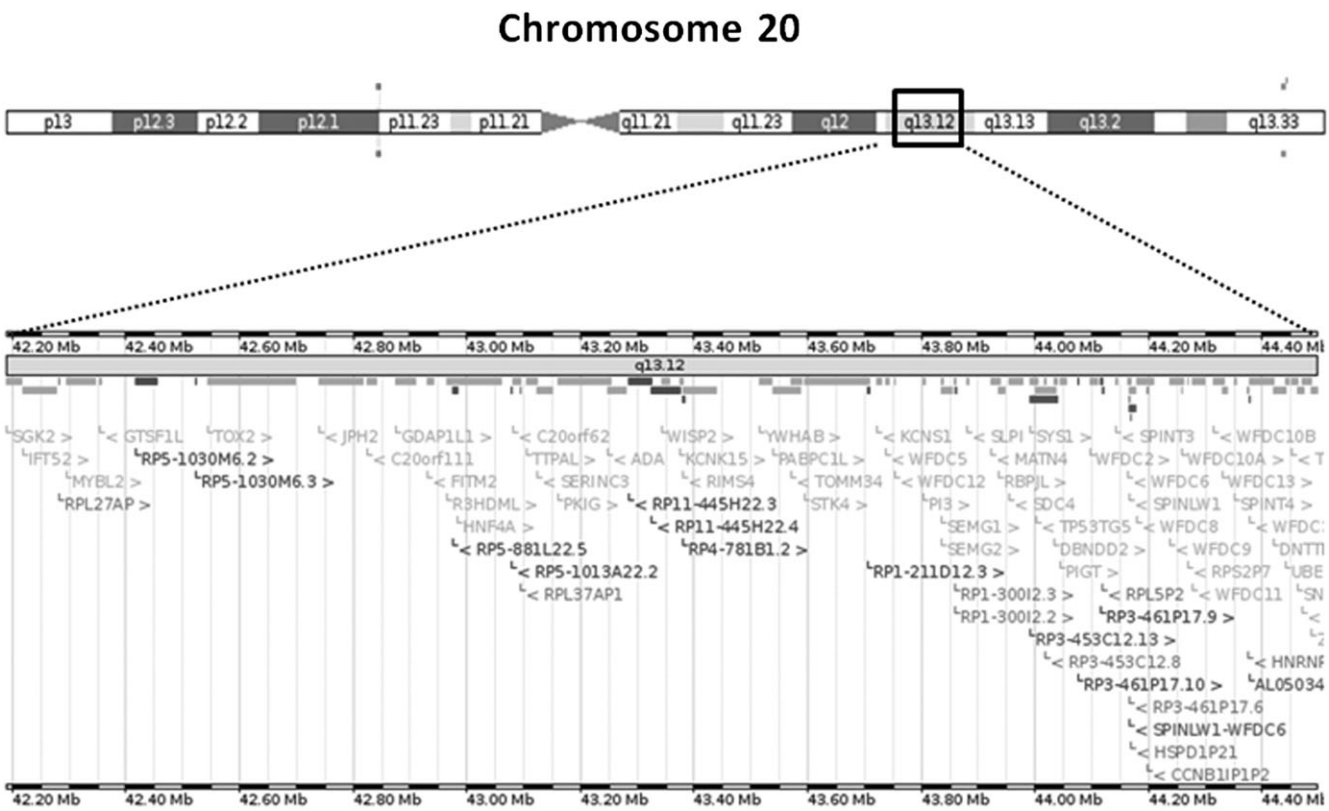


Figure 1. Integration of annotated genomic sequence with array comparative genomic hybridization data: common region of gain (CRG) on 20q (42188467–44495323 bp) showing the candidate genes (GRCh37, February 2009, hg 19).

Table 4. Characteristics of the CLL series (IgVH mutational status, number of aberrations, FISH subgroup and frequency of recurrent alterations detected by aCGH) according to the status of disease (at diagnosis versus progression)

Characteristics	Status of disease	
	At diagnosis (%)	Progression (%)
IgVH mutational status		
Mutated	63.6	46.7
Unmutated	36.4	53.3
Number of aberrations		
≤5	64	52.9
>5	36	47.1
FISH subgroup		
Normal FISH	38[a]	11.8
13q deletion	40	29.4
Trisomy 12	8	41.2[a]
17p deletion	6	17.6
11q deletion	6	11.8
t(14q32)	16	5.9
Recurrent alteration by aCGH		
Gains		
1q21.3–q22	24	17.6
1q31.2	8	29.4
6p21.31–p21.1	18	23.5
10q22.3	6	11.8
11q13.1	24	17.6
11q13.3	20	23.5
12	6	41.2[a]
16q23.2–q24.2	22	17.6
17q25.3	20	17.6
18q21.2	6	17.6
20q13.12	20	17.6
Losses		
5q13.3–q14.1	8	0
5q31.1	8	0
7q22	4	11.8
11q13.3	16	17.6
11q22.3–q23.1	8	11.8
13q14.2–q14.3	24	11.8
17p13.2–p13.1	10	11.8
18q21.2	8	5.9

[a]Statistically significant associations ($P < 0.05$).
aCGH, array comparative genomic hybridization; CLL, chronic lymphocytic leukemia.

involvement of this region, delineating the commonly gained region at 20q between 42 188 467–44 495 323 bp (2.31 Mb) (Figure 1).

gene expression profile confirmed the dosage effect of aCGH changes

In order to assess the relevance of the genomic imbalances in gene expression, a gene expression profile study was carried out. For this purpose, we grouped the cases by aCGH findings. The group of patients displaying trisomy 12 showed deregulation of 89 genes when compared with the rest of patients. A total of 76 of the 89 genes were overexpressed in relation to the other patients and 56% of them were located on chromosome 12.

It should be noted that overexpression of the 52 genes located in 20q13.12 (Figure 1), the 20q region gained by aCGH, was also observed ($P = 0.01$). Among these genes, we found well-known protein-coding cancer-related genes (supplemental Table S1, available at *Annals of Oncology* online) such as *PI3* (elafin), *SLPI* (secretory leukocyte peptidase inhibitor) and *WFDC2* [whey acidic protein (WAP) four-disulfide core domain 2], members of the WAP family; *PIGT* (phosphatidylinosotol glycan anchor biosynthesis, class T), a component of the glycosylphosphatidylinositol (GPI) glycan transamidase complex; *HNF4A* (hepatocyte nuclear factor 4, alpha) and *YWHAB* (tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, beta polypeptide), members of the SMAD and Ras signal transduction pathways, respectively. In addition, *ADA* (adenosine deaminase), a regulator of B-cell proliferation, overexpression was also present in CLL cases with gains on 20q.

Moreover, in patients with the 17p13 deletion, a significant proportion (83%) of the differentially underexpressed genes clustered in this region ($P < 0.05$). Among the downregulated genes were *GPS2* (G protein pathway suppressor 2)/*AMF1*, *SGSM2* (small G protein signaling modulator 2), *DRG2* (developmentally regulated GTP-binding protein 2), *SAT2* (spermidine/spermine N1-acetyltransferase family member 2) and *C17orf49* (chromosome 17 open reading frame 49). This gene dosage effect was also observed in CLL patients showing 11q-. Thus, all the genes located in the minimal region of deletion observed with aCGH on 11q22.3–q23.2 (108518932–112964950 bp) were downregulated when compared with the rest of patients ($P < 0.01$).

discussion

The presence of cytogenetic abnormalities is a hallmark of CLL. Indeed, these abnormalities have been associated with the prognosis or progression of the disease and for this reason the genetic changes have been extensively studied in CLL [30, 31]. The present study integrates genomic and gene expression profile analyses in a cohort of 67 CLL patients. Overall, the results enable us to detect hitherto undescribed recurrent alterations in CLL, such as gains on chromosome 20 and confirm the dosage effect, not only for the common cytogenetic abnormalities but also for this new genetic abnormality.

The present study found genomic copy number changes in 75% of the CLL patients. Our findings are similar to those previously reported in this disease [19, 22, 23, 32]. Detection rates of genomic alterations involving loci known to be associated with CLL occurred at expected frequencies [33] and overall, correlation between FISH and aCGH was observed except in the 13q- subgroup. Both FISH and aCGH revealed that 13q- was an heterogeneous group in size of the deletion and percentage of cells displaying the abnormality. Interestingly, when aCGH failed to demonstrate the presence of 13q deletion, FISH data revealed that most of these cases had <30%. This could justify, at least in part, the lack of correlation between both the techniques. We also confirmed that deletions are more abundant than gains in CLL: deletions in

Table 5. Correlation between the most frequent chromosomal imbalances identified by aCGH and the CLL cytogenetic subgroups

Aberration/%	Cytogenetic/FISH subgroup						Total (%)
	13q14.3	Trisomy 12	11q22.3	17p13.1	t(14q32)	Normal FISH	
1q21.3–q22							
Gain	28	27	20		11	24	22
1q31.1–q31.2							
Gain	8	27	40	17	22		13
5q13.3–q14.1							
Loss	16[a]			17	11		6
5q31.1							
Loss	16[a]			17	11		6
6p21.31–p21.1							
Gain	24	27	20		22	14	19
7q22.1							
Loss	12	9	20		11		6
10q22.3							
Gain	16	9	20		11		7.5
11q13.1							
Gain	28	18	20	33	22	19	22
11q13.3–q13.4							
Loss	12	18	60[a]		22	14	16
Gain	28	18		17	22	19	21
11q22.3–q23.1							
Loss		18	100			5	9
12							
Gain		91					15
13q14.2–q14.3							
Loss	56			17	11		21
16q23.2–q24.2							
Gain	24	27	20	17	22	19	21
17p13.2–p13.1							
Loss	16			83			10
17q25.3							
Gain	24	9	20	17	22	14	19
18q21.2							
Loss	16			33			7.5
Gain	12	18		17		9	12
20q13.12							
Gain	24	27	20		22	14	19

Results are expressed as percentages.

[a]Statistically significant associations ($P < 0.05$).

aCGH, array comparative genomic hybridization; CLL, chronic lymphocytic leukemia.

chromosomes 5, 7, 11, 13, 17 and 18 and gains in chromosomes 1, 6, 10, 11, 12, 16, 17, 18 and 20 were present in this series. Regarding other recently reported alterations, we observed gain on 2p [34] in one case.

Our study identifies a previously undescribed recurrent region of gain in CLL, located on 20q13 in 19% of CLL patients. This frequency is similar to other well-characterized abnormalities in CLL (+12, 11q- and 17p-). It should be noted that gains in 20q were not associated with any other cytogenetic abnormality, although no patients with loss on 17p displayed 20q gains. The presence of 20q gains was not associated with mutational status either. Abnormalities of chromosome 20 are frequently observed aberrations in cancer [35–37]. In addition, the presence of gains on 20q has been associated with aggressive tumor behavior and poor clinical prognosis [38]. By contrast, deletions of the long arm of chromosome 20 are a common chromosomal abnormality associated with myeloid malignancies and are rarely seen in lymphoid malignancies [39]. A detailed analysis of 20q gains in cancer revealed that the size and location of the alteration are both variable. A region of gain at 20q13 was identified in CGH studies in human breast tumors [40]. The region has been analyzed at higher resolution, enabling three independently amplified regions to be characterized, with 20q13.2 being the most common region of gain in breast cancer. In the present study, FISH studies identified a minimal region of gain on 20q13.12 of ~2.31 Mb. This region is located close to the 20q breast cancer amplicon but is not included in it.

The gain on 20q13 in CLL could be relevant to the pathogenesis and evolution of CLL because 11 protein-coding cancer-related genes have been identified in this region (supplemental Table S1, available at *Annals of Oncology* online). It should be noted that all of these genes were upregulated in

Table 6. Number of changes per patient in FISH groups and 20q13.12 cases

FISH	Number of aberrations (% cases)		Median of changes
	≤5	>5	
13q14.3	10 (40)	15 (60)	6
Trisomy 12	5 (45)	6 (55)	3
11q22.3	0	5[a] (100)	9
17p13.1	3 (50)	3 (50)	4
t(14q32)	5 (56)	4 (44)	3
Normal FISH	20[a] (95)	1 (5)	3
20q13.12 gain	3 (23)	10[a] (77)	7
Total (%)	4	33	

[a]Statistically significant associations ($P < 0.05$).

the CLL patients showing 20q gains in comparison with the other CLL cases. Thus, *PIGT*, *PI3*, *SLPI* and *WFDC2* could be potential candidate genes since they have been previously related to progression or tumor invasion. Phosphatidylinositol glycan (PIG) class T (*PIG-T*) is a component of the GPI transamidase complex and is amplified and overexpressed in human breast cancer cell lines and primary tumors [41]. Previous studies suggested that activation of the GPI transamidase complex could be a molecular mechanism underlying the progression of various human cancers [41, 42]. Interestingly, *GIP-S*, another GPI subunit, is located on 17p13.2, a region frequently deleted in cancer and in CLL. Therefore, further studies of these genes and their biological effects of all GPI transamidase complex subunits could be relevant in CLL. *PI3*, *SLPI* and *WFDC2* are members of the WAP family, a group of genes coding for proteins with a WAP motif. All of them have been identified as molecular markers for cancer and are clustered on chromosome 20q12–13.1. These genes are amplified and upregulated in several cancers [43]. The expression levels of all these genes were significantly higher in CLL cases with gains on 20q. Therefore, we suggest that 20q13.12 overexpressed genes may also be important in the evolution of CLL and warrant detailed study.

The present study also revealed a gene dosage effect in other chromosomal regions. Thus, CLL patients with trisomy 12 overexpressed genes located on chromosome 12, while patients with losses on 17p underexpressed genes located on 17p, as previously reported [16–18].

Gains in 20q13 in CLL did not occur as a single aberration because all CLL patients with gains in this region also had additional genetic changes. In fact, gains on 20q were associated with genomic complexity (Table 6). It is of note that genomic complexity has a significant impact on cancer prognosis and a number of studies have described the presence of several genomic changes as being predictors of disease progression and chemosensitivity in CLL [9, 44]. A significantly high level of genomic complexity in patients with loss on 11q was also observed. However, the CLL patients with losses on 17p did not have a large number of genomic alterations. This observation may indicate that the poor prognosis of patients with CLL exhibiting loss on 17p is unrelated to their genomic complexity [9]. The presence of a large number of genomic alterations in 20q13-gain patients suggest that this new genetic entity could be associated with a more advanced disease in CLL, as has been suggested in non-Hodgkin's lymphomas [45]. Genomic instability could therefore be another molecular feature of CLL progression, as has recently been suggested [46]. In order better to assess the clinical value of gain on 20q, a prospective study in a large series of CLL patients needs to be carried out.

Our results failed to demonstrate the presence of recurrent secondary genetic imbalances in the cytogenetic subgroups. In fact, only the group of patients with losses in 13q showed an association with losses in 5q13.3–q14.1 and 5q31. These changes had not been previously reported and could be examined further in subsequent studies.

In summary, our results demonstrated that submicroscopic 20q13.12 gains are common in CLL and confirmed that these gains result in an overexpression of the genes located on 20q13 [Figure 1, supplemental Table S1 (available at *Annals of Oncology* online)]. Furthermore, 20q gain is associated with great genomic complexity. These results suggest that the diversity of genomic aberrations in CLL is much greater than previously suggested. Further studies are needed to assess the prognostic significance of these alterations and how the genes located in these loci could contribute to the pathogenesis of CLL.

acknowledgements

We thank N Carter and H Fiegler (Sanger Center, Cambridge, UK) for providing us with the BACs library. We thank Irene Rodríguez, Sara González, Teresa Prieto, Mª Ángeles Ramos, Almudena Martín, Ana Díaz, Ana Simón, María del Pozo and Vanesa Gutiérrez of the Centro de Investigación del Cáncer, Salamanca, Spain, for their technical assistance.

funding

This work was partially supported by grants from the Spanish Fondo de Investigaciones Sanitarias (02/1041 and FIS 09/01543); Fondo Social Caja de Burgos de Investigación Clínica, Proyectos de investigación del SACYL (106/A/06) and by the 'Acción Transversal del Cáncer' project, through an agreement between the Instituto de Salud Carlos III (ISCIII), Spanish Ministry of Science and Innovation and the Cancer Research Foundation of Salamanca University and the Redes de Investigación RTIIC (FIS). AER is fully supported by an 'Ayuda predoctoral FIS de formación en investigación' by the Spanish Fondo de Investigaciones Sanitarias.

disclosure

The authors declare no conflicts of interest.

references

1. Chiorazzi N, Rai KR, Ferrarini M. Chronic lymphocytic leukemia. N Engl J Med 2005; 352: 804–815.
2. Dohner H, Stilgenbauer S, Benner A et al. Genomic aberrations and survival in chronic lymphocytic leukemia. N Engl J Med 2000; 343: 1910–1916.
3. Rozman C, Montserrat E. Chronic lymphocytic leukemia. N Engl J Med 1995; 333: 1052–1057.
4. Damle RN, Wasil T, Fais F et al. Ig V gene mutation status and CD38 expression as novel prognostic indicators in chronic lymphocytic leukemia. Blood 1999; 94: 1840–1847.

5. Orchard JA, Ibbotson RE, Davis Z et al. ZAP-70 expression and prognosis in chronic lymphocytic leukaemia. Lancet 2004; 363: 105–111.
6. Rassenti LZ, Huynh L, Toy TL et al. ZAP-70 compared with immunoglobulin heavy-chain gene mutation status as a predictor of disease progression in chronic lymphocytic leukemia. N Engl J Med 2004; 351: 893–901.
7. Hernandez JA, Rodriguez AE, Gonzalez M et al. A high number of losses in 13q14 chromosome band is associated with a worse outcome and biological differences in patients with B-cell chronic lymphoid leukemia. Haematologica 2009; 94: 364–371.
8. Finn WG, Kay NE, Kroft SH et al. Secondary abnormalities of chromosome 6q in B-cell chronic lymphocytic leukemia: a sequential study of karyotypic instability in 51 patients. Am J Hematol 1998; 59: 223–229.
9. Kujawski L, Ouillette P, Erba H et al. Genomic complexity identifies patients with aggressive chronic lymphocytic leukemia. Blood 2008; 112: 1993–2003.
10. Kipps TJ. Genomic complexity in chronic lymphocytic leukemia. Blood 2008; 112: 1550.
11. Grubor V, Krasnitz A, Troge JE et al. Novel genomic alterations and clonal evolution in chronic lymphocytic leukemia revealed by representational oligonucleotide microarray analysis (ROMA). Blood 2009; 113: 1294–1303.
12. de Leeuw RJ, Davies JJ, Rosenwald A et al. Comprehensive whole genome array CGH profiling of mantle cell lymphoma model genomes. Hum Mol Genet 2004; 13: 1827–1837.
13. Kohlhammer H, Schwaenen C, Wessendorf S et al. Genomic DNA-chip hybridization in t(11;14)-positive mantle cell lymphomas shows a high frequency of aberrations and allows a refined characterization of consensus regions. Blood 2004; 104: 795–801.
14. Rubio-Moscardo F, Climent J, Siebert R et al. Mantle-cell lymphoma genotypes identified with CGH to BAC microarrays define a leukemic subgroup of disease and predict patient outcome. Blood 2005; 105: 4445–4454.
15. Tyybakinoja A, Saarinen-Pihkala U, Elonen E et al. Amplified, lost, and fused genes in 11q23-25 amplicon in acute myeloid leukemia, an array-CGH study. Genes Chromosomes Cancer 2006; 45: 257–264.
16. Haslinger C, Schweifer N, Stilgenbauer S et al. Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status. J Clin Oncol 2004; 22: 3937–3949.
17. Porpaczy E, Bilban M, Heinze G et al. Gene expression signature of chronic lymphocytic leukaemia with trisomy 12. Eur J Clin Invest 2009; 39: 568–575.
18. Dickinson JD, Joshi A, Iqbal J et al. Genomic abnormalities in chronic lymphocytic leukemia influence gene expression by a gene dosage effect. Int J Mol Med 2006; 17: 769–778.
19. Pfeifer D, Pantic M, Skatulla I et al. Genome-wide analysis of DNA copy number changes and LOH in CLL using high-density SNP arrays. Blood 2007; 109: 1202–1210.
20. Gunn SR, Bolla AR, Barron LL et al. Array CGH analysis of chronic lymphocytic leukemia reveals frequent cryptic monoallelic and biallelic deletions of chromosome 22q11 that include the PRAME gene. Leuk Res 2009; 33: 1276–1281.
21. Patel A, Kang SH, Lennon PA et al. Validation of a targeted DNA microarray for the clinical evaluation of recurrent abnormalities in chronic lymphocytic leukemia. Am J Hematol 2008; 83: 540–546.
22. Schwaenen C, Nessling M, Wessendorf S et al. Automated array-based genomic profiling in chronic lymphocytic leukemia: development of a clinical tool and discovery of recurrent genomic alterations. Proc Natl Acad Sci U S A 2004; 101: 1039–1044.
23. Tyybakinoja A, Vilpo J, Knuutila S. High-resolution oligonucleotide array-CGH pinpoints genes involved in cryptic losses in chronic lymphocytic leukemia. Cytogenet Genome Res 2007; 118: 8–12.
24. Binet JL, Caligaris-Cappio F, Catovsky D et al. Perspectives on the use of new diagnostic tools in the treatment of chronic lymphocytic leukemia. Blood 2006; 107: 859–861.
25. Gonzalez MB, Hernandez JM, Garcia JL et al. The value of fluorescence in situ hybridization for the detection of 11q in multiple myeloma. Haematologica 2004; 89: 1213–1218.
26. Robledo C, Garcia JL, Caballero D et al. Array comparative genomic hybridization identifies genetic regions associated with outcome in aggressive diffuse large B-cell lymphomas. Cancer 2009; 115: 3728–3737.
27. Ghia P, Stamatopoulos K, Belessi C et al. ERIC recommendations on IGHV gene mutational status analysis in chronic lymphocytic leukemia. Leukemia 2007; 21: 1–3.
28. Gutierrez NC, Lopez-Perez R, Hernandez JM et al. Gene expression profile reveals deregulation of genes with relevant functions in the different subclasses of acute myeloid leukemia. Leukemia 2005; 19: 402–409.
29. Risueno A, Fontanillo C, Dinger ME et al. GATExplorer: genomic and transcriptomic explorer; mapping expression probes to gene loci, transcripts, exons and ncRNAs. BMC Bioinformatics 2010; 11: 221.
30. Di Bernardo MC, Crowther-Swanepoel D, Broderick P et al. A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. Nat Genet 2008; 40: 1204–1210.
31. Crowther-Swanepoel D, Broderick P, Di Bernardo MC et al. Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. Nat Genet 2010; 42: 132–136.
32. Ouillette P, Erba H, Kujawski L et al. Integrated genomic profiling of chronic lymphocytic leukemia identifies subtypes of deletion 13q14. Cancer Res 2008; 68: 1012–1021.
33. Gunn SR, Mohammed MS, Gorre ME et al. Whole-genome scanning by array comparative genomic hybridization as a clinical tool for risk assessment in chronic lymphocytic leukemia. J Mol Diagn 2008; 10: 442–451.
34. Jarosova M, Urbankova H, Plachy R et al. Gain of chromosome 2p in chronic lymphocytic leukemia: significant heterogeneity and a new recurrent dicentric rearrangement. Leuk Lymphoma 2010; 51: 304–313.
35. Yang SH, Seo MY, Jeong HJ et al. Gene copy number change events at chromosome 20 and their association with recurrence in gastric cancer patients. Clin Cancer Res 2005; 11: 612–620.
36. Zhu H, Lam DC, Han KC et al. High resolution analysis of genomic aberrations by metaphase and array comparative genomic hybridization identifies candidate tumour genes in lung cancer cell lines. Cancer Lett 2007; 245: 303–314.
37. Lassmann S, Weis R, Makowiec F et al. Array CGH identifies distinct DNA copy number profiles of oncogenes and tumor suppressor genes in chromosomal- and microsatellite-unstable sporadic colorectal carcinomas. J Mol Med (Berl) 2007; 85: 293–304.
38. Bar-Shira A, Pinthus JH, Rozovsky U et al. Multiple genes in human 20q13 chromosomal region are involved in an advanced prostate cancer xenograft. Cancer Res 2002; 62: 6803–6807.
39. Bench AJ, Nacheva EP, Hood TL et al. Chromosome 20 deletions in myeloid malignancies: reduction of the common deleted region, generation of a PAC/BAC contig and identification of candidate genes. UK Cancer Cytogenetics Group (UKCCG). Oncogene 2000; 19: 3902–3913.
40. Tanner MM, Tirkkonen M, Kallioniemi A et al. Independent amplification and frequent co-amplification of three nonsyntenic regions on the long arm of chromosome 20 in human breast cancer. Cancer Res 1996; 56: 3441–3445.
41. Wu G, Guo Z, Chatterjee A et al. Overexpression of glycosylphosphatidylinositol (GPI) transamidase subunits phosphatidylinositol glycan class T and/or GPI anchor attachment 1 induces tumorigenesis and contributes to invasion in human breast cancer. Cancer Res 2006; 66: 9829–9836.
42. Scotto L, Narayan G, Nandula SV et al. Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. Genes Chromosomes Cancer 2008; 47: 755–765.
43. Clauss A, Lilja H, Lundwall A. A locus on human chromosome 20 contains several genes expressing protease inhibitor domains with homology to whey acidic protein. Biochem J 2002; 368: 233–242.
44. Kay NE, Eckel-Passow JE, Braggio E et al. Progressive but previously untreated CLL patients with greater array CGH complexity exhibit a less durable response to chemoimmunotherapy. Cancer Genet Cytogenet 2010; 203: 161–168.
45. Carter SL, Eklund AC, Kohane IS et al. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat Genet 2006; 38: 1043–1048.
46. Stephens PJ, Greenman CD, Fu B et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell 2011; 144: 27–40.